State of AI Data Labs and Human Feedback

READY-FOR-ARCHIVE | research-2026-11-12.1 | Updated 2026-11-15

Complete research package frozen before archival publication. Check the release record for current DOI registration and publication status. Reserved version DOI (not registered): 10.5281/zenodo.24249277.

Coverage and publication status

Reviewed research	Current	Launch minimum
Existing organization inclusion audits	39	39
Qualifying suppliers	73	60
Full-source paper reviews	225	150
Dataset families	85	60
Benchmark families	87	20

Contents

1. Executive assessment 3

2. Weekly developments and edition discipline 14

3. Market definition, structure and provider categories 16

4. Technical landscape: from supervision to executable evaluation 21

5. Provider dossiers: offerings, delivery, differentiation and limits 147

6. Dataset and benchmark landscape: provenance, versions and use limits 203

7. Human expertise, recruitment, disagreement and labor 306

8. Commercial evidence and procurement economics 312

9. Rights, security, privacy and attribution governance 316

10. Buyer decision guides: use cases, pilot designs and acceptance 320

11. Research outlook: testable judgments and open questions 330

12. Methods, coverage, bibliography and reproducible appendices 334

Bibliography and source versions 398

Evidence and reproduction 482

1. Executive assessment

Method: ScienceToStartup publishes State of AI Data Labs and Human Feedback under the publisher identity Musa Ozdem. It is an evidence-linked analyst synthesis for buyers, research teams and contributors. Publication status, exact-edition citations and verified archive links are reported separately from the captured-source cutoff and substantive review dates. A scheduled date or a prepared package does not establish archival publication, and no human peer review is claimed. The source-qualified eligible core is frozen at 416 identities; the actual reference audit is accepted, while publication separately requires screening and all other gates. Canonical dataset 2.5.0 and historical source/review dates are unchanged. The purpose is to expose the source observation, the analyst’s reasoning, its confidence and the evidence that would change it.

Method: The current authoring input contains 2,548 approved assertions, 73 qualifying supplier groups, 225 distinct reviewed paper works, 85 dataset families and 87 benchmark families under coverage method 1.3.0. All 39 legacy organizations have completed bounded inclusion audits; 38 have conclusive dispositions and BenchFlow remains commercially pending. The graph retains 195 candidate and 39 historical assertions. Candidate records include unresolved or superseded material; their count is not a scientific-error count. These are selected-corpus counts. Within the separately source-qualified 416-unit frozen core, 389 are deeply dispositioned: 389/416 = 93.5096%, leaving 27 genuine deep-review gaps. This report covers every counted paper work and supplier group and all counted families without claiming a census; the methods preserve the full unresolved scope.

Analysis: The market is more intelligible when the comparison begins with the unit being bought. A participant platform, a managed annotation program, a licensed dataset, an evaluation service and an executable RL environment leave different responsibilities with the buyer. Prolific distinguishes participant access from managed collection; Snorkel describes refreshed non-exclusive products; Mercor describes expert task construction; Surge describes environments and verifiers. The analytical implication is to compare statements of work and accepted deliverables before comparing brand claims. This evidence supports a responsibility map within this sample, rather than a universal competitive ranking.

Sources: [684], [690], [720], [753]

Analysis: The word human conceals several different supervision arrangements. Human-written source material, human task design, independently collected judgments, correction of model drafts and human review of a model-selected subset are not interchangeable inputs. HelpSteer3, UltraFeedback, FineWeb-Edu and the described DeepSeek-R1 cold-start process demonstrate distinct combinations. A buyer should ask where a person intervenes and what that person can observe. Nothing in this sample establishes that one provenance class always produces better models; the defensible comparison is whether the supervision target and validation design fit the intended task.

Sources: [295], [454], [715], [734]

Analysis: Quality requires denominators and a theory of disagreement. HelpSteer3's retained consensus, HelpSteer2's closest-rating aggregation, InstructGPT's ranked sets and DICES' repeated safety judgments arise from different observation processes. A pair count can expand from rankings without adding independent people; a post-filter agreement result can improve while excluding difficult items; a majority can erase a meaningful minority. Procurement should preserve raw labels, exclusions and adjudication reasons. These are methodological judgments based on the documented workflows, not measured differences in supplier performance.

Sources: [132], [451], [715], [731]

Analysis: Executable outcomes improve observability but do not eliminate judgment. SWE-bench tests, OSWorld state checks, τ-bench database rewards and PaperBench rubric leaves expose concrete success conditions. Their papers also identify incomplete test coverage, damaging side effects, missed policy violations or judge limitations. A reliable acceptance plan therefore needs both outcome checks and a separate review of constraints, side effects and ambiguous failures. The evidence does not justify treating a passed verifier as a certification of deployment readiness or general professional competence.

Sources: [233], [254], [365], [733]

Analysis: Public financial and workforce numbers answer different questions. Appen's annual group revenue, Proximal's annualized pace, Scale's cumulative contributor payout and Prolific's active-participant disclosure do not form a common revenue, capacity or labor denominator. Labelbox financing combines equity and a credit facility. The report therefore offers a commercial evidence ledger and project-cost methodology rather than market-share estimates or a valuation league table. Missing disclosure remains unknown, and a large advertised pool does not establish current paid availability for a particular language, credential or task.

Sources: [72], [246], [713], [720], [736]

Analysis: Rights, privacy and security must follow the exact material and service boundary. Marketplace terms, a website privacy notice, a training-data license, a worker agreement and a security attestation can cover different subjects. Defined.ai separates marketplace datasets from new collection; AfterQuery excludes customer processor work from its website policy; Mercor directs work-platform users to a separate notice; MathVista retains underlying image and question rights while restricting training use. The analyst judgment is to obtain project-specific evidence for each boundary. Public documentation completeness is useful for diligence, but it is not itself a quality or compliance rating.

Sources: [59], [177], [208], [739]

Method: All nine executive finding identifiers from the previous unpublished synthesis are retained. The additional studies refine their evidence and limits: post-selection agreement differs from raw agreement; model-simulated feedback differs from live expert assistance; disclosure audits describe reporting; and observed costs depend on the counted work and time. These refinements do not create a published weekly comparison or rewrite a historical archive. Confidence remains attached to each scoped judgment and to the evidence that could change it.

Analysis: The added annotation studies strengthen a practical distinction between an improved retained dataset and an improved production process. HelpSteer2 raises agreement after selecting and filtering labels; the surgical-annotation study compares instruction conditions and subsequent QA; LabelAId measures a particular intervention against researcher-adjudicated labels. Those designs answer different questions. A buyer should require raw-input, excluded-item and final-acceptance denominators, plus independent error inspection, before interpreting a high agreement score or a QA stage as evidence of better delivered work.

Sources: [430], [432], [435]

Analysis: The added professional-work literature strengthens a measurement warning rather than a labor forecast. APEX-Agents distinguishes any-attempt and all-attempt success; APEX-Accounting contrasts mean criterion completion with all-eight-attempt reliability; DAYJOB separates strict pass from relaxed thresholds and reports creator-estimated healthcare time with a separately described finance mean; PLE combines rubric and hybrid executable scores. These are different questions about selected tasks. Confidence is high that their percentages cannot be pooled into a share of jobs automated. A useful buyer comparison requires a common task population, complete-attempt accounting, independently calibrated judges and observed practitioner baselines.

Sources: [561], [695], [697], [700]

Analytical judgment: delivery-unit-before-provider

In this reviewed sample, a useful supplier comparison begins with the delivery unit: participant access, licensed data, custom tasks or executable environments.

Buyer implication: Specify who owns task design, annotation, QA, rights clearance and acceptance before comparing proposals.

Limitations: First-party offering descriptions do not establish interchangeable services, capacity, realized quality or a market-wide supplier ranking.

Confidence: Moderate for this sample’s disclosed distinctions; low for prevalence outside it.

Reasoning: Prolific distinguishes participant access and managed collection; Snorkel describes refreshed non-exclusive products; Mercor describes custom task construction; Surge offers environments and verifiers. Their responsibilities differ even where capability labels overlap. The expanded corpus also distinguishes engineering teams, managed evaluation, hosted compute and contract-limited talent acquisition; the comparison remains task-specific. The current supplier set also contains schema-mapped web feeds and a Datatang/Nexdata group combining licensed resources and customization. A group identity is not proof that every brand or service is interchangeable.

What would change this assessment: Matched statements of work and independently checked deliveries could show that particular offerings meet the same requirement and permit narrower outcome comparisons.

Approved evidence IDs: candidate.supplier.prolific.platform-managed, repair.qualbinding.20261111.v1.candidate.supplier.snorkel_ai.data-series, repair.qualbinding.20261111.v1.candidate.supplier.mercor.task-construction, candidate.supplier.surge_ai.rl-environments, repair20261111.claim.current-reference192, candidate.cohort4.prime_intellect.professional-service-separation, candidate.mandatory14.tagx.collection-limits, candidate.mandatory14.tagx.delivery-formats, candidate.mandatory14.tagx.legal-footer-identity, candidate.mandatory14.tagx.managed-web-data-offer, candidate.mandatory14.datatang.crowdsourced-training-data-production, candidate.mandatory14.datatang.group-dedup-ownership, candidate.mandatory14.datatang.issuer-identity, candidate.mandatory14.datatang.japanese-brand-identity, candidate.mandatory14.datatang.licensed-resource-service, candidate.mandatory14.datatang.reported-revenue2026h1

Analytical judgment: human-and-model-provenance

Human content, human labels and human oversight are distinct provenance claims; the selected datasets contain materially different combinations.

Buyer implication: Require a stage-level provenance manifest and test the label target relevant to the intended task.

Limitations: The reviewed methods are source-reported and version-specific. No universal advantage of human or synthetic labels is established. Human review of a model draft does not establish wholly human authorship, and synthetic provenance does not establish inferior or superior quality.

Confidence: High for the documented provenance distinctions; unestablished for comparative training outcomes.

Reasoning: HelpSteer3 describes independent human preference annotations, UltraFeedback describes GPT-4 labels and FineWeb-Edu describes model-scored educational filtering. Calling these one annotation category would erase the origin of the supervision signal. The expanded methods also include human correction of model responses and model-generated process signals, reinforcing stage-level provenance. StarCoder separates human PII annotation from code-corpus content; the rubric study separates human taxonomy development from model-persona scoring. These additions strengthen stage-level rather than binary provenance. CharXiv separates model suggestions from human selection; ProgramBench separates source-visible construction agents from source-hidden reconstruction; GeneBench-Pro reports simulated problem creation with AI assistance and author oversight. Provenance must follow each stage.

What would change this assessment: Independent task-matched validation with raw annotations and generation logs could support outcome comparisons or expose errors in the reported provenance.

Approved evidence IDs: candidate.paperfull.helpsteer3.independent-human-labels, next.claim.ultrafeedback.model_feedback, candidate.paperfull.fineweb.synthetic-educational-labels, cohort5.claim.deepseekv3.human-check, cohort8.claim.prime.online-prm, cohort9.claim.recodeh.simulated-feedback, cohort9.claim.recodeh.metric-scope, cohort9.claim.ultrafeedback.simultaneous-rubrics, cohort9.claim.ultrafeedback.human-agreement, mandatory1.claim.arxiv.2305.06161.limitation, mandatory1.claim.arxiv.2305.06161.main_result, mandatory1.claim.arxiv.2305.06161.method, mandatory1.claim.arxiv.2305.06161.provenance, mandatory1.claim.arxiv.2305.06161.result,

mandatory2.claim.arxiv.2604.01375.limitation, mandatory2.claim.arxiv.2604.01375.method, mandatory2.claim.arxiv.2604.01375.provenance, mandatory2.claim.arxiv.2604.01375.result, candidate.normal18.charxiv.paper.chart-selection, candidate.normal18.charxiv.paper.documents-family, candidate.normal18.charxiv.paper.funding-roles, candidate.normal18.charxiv.paper.grading-and-limitations, candidate.normal18.charxiv.paper.mixed-human-model-question-construction, candidate.normal18.programbench.agent-construction, candidate.normal18.programbench.assertion-quality-method.precision-v2, candidate.normal18.programbench.baseline-configuration, candidate.normal18.programbench.documents-family, candidate.normal18.programbench.judge-ablation-scope, candidate.normal18.programbench.manual-review-scope, assertion.genebench.pro.agent-task.20261112, assertion.genebench.pro.decision-points.20261112, assertion.genebench.pro.drx1-case-workflow.20261112, assertion.genebench.pro.external-expert-pool.20261112, assertion.genebench.pro.external-review-scope.20261112, assertion.genebench.pro.external-review-workspace.20261112, assertion.genebench.pro.ldsc-review-example.20261112, assertion.genebench.pro.provenance-wording-discrepancy.20261112, assertion.genebench.pro.public-answer-separation.20261112, assertion.genebench.pro.public-card-integrity.20261112, assertion.genebench.pro.public-card-inventory.20261112, assertion.genebench.pro.public-card-license.20261112, assertion.genebench.pro.public-grader-semantics.20261112, assertion.genebench.pro.recoverable-target.20261112, assertion.genebench.pro.release-selection.20261112, assertion.genebench.pro.release-tiers.20261112, assertion.genebench.pro.review-revisions.20261112, assertion.genebench.pro.scope.20261112, assertion.genebench.pro.simulated-origin.20261112, assertion.genebench.pro.validation-loop.20261112, assertion.genebench.pro.version-transition.20261112

Analytical judgment: agreement-needs-denominators

Released agreement and preference-pair totals cannot be interpreted without the independent-judgment and exclusion denominators.

Buyer implication: Request raw-to-retained counts, invalid-item rules, repeat labels, disagreement slices and corrected acceptance.

Limitations: The reviewed workflows differ in task and sampling. They cannot rank provider quality, and no common reliability measurement was executed.

Confidence: High for the methodological distinction; low for transferring a numerical agreement result between workflows.

Reasoning: HelpSteer3’s retained consensus, HelpSteer2’s post-filter closest-rating aggregation, InstructGPT’s ranked sets and DICES’ repeated judgments have different observation and selection denominators. The added HelpSteer2 paper makes the exclusions explicit; the inference study limits normalized scores to within-pair comparisons; Gilardi excludes trained-rater disagreement from the accuracy reference. These distinctions make raw counts, exclusions and adjudication necessary before comparison. MACE reports coverage-dependent comparisons, MMRB2 excludes ambiguous pairs and Regulated Advice Bench retains translated rows under one original-item label. None is an independent-judgment count merely because it has many rows. The added sources also distinguish nested chart questions, predefined consensus criteria from individualized rubrics, and sampled judge agreement from review of every response. Source-stage headcount discrepancies remain unresolved.

What would change this assessment: A common held-out task with independent raw judgments and the same exclusion rules could make reliability comparisons meaningful.

Approved evidence IDs: candidate.paperfull.helpsteer3.postfilter-consensus, next.claim.helpsteer2.qa, candidate.paperfull.instructgpt.human-ranking-sets, candidate.paperfull.instructgpt.single-contractor-comparisons, cohort3.claim.dices.expert-crowd-difference, cohort3.claim.disagreement.agreement-not-superiority, cohort9.claim.helpsteer2.agreement-selection, cohort9.claim.inherent.normalization-limit, cohort9.claim.gilardi.gold-denominator, mandatory1.claim.acl.N13-1132.limitation, mandatory1.claim.acl.N13-1132.method, mandatory1.claim.acl.N13-1132.provenance, mandatory1.claim.acl.N13-1132.result, mandatory2.claim.arxiv.2512.16899.limitation, mandatory2.claim.arxiv.2512.16899.method, mandatory2.claim.arxiv.2512.16899.provenance, mandatory2.claim.arxiv.2512.16899.result, candidate.mandatory19.benchmark.regulated-advice-bench.agreement-denominators, candidate.mandatory19.benchmark.regulated-advice-bench.card-license, candidate.mandatory19.benchmark.regulated-advice-bench.policy-and-translation-limits, candidate.mandatory19.benchmark.regulated-advice-bench.synthetic-expert-label-process, candidate.mandatory19.benchmark.regulated-advice-bench.translated-item-units, candidate.normal18.charxiv.family.component-rights-and-use, candidate.normal18.charxiv.family.princeton-annotation-role.binding-correction-v1, candidate.normal18.charxiv.family.question-units-and-splits.binding-correction-v1, candidate.normal18.charxiv.family.release-access-and-known-error.binding-correction-v1, candidate.normal18.healthbench.paper.compensated-physician-cohort, candidate.normal18.healthbench.paper.conversation-construction, candidate.normal18.healthbench.paper.documents-family, candidate.normal18.healthbench.paper.grader-validation-scope, candidate.normal18.healthbench.paper.rubric-score-definition, candidate.normal18.healthbench.paper.workflow-outcome-limits, candidate.normal18.swe-atlas.documents-family, candidate.normal18.swe-atlas.expert-construction-and-qa, candidate.normal18.swe-atlas.grader-validation-scopes, candidate.normal18.swe-atlas.paid-author-reviewer-disclosure, candidate.normal18.swe-atlas.runtime-and-provider-roles, candidate.normal18.swe-atlas.three-trial-metric-definitions

Analytical judgment: artifact-role-is-narrow

Named artifact attribution is strongest when it preserves the exact contribution stage and release rather than extending a brand association across a family.

Buyer implication: Ask for source-supported roles for the specific delivered artifact; keep recruitment, annotation, validation and authorship separate.

Limitations: These examples do not reveal contract values, contribution shares, exhaustive supplier lists or causal provider performance.

Confidence: High for the explicit scoped roles in this sample; incomplete for the broader supplier graph.

Reasoning: HelpSteer3 names Scale and Translated for different subsets; the Prolific case establishes recruitment through a researcher-owned interface; U-MATH credits Toloka AI and Gradarius with validation of professor-authored problems. These relationships support narrower claims than sole production. The added papers document contractor demonstrations/comparisons, a PII-labeling project and expert benchmark construction as separate roles. The report does not extend them into complete corpus authorship or a causal provider-performance claim. CVDP thanks Turing for dataset collaboration and Cadence for tool licenses; SUPERChem names original-problem and compute contributors; HCAST identifies affiliated individuals' QA roles. These are different contributions and do not imply complete corpus ownership or an unstated organizational supply contract.

What would change this assessment: Additional artifact-specific acknowledgments, contributor records or contracts could establish further stages or revise the scope.

Approved evidence IDs: candidate.supplier.scale_ai.helpsteer3-contribution, next.claim.helpsteer3.translated, candidate.supplier.prolific.cohere-recruitment, candidate.supplier.toloka.umath-validation, mandatory1.claim.arxiv.2112.09332.limitation, mandatory1.claim.arxiv.2112.09332.method, mandatory1.claim.arxiv.2112.09332.provenance, mandatory1.claim.arxiv.2112.09332.result, mandatory1.claim.arxiv.2305.06161.limitation, mandatory1.claim.arxiv.2305.06161.main_result, mandatory1.claim.arxiv.2305.06161.method, mandatory1.claim.arxiv.2305.06161.provenance, mandatory1.claim.arxiv.2305.06161.result, mandatory2.claim.arxiv.2601.14242.limitation, mandatory2.claim.arxiv.2601.14242.method, mandatory2.claim.arxiv.2601.14242.provenance, mandatory2.claim.arxiv.2601.14242.result, candidate.normal18.cvdp.family.component-specific-license-declarations, candidate.normal18.cvdp.family.context-and-tooling-scope, candidate.normal18.cvdp.family.creator-reported-coalition-adoption.wording-v2, candidate.normal18.cvdp.family.original-problem-units, candidate.normal18.cvdp.family.public-release-access-limits, candidate.normal18.cvdp.family.turing-and-cadence-contribution-roles.wording-v2, candidate.normal18.cvdp.family.version-and-comparability, candidate.normal18.superchem.center-mass-problem-contribution, candidate.normal18.superchem.chemy-problem-contribution, candidate.normal18.superchem.compute-contribution, candidate.normal18.superchem.curation-workflow, candidate.normal18.superchem.curator-cohort, candidate.normal18.superchem.documents-family, candidate.normal18.superchem.published-comparability, candidate.normal18.superchem.rpf-method, candidate.normal18.superchem.tal-problem-contribution, candidate.next8.metr.hcast.baseline-recording-and-correction.v1, candidate.next8.metr.hcast.baseliner-recruitment-and-qualification.v1, candidate.next8.metr.hcast.container-and-score.v1, candidate.next8.metr.hcast.contamination-strata.v1, candidate.next8.metr.hcast.equistamp-affiliated-contribution.v1, candidate.next8.metr.hcast.human-task-construction.v1, candidate.next8.metr.hcast.manual-task-quality-control.v1, candidate.next8.metr.hcast.measured-and-forecast-human-times.v1, candidate.next8.metr.hcast.model-generated-appendix-summaries.v1, candidate.next8.metr.hcast.paid-human-incentives.v1, candidate.next8.metr.hcast.public-subset-license-and-access.v1

Analytical judgment: benchmark-protocol-before-score

Benchmark family names alone are insufficient for comparing scores or deciding deployment readiness.

Buyer implication: Freeze task version, candidate responses, aggregation and execution environment, then inspect errors on the buyer’s actual task.

Limitations: No benchmark was executed for this report. The reviewed sources do not establish a universal deployment-quality ranking. Reported task, release, harness, attempt budget and uncertainty conventions differ; unresolved source counts are not silently reconciled.

Confidence: High for protocol differences and stated limitations; unestablished for any new system’s performance.

Reasoning: RewardBench versions change scoring structure; TruthfulQA changes answer protocol and separates generation metrics; SWE-bench warns that passing tests does not ensure complete, efficient or readable solutions. The measures answer different questions. OSWorld and τ-bench additionally identify outcome-checking gaps for damaging side effects and policy violations. The new professional suites distinguish repeated any/all success, criterion averages, strict/relaxed passes and hybrid scores. They cannot be pooled into a professional-capability percentage or a job-automation rate. The added evidence spans median-of-three criterion scores,

equal-task all-field pass rates with hierarchical bootstrap, and fitted human-time horizons over overlapping task inputs. Release revisions and source discrepancies remain part of the measurement definition.

What would change this assessment: Reproducible matched-protocol results and independently checked target-task outcomes could support a narrower comparative claim.

Approved evidence IDs: next.claim.rewardbench.v1.scoring, next.claim.rewardbench.v2.scoring, next.claim.truthfulqa.metrics, next.claim.truthfulqa.binary_revision, candidate.paperfull.swebench.regression-test-scoring, cohort4.claim.osworld.side-effect-gap, cohort4.claim.taubench.reward-limit, mandatory2.claim.arxiv.2601.14242.limitation, mandatory2.claim.arxiv.2601.14242.method, mandatory2.claim.arxiv.2601.14242.provenance, mandatory2.claim.arxiv.2601.14242.result, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2607.27189.limitation, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2607.27189.method, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2607.27189.result, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.01306.limitation, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.01306.method, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.01306.result, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.30537.limitation, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.30537.method, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.30537.result, candidate.next8.apex.paper.criterion-majority-judge.v3, candidate.next8.apex.paper.cross-benchmark-comparability.v3, candidate.next8.apex.paper.document-input-and-access.v3, candidate.next8.apex.paper.economic-value-interpretation.v3, candidate.next8.apex.paper.human-judge-validation-scope.v3, candidate.next8.apex.paper.judge-repeatability-scope.v3, candidate.next8.apex.paper.original-model-results.v3, candidate.next8.apex.paper.rubric-error-and-retrieval-limits.v3, candidate.next8.apex.paper.thinking-comparison-limit.v3, candidate.next8.apex.paper.three-run-median.v3, assertion.genebench.pro.ai-human-oversight.20261112, assertion.genebench.pro.attempts-and-errors.20261112, assertion.genebench.pro.author-contributions.20261112, assertion.genebench.pro.binary-grading.20261112, assertion.genebench.pro.claude-max-result.20261112, assertion.genebench.pro.evaluation-scope.20261112, assertion.genebench.pro.metric-and-uncertainty.20261112, assertion.genebench.pro.notice-act-interpretation.20261112, assertion.genebench.pro.offline-environment.20261112, assertion.genebench.pro.preprint-rights.20261112, assertion.genebench.pro.reasoning-wording-discrepancy.20261112, assertion.genebench.pro.sol-max-result.20261112, assertion.genebench.pro.sol-pro-result.20261112, assertion.genebench.pro.wall-clock-condition.20261112, candidate.next8.metr.time-horizon.th11-growth-comparability.v2, candidate.next8.metr.time-horizon.th11-model-and-infrastructure-coverage.v1, candidate.next8.metr.time-horizon.th11-scaffold-comparison.v1, candidate.next8.metr.time-horizon.th11-task-version.v1

Analytical judgment: commercial-denominators-stay-separate

The selected commercial disclosures cannot be combined into supplier market shares, production-capacity rankings or a common unit-cost comparison.

Buyer implication: Retain period, currency, organizational scope, active-window definition and acceptance basis; obtain a project-specific quote and availability check.

Limitations: Coverage is selected and incomplete. Missing disclosure is unknown, not zero; the observed figures are not matched measures.

Confidence: High for the incompatibility of these disclosed denominators; no confidence claim is made about market-wide shares.

Reasoning: Appen reports full-year consolidated and geographic-segment revenue, Scale reports cumulative contributor payouts with unspecified period and dollar currency, and Prolific distinguishes active participants from a waitlist without defining the active window. None supplies a matched current paid-work or human-data revenue denominator across suppliers. Proximal's annualized disclosure and Labelbox's mixed financing add further incompatible commercial measures. Datatang's unaudited six-month group revenue and HCAST's task-specific human incentive schedule add useful but unmatched commercial units. Neither supplies a current common service price or a supplier-capacity denominator.

What would change this assessment: Comparable segment accounts, audited activity windows and accepted-unit cost records across the same scope could permit limited financial or operational comparisons.

Approved evidence IDs: candidate.supplier.appen.revenue-fy2025, candidate.supplier.appen.segments-fy2025, candidate.supplier.scale_ai.cumulative-payout, repair20261111.claim.current-reference221, repair.binding.v1.candidate.cohort4.proximal.annualized-revenue-disclosure, candidate.cohort2.labelbox.financing, cohort9.claim.lmcrowd.cost-boundary, cohort9.claim.earnings.work-unit-wages, cohort9.claim.earnings.causality-limit, candidate.mandatory14.datatang.crowdsourced-training-data-production, candidate.mandatory14.datatang.group-dedup-ownership, candidate.mandatory14.datatang.issuer-identity, candidate.mandatory14.datatang.japanese-brand-identity, candidate.mandatory14.datatang.licensed-resource-service, candidate.mandatory14.datatang.reported-revenue2026h1, candidate.next8.metr.hcast.baseline-recording-and-correction.v1, candidate.next8.metr.hcast.baseliner-recruitment-and-qualification.v1, candidate.next8.metr.hcast.container-and-score.v1, candidate.next8.metr.hcast.contamination-strata.v1, candidate.next8.metr.hcast.equistamp-affiliated-contribution.v1, candidate.next8.metr.hcast.human-task-construction.v1, candidate.next8.metr.hcast.manual-task-quality-control.v1, candidate.next8.metr.hcast.measured-and-forecast-human-times.v1, candidate.next8.metr.hcast.model-generated-appendix-summaries.v1, candidate.next8.metr.hcast.paid-human-incentives.v1, candidate.next8.metr.hcast.public-subset-license-and-access.v1

Analytical judgment: multilingual-intersections-before-language-counts

Multilingual suitability depends on language-domain-task intersections and native validity, rather than a language count alone.

Buyer implication: Specify every required intersection, preserve translation/review provenance and calibrate the judge in the target locale.

Limitations: These selected studies do not establish universal translation inferiority, global language coverage or a provider ranking.

Confidence: Moderate for these documented methodological distinctions; unmeasured for current supplier performance.

Reasoning: HelpSteer3 records missing intersections, M-IFEval changes language-inappropriate constraints, BenchMAX post-edits translations and MENLO measures localized dimensions. JEEM's local dialect judgments and StereoTales' narrower human reference panel expose different language-generation and normative-validation boundaries. Prompt translation, chemical image-to-text parity and culturally bounded harmfulness judgments are different validation problems. StereoTales also retains an unresolved agreement/acceptability wording-direction conflict, so its correlation is not treated as universal normative validity.

What would change this assessment: Matched native-authored, translated and post-edited task studies with independent target-locale evaluation could narrow or revise the judgment.

Approved evidence IDs: repair20261111.claim.current-reference145, cohort7.claim.mifeval.japanese-filter, cohort7.claim.benchmax.pipeline, cohort7.claim.menlo.dimensions, cohort9.claim.aya.dataset-composition, cohort9.claim.aya.quality-limit, cohort9.claim.helpsteer.language-cultural-limit, mandatory2.claim.arxiv.2503.21910.limitation, mandatory2.claim.arxiv.2503.21910.method, mandatory2.claim.arxiv.2503.21910.provenance, mandatory2.claim.arxiv.2503.21910.result, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2605.10442.limitation, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2605.10442.method, repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2605.10442.result, cohort4.dataset.claim.xp3.count-units, cohort4.dataset.claim.xp3.license, cohort4.dataset.claim.xp3.schema, cohort4.dataset.claim.xp3.translated-variant, repair.binding.v1.cohort4.dataset.claim.xp3.scope, candidate.mandatory19.dataset.xp3.composite-and-variant-units, candidate.mandatory19.dataset.xp3.evaluation-and-limits, candidate.mandatory19.dataset.xp3.language-table-units, candidate.mandatory19.dataset.xp3.license-and-artifact-scope, candidate.mandatory19.dataset.xp3.prompt-and-annotation-origin, candidate.normal18.superchem.center-mass-problem-contribution, candidate.normal18.superchem.chemy-problem-contribution, candidate.normal18.superchem.compute-contribution, candidate.normal18.superchem.curation-workflow, candidate.normal18.superchem.curator-cohort, candidate.normal18.superchem.documents-family, candidate.normal18.superchem.published-comparability, candidate.normal18.superchem.rpf-method, candidate.normal18.superchem.tal-problem-contribution

Analytical judgment: rights-follow-exact-material-and-service

Rights, privacy and security diligence must follow exact components, services and uses rather than a top-level policy or license label.

Buyer implication: Obtain component rights and the actual project processing/security scope before accepting data or environments.

Limitations: This is public-source evidence and methodological judgment, not legal advice, rights clearance or an assurance audit.

Confidence: High for the documented scope differences; unestablished for legal compliance of particular deployments.

Reasoning: AlpacaEval separates code/data terms, RewardBench retains component obligations, MathVista restricts training use and provider notices distinguish website from customer processing. Magpie retains conflicting/source-specific declarations, StarCoder retains detection/opt-out limits and Chartography distinguishes online provenance from redistribution permission. A public-data collection policy does not establish downstream rights; CVDP has component-specific declarations and unspecified entries; GeneBench-Pro’s paper and public package make different declarations for different objects.

What would change this assessment: Project-specific agreements, rights-holder permissions and independent scoped assurance evidence could resolve particular unknowns.

Approved evidence IDs: cohort2.dataset.claim.alpacaeval.component-license, next.claim.rewardbench.terms, cohort4.dataset.claim.mathvista.terms, candidate.cohort4.afterquery.processor-privacy-boundary, cohort9.claim.wildchat.consent-gate, cohort9.claim.wildchat.anonymization, repair.qualification.v1.claim.magpie.source-specific-license-labels, mandatory1.claim.arxiv.2305.06161.limitation, mandatory1.claim.arxiv.2305.06161.main_result, mandatory1.claim.arxiv.2305.06161.method, mandatory1.claim.arxiv.2305.06161.provenance, mandatory1.claim.arxiv.2305.06161.result,

repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2608.10677.limitation,
repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2608.10677.method,
repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2608.10677.result,
candidate.mandatory14.tagx.collection-limits, candidate.mandatory14.tagx.delivery-formats,
candidate.mandatory14.tagx.legal-footer-identity, candidate.mandatory14.tagx.managed-web-data-offer,
candidate.normal18.cvdp.family.component-specific-license-declarations,
candidate.normal18.cvdp.family.context-and-tooling-scope,
candidate.normal18.cvdp.family.creator-reported-coalition-adoption.wording-v2,
candidate.normal18.cvdp.family.original-problem-units, candidate.normal18.cvdp.family.public-release-access-limits,
candidate.normal18.cvdp.family.turing-and-cadence-contribution-roles.wording-v2,
candidate.normal18.cvdp.family.version-and-comparability, assertion.genebench.pro.agent-task.20261112,
assertion.genebench.pro.decision-points.20261112, assertion.genebench.pro.drx1-case-workflow.20261112,
assertion.genebench.pro.external-expert-pool.20261112, assertion.genebench.pro.external-review-scope.20261112,
assertion.genebench.pro.external-review-workspace.20261112, assertion.genebench.pro.ldsc-review-example.20261112,
assertion.genebench.pro.provenance-wording-discrepancy.20261112,
assertion.genebench.pro.public-answer-separation.20261112, assertion.genebench.pro.public-card-integrity.20261112,
assertion.genebench.pro.public-card-inventory.20261112, assertion.genebench.pro.public-card-license.20261112,
assertion.genebench.pro.public-grader-semantics.20261112, assertion.genebench.pro.recoverable-target.20261112,
assertion.genebench.pro.release-selection.20261112, assertion.genebench.pro.release-tiers.20261112,
assertion.genebench.pro.review-revisions.20261112, assertion.genebench.pro.scope.20261112,
assertion.genebench.pro.simulated-origin.20261112, assertion.genebench.pro.validation-loop.20261112,
assertion.genebench.pro.version-transition.20261112

Analytical judgment: verifiable-outcomes-need-side-effect-review

Executable success checks should be paired with independent review of constraints, side effects and verifier error.

Buyer implication: Test valid alternatives, plausible shortcuts, harmful extra actions and infrastructure failures in the acceptance pilot.

Limitations: No environment or provider was independently executed for this report, and no failure-prevalence estimate is claimed.

Confidence: High for the stated benchmark limitations; unmeasured for any new buyer deployment.

Reasoning: SWE-bench passing tests do not guarantee a complete solution; OSWorld identifies damaging-side-effect gaps; τ-bench notes missed policy violations; MBPP documents assertion gaming. PivotRL local checks, HANDBOOK.md prohibited-state criteria and PLE’s hybrid gate show why check coverage and final-task safety require separate inspection. Finite behavioral tests, no-op mutants and reference-isolated grading need independent falsification. Judge/verifier disagreement remains a disagreement measure rather than a human-verified error rate.

What would change this assessment: A matched evaluation showing that the project verifier reliably detects the buyer’s critical constraints could narrow the need for additional manual review in that scope.

Approved evidence IDs: candidate.paperfull.swebench.regression-test-scoring, cohort4.claim.osworld.side-effect-gap,
cohort4.claim.taubench.reward-limit, cohort4.claim.mbpp.test-overfit, mandatory2.claim.arxiv.2603.21383.limitation,
mandatory2.claim.arxiv.2603.21383.method, mandatory2.claim.arxiv.2603.21383.provenance,
mandatory2.claim.arxiv.2603.21383.result, mandatory2.claim.arxiv.2603.21383.result_ood,
repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2607.25398.limitation,

repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2607.25398.method,
repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2607.25398.result,
repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.30537.limitation,
repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.30537.method,
repair.qualbinding.20261111.v1.mandatory3.claim.arxiv.2610.30537.result,
candidate.normal18.programbench.available-evaluation-release,
candidate.normal18.programbench.difficulty-version-conflict.precision-v2,
candidate.normal18.programbench.inference-isolation, candidate.normal18.programbench.metric-meaning-and-limits,
candidate.normal18.programbench.task-unit-and-tests, candidate.normal18.programbench.test-release-and-rights,
candidate.normal18.swe-atlas.aa-variant-lineage, candidate.normal18.swe-atlas.artificial-analysis-contribution,
candidate.normal18.swe-atlas.inductive-ml-contribution, candidate.normal18.swe-atlas.memorization-screen-limits,
candidate.normal18.swe-atlas.qna-evaluation, candidate.normal18.swe-atlas.refactoring-evaluation,
candidate.normal18.swe-atlas.repository-rights-and-upstream-limit, candidate.normal18.swe-atlas.test-writing-evaluation,
candidate.normal18.swe-atlas.v11-protocol-distinction, candidate.normal18.swe-atlas.workflow-units,
candidate.next8.deepswe.paper.corpus-and-repository-selection.v2,
candidate.next8.deepswe.paper.cross-benchmark-spread-limit.v1,
candidate.next8.deepswe.paper.fixed-model-and-harness-conditions.v1,
candidate.next8.deepswe.paper.functional-binary-scope.v1,
candidate.next8.deepswe.paper.functional-verifier-and-regression.v1,
candidate.next8.deepswe.paper.human-review-and-production-roles.v1,
candidate.next8.deepswe.paper.later-v11-isolated-grading.v1,
candidate.next8.deepswe.paper.native-harness-pilot-scope.v2,
candidate.next8.deepswe.paper.original-task-and-contamination-scope.v1,
candidate.next8.deepswe.paper.release-and-captured-license-scope.v1,
candidate.next8.deepswe.paper.reported-headline-result-scope.v2,
candidate.next8.deepswe.paper.run-to-run-uncertainty.v1,
candidate.next8.deepswe.paper.sampling-and-failure-denominators.v1,
candidate.next8.deepswe.paper.structural-long-horizon.v1,
candidate.next8.deepswe.paper.task-macroaverage-and-four-attempts.v1,
candidate.next8.deepswe.paper.verifier-disagreement-and-judge-limit.v1

2. Weekly developments and edition discipline

Method: Edition status at this 12 November authoring receipt: prelaunch, bounded 416-identity eligible core frozen, archive not_published, DOI null and no published edition history. The proposed first edition is research-2026-11-12.1, edition date 12 November 2026, with the actual captured-source cutoff 2026-11-12T17:32:13.289438Z. A date and release identity do not establish publication. The first actual edition is intended as soon as all existing research, source, review, audit, privacy and technical gates pass under the authorized one-time launch calendar. Later comprehensive editions target Monday 09:00 America/Los_Angeles with the preceding Friday 17:00 captured-source cutoff. Daily preparation targets 16:00 Pacific. No capture, review or publication is backdated.

Method: The weekly change unit is a versioned source-supported claim, not an increase in word count or an older paper newly encountered by the research team. Source publication date, described event date, capture timestamp, assertion as-of date, review timestamp and edition publication date remain separate. A capture performed on 9 November in Los Angeles may be dated 10 November UTC. It does not retroactively satisfy a 6 November cutoff. Later independent review can approve the captured content without changing the time at which the document was captured.

Method: This authoring refresh extends the actual e55-reviewed unpublished input of 2,292 approved assertions, 225 reviewed paper works, 73 qualifying supplier groups, 68 dataset families and 47 benchmark families to the final selected input of 2,548 approved assertions, the same 225 reviewed works and 73 supplier groups, 85 dataset families and 87 benchmark families. Seventeen dataset and forty benchmark families add 57 substantive atlas entries; these include the five reasoning-distillation families and 52 further families. The full twelve chapters, nine stable executive finding IDs, ten original pilot identities, supplier dossiers and reviewed-work bibliography are retained. Exact unchanged paragraphs are preserved and every changed or added block is inventoried for new synthesis review. Counts describe approved source coverage and identity dispositions, not contemporaneous news, an automatically improved method or a new published weekly edition.

Evidence: The commercial event dates illustrate why review time must not become news time. Labelbox’s financing announcement is dated 14 October 2026; Proximal’s annualized-revenue announcement is dated 29 September 2026; Mercor’s 25 June investigation account describes a March 2026 LiteLLM incident. Their inclusion in this newly expanded preparation does not make them November events. The financing, annualization and incident qualifications remain part of the respective evidence records.

Sources: [72], [246], [740]

Method: The earlier initial calculation counted FineWeb-Edu as a separate family. Current identity review treats it as a FineWeb subset release, so current coverage counts one FineWeb family. The earlier frozen files retain their original exact bytes. The correction is an explicit change in identity accounting, not a disappearance of the educational subset or a retraction of its documented model-scoring workflow. Future change logs should report family identity corrections separately from newly added source observations.

Research gap: No prior published edition exists in this ledger, so a week-over-week market trend is not estimable from these records. The first publishable comparison will need an actual frozen predecessor, a current source cutoff, a claim-level change list and disclosure of whether each change is a new event, a newly captured older source, a corrected scope or a reclassified identity. Scheduled dates, package preparation and website visibility should be reported separately from archive confirmation.

Method: Correction accounting preserves original assertions and their actual prior positive/negative review history. The current report assesses 141 affected paragraphs with 370 stale binding occurrences and five affected executive findings; these resolve through 84 distinct approved successors. Some statements remain scientifically supported with narrowed qualifications, while others require explicit wording changes: DAYJOB healthcare and finance time bases differ, Snorkel timing is acceptance- and project-conditioned, TrustScale’s rebrand evidence is decoded application-bundle text, and paper-only APEX-Accounting counts are separate from a captured leaderboard conflict. Original APTO, Magpie, OpenAssistant and Dataocean scope restrictions remain. These changes neither relabel prior diagnostics nor complete the formal launch audits.

3. Market definition, structure and provider categories

Method: Operational market definition: an AI data lab or human-feedback supplier is included here when an approved public source documents an offering that creates, collects, annotates, curates, validates or evaluates data or executable tasks used in AI development. The category includes services, participant access, software-assisted workflows and environment products. Inclusion means that the scoped offering meets this research definition. It does not assert a legal industry classification, a minimum revenue level, independent ownership, a named downstream contribution or an exclusive business focus.

Analysis: Participant access is one delivery category. Prolific offers selected participants and distinguishes platform access from managed collection; Terac offers verified-expert access with an expert-payout-plus-platform-fee model. Handshake’s contractor agreement describes a Human Data Services program whose onboarding does not guarantee a project. These are useful starting points when the buyer can own task design and acceptance, but the degree of project management and supplier responsibility must be read in the contract. Access to a qualified population is not the same deliverable as an accepted dataset.

Sources: [221], [329], [712], [720]

Analysis: Managed data production combines personnel, instructions, tools and review. Appen describes expert alignment and evaluation; Sama describes an in-house workforce and human-verified data services; iMerit describes expert review and client audits; DataForce describes expert/native-speaker matching and adjudication. These examples are grouped by disclosed responsibility, not presumed equivalence. A managed-service comparison should identify who writes the rubric, who employs or contracts with contributors, who owns correction work and who can reject a disputed final unit.

Sources: [54], [62], [81], [82], [735]

Analysis: Licensed data products are a different procurement path. Snorkel describes non-exclusive dataset and environment products refreshed quarterly; Mercor, Surge, Datacurve and Invisible advertise prebuilt or ready-to-license collections. Defined.ai explicitly separates marketplace or third-party proprietary datasets from new collection and services. The relevant comparison is therefore version, license, overlap with existing training data, refresh entitlement and fitness on the buyer’s task. A catalog listing does not establish the right to redistribute every source component or the availability of exclusive data.

Sources: [59], [64], [217], [684], [690], [753]

Analysis: Environment builders deliver executable tasks, state and reward logic. Abundant describes turning customer workflows into Foundry environments; Refresh describes simulated coding and computer-use worlds; Bespoke describes company-scale simulations; HackerRank Bench offers resettable repository environments and hidden verifiers; Prime Intellect advertises SFT fine-tuning, hosted RL on managed Prime-RL, inference and compute alongside its environment tooling. A buyer should separate environment authorship from hosted execution and post-training services. Portability, reset determinism, observability, judge calibration and maintenance obligations can matter as much as the number of tasks in a package. The platform description does not establish a human-data workforce or bespoke task-authoring service.

Sources: [206], [251], [291], [305], [691]

Analysis: Evaluation and safety services may use expert data without selling a general annotation workforce. Giskard offers dynamic evaluation data and adaptive attacks; Alinia describes policy judges and comparison with customer experts; Trajectory describes expert-discovered safety attacks and automated testing of known attacks. Their documented value propositions concern discovery, assessment and iteration. A procurement team should specify whether the output is a test dataset, a report, an integrated judge, a remediation cycle or a continuous service before comparing prices or treating these providers as substitutes for custom training-data production.

Sources: [288], [301], [302], [331]

Analysis: Engineering teams and talent programs can occupy a distinct contractual position. A.Team offers builders for environment and evaluation engineering. G2i advertises engineering data programs and dual-developer review, yet its inspected human-data terms limit the service to talent acquisition and disclaim responsibility for candidate work. Proximal’s MSA instead keeps responsibility for subcontracted services with Proximal. These documents demonstrate why a shared engineering capability label cannot establish who warrants the final artifact. The contract should identify the accountable supplier, the accepted output and the path for defects.

Sources: [218], [245], [285], [717]

Analysis: Annotation software is another distinct responsibility model. Dataloop documents human/AI feedback nodes; Datasaur combines automated NLP pre-labeling with human verification; Kili lets customer specialists author and judge post-training examples and introduces external partners; V7 Darwin supports multistage workflows and advertised specialist access. These products can support human-data production without establishing that the software company directly employs the annotators or warrants every add-on’s delivery. Compare platform controls, expert access and managed-service responsibility as separate line items.

Sources: [341], [344], [359], [378], [379]

Analysis: Modality and domain are independent comparison dimensions. Encord offers physical-AI, multimodal and LLM annotation with role matching; APTO lists image, video, LiDAR, text and audio; Shaip describes credentialed professionals for healthcare, legal, finance and code preference work. A modality claim does not establish specialist expertise, and a credential claim does not establish coverage in every language or format. Buyers should intersect the required domain, locale, tool environment and output type, rather than add separate broad claims into an unsupported statement of complete coverage.

Sources: [135], [157], [290]

Method: Provider categories are overlapping analyst groupings of documented offerings, not mutually exclusive market segments. A supplier can appear in several use-case comparisons without being counted several times. A brand, product, division or contractor platform is not automatically a separate corporate supplier. RWS/TrainAI, Welocalize/Welo Data, HackerRank/Bench and Alegion/SanctifAI retain the scoped identity disclosed in their records. Historical aliases remain searchable without inventing a new ownership transaction or a second supplier.

Analysis: Named delivery strengthens the ability to ask focused diligence questions, but evidence type still matters. DataForce's IBM case, SuperAnnotate's Databricks case, Aya Data's Cydar case and Keymakr's SeeChange case are provider-authored descriptions with different artifacts and dates. They are not matched outcome trials. A supplier without a named public case in this graph may have confidential deliveries; a supplier with one has not thereby demonstrated superior quality. Disclosure completeness and provider performance are distinct objects of study.

Sources: [52], [84], [145], [716]

Research gap: The sample cannot support a commercial quadrant, market size, share estimate, concentration index, price curve or fastest-growing-provider list. Comparable segment revenue, current buyer demand, accepted-unit volume, active paid contributor windows and task-matched outcomes are missing or inconsistent. The report instead makes a narrower, falsifiable judgment: supplier comparison should follow responsibility and deliverable structure. A later study could test that judgment using matched contracts, acceptance logs and common pilot tasks.

Analysis: A managed annotation offer and an annotation method are different comparison objects. CrowdAgent describes routing and hidden-gold aggregation; LMCrowd describes staged model-label purchasing. Neither establishes that a named supplier currently implements those workflows or achieves their results. For managed work, compare responsibility for guidelines, routing, escalation, independent acceptance and full cost. For a method implementation, compare the exact data, models, selection rules and held-out checks. A paper-to-practice connection can suggest questions without becoming an unsupported supplier-contribution edge.

Sources: [423], [434]

Analysis: A second responsibility map concerns who creates the measurement, rather than who supplies a worker. WebGPT records contractor contributions to demonstrations and answer comparisons; StarCoder records a distinct PII annotation project; APEX-Agents records expert task/world/rubric construction; the rubric-quality study separates human taxonomy development from generated persona scoring. These source-specific roles are more informative than a blanket human-data label. Confidence is high in the documented distinctions and unestablished for their relative commercial value. A statement of work should identify deliverable authorship, independent review, acceptance authority and continuing maintenance for each stage.

Sources: [468], [476], [561], [565]

Evidence: BenchFlow - reviewed adjacent environment software; commercial qualification pending. BenchFlow documentation instructs users to check solvability, no-op failures, verifier robustness and task flakiness before trusting scores. BenchFlow calls itself a frontier environment lab and says it ships SkillsBench, ClawsBench, PostTrain and their runtime. BenchFlow documentation describes fresh task sandboxes and retained records of attempts, calls, settings and workspaces. BenchFlow describes building stateful environments with services, files, tools, verifiers, traces and replay. The BenchFlow repository LICENSE file at the same pinned commit contains the Apache License 2.0. The BenchFlow README at commit a61a26f4db4d566de16928565ea58abf3dbaf064 says it describes BenchFlow 0.10. The pinned README describes storing run rewards, token usage and trajectories in the job directory. The pinned BenchFlow README lists Docker, Daytona, Modal and AgentCore as task sandbox backends for evaluation and RL training. The pinned README links an export interface for JSONL rows containing rewards and messages for training. BenchFlow docs say their current documented release is 0.9.2 and explicitly flag pages requiring the 0.10.0rc2 prerelease.

Sources: [211], [212], [213], [335], [336]

Analysis: BenchFlow - reviewed adjacent environment software; commercial qualification pending. Research software, documented controls and owned benchmark publication do not establish a managed data-supply contract. Confidence is moderate for the disclosed runtime scope and unestablished for commercial data delivery. An explicit current supply offer and source-backed engagement terms would change that disposition.

Sources: [211], [212], [213], [335], [336]

Evidence: Quesma - adjacent analytics and verification scope. Quesma currently markets coding-agent analytics that attribute AI spending by project, person, task type and model. Quesma’s homepage says it built training environments for difficult multi-hour agent tasks. Quesma says its Blitzy audit reran solutions and checked trajectories for unauthorized network access, branches and leakage. Quesma’s website privacy policy names Quesma Poland Sp. z o.o. as website operator and data controller. Quesma publishes open coding-agent benchmarks covering puzzle solving, binary security analysis, instrumentation and compilation tasks. Quesma reports being contracted by Blitzy to audit and verify SWE-Bench Pro results.

Sources: [247], [248], [249], [681]

Analysis: Quesma - adjacent analytics and verification scope. The reported contracted Blitzy audit is a narrow evaluation role; it does not establish training-data authorship. Current analytics positioning does not prove all historic/private supply work ceased. Confidence is moderate for the disclosed role and low for unobserved business activity. A current task-production offering or independently confirmed delivery would change the supplier assessment.

Sources: [247], [248], [249], [681]

Evidence: AIXI Labs - theoretical research scope. Its official footer identifies AIXI Labs as a fiscally sponsored project of Principles of Intelligence, a US nonprofit. AIXI Labs states that it shares research openly where possible, including failures. AIXI Labs describes its focus as theoretical AI-risk and mitigation research using AIXI variants. AIXI Labs says it develops methods to apply theoretical findings to real AI agents. Its research overview describes creating model variants with specific risks or safety measures.

Sources: [286], [287]

Analysis: AIXI Labs - theoretical research scope. The fiscal-sponsorship statement and open-research goal support a research-organization description. Confidence is moderate for that scope and unestablished for commercial data services. A concrete current collection, annotation or evaluation service offer would change the inclusion interpretation.

Sources: [286], [287]

Evidence: Elorian - model access and illustrative annotation. Elorian’s privacy policy says free inputs/outputs are used for model training by default and API users are opted out by default. The October 16, 2026 Glimpse announcement offers the model through an API and chat interface. The announcement describes Glimpse using tools to write code, transform images and inspect video. Its robotics example shows Glimpse assigning object boxes and event timestamps to teleoperation material. Elorian’s homepage positions the organization around models for visual thinking.

Sources: [298], [299], [300]

Analysis: Elorian - model access and illustrative annotation. API/chat access and a robotics labeling example are model offerings, without a separately contracted data-production service. Confidence is moderate for that distinction and unmeasured for task effectiveness. A source-backed managed delivery offering and task-matched outputs would change the supplier assessment.

Sources: [298], [299], [300]

Evidence: Laude Institute - research support and software scope. Laude’s about page says individual technologists fund the institute. Laude says it operates and supports openly published research and open-source projects. Its website terms identify the operator as Laude Research Institute PBC. Laude Institute describes combining grantmaking with hands-on support for computing researchers. Laude’s terms list code competitions, grant applications and a researcher impact index as website services.

Sources: [307], [308]

Analysis: Laude Institute - research support and software scope. The website operator is a PBC; institute branding and individual funding do not establish nonprofit status. Confidence is moderate for the disclosed research-support role and unestablished for a commercial data-supply offering. A concrete current service and independently confirmed task delivery would change that boundary.

Sources: [307], [308]

4. Technical landscape: from supervision to executable evaluation

Method: This chapter separates stages that are often collapsed into RLHF: source collection; task and rubric authoring; demonstrations for supervised fine-tuning; outcome ratings, preference pairs or ranked sets; reward or judge construction; policy optimization; executable environment design; evaluation; and documentation. The reviewed-work atlas below summarizes approved observations from exact inspected manifestations. A paper’s work identity counts once even when several versions exist. Bounded full_source review covers methods plus main results or worked examples and limitations; it does not claim all appendices, proofs, raw data or experiments were reproduced.

Analysis: Classical human-preference pipelines expose several independent design choices. The original trajectory-comparison study alternates interaction, human clip comparisons and reward fitting. The summarization study uses comparison-trained rewards and PPO with a KL penalty. InstructGPT ranks sets of responses and uses a separate unlabelled prompt pool for PPO. These are not merely different names for pairwise annotation: they differ in task construction, candidate sampling, comparison independence and the distribution exposed to optimization. A data purchase should preserve those distinctions in its manifest.

Sources: [60], [158], [731]

Analysis: Preference optimization changes the training interface, rather than erasing data requirements. DPO fits a direct objective without a separate reward-model/RL loop; IPO studies a direct squared-loss formulation under a narrower experimental setting; ORPO combines chosen-response likelihood and odds-ratio loss without a frozen reference; KTO accepts desirable/undesirable labels. KTO experiments often transform existing pairs, so support for a binary interface is not evidence that all reported data were independently collected as binary judgments. Method comparison must retain label origin, model scale, candidate distribution and evaluation protocol.

Sources: [68], [69], [76], [732]

Analysis: Selection methods consume a candidate budget that is absent from a final retained-pair count. RAFT retains a high-reward fraction; statistical rejection sampling generates 64 candidates and subsamples eight in its standard experiment; SimPO’s instruct setting selects highest/lowest PairRM scores from five candidates; SPPO also uses five responses per prompt in its experiment. Their results cannot be explained by final training-row counts alone. A reproducible data and compute comparison should count generated candidates, judge calls, retained examples, rejected items, training updates and evaluation calls separately.

Sources: [315], [320], [325], [327]

Analysis: Process supervision targets intermediate steps, but step granularity and validation still differ. Let’s Verify Step by Step evaluates selection from a fixed generator and flags a large-scale comparison confound; Step-DPO contrasts next steps conditioned on earlier correct steps and warns that correct final answers can conceal bad intermediate reasoning; PRIME learns an online implicit process reward from outcome supervision. A buyer should distinguish human step labels from model-localized errors and from token rewards derived indirectly from final outcomes. None of these method labels alone proves faithful reasoning.

Sources: [78], [328], [413]

Analysis: Optimization papers also change normalization, sampling and infrastructure assumptions. DAPO filters all-correct or all-incorrect sampled groups and changes clipping and token aggregation; Dr. GRPO analyzes response-length and reward-standard-deviation biases; GSPO clips sequence ratios; VAPO adds value pretraining and length-adaptive advantages. These are specific recipes under their inspected experiments, not evidence of a universally best algorithm or a provider’s operational advantage. A buyer of training services should ask which recipe, reward definition, sampling budget and failure controls are actually implemented.

Sources: [402], [405], [407], [420]

Analysis: Test-time adaptation needs a different evaluation contract. TTRL uses agreement with a sampled consensus as reward without target ground-truth labels; ReliableTTRL distinguishes population-majority error from finite-rollout variation; SelectTTRL fixes a selection rule without target answers but uses a separately labeled calibration set. Reporting only a final score hides adaptation data access and evaluation-budget differences. A fair comparison should state which test questions drive updates, which labels enter calibration, the total rollout budget, checkpoint-selection rule and whether every target question remains in the denominator.

Sources: [416], [417], [419]

Analysis: Synthetic and hybrid pipelines require stage-level lineage. Constitutional AI mixes human helpfulness with model harmlessness labels; ToolBench combines expert seed examples with ChatGPT instruction and trajectory generation; RLAIF-V uses model claim checking while its evaluation references include human corrections; Llama 3 collects new annotations and synthetic data across rounds. The practical requirement is a manifest identifying generator, prompt, seed material, selection model, correction roles and independent evaluation. A statement that humans were involved somewhere cannot establish that a particular training target is human-authored.

Sources: [51], [155], [255], [310]

Analysis: Executable evaluation spans different abstractions. HumanEval checks short Python functions; SWE-bench checks issue-resolution patches; WebArena combines semantic and state checks; OSWorld uses desktop state evaluators; τ-bench checks database outcomes under interaction policies. Agents can receive different observation modes, budgets and simulator affordances. A single leaderboard column cannot remove these differences. The comparison should first ask whether the measured task and environment resemble the buyer’s deployment, then assess whether the scoring protocol detects the failures the buyer cares about.

Sources: [91], [224], [233], [254], [733]

Analysis: Research and professional-work benchmarks add substantial human design and review labor. RE-Bench records expert runs with tools; ScienceAgentBench uses subject-matter validation and separate execution verification; PaperBench aggregates binary rubric leaves and prohibits use of original authors’ code; TheAgentCompany records task-creation effort but lacks a professional human comparison in that iteration. These benchmarks make particular research activities observable, but their human baselines, time budgets and allowed resources differ. They cannot establish a general percentage of jobs that an agent can perform.

Sources: [334], [365], [366], [368]

Analysis: Reward hacking and contamination should be studied as separate mechanisms. A learned reward can be overoptimized while human preference worsens; a program can pass supplied assertions by hard-coding tested inputs; a reasoning monitor can lose recall under direct optimization pressure; a public benchmark can be detected in training only under particular access and watermark assumptions. These examples motivate separate checks for objective misspecification, verifier gaming, monitor adaptation and training/test exposure. They do not establish the incidence of those failures in any provider's current production service.

Sources: [158], [227], [393], [418]

Analysis: Feedback-format comparison - ordinal ratings, rankings and binary signals preserve different information. HelpSteer2's attribute ratings can retain multiple dimensions; InstructGPT's ranked sets relate several responses under one prompt; Llama 2 adds preference strength; KTO can use one desirable/undesirable output. Converting all of them to binary pairs can discard attribute structure or strength and create correlated observations. The buyer should decide which information is required by the intended objective before paying for annotation. The reviewed studies do not establish one universally most efficient format, and the matched disagreement study specifically cautions against assuming paired preferences are inherently less ambiguous.

Sources: [69], [133], [309], [451], [731]

Analysis: Feedback-format comparison - rationale text and process labels answer different questions. The natural-language-feedback study uses author-written feedback to select refinements by embedding similarity; Innodata describes explanatory reasons for response scores; Step-DPO localizes the next erroneous step; DeepSeek-R1-Zero uses final accuracy and format rewards without a neural process reward. A rationale can be useful context without being a validated causal explanation or a correct intermediate proof. Procurement should distinguish the requested text artifact, its validation target and whether downstream training actually consumes it. These examples do not show that longer reasoning or more explanation is always better.

Sources: [142], [295], [306], [328]

Analysis: Distribution comparison - offline preferences and current-policy rollouts can differ even when the file schema is identical. SLiC-HF notes mismatch between off-policy human-feedback data and SFT decodes; WPO reweights preference losses using current-policy response weights; REBEL permits a fixed or current-policy response distribution but samples both responses from the current policy in its described experiments. These methods make the candidate-generation distribution a material part of the recipe. A licensed preference dataset should therefore identify generating models, sampling settings and collection period, and a training-service proposal should explain whether fresh responses or labels are required.

Sources: [316], [326], [333]

Analysis: Objective comparison - a scalar reward is a modeling choice, not a property guaranteed by human judgment. SPPO explicitly uses a two-player constant-sum preference game rather than requiring one transitive scalar ordering; WPO's conceptual labeling construction assumes no conflicting pairs for a prompt; the instruction-tuning disagreement experiment observes task-dependent ambiguity. The analytical lesson is to state the aggregation assumptions and preserve counterexamples before fitting a reward. This does not prove that scalar models are inappropriate in every task; it identifies a condition to test when disagreements are central to the buyer's use case.

Sources: [133], [327], [333]

Analysis: Budget comparison - optimization improvements can be inseparable from sample selection and validation conditions. Llama 3 rejection sampling generates 10-30 outputs per human-annotation prompt; DeepSeekMath samples 64 outputs per question in its described run; TTRL uses 64 voting responses and a 32-response training subsample. These are different compute and information budgets. A comparison based on retained examples, final model size or optimizer name alone is incomplete. Report total generations, token caps, judge calls and human checking, then distinguish training expenditure from evaluation expenditure. No cross-paper cost-normalized winner is computed here.

Sources: [294], [310], [419]

Analysis: Systems comparison - orchestration describes how computation moves, not whether feedback is valid. OpenRLHF uses separate rollout/training roles with vLLM and DeepSpeed and limits the inspected paper's support to language models. HybridFlow routes data between parallel layouts and generally executes colocated roles sequentially to avoid contention or memory failures. A platform's architectural support is therefore separate from a proof of label provenance, a task's reward soundness or the performance of a trained model. For a systems purchase, ask for reproducible resource and data-flow measurements under the actual recipe while auditing feedback quality through a different protocol.

Sources: [408], [412]

Analysis: Curation comparison - dataset selection can change task composition as well as nominal quality. AlpaGasus retains a model-scored subset of Alpaca and flags disproportionate removal of keyword-defined coding examples; DCLM's classifier retains a top-scoring fraction under a defined positive/negative training set; DataComp permits filtered size to vary under fixed samples-seen compute. The defensible interpretation is that filters select a distribution according to their targets. A buyer should inspect what was removed and the resulting task mix, rather than infer that every discarded example was objectively bad or that a smaller selected set is universally better.

Sources: [121], [130], [131]

Analysis: Evaluation comparison - win rate depends on who judges, what counts as a win and how ties are handled. The natural-language-feedback experiment counts ties as half a win; LIMA reports separate wins, ties and losses against dated responses; Llama 3 excludes slightly-better ratings from wins; RSO uses a three-rater majority. These figures cannot be pooled into a common preference advantage without the underlying protocol. A pilot should freeze the response population, judge rubric, blinding, tie rule and aggregation before comparing candidates, and report enough raw counts to reconstruct the stated metric.

Sources: [149], [306], [310], [320]

Analysis: Human-baseline comparison - the reference population and tools define the comparison. MMLU's approximate expert level is an estimate rather than a measured panel across all subjects; RE-Bench allows human internet and LLM use; VisualWebArena avoids assigning task creators to their own tasks; WebShop gives humans HTML observations while models use simplified text. These are substantive design choices. A claim that an agent exceeds humans should identify which humans, prior task exposure, allowed tools, time limits and observation mode. The reviewed record does not support a general human-versus-AI capability ranking.

Sources: [256], [257], [366], [714]

Analysis: Contamination comparison - tests detect particular kinds of overlap. MBPP's normalized line-overlap analysis concerns eligible reference-program lines; BIG-bench's canary is an intended training-exclusion signal; DeepSeekMath removes benchmark n-grams under its stated matching rules; STAMP requires pre-release watermarking and token-probability access. None of these procedures proves all forms of semantic or strategy-level independence. Buyers using public tasks should record exposure assumptions and run private or freshly authored holdouts where appropriate, while acknowledging that private construction also needs verification and rights review.

Sources: [214], [227], [294], [418]

Analysis: Generalization comparison - different kinds of difficulty should remain separate. GPQA explicitly disclaims representativeness of scientific practice; HLE states that a high score alone would not demonstrate autonomous research; ScienceAgentBench selects Python partly because of annotator familiarity; RE-Bench warns that short clear tasks with supplied resources are unlike open-ended research. A buyer should use these benchmarks to locate specific weaknesses and design target-workflow tests. They do not justify extrapolating an exam score or a short-horizon optimization result into complete professional capability, employment substitution or general intelligence.

Sources: [351], [354], [366], [368]

Analysis: The additional annotation literature changes how the earlier technical methods should be read. Snow's simplified temporal task and Gilardi's agreed-human subset show that the reference population can change a model-versus-human comparison. HelpSteer2's filtering changes the retained label population, while the label-error study shows that even corrected benchmark labels remain uncertain. A method comparison therefore needs three explicit objects: the items admitted, the rule defining reference truth and the items excluded before the headline result. This report treats those as experimental scope, rather than evidence of a universal labor or model advantage.

Sources: [428], [430], [433], [437]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Consent in Crisis: The Rapid Decline of the AI Data Commons. The audited corpus sources are C4, RefinedWeb and Dolma. The head sample unions the top 2,000 domains by token count from each corpus into 3,950 domains. The random sample draws 10,000 domains from the 10,136,147-domain intersection and uses a random 2,000-domain subset for human annotation. The study reports inconsistencies between robots.txt signals and Terms of Service statements. Reading boundary: The study describes their differing crawl windows and filtering; paper version 2407.14933v1 is historical, not a current web-policy census.

Sources: [383]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. The toxicity-labeling case study is explicitly hypothetical and presents its documentation responses as non-prescriptive and non-comprehensive. The framework recommends a dataset maintenance plan and consideration of when annotations can become outdated as language or standards change. CrowdWorkSheets organize documentation questions around stages of dataset construction, from task formulation to dissemination. The framework asks developers to connect annotator-pool selection to intended dataset uses and potentially impacted communities. Reading boundary: Its pay, risk disclosures and platform capabilities are illustrative assumptions, not observed operational facts or supplier relationships.

Sources: [401]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. Data Cards document critical dataset information across its life cycle and allow tailored templates rather than one fixed schema for every dataset. The design principles organize cards into self-contained modular sections and support multiple granularities and content types. The authors develop a participatory workshop approach that engages cross-functional stakeholders when creating transparent dataset metadata schemas. In the reported cases, Google Docs card templates supported collaboration but field-level iteration caused fragmentation and limited automation of collection. Reading boundary: Framework design; flexibility does not establish that a particular card is complete or accurate.

Sources: [403]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. The proposed Annotator Demographic field includes relevant training of annotators and guideline developers. The proposed Language Variety field combines a BCP-47 tag with a prose description. For datasets assembled from existing datasets, the schema proposes an appendix containing the source datasets' data statements. The proposed short-form data statement is a 60-100 word summary pointing to the long form. Reading boundary: Demographic and educational attributes are proposed documentation items; privacy and availability constrain collection/disclosure.

Sources: [386]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Datasheets for Datasets. The authors propose documenting a dataset's motivation, composition, collection process and recommended uses. The questions and workflow were refined over roughly two years with multiple rounds of feedback. The paper states that datasheets alone are not a complete mitigation for unwanted societal bias or potential dataset harms. The authors intend datasheet creation to involve reflective human documentation rather than an automated creation process. For infrequently changing datasets, the authors recommend accompanying updated dataset versions with updated datasheets. The proposed collection-process documentation asks dataset creators to identify collection participants and their compensation. Reading boundary: Subsequent paragraphs describe example datasheets, product teams, researcher/practitioner/policy comments and legal feedback; this is author-reported development, not a controlled effectiveness trial.

Sources: [721], [730]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. The contamination analysis uses exact matches normalized for capitalization and punctuation and treats multi-component benchmark inputs separately. The authors leave quantification or removal of likely personally identifiable and copyrighted text in C4 for future work. C4 is derived from the April2019 Common Crawl snapshot with English identification and content filters, including a blocklist. The audit recommends documenting web metadata, the text included in a corpus, and the text excluded by filtering. Reading boundary: Exact input overlap does not alone establish that a benchmark label was present or that a trained model memorized it.

Sources: [404]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Extracting Training Data from Large Language Models. The evaluation generates three pools of 200,000 samples, each sample 256 tokens long. The study identifies 604 unique memorized examples among its 1,800 selected candidates, reported as a 33.5% aggregate true-positive rate. For each of 18 generation-strategy/metric configurations, the study selects 100 samples from the metric's top 1,000. The paper distinguishes training-data extraction of verbatim examples from inversion that reconstructs representative examples. Reading boundary: GPT-2 XL (1.5B parameters); strategies are top-n empty-sequence sampling, temperature sampling and Internet-conditioned sampling. Additional model identity support: physical p3 lines1-9.

Sources: [387]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: How do data owners say no? A case study of data consent mechanisms in web-scraped vision-language AI training datasets. Copyright-notice detection applies regular expressions to captions and PaddleOCR-extracted image text. The study operationalizes a data owner as the actor embedding the image on a webpage. Two coders annotate domain category, license type and scraping policy, resolving disagreements through annotation or codebook changes. The study audits CommonPool small and medium scales, restricting to English captions using fastText. Reading boundary: Signal-detection method; notices and their absence are not a legal authorization determination.

Sources: [384]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Improving Task Instructions for Data Annotators: How Clear Rules and Higher Pay Increase Performance in Data Annotation in the AI Economy. The experiment uses a 2×3 between-subjects design crossing instruction type with presence of an incentive. Besides image-label accuracy, the study measures perceived instruction helpfulness and perceived task difficulty separately. The study recruits 307 participants who completed the experiment through Prolific Academic. The payment design compares £4.50 base pay with a condition offering an additional single £3 incentive. Reading boundary: Instructions: rules, incomplete rules, standards. Task: classify 100 building-entrance images as barrier-free or not for a wheelchair user.

Sources: [389]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Minority Reports: Balancing Cost and Quality in Ground Truth Data Annotation. The pruning rule retains assignments involving unseen workers or tasks; otherwise it prunes when estimated minority-report probability exceeds a fixed threshold. Simulation quality is measured by comparing post-pruning majority votes against the counterfactual unpruned majority, using yes/no accuracy and F1. The paper defines a minority report as an annotator response that differs from the final majority-vote consensus for an image-crop question. The discussion cautions that RLHF ranked preferences can have high variability without a clear majority or a single notion of ground truth. Reading boundary: Prediction uses previously observed assignments and is recalibrated over time; a pruning decision is not a general worker competence judgment.

Sources: [410]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Model Cards for Model Reporting. The proposed quantitative analysis reports results by individual factors and their intersections. The proposed Intended Use section records primary uses, primary users and out-of-scope uses. The proposed Metrics section includes performance measures, decision thresholds and variation approaches. Where detailed training-data disclosure is infeasible, the paper advocates basic group-distribution information. Reading boundary: Section 4.7 also recommends confidence intervals where possible and metric variation; no universal fairness guarantee follows.

Sources: [392]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: Sources of Disagreement in Data for LLM Instruction Tuning. On the full human dataset, quadratic-weighted kappa is 0.48 for SINGLE ratings and 0.43 for PREFERENCE combined scores; the authors find no evidence in this setup that preference annotation is easier or less ambiguous. Across the two human task formulations, the study reports instance-level disagreement correlation R=0.2 and ambiguity-set overlap 0.25, despite using the same annotators and data points. The zero-shot annotation experiment uses four instruction-tuned models from 1B to 8B; the 1B model has no reported human-label correlation, while the other three have correlation in 0.2-0.25. The human experiment selects720 Anthropic context/two-response instances, evenly split between helpful and harmless, and recruits 33 graduate-course annotators to perform both individual 1-5 quality ratings and paired preference/magnitude tasks. Reading boundary: This is a result for the studied population, generic quality instructions and sample; it does not establish all disagreement as noise.

Sources: [133]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. The initial audit covers 44 instruction/alignment finetuning collections comprising 1,858 individual datasets selected by experts for widespread use; overlapping datasets are intentionally not deduplicated to preserve collection design choices. The license protocol searches for creator-attributable licenses explicitly applying to data, distinguishes them from code licenses, and records permitted-use, attribution and share-alike categories with the strictest feature across multiple licenses. The provenance audit combines linked public metadata, human legal-expert annotation and GPT-4-assisted annotation; its topic annotation uses GPT-4 on a sample of 100 examples for up to 10 topics. The authors explicitly characterize their collection as best-effort categorization of self-reported licenses for research/transparency, not legal advice or a determination that the listed licenses are enforceable.

Sources: [153]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: The Data-Production Dispositif. Both authors analyze the instruction documents through structural, detailed-fragment and synoptic discourse-analysis stages. The investigation combines 55 interviews, participant observation and analysis of 210 instruction documents collected during 2019-2021 fieldwork. The Venezuela worker recruitment starts with convenience sampling and expands through snowball sampling. Fieldwork covers three crowdsourcing platforms operating in Venezuela and a data-work BPO in Buenos Aires. Reading boundary: Iterative interpretive analysis is contrasted with interview/observation material; not a controlled experiment or automated annotation-quality measurement.

Sources: [385]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. The paper warns that the validity of a ground-truth comparison module depends on its reference dataset and can be harmed by a flawed comparator. The label uses stand-alone modules that can be combined to fit dataset type, information availability and specialist expertise. The Dollars-for-Docs demonstration combines metadata, provenance and variable information with statistical and interactive comparison modules. The initial Nutrition Label prototype restricts exploration to tabular datasets with fewer than ten thousand rows. Reading boundary: Labels for comparator datasets and community feedback are suggested mitigations, not proof that bias has been eliminated.

Sources: [411]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: The Problems with Proxies: Making Data Work Visible through Requester Practices. The authors distinguish proxies for worker identity/suitability from proxies for ability to produce desired data, observing combined filtering in requester accounts. The study interviews 52 requesters from academic and industry settings between November 2022 and January 2023. The paper limits its findings to US requester perspectives and warns that reported experiences may differ from behavior in practice. Interviews include a task walkthrough and questions on payment calibration, worker perceptions and qualification metrics, followed by demographic and follow-up-consent forms. Reading boundary: Qualitative thematic finding in this US requester sample; no prevalence estimate or global ranking of platforms is asserted.

Sources: [414]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: The digital labour of artificial intelligence in Latin America: a comparison of Argentina, Brazil, and Venezuela. Recorded worker interviews are transcribed, accompanied by debriefing reports and coded using both descriptive categories and literature-informed theoretical categories. Spanish-language questionnaires and interviews are fielded as paid platform tasks, with later follow-up and cross-platform re-fielding to examine pandemic-specific conditions. The authors state that statistical representativeness is difficult to apply because the boundaries and characteristics of platform populations are unknown. The article combines a Spanish-language platform-worker study with a Portuguese-language Brazilian study, including questionnaires, follow-up interviews and online ethnography. Reading boundary: Topics include task types, trajectories, informality and social protections; this methodology does not itself verify reported pay or any named supplier’s contribution.

Sources: [409]

Evidence: Reviewed-work atlas - Documentation, labor and evidence quality: “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. COVID-era video/phone interviews prevent workflow shadowing and contextual inquiry; the findings are self-reported practices in high-stakes projects, with sampling and gender-distribution limitations. The study groups cascades into physical-world brittleness, inadequate application-domain expertise, conflicting reward systems and poor cross-organizational documentation. Between May and July 2020, the authors conduct semi-structured interviews with 53 high-stakes AI practitioners: 23 in India, 16 in the US and 14 in East/West African countries, using purposive and snowball recruitment. The authors report that 92% of their 53 interview participants experienced at least one data cascade, defined as compounding data-related events producing negative downstream effects and technical debt. Reading boundary: Percentages are coded self-reports in the selected interview sample and are not population prevalence.

Sources: [127]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: A General Theoretical Paradigm to Understand Learning from Human Preferences. IPO is an identity-mapping special case of ΨPO that learns directly from preferences without a reward-modeling stage or the Bradley-Terry pairwise-to-pointwise assumption. The sampled-preference comparisons use three-action bandits, softmax parameterizations, 18,000 Adam steps and ten seeds; reported curves include means and 95% confidence intervals. The paper leaves scaling IPO experiments to more complex settings, including training language models on human preference data, to future work. IPO optimizes an empirical squared loss on the policy/reference log-likelihood ratio of preferred versus dispreferred actions, targeting one-half the inverse regularization coefficient. Reading boundary: These are illustrative bandit experiments, not LLM benchmark results.

Sources: [68]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. The study does not measure correlation between its simulated preference judgments and final human evaluation preferences. RLOO uses the other k−1 sampled-completion rewards as the baseline for each completion in a prompt's group. Across the reported datasets/models, online rollout uses the same prompts as the SFT stage. For Pythia 6.9B TLDR, Table 1 reports GPT-4-simulated win rates 77.9 for RLOO(k=4) and 67.6 for PPO against original completions. REINFORCE/RLOO treat the entire completion as one action in the paper's deterministic bandit formulation. Reading boundary: Values are percentages. Each method uses its checkpoint with highest test reward; test-reward selection is part of this historical comparison.

Sources: [318]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: DAPO: An Open-Source LLM Reinforcement Learning System at Scale. DAPO separates lower and upper clipping ranges, increasing the upper range while retaining the lower range to encourage low-probability-token exploration. Dynamic sampling oversamples and excludes prompts whose sampled response group has all-correct or all-incorrect answers, continuing until the effective batch is filled. Soft overlong punishment adds a length-dependent penalty to correctness rewards within a configured interval near the generation limit. DAPO aggregates the policy loss over tokens across the sampled batch, replacing equal weighting of response-level average losses. Reading boundary: Mechanism described in the selected version; the candidate makes no universal entropy or quality guarantee.

Sources: [402]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Deep Reinforcement Learning from Human Preferences. The method asynchronously alternates environment interaction/policy updates, eliciting human trajectory-pair comparisons, and supervised fitting of a learned reward predictor. Human raters compare 1-2 second trajectory clips and may prefer either clip, mark them equal, or mark them incomparable; incomparable judgments are omitted from the comparison database. The evaluation separates real human comparisons, synthetic oracle comparisons derived from environment reward, and RL with the real reward. Real-human MuJoCo and Atari curves each represent a single run. In the offline-only reward-predictor ablation, optimizing a predictor fitted only at the start can exploit incomplete reward estimates, including Pong policies that avoid losses without scoring. Reading boundary: Curve averaging differs by condition: five runs for other MuJoCo curves and three for other Atari curves.

Sources: [60]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Table 6 reports AIME 2024 pass@1 of 72.6 for DeepSeek-R1-Distill-Qwen-32B versus 47.0 for DeepSeek-R1-Zero-Qwen-32B. The six described Qwen/Llama distilled models are fine-tuned with the approximately 800,000 curated samples using SFT, without a distillation-stage RL step. R1 cold-start collection includes model-generated or R1-Zero responses refined through post-processing by human annotators. The rejection-sampling stage retains correct responses and collects approximately 600,000 reasoning training samples. DeepSeek-R1-Zero uses accuracy and format rewards and does not use an outcome or process neural reward model in the described training. Reading boundary: Scores are percentages estimated with nonzero-temperature sampling, not cons@64. The RL-only comparison is a specific Qwen 32B training run exceeding 10,000 steps (p14).

Sources: [295]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: DeepSeek-V3 Technical Report. In the DeepSeek-V2.5 distillation ablation, MATH-500 pass@1 rises from 74.6 to 83.2 while average response length rises from 769 to 1,510. DeepSeek-V3 uses GRPO, estimating the baseline from group scores instead of a separate critic model. For non-reasoning SFT data, DeepSeek-V2.5 generates responses and human annotators verify accuracy and correctness. DeepSeek-V3 instruction tuning uses 1.5 million curated instances with domain-specific creation procedures. Table 6 reports DeepSeek-V3 MATH-500 exact match of 90.2 under greedy decoding and an 8,192-token output cap. Reading boundary: This is the report’s V2.5 ablation, not a direct causal estimate for the final V3 system. Length units are not named in the cited table.

Sources: [296]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. Math-corpus decontamination removes segments matching benchmark 10-grams; shorter benchmark texts with at least three grams use exact matching. Outcome-supervised GRPO normalizes sampled response rewards by their group mean and standard deviation and assigns that normalized reward to all response tokens. DeepSeekMath-RL 7B reports chain-of-thought accuracy of 88.2% on GSM8K and 51.7% on MATH. The reported DeepSeekMath RL run samples 64 outputs per question with maximum length 1,024 and one policy update after each exploration stage. The authors identify geometry and theorem proving as relatively weak areas and suggest pretraining/fine-tuning data-selection bias as a possible contributor. The mathematical SFT collection contains 776,000 English/Chinese examples with chain-of-thought, program-of-thought or tool-integrated solutions. Reading boundary: These are benchmark-specific no-tool CoT results, not tool-integrated accuracy or pass@multiple-samples.

Sources: [294]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Direct Preference Optimization: Your Language Model is Secretly a Reward Model. The authors derive a classification objective for learning from preferences. DPO fits an implicit reward through a direct policy objective without fitting a separate reward model and running an RL training loop. The summarization experiment reuses human preference annotations collected by Stiennon et al. The controlled IMDb sentiment experiment generates preference pairs using a pretrained sentiment classifier. The paper's empirical evaluation covers models up to 6 billion parameters. The GPT-4 summary-judgment validation study recruits 25 volunteer human raters. Reading boundary: The equivalence uses the paper's KL-constrained objective and preference-model assumptions. Preference-labelled pairs are still required.

Sources: [725], [732]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Evaluating Large Language Models Trained on Code. The paper finds overlapping BLEU-score distributions for unit-test-correct and incorrect HumanEval completions and cautions that higher BLEU need not imply better functional correctness. The study generates 200 samples per HumanEval task and estimates pass@k for k at most 100 using the count of samples that pass the unit tests. The authors report that the studied code model struggles with increasingly long, higher-level or system-level specifications, limiting interpretation of success on the short HumanEval tasks. HumanEval contains 164 hand-written Python programming problems, each with a function signature, docstring, body and unit tests; the paper reports a mean of 7.7 tests per problem. Reading boundary: The estimator and its proof were read descriptively; no proof verification or experiment reproduction is claimed.

Sources: [224]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. Model-based screening does not autonomously alter tasks; human experts retain responsibility for accuracy and completeness. The experimental automated grader agrees with expert graders 66% of the time, compared with 71% human inter-rater agreement. The authors report an average of five human reviews per task and a minimum of three across all 1,320 tasks. The authors identify precisely specified, one-shot tasks as a limitation relative to contextual, interactive work. Reading boundary: Reported gold-subset pairwise-grading agreement, not task-completion accuracy.

Sources: [350]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: GPT-4 Technical Report. The report states that GPT-4 uses public and third-party licensed pretraining data but withholds dataset construction, model size and training-compute details. Formal-essay exam responses were graded by one or two qualified third-party contractors with relevant essay-grading experience. The authors report GPT-4 accuracy 19 percentage points above their latest GPT-3.5 on nine internal adversarial factuality evaluations. In the system-card pipeline, human demonstrations train SFT, human rankings train a reward model, and PPO optimizes the SFT policy using that reward. Rule-based reward models are zero-shot GPT-4 classifiers conditioned on human-written rubrics and supply an additional PPO reward on selected prompts. Reading boundary: Accuracy is agreement with human ideal responses; per-category sample denominators are not specified in the cited passage. This historical internal evaluation is not a general provider-performance estimate.

Sources: [304]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Group Sequence Policy Optimization. The authors propose using inference-engine sequence likelihoods directly as a possible way to avoid old-policy likelihood recomputation in the training engine. The GSPO comparison begins with a cold-start model fine-tuned from Qwen3-30B-A3B-Base and evaluates AIME24, LiveCodeBench from October 2024 to February 2025, and CodeForces. GSPO clips entire responses, with length normalization intended to control ratio variance and numerical range across response lengths. GSPO defines its importance ratio from sequence likelihood, using a length-normalized likelihood ratio rather than individual token ratios. Reading boundary: Presented as an engineering possibility based on tolerance of precision discrepancies, not a demonstrated universal guarantee for every engine or workflow.

Sources: [407]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages. The approved record attributes the specific documents relationship to HelpSteer3.

Sources: [728]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: If in a Crowdsourced Data Annotation Pipeline, a GPT-4. Against Bio Expert reference labels, combining GPT4 at temperature 0.2 with advanced-interface crowd labels surpasses GPT4 alone under the One-Coin Dawid-Skene and MACE settings. The advanced crowd interface includes visual click feedback, a color-coded overview and a time lock, while the basic interface emphasizes task simplicity. The study uses CODA19 to label abstract sentence segments as Background, Purpose, Method, Finding/Contribution or Other. The test set samples 200 papers from an English-language CORD19 pool restricted to 2022 or later publications, producing 3,177 abstract segments. Reading boundary: Only those two tested aggregation settings beat GPT4 in this study; outcomes depend on cleaning/aggregation and task, and worker LLM use cannot be excluded.

Sources: [406]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Instruction-Following Evaluation for Large Language Models. Table 3 reports GPT-4 prompt-level strict accuracy of 76.89%. The authors warn that loose response transformations can introduce false positives. Humans manually check and edit every rephrased prompt after model-assisted generation, screening and rephrasing. Prompt-level strict accuracy requires every verifiable instruction in a prompt to be followed. Reading boundary: November 2023 API responses; approximately 500 prompts in the inspected v1. This is constraint compliance rather than holistic answer quality.

Sources: [355]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Is It Thinking or Cheating? Detecting Implicit Reward Hacking by Measuring Reasoning Effort. TRACE truncates a response’s CoT,forces an answer and computes the area under verifier pass rate versus retained-CoT percentage. A hacking response is defined as passing the verifier with the loophole present but failing when it is removed. The authors state that TRACE is designed for reasoning tasks requiring inference-time exploration. The study trains hacking and nonhacking instruction-tuned models with RLOO. Reading boundary: Instance-level early-answering evaluation generalizes same-answer faithfulness checks to verifier passing; high area is interpreted as a possible low-effort shortcut,not definitive intent.

Sources: [398]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: KTO: Model Alignment as Prospect Theoretic Optimization. KTO directly uses a binary signal indicating whether a given output is desirable or undesirable for its input, instead of requiring a preferred/dispreferred pair. The paper's PPO comparator is an explicitly offline variant with a fixed reference/old policy and dummy +1/-1 rewards, differing from standard online reward-model-based PPO. For most experiments, the authors convert existing HH, SHP and OpenAssistant preference pairs into desirable and undesirable examples; separate experiments subsample one output per input. KTO can ignore examples that are difficult to learn, which may help with noisy feedback but can also discard necessary information and produce underfitting. Reading boundary: The data requirement is binary; these principal experiments nevertheless reuse preference datasets.

Sources: [69]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Learning from Natural Language Feedback. Two authors wrote feedback on initial summaries while seeing the title, post and initial summary. One author conducted the human evaluation for the fine-tuned-model comparisons in Figures 2 and 5. Fine-tuning GPT-3 on feedback-selected refinements reports 51±5.0% win rate against Reddit-user reference summaries. The method chooses among 20 generated refinements by maximum embedding cosine similarity to the natural-language feedback. The summarization test uses 100 unseen Reddit posts, ranks summaries with ties allowed and counts a tie as half a win. Reading boundary: The± value is standard error;100 test posts and the one-author evaluation apply. This does not establish parity with professional summarizers or humans generally.

Sources: [306]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Learning to summarize from human feedback. The authors report about 320 GPU-days for 6.7B RL fine-tuning and thousands of labeler hours, and state that they did not collect an equivalent amount of high-quality supervised human demonstrations as a baseline. After controlling for summary length, the paper reports that its 6.7B human-feedback model summaries remain preferred to the TL;DR reference summaries about 65% of the time. An experiment using an earlier reward-model version finds that more PPO optimization initially improves human preferences but eventually worsens them even as the reward-model prediction rises. The paper fine-tunes a TL;DR summarization baseline, collects human comparisons of candidate summaries, trains a scalar reward model from those comparisons and optimizes a policy with PPO plus a KL penalty. The study explicitly states that its summary-comparison labelers are hired from Upwork and the labeling services Scale and Lionbridge, with paid onboarding and researcher calibration. Reading boundary: The outcome is this experiment’s human preference against dataset reference summaries, not objective universal summary accuracy.

Sources: [158]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Let's Verify Step by Step. The study evaluates reward models by best-of-N selection from a fixed generator and deliberately does not improve that generator with reinforcement learning. The evaluation uses a uniformly selected 500-problem subset of MATH's test split; 4,500 other test-split problems are included in reward-model training. Large-scale PRM and ORM training sets are not directly comparable: active PRM selection biases toward wrong answers and uses an order of magnitude fewer examples; small-scale ablations use model-generated supervision on matched data. Reading boundary: Do not describe these results as evaluation on all 5,000 MATH test problems.

Sources: [78]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Llama 2: Open Foundation and Fine-Tuned Chat Models. The authors state that their human evaluation prompt set contains no coding- or reasoning-related prompts. On the paper's helpfulness prompt set, Llama 2-Chat 70B has a reported 36% win rate and 31.5% tie rate relative to ChatGPT. Preference annotators write prompts, choose between responses from two model variants, and label four levels of preference strength. Iterative reward-ranked sampling selects the best-scoring answer per prompt; later iterations reuse top samples from earlier rounds to mitigate observed forgetting. The authors stopped their vendor-based SFT annotation collection at 27,540 annotations and state that it excludes Meta user data. Reading boundary: ChatGPT is gpt-3.5-turbo-0301; approximately 4,000 helpfulness prompts, three raters per prompt. This is a historical prompt-set comparison.

Sources: [309]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: On the Reliability of Majority-Vote Pseudo-Labels in Test-Time Reinforcement Learning. When all sampled rollouts for a prompt share one answer, the method caches them; if a later majority conflicts with that answer, it discards the new pseudo-pairs and uses a cache loss. PC-Dr.DPO sets each prompt's robust temperature from its observed majority fraction, with weaker majorities receiving sharper within-prompt weighting. The analysis separates population-majority bias, where the policy's modal answer is wrong, from finite-rollout variation, where sampled majorities differ from that mode. The evaluation adapts on each full evaluation set or official test split using fixed benchmark budgets and reports the final checkpoint rather than the best intermediate checkpoint. Reading boundary: The safeguard preserves earlier self-consistency, which can itself be wrong; it does not correct population-majority bias.

Sources: [416]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Process Reinforcement through Implicit Rewards. PRIME can initialize its PRM from the starting SFT/base policy and filters prompts by their sampled accuracy range during training. PRIME updates an implicit PRM online using policy rollouts and outcome supervision and uses token-level dense rewards for policy training. The instantiated advantage estimator computes implicit-process and outcome returns separately before combining them, using leave-one-out baselines. The implicit PRM parameterizes reward with the log likelihood ratio of a reward language model to a reference model, yielding token-level process rewards. Reading boundary: The described filtering aims at intermediate difficulty and PRM data balance; exact selected accuracy bounds are not specified by this claim.

Sources: [413]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Proximal Policy Optimization Algorithms. On 49 Atari games with metrics averaged over three trials, PPO wins 30 games by average reward across training and 19 by average reward over the final 100 episodes; ACER wins 18 and 28 respectively. PPO proposes a surrogate that takes the minimum of an unclipped probability-ratio advantage term and its clipped counterpart, removing incentive for some excessively large policy updates. The continuous-control surrogate comparison uses seven Gym MuJoCo -v1 tasks, one million timesteps per run, three random seeds per task, and a normalized aggregate over 21 runs. The actor-critic PPO algorithm collects rollout data from parallel actors, computes advantage estimates, then optimizes a surrogate on those samples for multiple minibatch epochs. Reading boundary: Game-win counts depend on the specified metric and compared A2C/ACER implementations.

Sources: [77]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. RAFT selects the batch’s highest-reward fraction for supervised fine-tuning at each stage. For IMDB, the study measures perplexity on 500 test samples and constrains its increase to at most 5%. The IMDB table reports LLaMA 7B-RAFT mean sentiment reward 0.4293 and mean perplexity 9.9375, versus the unaligned row’s 0.0993 and 9.5. The authors identify sparse desirable samples as a failure mode in which RAFT may stall without enough favorable outputs to select. The RAFT algorithm optionally mixes an expert generator with the current generator during data collection. Reading boundary: The reward is a learned sentiment-model score; this result is not a general helpfulness/safety gain.

Sources: [315]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: REBEL: Reinforcement Learning via Regressing Relative Rewards. REBEL permits its base response distribution to be either a fixed offline dataset or the current policy. The TLDR evaluation compares generated summaries against human references using GPT-4 on a randomly sampled 600-example test subset. In the described experiments, both responses in each REBEL regression pair are generated by the current policy. REBEL regresses response-pair reward differences against policy log-ratio differences, removing the prompt-only partition term. The 6.9B TLDR table reports 78.0% win rate for REBEL and 77.6% for PPO. Reading boundary: The PPO result is imported from Huang et al.; the table is not a jointly rerun paired comparison. GPT-4 is the preference judge, not a human panel.

Sources: [316]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: REINFORCE++: Stabilizing Critic-Free Policy Optimization with Global Advantage Normalization. REINFORCE++ replaces prompt-local advantage normalization with global normalization and describes a single-response configuration for general RLHF. The baseline variant first subtracts the within-prompt mean reward and then normalizes using global batch statistics. The baseline variant uses a separate KL regularization loss with the k2 estimator. The best-practice discussion distinguishes baseline variants that support binary 0/1 or symmetric −1/1 rewards from plain REINFORCE++, which the authors recommend with symmetric rewards. Reading boundary: Single-response mode means one sampled response per prompt. Selected v9 also evaluates plain REINFORCE++ with k > 1 for reasoning (section 4.2). The baseline variant adds group-mean reshaping and is particularly advised for sample filtering and complex tasks such as multi-turn tool use (sections 3.2 and 5.1); multiple responses alone do not require it.

Sources: [415]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: RLAIF-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness. RLAIF-V generates candidate responses with the same model/distribution, decomposes factual response content into atomic claims, queries an open-source multimodal labeler with yes/no questions and constructs preference pairs for iterative DPO training. RefoMB descriptions start with GPT-4V drafts merged by GPT-4, then human additions/corrections and verification by at least two graduate students; GPT-4 uses descriptions to judge response pairs against GPT-4V. The authors acknowledge training costs and remaining hallucination despite improvements, and leave further trustworthiness gains for future methods. For the 146-sample MHumanEval response-level hallucination measure in v1, Table 1 reports 50.7% for OmniLMM 12B, 29.5% after RLAIF-V, and 45.9% for the compared GPT-4V condition. The 7B setting uses LLaVA 1.5 as instruction model and LLaVA-NeXT as labeler; the 12B setting uses OmniLMM as both instruction model and labeler, with four 4k-instruction iterations described for each. Reading boundary: Open-source training feedback does not mean the complete benchmark/evaluation workflow excludes humans or proprietary models. Table 4 defines99 dev and 261 test instructions; the prose’s images wording is inconsistent.

Sources: [155]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. In this v1 summarization study, PaLM2 Large generates preference labels and PaLM2 Extra-Small initializes the models; RL uses modified A2C rather than PPO. Human policy evaluation uses 1,200 ratings: evaluators rank RLAIF, RLHF, SFT and human-reference summaries without ties for held-out TL;DR posts. For each summary pair, the AI labeler is queried twice with reversed candidate order and the resulting preference distributions are averaged to mitigate position bias. Human evaluators prefer RLAIF summaries to SFT 71% of the time and RLHF to SFT 73%; the paper reports the difference as not statistically significant and RLAIF versus RLHF as 50%. Reading boundary: The study is summarization-only v1; longer outputs are a reported confound and comparative monetary costs were not estimated (conclusion p9).

Sources: [79]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: RRHF: Rank Responses to Align Language Models with Human Feedback without tears. The experiments use the HH preference corpus with the Dahoas/gptj-rm-static proxy reward model. RRHF scores each response with length-normalized conditional log probability. Table 2 reports Alpaca-RRHF with diverse-beam sampling mean proxy reward−1.02 and GPT-2-medium perplexity 14.75 on HH. The pinned RRHF study does not experiment with refreshing the sampling policy at every training step. RRHF adds ranking loss to cross-entropy training on the highest-reward response. Reading boundary: Reward units are model-specific; a high proxy reward is not an independent measure of helpfulness or safety.

Sources: [319]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: ReMax: A Simple, Effective, and Efficient Method for Aligning Large Language Models. The reported GPT-4 model comparison uses 300 LIMA test samples. ReMax subtracts the reward of a current-model greedy response from the reward of its sampled response when forming the policy-gradient loss. ReMax does not resolve reward bias or reward hacking; the authors identify data collection and reward-learning design as remaining requirements. The authors report no substantial training-speed acceleration for models on the scale of OPT 1.3B after DeepSpeed optimization. The authors report about 2.2× training acceleration for ReMax in a GPT-2-137M IMDB positive-sentiment continuation experiment without DeepSpeed. The rm-static experiments randomly split the public dataset 20% for SFT,40% for reward learning and 40% for reward maximization. Reading boundary: OPT win/tie/loss counts in one displayed comparison sum to 299 rather than 300; that mismatch is retained as an issue and no corrected count is inferred.

Sources: [317]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Reference-free Monolithic Preference Optimization with Odds Ratio. The paper uses GPT-4 as judge in AlpacaEval1.0 and MT-Bench, and GPT-4-turbo in AlpacaEval2.0, with text-davinci-003 and GPT-4 references respectively for the two AlpacaEval versions. ORPO combines supervised chosen-response NLL loss with a weighted log-sigmoid odds-ratio loss comparing chosen and rejected responses. ORPO updates the model directly without retaining a separate frozen reference model; the paper compares this computation to its stated RLHF/DPO training setup. The paper leaves scaling beyond 7B parameters, broader preference-method comparisons and more diverse fine-tuning domains to future work. Reading boundary: These are automated evaluation outcomes, distinct from human preference annotation.

Sources: [76]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: SLiC-HF: Sequence Likelihood Calibration with Human Feedback. The sample-rank experiments generate eight SFT-model decodes per example at temperature 0.7 and top-k 40. SLiC-HF-direct calibrates directly on preferred/rejected human-feedback sequences without training a ranking or reward model. SLiC-HF-direct can face mismatch between the off-policy human-feedback distribution and the SFT model’s decode distribution. In a 100-validation-example four-way human comparison, the sample-rank SLiC-HF model is chosen best 73% of the time. SLiC-HF combines a margin-based sequence ranking calibration loss with cross-entropy regularization. Reading boundary: Three raters per task; choice uses majority vote, with 4% same/no-preference. This is a four-way selection rate, not pairwise win rate against every baseline.

Sources: [326]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: STAMP Your Content: Proving Dataset Membership via Watermarked Rephrasings. STAMP requires watermarking before online release and access to model token probabilities; it does not retroactively mark previously published content. The method uses a one-sided paired t-test for lower public-version perplexity relative to private versions. STAMP compares a public keyed paraphrase collection with privately held differently keyed paraphrases using paired perplexity differences. STAMP can compare each public paraphrase against the average perplexity over multiple differently keyed private paraphrases to reduce individual-document bias. Reading boundary: The paper also warns about paraphrase errors and evaluates continual pretraining rather than models trained from scratch.

Sources: [418]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: SWE-Lancer: Can Frontier LLMs Earn $1 Million from Real-World Freelance Software Engineering?. SWE-Lancer Diamond contains 502 tasks: 237 implementation tasks and 265 managerial tasks. Table 1 reports Claude 3.5 Sonnet passing 26.2% of the 237 Diamond implementation tasks with the user tool enabled. The implementation-task prompt tells the agent to edit code without running the app. The authors state that evaluated models cannot ask clarifying or follow-up questions. The task-selection process reports 100 professional software engineers reviewing tasks, proposals and code for clarity and executability. Reading boundary: Historical single-attempt task success; not the dollar-weighted earnings rate and not the managerial score.

Sources: [373]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: SelectTTRL: Representation-Guided Data Selection for Efficient Test-Time Reinforcement Learning. Percentile selection is fixed before adaptation without target ground-truth answers, while all target questions remain in the evaluation set. SelectTTRL constructs a frozen correctness-related activation direction from paired correct and incorrect answers in a separate TruthfulQA-derived labeled calibration set. In the reported experiment, subset baselines retain the same question fraction and vary the selection rule while keeping the TTRL procedure, budget and hyperparameters fixed. Initial target responses are scored and ranked, then only a prescribed percentile interval receives standard TTRL multi-response sampling and policy updates. Reading boundary: The selected subset is defined by floor-based percentile ranks and fixed-identifier tie breaking, not post-hoc selection of only correctly answered targets.

Sources: [417]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Self-Play Preference Optimization for Language Model Alignment. The authors report that general leaderboard performance for DPO, IPO and SPPO declines after the first or second alignment iteration. SPPO formulates alignment as a two-player constant-sum preference game rather than assuming one transitive scalar reward for all comparisons. The experiment uses three SPPO iterations, advancing a first-epoch model trained on 20,000 prompts each iteration. Table 1 reports 28.53% length-controlled AlpacaEval 2 win rate for SPPO Iter 3 versus 17.11% for Mistral 7B-Instruct-v0.2. The experiment samples five responses per prompt and retains the highest- and lowest-PairRM-scoring responses for preference data. The SPPO convergence theorem assumes realizability of the ideal optimization problem and concerns the mixture policy across iterations. Reading boundary: 805 prompts; GPT-4-Turbo judging/reference protocol. These are ordinary decoding rows, separate from best-of 16 reranking. Table score 28.53 differs from prose 28.52.

Sources: [327]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Self-Refine: Iterative Refinement with Self-Feedback. The reported blind human A/B evaluation is conducted by the paper authors; Table 6 describes150 examples for each evaluated dataset. The paper limits its main experiments to English and proprietary base models, requires sufficient few-shot/instruction-following ability, and reports that erroneous feedback dominates the analyzed failed refinements. In Table 1, GPT-4 code-optimization performance rises from 27.3% to 36.0%, while math solve rate changes from 92.9% to 93.1%; the paper uses task-specific, human-preference and GPT-4-preference metrics rather than one common accuracy measure. Self-Refine uses the same language model for initial generation, natural-language feedback and refinement, appending prior feedback/output history without additional model training. Reading boundary: These values are reported under this paper’s task setups and are not a general hallucination or correctness improvement.

Sources: [156]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: SimPO: Simple Preference Optimization with a Reference-Free Reward. For Instruct setups, five responses per UltraFeedback prompt are generated from the starting model and PairRM selects the highest/lowest-scoring pair. SimPO introduces a positive target reward margin in the Bradley-Terry preference objective. For Llama 3-8B-Instruct, Table 4 reports AlpacaEval 2 length-controlled win rate 44.7% for SimPO versus 40.3% for DPO. SimPO uses length-normalized policy log-likelihood as an implicit reward and removes the reference-policy term. The authors state that their experiments focus on helpfulness and do not evaluate safety, honesty and fairness comprehensively. Reading boundary: 805 questions and the benchmark’s automatic evaluation protocol;44.7 is length-controlled, not raw win rate (40.5).

Sources: [325]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Statistical Rejection Sampling Improves Preference Optimization. The standard RSO experiment generates 64 SFT-policy candidates and rejection-subsamples eight. The CNN/DailyMail experiment assumes no training access to CNN/DailyMail targets or preference texts, testing transfer from TLDR training. The HH experiments use the helpfulness task rather than the entire mixed helpfulness/harmlessness task set. Human evaluation uses three raters per task, anonymized/shuffled responses and majority choice;47 distinct raters participated. For TLDR sigmoid-normalized optimization, the three-way human comparison selects rso-sample-rank 48% of the time versus 21% for direct optimization. The empirical RSO sampler accepts candidates according to an exponential function of reward relative to the maximum remaining candidate reward. Reading boundary: All-same outcomes mean percentages need not sum to 100; this is three-way best-choice frequency, not a 48% pairwise win rate.

Sources: [320]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Step-DPO: Step-wise Preference Optimization for Long-chain Reasoning of LLMs. On MATH, the Qwen2-7B-SFT ablation reports 54.8% baseline,55.0% with DPO and 55.8% with Step-DPO. The proposed localization procedure permits humans or GPT-4 to find the first erroneous reasoning step. The described data construction produces 10,000 pairwise preference records for Step-DPO. The authors caution that outputs with correct final answers can still contain erroneous intermediate steps, requiring additional preferred-step filtering. Step-DPO contrasts preferred and rejected next reasoning steps conditioned on the prompt and preceding correct steps. Reading boundary: Both preference methods use 5,000 training records in this ablation; MATH evaluation has 5,000 problems. These values do not describe the full 10,000-record main run.

Sources: [328]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: TTRL: Test-Time Reinforcement Learning. The example reward function extracts final answers and gives reward one for matching the most frequent sampled answer and zero otherwise. TTRL samples multiple candidate outputs, derives a consensus answer and uses agreement with that answer to construct its reward signal. TTRL is defined as reinforcement-learning adaptation of a pretrained model during test time without ground-truth labels for the target data. The implementation samples 64 responses for voting and downsamples 32 per prompt for policy training. Reading boundary: This exact rule is answer-frequency agreement; it does not validate the intermediate reasoning.

Sources: [419]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: The Llama 3 Herd of Models. The Llama 3 DPO recipe masks header and termination tokens from both chosen and rejected response losses to stabilize training. Some preference records contain three ranked responses: human-improved edited response, chosen response, and rejected response. The authors report that Llama 3 405B and GPT-4-0125 win rates are within the margin of error on nearly all capabilities in their human evaluation. The human evaluation counts a model win only when an annotator rates its response better or much better, excluding slightly-better ratings from wins. Llama 3 post-training applies six rounds, collecting new preference annotations and SFT data and sampling synthetic data from the latest models each cycle. Rejection sampling typically generates 10 to 30 outputs per human-annotation prompt and uses the reward model to choose one candidate. Reading boundary: About 7,000 prompts across capability sets; pairwise seven-point preference ratings. This is not an across-provider or present-day ranking.

Sources: [310]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. ToolEval measures pass rate through completion and exclusion rules and win rate through ChatGPT path comparisons; its judge-validation study uses four human preference annotations for each of 600 solution pairs. For API responses exceeding 2,048 tokens, ToolBench applies a ChatGPT-derived compression schema and, if necessary, retains only the first 2,048 tokens. The ToolBench construction pipeline uses ChatGPT gpt-3.5-turbo-16k for API collection, instruction generation and solution-path annotation, while human experts provide in-context seed examples. After filtering, the pinned ToolLLM paper reports 3,451 retained tools containing 16,464 APIs; these are different counting units.

Sources: [255]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Training Verifiers to Solve Math Word Problems. For the paper's 6B verifier, increasing candidate completions helps up to 400 and then reduces performance; the authors attribute this to the risk of finding solutions that fool the verifier. Verifier targets label model-generated solutions by whether their final answer is correct; the paper explicitly notes false positives when flawed reasoning reaches the right answer. Reading boundary: This is the tested model/setup; the usual reported verification evaluation uses 100 completions.

Sources: [90]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. For RLHF training, the paper combines 137k prompts from the static human-feedback dataset with 369k model-generated prompts; the synthetic prompts are distinct from the human comparison labels used to train preference models. The authors emphasize average-case rather than worst-case behavior, unresolved honesty and subtle-bias issues, and RL stability limitations. A controlled 52B helpfulness experiment uses equal-sized data and identical hyperparameters and finds that a base/rejection-sampling/online mixture is preferred by crowdworkers to all-base data. The study collects helpfulness dialogues by choosing the more helpful response and red-team dialogues by choosing the more harmful response; retained comparisons are treated as equally weighted binary labels without ties. Workers are recruited through US MTurk and Upwork; the authors select workers using writing/dialogue spot checks rather than agreement-based label filtering, and explicitly note that the worker distribution changes over the project. Reading boundary: In this paper, online iterations retrain new models rather than continually updating one model. This is not a guarantee of worst-case safety.

Sources: [138]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Training language models to follow instructions with human feedback. The authors describe collecting human rankings of model outputs as part of their instruction-following training method. Reward-model data collection presents human labelers with 4 to 9 model responses to rank for a prompt. The paper’s prompt dataset primarily uses prompts submitted to early InstructGPT through the OpenAI API Playground. Most comparisons in this workflow are labelled by one contractor for cost reasons. The PPO prompt dataset is used as RLHF input without per-prompt human labels. Reading boundary: Pairwise comparisons derived from a single ranked set are correlated; they are not independent human judgements for each pair.

Sources: [724], [731]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: TruthfulQA: Measuring How Models Mimic Human Falsehoods. The human generation-evaluation procedure blinds evaluators to the model name and prompt that produced an answer. The paper recruits two external researchers for separate reference-label validation and a human-answer baseline. GPT-judge training includes about 6,900 true/false-labelled author reference answers and about 15,500 human-evaluated model answers. Reading boundary: Section 3.2 (p.6) says the authors carried out all main-generation evaluations. A predefined qualitative rubric maps to scalar truth scores; it does not remove all judgment or disagreement.

Sources: [675]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: Understanding R1-Zero-Like Training: A Critical Perspective. Dr. GRPO replaces response-dependent loss normalization with a constant and uses a Monte Carlo return baseline in the recovered objective. The authors identify within-question reward-standard-deviation normalization as reweighting questions and introducing difficulty bias. The authors identify response-length normalization as a bias favoring shorter correct responses and longer incorrect responses in their GRPO objective analysis. The optimizer experiment assigns reward one when Math-Verify accepts the final answer and zero otherwise, training on MATH questions and evaluating on AIME2024, AMC, MATH500, Minerva Math and OlympiadBench. Reading boundary: The preceding method also removes group standard-deviation normalization (physical p6 lines150-151); unbiasedness is reported by the paper and not independently proved.

Sources: [405]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks. VAPO separates GAE parameters for value targets and policy advantages, using lambda one for the value-network returns. VAPO adjusts policy GAE lambda as a function of output length and a bias-variance control parameter. VAPO combines policy-gradient loss with an additional negative-log-likelihood loss on correctly sampled outcomes. VAPO pretrains the value model on Monte Carlo returns from a fixed policy before saving and loading that value checkpoint for policy training. Reading boundary: Policy lambda is separately length-adaptive in this recipe; no general unbiasedness theorem is asserted by this candidate.

Sources: [420]

Evidence: Reviewed-work atlas - Feedback, rewards and policy optimization: WPO: Enhancing RLHF with Weighted Preference Optimization. The conceptual WPO preference-labeling construction assumes there are no conflicting pairwise preferences for a prompt. The authors state that their preference dataset does not include safety aspects and that they do not study preference optimization for multi-turn conversations. The hybrid experiment labels five SFT-model outputs plus one GPT-4-Turbo output using GPT-4-Turbo additive scoring. In the off-policy Llama 3-8B-Instruct setting, WPO reports AlpacaEval 2 length-controlled win rate 33.8(1.3)% versus DPO 28.2(0.5)%. WPO weights each preference-pair loss by the product of current-policy response weights, detached from backpropagation; sequence probabilities are length-normalized in practice. Reading boundary: Parentheses are standard deviations across five random-seed runs;805 evaluation instructions. This is the off-policy row, separate from hybrid results.

Sources: [333]

Evidence: Reviewed-work atlas - Multilingual supervision and evaluation: BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models. For IFEval, BenchMAX removes language-specific instructions and retains 429 samples. BenchMAX selects 17 languages spanning multiple language families and writing systems. The construction pipeline translates English data, has three human annotators post-edit each sample, then selects a final translation. All benchmark translations except long-context data are post-edited by native human experts. Reading boundary: Selection applies to the BenchMAX IFEval component, not all tasks or the original IFEval corpus.

Sources: [381]

Evidence: Reviewed-work atlas - Multilingual supervision and evaluation: How Reliable is Multilingual LLM-as-a-Judge?. Cross-language consistency is measured using Fleiss' Kappa,treating each language-specific output as a rater's judgment. The experiments use datasets with parallel data and treat ground-truth answers as the answers being judged. The experiments use pointwise judgments with both binary Yes/No and1-5 score criteria. The main setup uses an English instruction prompt with the target language specified through an eval_language placeholder. Reading boundary: The paper separately considers accuracy/grade because high consistency alone can reflect uniformly incorrect judgments. No kappa-to-accuracy conversion is made.

Sources: [396]

Evidence: Reviewed-work atlas - Multilingual supervision and evaluation: M-IFEval: Multilingual Instruction-Following Evaluation. For Japanese,the construction removes word-count length constraints and uppercase/lowercase instructions from the English IFEval list. The benchmark selects French,Japanese and Spanish with one native-speaker language lead for each. Language leads design objectively evaluable instructions that include language-specific linguistic and textual controls. The constructed prompt sets contain 115 Spanish,172 Japanese and235 French prompts. Reading boundary: Writing-system-specific adaptation, not evidence that all translated English constraints transfer unchanged.

Sources: [391]

Evidence: Reviewed-work atlas - Multilingual supervision and evaluation: M-RewardBench: Evaluating Reward Models in Multilingual Settings. The translation component draws from MAPLE, which supplies five human-rated translations per source with scores 1-6. The experiment setup describes selecting 25 reward models spanning 3 B-104 B parameters and generative, classifier and implicit-DPO types. General-task instances are translated from RewardBench into 23 languages with Google Translate and human-reviewed filtering. The benchmark constructs 100 easy and 100 hard chosen/rejected translation pairs for each of four translation directions. Reading boundary: MAPLE derives from WMT20/21 and covers de→en,zh→en,en→de,en→zh. Describes the paper’s source material, not a separate dataset review.

Sources: [394]

Evidence: Reviewed-work atlas - Multilingual supervision and evaluation: MENLO: From Preferences to Proficiency -- Evaluating and Modeling Native-like Quality Across 47 Languages. Each generated response pair is shown in randomized order and rated by at least three annotators. MENLO measures fluency, tone, localized tone and localized factuality. MENLO contains 6,423 prompt-response preference pairs and reports 81,014 human annotations. The design selects 47 language varieties and uses human-written parametric English prompt templates to evoke local contexts. Reading boundary: Responses use GPT-4o, Llama4-Maverick, Llama4-Maverick with Search and Gemini1.5 with Search; 1-5 Likert ratings allow ties and final scores use majority vote. Expert annotators provide language-specific feedback and subset gold labels.

Sources: [390]

Evidence: Reviewed-work atlas - Safety, feedback and verification: Constitutional AI: Harmlessness from AI Feedback. For the current comparison tests, workers were instructed to prefer nuanced, transparent and thoughtful responses over evasive ones when similarly harmless, changing the harmlessness evaluation from prior work. The RL-CAI preference model mixes human-feedback labels for helpfulness with model-feedback labels for harmlessness; the RL pipeline then optimizes this hybrid preference model. The supervised CAI stage samples responses to red-team prompts, asks the model to critique and revise those responses under constitutional principles, and finetunes on revised responses. The paper explicitly reports that current comparison tests were performed with Surge AI workers; reused preference-model data came from an earlier period working primarily with Upwork and MTurk. Reading boundary: Current Elo comparisons inherit this preference rubric and are not directly interchangeable with the prior paper's Elo results.

Sources: [51]

Evidence: Reviewed-work atlas - Safety, feedback and verification: Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. The approach uses separate input-only and streaming output-only classifiers. The authors identify preset chatbot queries, incomplete adversary coverage and rubric difficulty distinguishing benign knowledge from actionable uplift as evaluation limitations. The improved chemistry-focused classifier system reports an absolute 0.38 percentage-point production refusal-rate increase on 5,000 sampled Claude.ai conversations. None of the 113 submitted prototype jailbreak reports answered all ten target questions at the half-helpful-only-detail rubric threshold. Constitution-guided data generation uses a helpful-only model for queries/outputs and a prompted model to remove inadvertent refusals. The automated-evaluation classifiers omit automated-red-teaming training data, unlike the bug-bounty prototype classifiers. Reading boundary: This is the improved system guarding Claude 3.5 Sonnet(New), not the earlier human-red-teamed prototype.

Sources: [293]

Evidence: Reviewed-work atlas - Safety, feedback and verification: DICES Dataset: Diversity in Conversational AI Evaluation for Safety. For DICES-350, a trust/safety gold rating is first assigned by one rater and verified by a pool of five experts; the paper reports disagreement with the crowd-majority label in 34% of examples. The paper identifies small conversation count, restricted demographic categories and unresolved separation of sincere belief differences from task/rater noise as limitations, despite high replication of ratings. DICES-350 rater quality assessment considers unusually short annotation time and invariant conversation labels; the paper removes 19 raters before its 104-rater analysis. DICES-990 recruits 173 US/India raters with partial conversation coverage; DICES-350 recruits a different 123-rater US pool, then excludes 19 raters and reports results from 104 raters who each annotate all 350 conversations. Reading boundary: Counts of raters recruited, retained, ratings and safety-question annotations must not be conflated.

Sources: [132]

Evidence: Reviewed-work atlas - Safety, feedback and verification: Deliberative Alignment: Reasoning Enables Safer Language Models. The StrongREJECT human review uses at least two reviewers per prompt/answer pair, with an expert third reviewer arbitrating disagreements. AppendixA reports mean human assistance grade 1.020 for o1 and 1.335 for GPT-4o on its StrongREJECT review. For SFT filtering, completions are judged repeatedly and assigned the minimum judge score before high-score retention. SFT trains on original prompts with specification text removed, while keeping the generated reasoning/output as targets. During RL, the safety judge sees the policies but not the model’s chain-of-thought; this particular RL reward is used for o1 and o3-mini. Training completions are generated by a specification-agnostic base reasoning model prompted with category-relevant safety specifications. Reading boundary: 2,100 adversarial prompts per model,35 trained reviewers, assistance grades 1/3/5. Gemini API-blocked prompts are scored 1.

Sources: [297]

Evidence: Reviewed-work atlas - Safety, feedback and verification: Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. For two systemic hacks,the study uses string-regex detections as retrospective ground truth for monitor evaluation. A prompted GPT-4o monitor looks for reward hacking in coding environments where an agent implements functionality to pass unit tests. In the direct CoT-pressure experiment, a non-frontier agent still hacks while its GPT-4o monitor recall falls near zero. The CoT monitor observes reasoning,tool calls and final outputs,while the action monitor sees only tool calls and final outputs. Reading boundary: Hacks are exit(0) early termination and raise SkipTest outside the testing framework. Ground truth covers these identifiable hacks,not every potential exploit.

Sources: [393]

Evidence: Reviewed-work atlas - Safety, feedback and verification: Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. The authors state that they do not know whether crowdworkers have expertise needed to assess some technical harms and that attack coverage is incomplete; the review workflow includes warnings, opt-out previews, support communication and a voluntary well-being survey. Self-reported attack success shows no clear model-size trend, while minimum harmlessness-model scores show RLHF models becoming harder to red-team with scale; harmful outputs still occur. At each dialogue turn, participants see two model responses and select the more harmful; after the conversation they rate attack success on a 0-4 scale, separately from the learned preference-model harmlessness score. A follow-up sample of 500 attacks each on 52B prompted and RLHF models receives three reviewer ratings per transcript; agreement including the attack author is Fleiss kappa0.32 on Likert ratings and 0.49 after binarization. The study’s red team comprises 324 US-based crowdworkers: 307 recruited through MTurk and 17 through Upwork; about 80% of attacks come from roughly 50 workers. Reading boundary: Worker-support descriptions do not establish absence of harm, and reduced average harmfulness does not establish a perfectly safe model.

Sources: [154]

Evidence: Reviewed-work atlas - Source curation and demonstrations: AlpaGasus: Training A Better Alpaca with Fewer Data. The study evaluates on Koala, Vicuna, WizardLM and Self-Instruct prompts with an API LLM judge such as GPT-4, comparing both response orders to mitigate position bias. The initial study is limited to 7B/13B models and the Alpaca instruction dataset, and explicitly states that it performs no human evaluation. The initial AlpaGasus study grades Alpaca instruction-input-response triplets with ChatGPT for accuracy and retains scores at least 4.5, selecting 9,229 of 52,002 samples. The 7B WizardLM-test analysis reports better-or-equal performance on 22 of 29 skills, but flags weaker coding categories and a keyword-defined coding-data removal rate of 88.16%, versus 82.25% overall. Reading boundary: This evaluation is automatic preference scoring, not human preference ground truth.

Sources: [121]

Evidence: Reviewed-work atlas - Source curation and demonstrations: DataComp-LM: In search of the next generation of training sets for language models. For the reported 2.6T-token 7B model, the authors augment DCLM-Baseline with StarCoder and ProofPile2, use a math-enriched annealing phase and extend context; Table 9 reports 63.7% MMLU. The human-quality study samples 499 rule-filtered/deduplicated documents and obtains three labels per document from 16 English-speaking AI graduate students/professors; 281 documents have unanimous labels. DCLM's first evaluation suite omits fairness, multilinguality and safety; compute limits prevent joint exploration of all curation dimensions, sufficient random-seed variation and testing beyond 7B models. The DCLM-Baseline quality filter uses fastText trained with OpenHermes 2.5 plus high-scoring ELI5 positives and RefinedWeb-sampled negatives, retaining the top 10% by classifier score. DCLM provides an unfiltered Common-Crawl-derived pool of 240T GPT-NeoX tokens and five fixed-recipe compute scales; evaluation separates 53-task extended performance, 22-task core performance and five-shot MMLU. Reading boundary: The result is not attributable to raw Common Crawl or the baseline filter alone. The paper's headline64% is rounded.

Sources: [131]

Evidence: Reviewed-work atlas - Source curation and demonstrations: DataComp: In search of the next generation of multimodal datasets. DataComp's filtering track fixes the candidate pool and training compute by total samples seen while allowing the filtered dataset size to vary; it also offers a separate bring-your-own-data track and four compute scales. The paper warns that ImageNet performance is not representative of every task and reports demographic biases in face classification; some tested BYOD sources can increase misclassification risk. At xlarge scale, the image-based/CLIP-score intersection baseline named DataComp-1B contains 1.4B samples and reports 79.2% zero-shot ImageNet accuracy after 12.8B samples seen, versus 72.3% for no filtering under the table's matched scale. CommonPool generation starts from Common Crawl image URLs with nonempty alt-text, uses unsafe-content filtering and evaluation-image near-duplicate removal to construct the 12.8B xlarge pool, and provides face bounding-box metadata for default blurring during participant image download. Reading boundary: These are described preprocessing steps, not a verified guarantee of absence of unsafe content or personal information.

Sources: [130]

Evidence: Reviewed-work atlas - Source curation and demonstrations: Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. The paper introduces Dolma as approximately 3T tokens across 5B documents from seven public data sources, alongside an open-source curation toolkit; its design primarily targets English text. Dolma's general deduplication is within each source subset rather than across sources; its web pipeline uses URL, exact-document and exact-paragraph Bloom-filter stages, while code starts from The Stack's prior near-deduplicated version. The paper's Olmo-1b comparison is trained on roughly 3T tokens from preliminary Dolma v1.5; the table explicitly notes unequal training-token budgets for some comparator models and unknown StableLM2 data composition. The web pipeline uses regular-expression PII detection for email addresses, IP addresses and phone numbers; its 1B-model ablation reports no discernible performance difference from the studied masking/removal strategies. The release discussion acknowledges residual incorrect, toxic, hateful, personal and copyrighted content in web-derived corpora despite mitigation, and says the project's decisions may be revisited. Reading boundary: This is a model-performance observation, not a privacy guarantee or a full PII-recall measurement. The prose and Figure 4 differ on the boundary count, so no exact boundary count is proposed here.

Sources: [134]

Evidence: Reviewed-work atlas - Source curation and demonstrations: LIMA: Less Is More for Alignment. The quality/diversity/quantity ablations fine-tune 7B models and use ChatGPT scores; the paper explicitly acknowledges possible confounding when using Stack Exchange versus wikiHow as a diversity proxy. In the paper's human preference evaluation over 300 test prompts, LIMA versus April-2023 GPT-4 responses records 18% wins, 25% ties and 57% losses. The discussion identifies substantial manual curation effort and limited robustness: unlucky decoding samples or adversarial prompts can produce weak LIMA responses. LIMA fine-tunes pretrained LLaMa 65B on exactly 1,000 curated demonstration sequences; Table 1 separates community-derived examples, 50 Natural Instructions examples and 200 paper-author training examples. Reading boundary: The source comparison does not isolate diversity alone, and the result does not establish universal diminishing returns to instruction data.

Sources: [149]

Evidence: Reviewed-work atlas - Source curation and demonstrations: SELF-INSTRUCT: Aligning Language Model with Self Generated Instructions. For user-oriented evaluation, the authors create 252 new instructions with one instance each, and the instruction authors judge model outputs on a four-level quality rubric. Reading boundary: This differs from the SuperNI ROUGE-L evaluation and is not a representative or independently recruited user sample.

Sources: [83]

Evidence: Reviewed-work atlas - Source curation and demonstrations: Scaling Instruction-Finetuned Language Models. The CoT mixture reuses nine prior datasets with human-written reasoning annotations and manually composed instruction templates; its ablation reports that omitting CoT data degrades performance on held-out CoT evaluations. The paper’s finetuning collection comprises 473 datasets, 146 task categories and 1,836 dataset/task-category tasks from Muffin, T0-SF, NIV2 and a CoT mixture; templates do not create separate tasks. The Flan-T5 model card states that it has not been tested in real-world applications; the responsible-AI appendix also notes limited identity/language coverage and noise/bias in automated toxicity measures. At 540B parameters, Table 3 reports a normalized six-score held-out average of 49.1 without finetuning and 58.5 with all 1,836 tasks, a 9.4-point difference. Reading boundary: The normalized macro-average is not raw overall accuracy. Section2.2 selects checkpoints using periodic evaluations of held-out tasks.

Sources: [137]

Evidence: Reviewed-work atlas - Source curation and demonstrations: The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. Common Pile v0.1 is described as an 8 TB text corpus curated from 30 sources. The curation definition excludes Creative Commons NonCommercial and NoDerivatives material. Source inclusion requires confidence that license information comes from the copyright holder. The curators take a conservative stance against including LLM-generated synthetic content. Reading boundary: Raw corpus description in version 2506.05209v1;8TB is not a token count or a claim that every component independently meets a legal standard.

Sources: [382]

Evidence: Reviewed-work atlas - Source curation and demonstrations: WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. Evol-Instruct separates instruction evolution, response generation and elimination of failed evolutions; in-depth prompts add difficulty while in-breadth prompts expand instruction diversity. The WizardEval human study uses 218 instructions spanning 29 skills and domains and recruits 10 annotators for source-hidden, shuffled-response comparison using five listed evaluation aspects. The paper acknowledges scalability and reliability limitations of its GPT-4 and human evaluations and warns that its test set may not cover every application scenario or domain. WizardLM v3 starts from 52k Alpaca instructions, runs four evolution rounds using Azure OpenAI ChatGPT, obtains a 250k instruction pool, and samples 70k examples for main LLaMA 13B fine-tuning.

Sources: [159]

Evidence: Reviewed-work atlas - Task construction and evaluation: τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. Telecom task success is evaluated only with assertion functions. The authors manually refine LLM-generated telecom implementations until their unit tests pass. The reported GPT-4.1 pass^1 falls to 34% on telecom from 74% on retail and 56% on airline. Table 2 records three critical simulator errors in 50 telecom conversations. The telecom evaluation subsamples 114 tasks from 2,285 programmatically composed tasks. Reading boundary: Historical four-run temperature-zero function-calling configuration with GPT-4.1 user simulation; not live customer accuracy.

Sources: [376]

Evidence: Reviewed-work atlas - Task construction and evaluation: AgentBench: Evaluating LLMs as Agents. AgentBench's overall score averages heterogeneous task metrics after weighting each by the reciprocal of its mean across the evaluated models. The initial AgentBench evaluation includes 25 API or open models; the open-model population is limited to models no larger than 30B. Reading boundary: The aggregate is population-dependent normalization, not percent of all tasks completed.

Sources: [48]

Evidence: Reviewed-work atlas - Task construction and evaluation: AgentGym: Evolving Large Language Model-based Agents across Diverse Environments. For its main evolution experiments, AgentGym converts rewards below one to zero and retains reward one as success, even when an environment originally provides dense rewards. AgentGym collects model-generated and crowdsourced trajectories, filters them by reward or correctness, and forms the 6,130-trajectory AgentTraj training collection. AgentGym describes 14 environments and 89 task types, while its main self-evolution experiments use 11 environments. AgentGym evaluates models without trajectory fine-tuning with few-shot prompts and trajectory-trained models with zero-shot prompts; evaluation uses temperature zero and exploration uses 0.7.

Sources: [210]

Evidence: Reviewed-work atlas - Task construction and evaluation: BenchGuard: Who Guards the Benchmarks? Automated Auditing of LLM Agent Benchmarks. The intended workflow surfaces findings with evidence for expert adjudication rather than autonomously deciding benchmark truth. The five-model union exactly aligns with 20 of 24 atomic BIXBench issues, reported as 83.3%. BenchGuard cross-verifies task instructions, reference programs, evaluation scripts and environment configurations. The authors identify hallucinated audit findings as a reason to retain expert review. ScienceAgentBench precision is conditioned on findings inside tasks with confirmed defects. Reading boundary: The 24 issues come from 17 expert-revised tasks within Verified-50; this is neither 83.3% of all tasks nor all possible defects.

Sources: [337]

Evidence: Reviewed-work atlas - Task construction and evaluation: Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. BIG-bench task-definition files include a canary identifier intended to help exclude benchmark text from web-scraped training corpora and diagnose training-on-test use. The study reports mean and best average task scores from expert evaluators but cautions that heterogeneous expertise, task subsampling and timed sessions constrain interpretation of these human reference scores. BIG-bench Lite is the paper's 24-task lightweight subset and contains only JSON tasks. The paper describes at least 204 BIG-bench tasks, intended principally for zero- and few-shot evaluation without task-specific fine-tuning, with JSON and Python programmatic task interfaces. Reading boundary: The mechanism does not establish exclusion by a particular model.

Sources: [214]

Evidence: Reviewed-work atlas - Task construction and evaluation: BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. The grading prompt instructs the judge to compare against the supplied answer rather than solve the question or argue for alternatives. The paper states that BrowseComp questions were collected entirely by human trainers. Human trainers solved 367 of the 1,255 attempted questions, reported as 29.2%. The authors cannot guarantee that every valid answer is represented by the reference answer. Reading boundary: Creators could not attempt their own questions, and AI assistance was disallowed; 11 of 1,266 questions were not attempted.

Sources: [338]

Evidence: Reviewed-work atlas - Task construction and evaluation: Fair human-centric image dataset for ethical AI benchmarking. Most annotations are self-reported by image subjects; observable labels such as pixel annotations,head pose and camera distance are exceptions. Vendors must obtain signed informed consent from primary and secondary image subjects and a copyright agreement for each image. Reported dataset terms restrict use to fairness/bias evaluation and mitigation,with training allowed only for bias-mitigation tools. Image subjects may withdraw consent for inclusion without losing their compensation. Reading boundary: Demographics are not inferred from images in this protocol. Annotation-guideline and raw-data compliance remain unaudited.

Sources: [388]

Evidence: Reviewed-work atlas - Task construction and evaluation: GAIA: A Benchmark for General AI Assistants. GAIA is designed for zero-shot concise factoid answers and grades final answers using type-dependent normalization followed by quasi-exact match to a unique reference answer. GAIA question creators provide answers and tool/step metadata; two new annotators independently answer each question, with ambiguous items repaired or removed. The paper calls its manually selected GPT-4-plugin configuration an oracle estimate because plugins changed and no stable automatically selected set was available. The pinned GAIA paper evaluates the final answer and explicitly leaves evaluation of the reasoning or action trace to future work. Reading boundary: This is a historical evaluation-setup limitation, not a statement about current products or current performance.

Sources: [220]

Evidence: Reviewed-work atlas - Task construction and evaluation: GPQA: A Graduate-Level Google-Proof Q&A Benchmark. The authors hire 61 contractors through Upwork to write and validate the dataset. GPT-4 with few-shot chain-of-thought scores 39.7% on the 448-question main split. The authors make no claim that GPQA represents questions arising in scientific practice. The authors flag expert and non-expert validator accuracies on main and Diamond splits as selection-biased. Reading boundary: Historical baseline; the separate search-plus-backoff configuration scores 41.0% and is not this result.

Sources: [351]

Evidence: Reviewed-work atlas - Task construction and evaluation: HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. The pinned HarmBench paper describes 510 behaviors: 400 textual and 110 multimodal, split into 100 development/validation and 410 test behaviors. The paper identifies target-model generation length as a major ASR comparison confounder and standardizes evaluation output length to 512 tokens. The HarmBench authors used GPT-4 to summarize several providers' acceptable-use policies and manually designed behaviors with that summary as guidance. HarmBench separates development and test classifiers and prohibits direct optimization on its test metric; the development classifier uses a different base model and half the fine-tuning set used for the test classifier. Reading boundary: The prose labels this output budget N while the equation uses N for test-case count; this assertion describes the prose-defined output condition.

Sources: [223]

Evidence: Reviewed-work atlas - Task construction and evaluation: Humanity's Last Exam. The authors say high HLE accuracy would not alone demonstrate autonomous research or artificial general intelligence. Table 1 reports o1 accuracy of 9.1% on the inspected HLE evaluation. Approximately 13,000 questions that stumped the checked models were forwarded to expert human review. The appendix describes two rounds of human review for submitted questions. Reading boundary: The inspected v1 describes 3,000 questions; this is a January 2025 snapshot, not current leaderboard performance.

Sources: [354]

Evidence: Reviewed-work atlas - Task construction and evaluation: JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. The JailbreakBench paper states that proper test-time-defense evaluation should use attacks adapted to the defense; transfer attacks from undefended models give only a lower bound on worst-case attack success. Appendix H of the pinned JailbreakBench paper states that all code and data are released under the MIT license. The study compares automatic jailbreak judges on 200 jailbreak prompt-response pairs labeled by three experts, using majority labels, plus 100 benign XS-Test examples. JBB-Behaviors contains 100 harmful behaviors and 100 corresponding topic-matched benign behaviors used to assess excessive refusal. Reading boundary: This records the paper's statement; separate current repository and upstream-rights checks were not performed.

Sources: [225]

Evidence: Reviewed-work atlas - Task construction and evaluation: MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. The authors state that reuse of high-level solution strategies is difficult to detect. Manual screening selects 75 Kaggle competitions, with seven additional development competitions. The headline metric is the proportion of attempts earning at least a bronze medal under the applied private-leaderboard thresholds. o1-preview with AIDE earns any medal on 16.9% ± 1.1 standard error of attempts. Submissions with Dolos similarity above 60% are surfaced for human plagiarism review. Reading boundary: 75 competitions with 16 seeds; 24-hour/A10 main setup. This is a historical normalized competition comparison, not awarded Kaggle medals.

Sources: [363]

Evidence: Reviewed-work atlas - Task construction and evaluation: MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. GPT-4o (0513) scores 51.9 overall, averaging 54.0 in ten-option standard form and 49.7 in vision form. Human experts use GPT-4o assistance to augment answer options and review original question-image relevance. The final 3,460 presentations comprise 1,730 standard questions and 1,730 screenshot/photo forms. The main results select the higher score from Direct and chain-of-thought prompting. Reading boundary: Paired underlying questions; the authors report the higher result from Direct and CoT prompts, so this is not a single fixed-prompt evaluation.

Sources: [364]

Evidence: Reviewed-work atlas - Task construction and evaluation: MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. The MMMU study extracts answers using response-processing workflows; if no valid answer is found, it randomly selects an option for multiple-choice questions and marks open-ended responses incorrect. MMMU annotation guidance favors questions whose answers are less directly accessible, such as in separate documents or textbook endings, as a contamination-risk precaution. MMMU flags potential duplicates using lexical overlap and source-URL similarity, has authors review them, checks formatting and typos, and excludes questions classified as very easy. MMMU selects thirty subjects across six disciplines and recruits over fifty university students specializing in those majors to collect multimodal questions and create new questions where needed. Reading boundary: The guidance is a precaution, not an audit establishing absence of training contamination.

Sources: [232]

Evidence: Reviewed-work atlas - Task construction and evaluation: MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. MathVista uses heuristic filtering of nineteen VQA datasets followed by three expert annotators and majority voting to identify mathematical-reasoning examples, capping each source at 400 examples. MathVista uses an LLM-based answer extractor such as GPT-4, normalizes the extracted answer into the required format, and computes accuracy against reference answers. Each MathVista testmini question is assigned to five AMT annotators with over 5,000 HITs and acceptance above 0.99; included workers must pass qualification questions and have at least a high-school diploma. The pinned MathVista paper defines 6,141 examples partitioned into a 1,000-example testmini and a 5,141-example test, and states that test answer labels will be withheld to reduce contamination.

Sources: [226]

Evidence: Reviewed-work atlas - Task construction and evaluation: Measuring Massive Multitask Language Understanding. The paper estimates expert-level accuracy at approximately 89.8%. GPT-3 evaluation selects the A/B/C/D answer token with the highest model probability. The paper reports 34.5% accuracy for unspecialized human respondents from Amazon Mechanical Turk. Reading boundary: This is an author estimate combining 95th-percentile examination scores with educated guesses where unavailable, not a measured expert panel across every subject.

Sources: [714]

Evidence: Reviewed-work atlas - Task construction and evaluation: MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs. MultiChallenge contains 273 test conversations across four challenge categories. Construction first generates synthetic candidate data and then has human experts review and edit it. Human raters provide an instance-level binary rubric question answerable using the final model response. Self-coherence tests whether a model remains reasonably consistent with its earlier responses. Reading boundary: Categories/counts:inference memory113,instruction retention69,reliable version editing41,self-coherence50; average5 turns and1231.7 words per conversation. Counts describe this benchmark version.

Sources: [395]

Evidence: Reviewed-work atlas - Task construction and evaluation: OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. OSWorld tasks pair instructions with initial-state setup and manually crafted evaluation scripts; the paper reports nine computer-science student authors spending over three months and about 1,800 person-hours on construction and checking. The paper reports 134 unique evaluation functions that use getters to extract final-state information and evaluators to assess task completion. The OSWorld human evaluation uses computer-science-major college students with basic software skills who had not previously seen the samples or software; their time and correctness are measured for each example. The paper states that its evaluation functions mainly assess task-completion correctness and do not efficiently detect latent unnecessary damaging side effects. Reading boundary: A successful task score is not a safety certification.

Sources: [233]

Evidence: Reviewed-work atlas - Task construction and evaluation: PaperBench: Evaluating AI's Ability to Replicate AI Research. Claude 3.5 Sonnet with BasicAgent obtains a mean Replication Score of 21.0% ± 0.8 standard error. The paper reports that Code-Dev performance is weakly correlated with full PaperBench performance. JudgeEval reports o3-mini-high SimpleJudge macro F1 of 0.83 overall and 0.72 for Code Development. PaperBench prohibits agents from using or viewing the original paper authors' codebases. The final Replication Score is computed by propagating weighted averages from binary rubric leaf decisions to the root. Reading boundary: 20 papers, three runs each, at most 12 hours per run; score measures weighted rubric credit, not percentage of fully replicated papers.

Sources: [365]

Evidence: Reviewed-work atlas - Task construction and evaluation: Program Synthesis with Large Language Models. The authors manually inspected, edited and pruned MBPP into a 426-question edited subset; their original-versus-edited evaluation compares 100 overlapping questions. MBPP was constructed by an internal pool of crowdworkers with basic Python knowledge who supplied a problem statement, a self-contained reference function and three semantic test cases per problem; the original collection has 974 problems. The authors' whitespace-normalized line-overlap analysis found 32 of 974 MBPP reference programs with more than half their eligible lines matched in a pretraining document. The paper documents a generated MBPP program that hard-codes a tested input and passes the supplied assertions without implementing the intended function. Reading boundary: The edited subset and original dataset are separate manifestations within one benchmark family.

Sources: [227]

Evidence: Reviewed-work atlas - Task construction and evaluation: RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. The authors report o1-preview with AIDE matching the 36th percentile of human experts in their best-of-k comparison. The human baseline includes 71 eight-hour runs by 61 distinct experts. Human participants are allowed to browse the internet and use LLMs and other tools. The authors warn that clear instructions, short progress horizons and supplied resources make these environments less representative of real research. The normalization sets the starting solution's score to zero and the author reference solution's score to one. Reading boundary: Seven environments; sampled shorter attempts use a shared total time budget. This percentile does not establish end-to-end research automation.

Sources: [366]

Evidence: Reviewed-work atlas - Task construction and evaluation: Reinforcement Learning on Web Interfaces Using Workflow-Guided Exploration. The Alaska flight-form experiment ports a web page into the MiniWoB framework, replaces its server backend with a JavaScript surrogate and fixes the environment date to March 1, 2017. The paper introduces MiniWoB++ tasks to test longer horizons, soft natural-language reasoning and stochastically varying layouts. The workflow-guided-exploration study selects forty tasks from the then-public eighty-task MiniWoB distribution and collects ten demonstrations per selected task using Amazon Mechanical Turk. For the paper main MiniWoB experiments, partial rewards are disabled so tasks use sparse end-of-episode success or failure feedback. The method induces workflows from demonstrations using a compositional language that constrains click/type actions and adds shortcut steps for unnecessary or collapsible demonstration actions. Reading boundary: This experiment does not demonstrate successful live flight booking or purchase.

Sources: [231]

Evidence: Reviewed-work atlas - Task construction and evaluation: RewardBench 2: Advancing Reward Model Evaluation. Non-Ties subsets score selection of one correct response among four completions, with a 25% random baseline. The authors manually verify every example in the math subset after model-based answer selection and matching. The authors conclude that reward-model benchmark accuracy is not sufficient for strong RLHF training. Section 3 construction prose reports that manual verification and filtering reduce an initial pool of about 3,000 prompts to 1,876 final prompts. Table 3 reports QRM-Gemma-2-27B averaging 76.7 on RewardBench 2. Reading boundary: The same version’s Table 2 caption reports 1,865 prompts, and its six printed domain counts sum to 1,865 (475+160+183+450+495+102). This is an unreconciled source-internal discrepancy. The 1,876 figure is attributed specifically to construction prose, not an independently recounted or resolved corpus total.

Sources: [367]

Evidence: Reviewed-work atlas - Task construction and evaluation: SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks?. Human augmentation adds task requirements grounded in original sources and gold-patch/test expectations. Table 1 reports GPT-5 resolving 23.3% of the 731 public tasks with SWE-Agent. The inspected version releases a public split of 731 problems. The authors acknowledge that valid software solutions may fail the original verification tests. The reported test-verification pipeline removes tests judged irrelevant to the task or too broad. Reading boundary: Historical Pass@1 result with augmented problem statements, requirements and interface; no present-day ranking.

Sources: [375]

Evidence: Reviewed-work atlas - Task construction and evaluation: SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains?. The authors report an average of ten hours of manual environment construction per repository. Table 3 reports SWE-agent M with GPT-4o resolving 12.2% of task instances. Human participants manually inspect and categorize all images using categories derived from a random 50-image sample. The limitations identify expansion beyond JavaScript as future work rather than measured coverage. Appendix B.2 describes running tests five consecutive times and filtering inconsistent task instances. Reading boundary: The body describes 517 test tasks and specifies gpt-4o-2024-08-06; total collection counts conflict elsewhere in the source.

Sources: [374]

Evidence: Reviewed-work atlas - Task construction and evaluation: SWE-bench: Can Language Models Resolve Real-World GitHub Issues?. The authors manually create per-release executable contexts by examining repository source and documentation. The authors state that SWE-bench collection, evaluation and experiments do not use crowdsourcing or recruited human task workers. Reading boundary: This manual configuration supports the automated collection/evaluation pipeline. It is not crowdsourced task annotation or an external supplier contribution.

Sources: [733]

Evidence: Reviewed-work atlas - Task construction and evaluation: ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. Nine subject-matter experts validate the tasks before a separate annotator execution-verification round. The authors select Python programs because their annotators are not familiar with R, Stata or Matlab. Table 3 reports Claude 3.5 Sonnet Self-Debug success of 32.4% without expert knowledge and 34.3% with it. Success Rate is zero when a generated program fails execution or does not save the required output. Reading boundary: 102 task goals; SR differs from valid-execution rate and rubric score; historical model/framework comparison.

Sources: [368]

Evidence: Reviewed-work atlas - Task construction and evaluation: ScienceWorld: Is your Agent Smarter than a 5th Grader?. The study reports all agent results using easy-mode simplifications and provides gold trajectories from thirty hand-coded oracle agents representing canonical solution methods. ScienceWorld tasks combine method-agnostic required goals with two to fifteen optional reward-shaping subgoals; task scores are normalized between zero and one. ScienceWorld is an object-oriented interactive text simulator implemented in Scala with a Python interface, with explicit object, action and simulation-engine abstractions. Within each ScienceWorld subtask, generated variations are split 50% training, 25% development and 25% test, with critical unseen variations placed in development and test.

Sources: [253]

Evidence: Reviewed-work atlas - Task construction and evaluation: ShieldAgent: Shielding Agents via Verifiable Safety Policy Reasoning. At inference,the system extracts the invoked action and retrieves corresponding action-rule circuits for verification. Circuit weights are learned from safety-labeled real-world data or simulated pseudo-learning. The policy-construction process extracts and refines rules to make them concrete,accurate and atomic. The decision thresholds the probability difference between invoking an action and taking no action,with an adjustable safety threshold. Reading boundary: A shielding plan assigns Boolean state-predicate values through tools/operations; correctness depends on those assignments.

Sources: [397]

Evidence: Reviewed-work atlas - Task construction and evaluation: SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks. The authors say observed skill gains may partly reflect added context and call for length-matched controls. The complete task list contains 86 tasks, with two exclusions leaving 84 evaluated tasks. Table 3 reports Gemini CLI with Gemini 3 Flash at 31.3% without Skills and 48.7% with curated Skills. The quality criteria reject tasks judged solvable without procedural guidance. The task-quality rules require deterministic programmatic verification rather than an LLM judge. Reading boundary: 84-task evaluation, five trials per task; result belongs to the model-harness pair.

Sources: [369]

Evidence: Reviewed-work atlas - Task construction and evaluation: Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. During auditing, an author manually inspects the exploit agent trajectories regardless of pass or failure. The paper identifies internet access as a source of changing dependencies and potential oracle-solution leakage. Each task has an oracle solution script used to demonstrate that its tests can pass. The reported construction process selects 89 Terminal-Bench 2.0 tasks from 229 contributions by 93 contributors. The experiments report 32,155 single-task trials, with at least five benchmark runs per supported model-agent combination.

Sources: [377]

Evidence: Reviewed-work atlas - Task construction and evaluation: The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. The paper’s controlled data ablation models have 1.71 billion parameters. Reading boundary: They are trained on equal token budgets, commonly 28B or 350B, and compared on selected academic benchmarks. The conclusion (p.10) notes scale and unaligned benchmark-only evaluation limitations.

Sources: [734]

Evidence: Reviewed-work atlas - Task construction and evaluation: TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. Claude 3.5 Sonnet with the common OpenHands CodeAct scaffold achieves 24.0% full task success. The authors report approximately 3,000 person-hours from 20 task creators over two months. Full completion is a binary score requiring every checkpoint to pass. The authors state that resource limits prevented a human-professional performance comparison in this iteration. Reading boundary: 175 tasks; this is full completion, not the 34.4% partial-credit aggregate. Simulated colleagues and evaluators use a fixed Claude backbone.

Sources: [334]

Evidence: Reviewed-work atlas - Task construction and evaluation: VerifiAgent: a Unified Verification Agent in Language Model Reasoning. The framework returns a Correct/Incorrect result and can provide an error reason and a possible revision method. VerifiAgent uses frozen LLMs and does not require a verifier training process. Meta verification rewrites a solution into known conditions,the final objective and explicit logical steps. Adaptive verification selects external tools,solves the problem and compares resulting answers against the original solution. Reading boundary: A separate confidence score is described but its mathematical derivation is outside this extraction. Feedback incorporates tool observations when requested by the prompt.

Sources: [399]

Evidence: Reviewed-work atlas - Task construction and evaluation: Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. Gemini 1.5 Pro scores 75.7% without subtitles and 81.6% with subtitles on the overall Video-MME evaluation. Authors annotate three four-option questions for each of 900 videos, producing 2,700 question-answer pairs. Gemini 1.5 Pro receives all subtitles while other evaluated models use subtitles corresponding to sampled frames. A different annotator reviews each question-answer pair before model-based question-only filtering and revision. Reading boundary: 2,700 questions; the model has denser video-frame sampling and full subtitles under the reported settings. This is not an equal-input causal comparison to other models.

Sources: [380]

Evidence: Reviewed-work atlas - Task construction and evaluation: VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. The VisualWebArena human evaluation uses seven college students familiar with commercial site analogues; participants who helped create tasks were not assigned their own tasks. VisualWebArena tasks and reward functions were curated by six graduate-student coauthors from intent templates; the paper reports 314 templates and includes 46 unachievable tasks. VisualWebArena reward primitives combine exact/inclusion/exclusion string checks, LLM semantic matching, VLM question answering, image similarity and final-webpage-state evaluation. VisualWebArena baseline settings vary observation budgets: the study generally uses 1280-by-2048 viewports and 3,840 text tokens, but shorter-context models use 1280-by-720 and 640 tokens.

Sources: [256]

Evidence: Reviewed-work atlas - Task construction and evaluation: WebArena: A Realistic Web Environment for Building Autonomous Agents. Information-seeking evaluation uses exact-match, required-key-concept and GPT-3.5-turbo-0613 fuzzy-match functions; navigation/content evaluation checks relevant website states programmatically. Reading boundary: Functional evaluation is not uniformly model-free, and information-answer scoring emphasizes key-concept recall.

Sources: [91]

Evidence: Reviewed-work atlas - Task construction and evaluation: WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. WebShop instructions are written by human annotators from target-product attributes, options and price constraints; reward compares the selected product against these fields. The WebShop paper divides 12,087 instructions into an iid train/development/test split of 10,587, 1,000 and 500 instructions. In the pinned WebShop study, human scores are collected in HTML mode while models are trained and evaluated in simplified text mode. WebShop manual reward verification finds that automatic exact matching can over-penalize lexically similar or synonymous product attributes compared with human re-scoring.

Sources: [257]

Evidence: Reviewed-work atlas - Task construction and evaluation: World of Bits: An Open-Domain Platform for Web-Based Agents. In FormWoB, an unseen HTTP request terminates the episode with zero reward. The WoB agent observes rendered pixels, localized DOM elements and a scalar reward, and acts through keyboard or pointer events. MiniWoB comprises 100 web tasks with manually constructed reward functions. QAWoB contains 11,650 queries instantiated from 521 templates. Reading boundary: Proxy records demonstration HTTP requests/responses; reward otherwise measures the fraction of request key-value pairs matching demonstrations. This is an offline approximation, not a live-site task success guarantee.

Sources: [400]

Evidence: Reviewed-work atlas - Task construction and evaluation: τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. τ-bench combines manually designed schemas, APIs and policies with GPT-4-generated data-sampling code, human code refinement, and manually annotated tasks iteratively checked with agent runs. τ-bench agents interact with database APIs and a simulated user under a domain-specific policy; database transitions are deterministic Python functions. τ-bench defines pass-hat-k as the probability that all k independent trials succeed, averaged across tasks, whereas pass@k measures at least one successful trial. The τ-bench rule-based reward checks the final database and required output strings, but the paper notes that a positive reward can still miss policy violations such as absent user confirmation.

Sources: [254]

Evidence: Reviewed-work atlas - Training systems and orchestration: HybridFlow: A Flexible and Efficient RLHF Framework. The single controller passes data futures between roles, while required output data is transferred directly between GPUs according to parallel ranks. Although ResourcePool supports parallel execution of colocated roles, HybridFlow generally uses sequential execution to avoid contention or out-of-memory failures. HybridFlow's ResourcePool virtualizes GPU sets; models sharing a pool are colocated, while different pool instances map to distinct GPU sets. HybridFlow attaches collect/distribute transfer protocols to model operations so data can be reshaped between models with different parallelism layouts. Reading boundary: The example is actor generation to critic inference; actual transfer and scheduling behavior depends on the chosen protocols.

Sources: [408]

Evidence: Reviewed-work atlas - Training systems and orchestration: OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework. The selected OpenRLHF paper limits its support scope to language models and states that vision-language or other multimodal architectures are not supported. OpenRLHF uses Ray to schedule separate response-generation and model-training roles and manage their workflow and data flow. OpenRLHF uses ring attention for sequence parallelism, distributing long-sequence attention computation across GPUs. The described pipeline uses vLLM for rollouts and DeepSpeed for actor/value training, transferring model slices between training and rollout engines. Reading boundary: This is a dated limitation of the pinned version, not a live audit of current source code; Ray/vLLM/DeepSpeed updates may also require compatibility work.

Sources: [412]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: AutoLibra: Agent Metric Induction from Open-Ended Human Feedback. AutoLibra turns open-ended feedback into a metric-and-judge pipeline. The inspected v3 uses 197 CoGym end-user trajectories with feedback and paper-author experts elsewhere; five authors act as expert annotators. Sotopia, WebArena and WebVoyager each contribute 100 GPT-4-based trajectories, with 20% held out for the Section 4 analysis. GPT-4o grounds feedback into signed behavioral aspects, o3-mini high clusters those aspects, and o3-mini medium rates trajectories +1, −1 or N/A against definitions and positive/negative examples.

Sources: [421]

Evidence: Study result - AutoLibra: Agent Metric Induction from Open-Ended Human Feedback. The human-alignment check samples 40 instances per validated step and task for binary judgments of complete correctness. Table 1 reports averages across five environments of 0.95 for grounding, 0.92 for LLM judging and 0.90 for meta-evaluation. These are step-specific validation summaries. Behavior clustering itself was not directly manually validated because reviewing more than 400 aspects was considered too costly.

Sources: [421]

Analysis: Study limitations and buyer implications - AutoLibra: Agent Metric Induction from Open-Ended Human Feedback. For a buyer, the deliverable is an induced rubric and trajectory judge whose intermediate transformations need their own checks. The demonstrated protocol collects one feedback element per complete text-rendered trajectory. Live or direct step feedback, generated programmatic evaluators and dense-reward RL remain future directions. The validation figures therefore cannot be transferred to those proposed products or read as an end-to-end accuracy guarantee for every induced metric.

Sources: [421]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. The inspected Aya v1 describes a sequence of original contribution, re-annotation and annotation-feedback tasks. Original examples receive review during re-annotation, and the resulting re-annotations receive feedback. Contributor leaderboard quality averages reviewer ratings. Inclusion also has distinct rules: all original annotations enter, while re-annotations need a summed prompt/completion character-level Levenshtein distance of at least five; a language needs at least 50 contributions.

Sources: [422]

Evidence: Study result - Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. The paper reports 204,114 prompt/completion instances in 65 languages: 138,844 original annotations and 65,270 re-annotations. Its resource grouping is 22 high-resource, 12 mid-resource and 31 low-resource languages. These are statistics of the dataset described by that exact paper, rather than a count independently recomputed from a current downloadable artifact. Original contributions and revised versions remain distinguishable components of the total.

Sources: [422]

Analysis: Study limitations and buyer implications - Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. Multilingual breadth does not establish equal validation depth in each language. The authors identify contributor concentration in some languages, no dedicated toxic/offensive-speech flags and no language-relabeling component; peer review does not guarantee removal of every unsafe or misclassified example. A proposed purchase should inspect language-specific contributor concentration and error samples. The paper review adds one reviewed work here without independently adding another inspected dataset family or implying a named commercial supplier.

Sources: [422]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: CrowdAgent: Multi-Agent Managed Multi-Source Annotation System. CrowdAgent combines cheap model annotation, small-model denoising and human routing. Core-Set selection sends low-confidence, diverse cases to humans, while budget, round and confidence conditions govern stopping. A hidden golden set informs the QA agent’s Bayesian confusion-matrix estimates; the source says these labels are neither exposed to annotators nor supplied as in-context examples. Difficult residual cases are flagged for final human verification.

Sources: [423]

Evidence: Study result - CrowdAgent: Multi-Agent Managed Multi-Source Annotation System. The main table places CrowdAgent above its listed baselines on six multimodal classification tasks at fixed human-label proportions: 5% on MM-IMDb and COV-CTR, and 15% on the three CrisisMMD subtasks and V-SNLI. The matched quantity is the human annotation proportion. It is not a controlled equality of total operating cost, model computation or all system-selection effort; appendix training and budget details were not fully examined in this bounded review.

Sources: [423]

Analysis: Study limitations and buyer implications - CrowdAgent: Multi-Agent Managed Multi-Source Annotation System. A buyer can use this design to ask how uncertain cases reach people and how hidden checks remain hidden. Confidence-threshold convergence is an internal stopping condition, rather than proof of error-free output. The authors acknowledge residual errors and dependence on underlying models. Acceptance should retain independent error inspection after convergence and a separately stated total-cost boundary before any efficiency claim is adopted.

Sources: [423]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. CrowdTruth 2.0 represents a worker response as a binary annotation vector. Closed tasks share a predefined answer space; an open task instead builds each media unit’s answer space from the workers’ unique keywords. Its worker, media-unit and annotation scores are mutually dependent, start at one and are repeatedly recalculated until total variation falls below a selected threshold, using disagreement-weighted cosine agreement.

Sources: [424]

Evidence: Study result - CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. The worker-quality definition multiplies worker-worker agreement by worker-media-unit agreement. Peer consistency and consistency with the aggregate of other workers both affect the score. The inspected paper supplies this operational definition rather than a new controlled accuracy improvement. In open-ended tasks, annotation quality cannot be estimated across multiple media units and is set to one; the model therefore has different information in open and closed settings.

Sources: [424]

Analysis: Study limitations and buyer implications - CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. The useful comparison is between assumptions about disagreement. A low consistency score can prompt examination of the worker, the item or the answer space; the formula alone cannot decide which is defective. Further comparative validation across tasks remains future work in this source. Buyers should test whether a proposed weighting rule preserves intended minority interpretations and improves independently checked outcomes before using it to exclude contributors or certify labels.

Sources: [424]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. This earnings study uses plugin logs from September 2014 through January 2017: 3,808,020 HIT records from 2,676 workers before duration-outlier removal, and 3,471,580 HITs from 2,666 workers afterward. Zero-reward qualification tasks remain included. Its accounting compares reward over accept-to-submit intervals with temporal clustering at D=0 and D=1 minute to address overlapping queued work and between-task time; returned HITs contribute zero reward.

Sources: [425]

Evidence: Study result - A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. For the filtered data, Table 1 reports median work-unit hourly rates of $2.54 per HIT, $3.18 per D=0 cluster and $1.77 per D=1-minute cluster. Their denominators are respectively 3,471,580 HITs, 2,560,066 clusters and 635,198 clusters. These are medians across task or cluster units, not a median across workers. They are historical study figures and do not describe current compensation at any supplier in this report.

Sources: [425]

Analysis: Study limitations and buyer implications - A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. The operational lesson is to expose the time denominator, including unpaid qualification, returned work and between-task intervals where the project records them. Plugin selection, unknown demographics and script use limit population inference, while elapsed intervals need not equal active effort. Associations between work patterns and earnings are not causal. Neither the historical rates nor a selected plugin sample establishes current worker welfare or a commercial service’s labor cost.

Sources: [425]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Experts Rise Where LLMs Disagree: Using Cross-Model Disagreement to Target Expert Effort in LLM Codebook Revision for Large-Scale Annotation. Experts Rise Where LLMs Disagree prioritizes expert attention using normalized vote entropy from six model predictions and a vote-weighted mean by label across 6,595 unlabeled tutoring utterances. Disagreement is a routing signal, not a determination that any label is wrong. Two experts each supply rationales for 60 selected utterances; GPT 5.6 Sol uses the resulting 120 annotations to revise 24 codebook entries. Mean reported Rationale Labeling time is 169.0 minutes per expert.

Sources: [426]

Evidence: Study result - Experts Rise Where LLMs Disagree: Using Cross-Model Disagreement to Target Expert Effort in LLM Codebook Revision for Large-Scale Annotation. On the same 4,032 expert-consensus held-out utterances, Rationale Labeling reaches 64.9% accuracy and 0.658 weighted F1, compared with 57.8% and 0.583 for the final expert-only codebook. The summary averages eight alternative aggregation methods using one temperature-zero run per codebook from each of six models. Aggregators reuse predictions, so they are not independent trials; 6.1% multi-label references are reduced to the first recorded label.

Sources: [426]

Analysis: Study limitations and buyer implications - Experts Rise Where LLMs Disagree: Using Cross-Model Disagreement to Target Expert Effort in LLM Codebook Revision for Large-Scale Annotation. The result supports a scoped codebook-revision comparison against the chosen reference label. One longitudinal project, two experts, unequal workflow effort/exposure, differing target-label assignments and possible carryover prevent isolation of input format as the cause. Transcript errors were not systematically measured or corrected. A buyer should preserve selection rules, rationale time and held-out label conventions before expecting similar returns from expert disagreement review in another domain.

Sources: [426]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?. This reporting audit began with five students independently coding the same papers and weekly discussion of disagreements; the team leader decided unresolved cases. After schema changes, at least two team members independently re-examined each of the 164 papers and the team reconciled them again. The repeated examination documents how this audit established its own coded reporting reference rather than assuming an initial extraction was gold.

Sources: [427]

Evidence: Study result - Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?. Within the 74 papers coded for original human annotation, Table 5 finds no original-annotator identity information in 18 (24.32%) and paper-author annotation in 22 (29.73%). Table 7 finds no formal-instruction information in 35 (47.30%), formal definitions/examples in 32 (43.24%), and only question text as instructions in seven (9.46%). These are categories of publication disclosure in the sampled literature.

Sources: [427]

Analysis: Study limitations and buyer implications - Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?. The study audits what publications report, rather than observing the underlying labor process. Its Twitter-classification focus and mostly arXiv sampling limit transfer to other fields; it did not record actual agreement values or adjudication/reconciliation processes in those projects. The procurement implication is a request for missing evidence, not an inference that undocumented instruction, compensation or QA never occurred. Disclosure completeness and delivered label quality require different measurements.

Sources: [427]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. The inspected v2 compares the same codebook wording across trained political-science students, screened US MTurk workers and gpt-3.5-turbo. Model annotations were collected in March-May 2023, using a fresh chat and one tweet per query, with two runs at each temperature of 1 and 0.2. Crowd recruitment required Masters status, over 90% approval, at least 50 approved HITs, two workers per tweet and no worker labeling over 20% of a task’s tweets.

Sources: [428]

Evidence: Study result - ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Across the four study datasets, the paper reports ChatGPT accuracy about 25 percentage points above MTurk on average, defining accuracy against trained-human labels only where both trained annotators agreed. The difficult items with trained-rater disagreement are outside that gold denominator. Mean within-group agreement is about 56% for MTurk, 79% for trained annotators, 91% for ChatGPT at temperature 1 and 97% at 0.2.

Sources: [428]

Analysis: Study limitations and buyer implications - ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. This is a dated, task-specific model-versus-recruitment comparison, not a universal replacement rule for annotation labor. Repeated model consistency does not establish truth or independence. The source records a relevance-task failure caused by missing prompt examples and leaves other languages and models for future evaluation. A buyer should inspect excluded disagreement cases and prompt sensitivity alongside the retained-set accuracy before deciding how much human review can be removed.

Sources: [428]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. HelpSteer v1 independently rates each response on helpfulness, correctness, coherence, complexity and verbosity using a 0-4 Likert scale. It collects scalar attributes rather than pairwise preferences. The paper describes approximately 200 US-based annotators engaged through Scale AI, English screening, training ending in 35 sample responses, at least two human QA reviews per annotation, automated checks and later research-team QA. This documents a reported workflow rather than an independent supplier audit.

Sources: [429]

Evidence: Study result - HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. Within the final 37,120 annotated samples, Table 2 reports Pearson correlation with helpfulness of 0.8525 for correctness and 0.6348 for coherence. These are descriptive associations among the collected 0-4 ratings. They do not estimate how helpfulness would change if a buyer causally changed correctness or coherence, and they should not be substituted for an independently measured improvement from any individual QA stage.

Sources: [429]

Analysis: Study limitations and buyer implications - HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. A multi-attribute contract can retain distinctions hidden by a single preference label, but its scales still need task and population calibration. All annotations here concern English prompts/responses and all annotators are US-based; the authors state uncertainty about other cultural judgments and non-English gains. The exact recruitment statement can support the scoped role it names, without extending it to every HelpSteer release, sole artifact authorship or current provider performance.

Sources: [429]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: HelpSteer2: Open-source dataset for training top-performing reward models. HelpSteer2 v1 initially assigns three raters per response, adding two if the helpfulness range exceeds two points; it reports 3.41 annotators per sample and approximately 1,000 US-based annotators engaged through Scale AI. The pipeline retains the three most-agreeing ratings from up to five, adds QA and removes responses still exceeding the spread rule. About half of collected annotations are excluded overall, leaving 21,362 responses for 10,681 prompts.

Sources: [430]

Evidence: Study result - HelpSteer2: Open-source dataset for training top-performing reward models. Helpfulness quadratic-weighted Cohen κ rises from 0.465 initially to 0.791 after postprocessing, a multistage selection-and-QA comparison. The two-epoch regression reward models use batch size 128, validation-loss checkpoint selection and model-specific learning-rate search. Reported RewardBench overall accuracy is 88.8% for Llama 3 70B and 92.0% for Nemotron-4 340B: the mean of four categories over 2,985 tasks, excluding optional Prior Sets. These are unreproduced results of the selected version, not current ranks.

Sources: [430]

Analysis: Study limitations and buyer implications - HelpSteer2: Open-source dataset for training top-performing reward models. Higher retained-set agreement cannot be read as a randomized estimate of one QA intervention or evidence that excluded preferences were invalid. English-only, exclusively US-based annotation, possible cultural bias, uneven attribute combinations and compute-dependent tuning narrow transfer. Buyers should cost the discarded annotations, preserve disagreement reasons and validate a new target population. Conclusions from the large models do not automatically extend to smaller models.

Sources: [430]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Inherent Disagreements in Human Textual Inferences. This study samples 100 short premise/hypothesis pairs from each of RTE2, SNLI, MNLI, JOCI and DNC, requesting 50 independent AMT ratings per pair on a −50 to 50 likelihood slider. Sentences have at most 20 words. After removing specified sparse/outlier workers, does-not-make-sense responses and pairs with fewer than 15 judgments, it retains 496 pairs averaging 39 labels each. Ratings are z-normalized within worker assignments.

Sources: [431]

Evidence: Study result - Inherent Disagreements in Human Textual Inferences. Using ten held-out judgments per pair and averaging ten random splits, the Gaussian-mixture analysis finds a second component with weight above 0.2 for about 20% of pairs; most pairs remain consistent with a single component. That is a model-based description of these collected, normalized judgments. It is not a universal estimate of linguistic ambiguity or an accuracy rate against an independently established single true interpretation.

Sources: [431]

Analysis: Study limitations and buyer implications - Inherent Disagreements in Human Textual Inferences. A buyer can preserve distributions for ambiguous tasks rather than force all cases into a single scale. The authors explicitly warn that normalized scores are not comparable across batches containing different pairs; the meaningful analysis compares raters within the same pair. A data pipeline that merges these normalized numbers across batches without the assignment structure would change the measurement. Recruitment and filtering rules also remain part of the population being described.

Sources: [431]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems. LabelAId combines domain-coded labeling functions in a Snorkel probabilistic label model, pretrains an FT-Transformer and fine-tunes on expert-validated city labels to infer image-label correctness. Its proof of concept concerns urban-accessibility imagery. The 34-person university/snowball study obtains consent and pays $30 per hour; two researchers independently validate all 3,574 participant labels, with a third resolving 74 disagreements to define the precision reference.

Sources: [432]

Evidence: Study result - LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems. The between-subjects Project Sidewalk study reports higher overall precision in the intervention group by Mann-Whitney U ($p \leq 0.01$). Route-completion time does not differ significantly at $p=0.693$; reported means are 2,303.3 seconds for control and 2,801.4 for intervention. The nonsignificant test does not establish equal speed. The experiment and its label adjudication were not reproduced for this report.

Sources: [432]

Analysis: Study limitations and buyer implications - LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems. The useful intervention target is a likely labeling mistake at the moment of work. The implementation cannot identify the exact cause and supplies neither explainability features nor global accuracy differences by label type. Its small university/snowball sample may differ from the wider user population. A buyer should separately test precision, time, error categories and participant representativeness rather than infer a general productivity gain from this one proof of concept.

Sources: [432]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. The inspected v4 first screens possible benchmark-label errors using confident learning with out-of-sample class probabilities and observed labels. Its noise model assumes class-conditional noise; datasets without test splits use whole-dataset cross-validation. Five MTurk workers then independently choose the given label, predicted label, both or neither. Fewer than three accepting the given label triggers the study's error rule; QuickDraw and Amazon Reviews use random samples of flagged cases, while the other flagged sets are fully checked.

Sources: [433]

Evidence: Study result - Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. For ILSVRC 2012, unavailable ImageNet test labels are replaced by the 50,000-image validation set. Table 1 reports 5,440 screened and checked examples, of which 2,916 meet the five-worker rule, or 5.83% of that validation set. These validated issues include ambiguous categories. They are not all cases with a uniquely recoverable replacement label, and they do not give an error estimate for a newly purchased dataset.

Sources: [433]

Analysis: Study limitations and buyer implications - Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. Gold-label review is itself a measurement process. The authors caution that even unanimous five-worker corrections can be wrong. Corrected-accuracy experiments remove unknown-label cases and change only the consensus alternative-label set, producing a different denominator from initial validated issues. Buyers should retain the screening candidate, uncertainty, adjudication and final usable correction as separate states rather than present every flagged example as a confirmed, correctable error.

Sources: [433]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Efficient Data Labeling by Hierarchical Crowdsourcing with Large Language Models. LMCrowd starts with Llama2-7B, FLAN-T5-base, Zephyr-7B and Qwen-7B labels, then buys GPT-4 relabeling for uncertain distributions. The evaluated models use four-shot prompts, and the paid labels remain uncertain inputs rather than unquestioned gold. The main RTE, BoolQ and COPA experiments use majority voting, approximate query selection and two relabeled tasks per iteration; BoolQ is sampled to 1.5K cases.

Sources: [434]

Evidence: Study result - Efficient Data Labeling by Hierarchical Crowdsourcing with Large Language Models. The paper plots annotation accuracy against budget and describes its reported results as maximum achieved accuracy across configurations, rather than repeated-trial averages. Tariff comparisons count API or annotation charges, excluding model selection, GPT-4 prompt selection and worker-selection overhead. The human comparators are three university and six graduate students. Historical tariff assumptions are not current verified prices and do not establish compute-cost parity for local deployment.

Sources: [434]

Analysis: Study limitations and buyer implications - Efficient Data Labeling by Hierarchical Crowdsourcing with Large Language Models. The study can inform a staged-label purchasing rule, but its cost boundary cannot be used as a complete service quote. Wider domains, datasets and model choices remain future evaluation beyond the three tasks and limited generic-model set. A buyer should reproduce selection overhead, human qualifications and uncertainty handling within the intended workload, then compare accepted-unit costs under a common end-to-end scope.

Sources: [434]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Quality Assured: Rethinking Annotation Strategies in Imaging AI. The inspected surgical-instrument segmentation study collects 57,648 instance masks on 4,050 unique laparoscopic frames, using 216 annotators and 34 QA workers at four unnamed annotation companies plus 708 MTurk workers. Masks are repeated annotation units, not distinct images. Its three instruction conditions contain 168 words/seven slides, 446 words/ten slides and 961 words/16 slides with pictures. Annotators and company QA workers see the same instructions.

Sources: [435]

Evidence: Study result - Quality Assured: Rethinking Annotation Strategies in Imaging AI. The researchers compare outputs before and after QA with Dice, normalized surface distance and severe-error counts. Under extended text or text-plus-picture instructions, two companies have no median-Dice change after QA and the other two have relative gains below 2%; the main figure describes larger gains from better instructions than additional QA in this project. Figure pixels were not separately inspected in this bounded review.

Sources: [435]

Analysis: Study limitations and buyer implications - Quality Assured: Rethinking Annotation Strategies in Imaging AI. QA headcount alone is an inadequate comparison when instructions and acceptance constructs differ. The reported effects are conditional on one safety-critical domain and these project conditions; costs can change elsewhere. Confidential company identities prevent any mapping to the named suppliers profiled here. The study cannot establish their service quality or working conditions, and it does not show that QA is generally unnecessary.

Sources: [435]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: RECODE-H: A Benchmark for Research Code Development with Interactive Human Feedback. RECODE-H uses 26 PhD-level researchers with relevant top-conference publications to review Gemini-2.5-Pro drafts of code comments, instructions and tests. Included research implementations must execute and require less than 24 GB GPU memory. At evaluation time GPT-o4-mini simulates an expert using execution results, canonical code and annotated comments. Its five feedback levels progress from logs through diagnosis and repair guidance to the correct code snippet.

Sources: [436]

Evidence: Study result - RECODE-H: A Benchmark for Research Code Development with Interactive Human Feedback. The seven-model study fixes temperature 0 and top-p 1, permits at most ten feedback rounds and three agent actions per turn, and reports correctness and task-solve gains with richer feedback. Those gains are conditional on the added information, including answer-code snippets at the highest level. MRR depends on the first test-correct turn; Recall@n counts tasks solved within n turns; CodeBLEU and CodeBERTScore compare to canonical implementations.

Sources: [436]

Analysis: Study limitations and buyer implications - RECODE-H: A Benchmark for Research Code Development with Interactive Human Feedback. Expert-reviewed task construction and live expert feedback are distinct provenance claims. The evaluated feedback is model-generated, and the conclusion leaves actual human-in-the-loop feedback and broader research stages for future work. Temperature zero is a decoding setting, not proof of deterministic remote responses. A buyer should specify what the feedback provider can see, whether it reveals solution material, and which task-success metric is being purchased before comparing assistance levels.

Sources: [436]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Across five selected NLP tasks, this study obtains ten independent annotations per item with multiple-choice or fixed-range numeric responses. Its temporal-order experiment deliberately narrows the original noun-and-verb, fourteen-relation problem to verb events and two relations. Expert interannotator-agreement figures are unavailable for that simplified task. The restriction changes the comparison that a headline expert-versus-crowd result might otherwise suggest.

Sources: [437]

Evidence: Study result - Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. On 462 TimeBank verb-event pairs restricted to strictly before versus strictly after, ten-label voting yields 0.94 agreement with dataset labels from 4,620 annotations. The cited 0.77 expert agreement comes from the broader task and is not a matched comparator. Separately, a unigram affect predictor trained on 100 headlines and tested on 900 reaches mean Pearson correlation 0.191 with ten-annotator-averaged non-expert training labels versus 0.122 with one expert, using five other experts as reference.

Sources: [437]

Analysis: Study limitations and buyer implications - Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Redundancy, task simplification and reference construction explain the actionable scope. These results support testing pooled labels for similarly bounded tasks; they do not establish that ten general annotators outperform specialists on arbitrary work. A procurement comparison should hold the ontology and evaluation reference fixed and cost all repeated labels. The historical experiment is not a current provider or platform performance measurement.

Sources: [437]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: UltraFeedback: Boosting Language Models with Scaled AI Feedback. UltraFeedback v2 samples four different models per instruction from a 17-model pool, yielding 255,864 completions for 63,967 instructions. Family/size diversity and varied system principles are intended to reduce style-quality confounding. GPT-4 jointly scores each prompt’s four completions on 1-5 instruction-following, truthfulness, honesty and helpfulness rubrics with rationales; a separate tutor-style prompt creates an overall critique per completion.

Sources: [438]

Evidence: Study result - UltraFeedback: Boosting Language Models with Scaled AI Feedback. For a 400-comparison check-100 each from UltraFeedback, AlpacaEval, Evol-Instruct and UltraChat-three independent student judges give win/tie/lose labels. GPT-4 agreement averages 59.7% with individual judges and 68.6% with their majority, including ties. Agreement assesses consistency with these judgments rather than verified factual correctness. A later 266-pair evaluation uses a separate filtered denominator and should not be merged with this check.

Sources: [438]

Analysis: Study limitations and buyer implications - UltraFeedback: Boosting Language Models with Scaled AI Feedback. Rubric-rich synthetic feedback still needs independent calibration. The impact statement identifies answer-position, length and style biases that can cause inaccurate or unfair annotation and evaluation. Model diversity is an intended safeguard rather than a proof those biases disappear. Buyers should retain the joint-scoring context and test order/style perturbations; neither the label volume nor majority agreement establishes universal truthfulness or equivalence to independent human judgments.

Sources: [438]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: Who Annotates in NLP? A Large-scale Assessment of Human Annotation Reporting between 2018 and 2025. The inspected study extracts one structured record per annotation experiment using a 26-category taxonomy from keyword-selected 2018-2025 ACL-family main/Findings papers. The selected corpus has 1,603 papers and 2,667 tasks, separate from its human-adjudicated gold set. Gold construction compares two independent annotations, discusses disagreements and uses a third annotator for unresolved cases before evaluating extraction.

Sources: [439]

Evidence: Study result - Who Annotates in NLP? A Large-scale Assessment of Human Annotation Reporting between 2018 and 2025. For the final Gemini-3.1-Pro-extracted corpus, reported information appears for recruitment in 90.4% of tasks, expertise in 86.5% and item volume in 86.0%, compared with training in 18.7%, language proficiency in 24.0% and released guidelines in 34.1%. These percentages describe automated extraction of publication reporting. Their denominator is annotation tasks in the selected corpus, rather than workers, suppliers or all ACL research.

Sources: [439]

Analysis: Study limitations and buyer implications - Who Annotates in NLP? A Large-scale Assessment of Human Annotation Reporting between 2018 and 2025. Documentation omissions are prompts for follow-up, not findings that the underlying practices were absent. Keyword selection is not fully representative, model extraction can misclassify papers or categories, and the released records lack calibrated per-instance confidence. The study makes disclosure comparison useful while preserving uncertainty about the actual work. Its reconciliation design also illustrates why an independently established reference should precede scoring a research extraction system.

Sources: [439]

Evidence: Reviewed-work atlas - Annotation, feedback and evidence design: WildChat: 1M ChatGPT Interaction Logs in the Wild. The inspected WildChat v1 describes public GPT-3.5-Turbo/GPT-4 chat services collected from 9 April 2023 to 1 May 2024. Users must accept a collection/use/sharing agreement and a second confirmation before opening the chat interface. The authors report multilingual PII removal using Presidio, spaCy and custom rules, followed by GeoLite2 country/state mapping and hashed IP release. This is a description of their process, not an independently tested de-identification guarantee.

Sources: [440]

Evidence: Study result - WildChat: 1M ChatGPT Interaction Logs in the Wild. The exact v1 PDF contains an unresolved count discrepancy: its public-release statement gives 1,039,785 conversations after reserving 14,743 for WildBench, while the following Basic Statistics paragraph gives 1,009,245. This report preserves both statements instead of choosing a canonical artifact count. Neither figure has been recomputed from corpus bytes here, and this paper review does not independently add a dataset-family inspection.

Sources: [440]

Analysis: Study limitations and buyer implications - WildChat: 1M ChatGPT Interaction Logs in the Wild. Consent gates and filtering do not establish population representativeness or zero privacy risk. The authors warn that hashed IPs and request headers can still link conversations from the same user; Hugging Face hosting and subreddit dissemination can favor IT communities, while anonymous access may select more toxic discourse. Buyers should inspect the actual release and linkage risk under their use case rather than treat conversational realism, anonymization or a headline million count as a certification.

Sources: [440]

Evidence: Reviewed-work atlas - Annotation engineering and labor: Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions. The 2021 curriculum organized data labeling into three project rounds focused on annotation quality, budget constraints, and changing data over time; students collected data for a specified model. The course originated inside Toloka and combined practitioners’ case studies with hands-on homework; the pipeline components included task decomposition, interface/instruction design, quality control, pricing, and answer aggregation.

Sources: [457]

Evidence: Study result - Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions. For third-year respondents, reported career usefulness rose from 3.16 in 2020 to 3.8 in 2021 on a 1-5 survey scale; the paper reports 35 third-year respondents and also reports mixed fourth-year feedback.

Sources: [457]

Evidence: Study limitations - Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions. The course requires programming and ML/deep-learning preparation; confidential business algorithms and the need to train and coordinate practitioner lecturers constrain transfer of the approach.

Sources: [457]

Analysis: Study assessment and change conditions - Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions. The operational unit is a pipeline-design exercise rather than an annotation-hour purchase. A buyer training internal staff could adapt the three rounds to a live but nonproduction labeling brief and require a reproducible budget, adjudication log and changed-data response. Confidence is moderate in the usefulness of this training design and low in any claimed career effect beyond the surveyed cohorts. A controlled assessment of retained skills and independently scored project outputs would change that judgment.

Sources: [457]

Evidence: Reviewed-work atlas - Annotation engineering and labor: Quantifying and Avoiding Unfair Qualification Labour in Crowdsourcing. The study compared worker performance across 15 combinations of approved-HIT counts and approval percentages on coreference and sentiment tasks, restricting recruitment to U.S.-based MTurk workers. Coreference used a 244-word OntoNotes document with parsed noun phrases; sentiment followed Stanford Sentiment Treebank instructions. The authors targeted $12 per hour and reported worker-reported hourly rates separately.

Sources: [458]

Evidence: Study result - Quantifying and Avoiding Unfair Qualification Labour in Crowdsourcing. In the follow-up coreference condition requiring 99-100% approval and 1,000+ approved HITs, 60 of 92 workers scored at least 80 F-score; the task-level qualification patterns differed between coreference and sentiment.

Sources: [458]

Evidence: Study limitations - Quantifying and Avoiding Unfair Qualification Labour in Crowdsourcing. Qualification effects were task dependent, and workers self-selected by returning HITs; the paper recommends calibrating the lowest suitable threshold and warns that higher approval-percentage requirements can increase vulnerability to unfair rejection.

Sources: [458]

Analysis: Study assessment and change conditions - Quantifying and Avoiding Unfair Qualification Labour in Crowdsourcing. Qualification thresholds should be calibrated to the task rather than inherited as a universal proxy for competence. A paid pilot can compare a low-threshold route plus task-specific training with a stricter route, measuring output quality, exclusions and uncompensated preparation separately. Confidence is moderate for this task-specific caution; it does not establish a globally optimal threshold. Independent replication across languages, worker pools and payment rules would make a broader recommendation possible.

Sources: [458]

Evidence: Reviewed-work atlas - Safety, feedback and verification: Red Teaming Language Models with Language Models. The pipeline generates test inputs with a red language model, obtains target-model replies, and uses a classifier to flag harmful outputs; the offensive-dialogue study used 280B Gopher for red and target models and a 1.4B Gopher classifier fine-tuned on BAD. Zero-shot generation sampled 0.5 million unique valid conversation-starting questions; for method comparison the study used 2,598 questions per method to match the BAD human-adversary reference set.

Sources: [459]

Evidence: Study result - Red Teaming Language Models with Language Models. The 0.5 million zero-shot questions elicited classifier-flagged offensive replies in 3.7% of cases, yielding 18,444 failures; RL with KL penalty 0.3 elicited over 40%, but 78% of its test cases contained “invisible”.

Sources: [459]

Evidence: Study limitations - Red Teaming Language Models with Language Models. Biases in the red model can restrict diversity and flawed red-team classifiers can create false positives or miss failures; the authors present automated red teaming as complementary to manual testing rather than exhaustive discovery.

Sources: [459]

Analysis: Study assessment and change conditions - Red Teaming Language Models with Language Models. An attack generator and its classifier form one coupled measurement system. More flagged failures can come with narrower attack diversity, so procurement should require failure-family counts and independent validation as well as a headline success rate. Confidence is high that classifier error and diversity are distinct acceptance concerns, and unestablished for a current defense. A private human-authored holdout and manual review of both flagged and unflagged outputs would change the practical assessment.

Sources: [459]

Evidence: Reviewed-work atlas - Multilingual and multimodal supervision: Findings of the WMT’22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages. Human evaluators scored two system translations per source sentence on a 0-100 scale combining accuracy and fluency, with annotator-level Z-score normalization. Human evaluation ran on Toloka with Africa-IP and declared-language filters, English/French platform language tests when available, five-item exams, and bans after repeated very fast submissions; non-English/French exam pass threshold was 80%, compared with 100% when a platform language test was available.

Sources: [460]

Evidence: Study result - Findings of the WMT’22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages. Across the 72 language pairs comparable with the previous shared task, the reported average BLEU increased from 15.09 to 22.60, about 7.5 points, with 67 pairs improving.

Sources: [460]

Evidence: Study limitations - Findings of the WMT’22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages. Human-score averages cover a smaller annotator-available subset of the 100 evaluation pairs; the paper also notes that the previous year used the FLORES-101 sentencepiece tokenizer while this year used FLORES-200.

Sources: [460]

Analysis: Study assessment and change conditions - Findings of the WMT’22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages. Human availability, normalization and tokenizer version define separate comparison boundaries. A translation-evaluation purchase should report per-pair human coverage and recruitment criteria instead of allowing a large language-pair count to imply equal human scrutiny everywhere. Confidence is moderate for this design lesson and low for inferring a supplier advantage from the shared-task result. Matched language-pair samples with stable normalization and independently checked reference quality would support a stronger conclusion.

Sources: [460]

Evidence: Reviewed-work atlas - Human simulation and measurement: LLMs Simulate Big5 Personality Traits: Further Evidence. The study administered the 120-item IPIP-NEO questionnaire to GPT-4, Llama-2-70 b-chat and Mixtral-8 x 7B-Instruct using two prompt headers and three model-specific temperature settings, repeating each treatment five times. Questionnaire items elicited one numerical Likert response from each model; one prompt variant added “Answer as if you were a person”, and Llama-2 refused two items in both variants.

Sources: [461]

Evidence: Study result - LLMs Simulate Big5 Personality Traits: Further Evidence. In Table 2, GPT-4 at first-prompt temperature 1 had mean neuroticism 2.18 ± 0.10 and conscientiousness 4.21 ± 0.00 on the 1-5 scale; these are simulated questionnaire scores, not measured human personalities.

Sources: [461]

Evidence: Study limitations - LLMs Simulate Big5 Personality Traits: Further Evidence. The authors call the study a particular empirical case study, say generalization is difficult, and state that the temperature results do not allow conclusive claims of a temperature effect.

Sources: [461]

Analysis: Study assessment and change conditions - LLMs Simulate Big5 Personality Traits: Further Evidence. A model can produce a stable-looking questionnaire profile without representing the lived preferences of a human panel. The relevant buyer test is whether simulated responses predict held-out decisions by the target population, including refusals and prompt sensitivity. Confidence is high in this distinction and unestablished in substituting these model responses for participants. Direct, preregistered comparison against an appropriate consented human sample would change the substitution judgment.

Sources: [461]

Evidence: Reviewed-work atlas - Annotation engineering and labor: Learning Whom to Trust with MACE. MACE models a binary spamming state per annotator-instance and an annotator-specific trust parameter, estimating unobserved true categorical labels with EM or variational Bayes; nonspamming annotations are assumed to copy the true label. Natural-data evaluation reused categorical AMT datasets from Snow et al.: word-sense disambiguation (177 items, 34 annotators), textual entailment (800,164), and temporal relations (462, 76), with expert gold labels.

Sources: [462]

Evidence: Study result - Learning Whom to Trust with MACE. At full coverage, Table 2 reports MACE-EM/VB accuracy 0.93 on RTE, 0.94 on temporal relations and 0.99 on WSD, compared with majority-vote 0.90, 0.93 and 0.99 respectively; lower-coverage thresholded results are explicitly not comparable at full coverage.

Sources: [462]

Evidence: Study limitations - Learning Whom to Trust with MACE. The simple categorical model assumes an annotator is correct whenever not spamming and learns spamming biases independent of the true label; the authors acknowledge that this does not reflect the full reality of AMT, and do not model the continuous-label tasks.

Sources: [462]

Analysis: Study assessment and change conditions - Learning Whom to Trust with MACE. Aggregation gains should be evaluated at a fixed coverage level, with abstentions visible. The model is useful when one categorical target is defensible; its spamming assumptions may misclassify legitimate disagreement when that condition fails. Confidence is moderate for testing it as a baseline and low for treating inferred trust as a worker-quality verdict. Blinded error review, coverage curves and a target-sensitive disagreement analysis would change the recommendation.

Sources: [462]

Evidence: Reviewed-work atlas - Human and model supervision: Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. PALMS specified desired behavior for eight topic categories, created 80 training question-answer pairs, fine-tuned a language model, evaluated it and iterated on weaknesses. A contracted professional writer created 40-340-word completions; author-written answer outlines guided targeted prompts, and authors checked and modified broad-topic answers. Three blinded evaluators rated generated answers from 1 to 5 against desired behavior.

Sources: [463]

Evidence: Study result - Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. On weakness-targeting question-answer evaluations with three completions per prompt at temperature 0.7, the values-targeted models had lower mean toxicity and higher mean human intended-sentiment ratings than base models; the authors report stronger improvements for larger models.

Sources: [463]

Evidence: Study limitations - Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. The study was limited to U.S. English and selected prompt formats, and the authors warn that underrepresented cultural contexts and other downstream tasks may not generalize; broad training prompts/completions could not be shared under their data policy.

Sources: [463]

Analysis: Study assessment and change conditions - Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. The useful deliverable is a documented values specification, writer contribution and revision loop, rather than eighty context-free examples. A pilot should separate adherence to the specified behavior from toxicity proxies and sample cultural cases outside the initial authoring assumptions. Confidence is moderate in this workflow interpretation and unestablished for broad social representativeness. Independent evaluation by affected populations on novel contexts would strengthen or narrow it.

Sources: [463]

Evidence: Reviewed-work atlas - Multilingual and multimodal supervision: CrowdSpeech and VoxDIY: Benchmark Datasets for Crowdsourced Audio Transcription. CrowdSpeech crowdsourced transcriptions of English LibriSpeech audio; VoxDIY synthesized Russian news text with Yandex SpeechKit before human transcription. Each recording was annotated by seven workers. Workers self-reported the task language and took a ten-recording entrance exam requiring WER at most 40%; ongoing controls blocked repeated sub-ten-second tasks and poor recent gold-question WER. Attempted exams and pre-block completed tasks were compensated.

Sources: [464]

Evidence: Study result - CrowdSpeech and VoxDIY: Benchmark Datasets for Crowdsourced Audio Transcription. On CrowdSpeech test-clean, Table 4 reports mean WER 7.29 for ROVER versus 4.32 for a gold-aware oracle; on test-other the corresponding WERs are 13.41 and 8.50. The oracle selects the best submitted transcription and does not establish a general upper bound for methods that can edit text.

Sources: [464]

Evidence: Study limitations - CrowdSpeech and VoxDIY: Benchmark Datasets for Crowdsourced Audio Transcription. Speech synthesis can change the distribution of human transcription errors and limit transfer from VoxDIY; the stringent industry QC pipeline also limits coverage of harder or lower-quality annotation settings.

Sources: [464]

Analysis: Study assessment and change conditions - CrowdSpeech and VoxDIY: Benchmark Datasets for Crowdsourced Audio Transcription. The two corpora distinguish naturally read source speech from synthesized speech before human transcription. A buyer should stratify recording origin and difficulty, compare aggregation with individual transcripts and account for paid entrance attempts. Confidence is high in the provenance distinction and moderate in its procurement relevance. Matched natural and synthetic audio with independently corrected references would change the transfer assessment; an oracle selecting submitted transcripts is not the same comparator as a method allowed to edit them.

Sources: [464]

Evidence: Reviewed-work atlas - Curation and provenance: Deduplicating Training Data Makes Language Models Better. ExactSubstr uses suffix arrays to remove repeated substrings with a 50-BPE-token threshold; NearDup uses MinHash on 5-grams plus edit similarity greater than 0.8 to identify near-duplicate documents. The duplication analysis uses the English part of Wiki-40B, LM1B, C4 and RealNews; these corpora had differing prior deduplication procedures, so prior cleaning did not ensure equivalent duplication rates.

Sources: [465]

Evidence: Study result - Deduplicating Training Data Makes Language Models Better. For unprompted top-k sampling with k 50, the authors generated 100,000 sequences up to 512 tokens and marked tokens within 50-token verbatim training matches as memorized; original-C4 models emitted over 1% memorized tokens, about ten times the deduplicated-model rate.

Sources: [465]

Evidence: Study limitations - Deduplicating Training Data Makes Language Models Better. The analysis did not distinguish harmful from innocuous or desirable memorization and did not evaluate negative effects of deduplication, including loss of attribution or tasks requiring memorization; deduplication alone cannot remove all privacy-sensitive data.

Sources: [465]

Analysis: Study assessment and change conditions - Deduplicating Training Data Makes Language Models Better. Deduplication is a specified matching procedure with retention tradeoffs. A pilot should retain duplicate-cluster membership, removal reasons and source attribution, and evaluate both memorized-string exposure and task coverage after filtering. Confidence is moderate that this is more informative than a single deduplicated badge; complete privacy protection remains unestablished. An exact-payload audit with sensitivity analyses for thresholds, rare content and attribution loss would change the buyer’s risk estimate.

Sources: [465]

Evidence: Reviewed-work atlas - Annotation engineering and labor: Learning from Crowds with Crowd-Kit. Crowd-Kit implements crowdsourced-answer aggregation and learning-from-crowds methods in Python, with pandas/scikit-learn-compatible interfaces and PyTorch modules for direct learning from raw crowd labels. Categorical evaluation reused four real-world noisy-label datasets and synthetic datasets from prior work; the direct-learning experiments used IMDB reviews and CIFAR-10, with three annotators per object and 100 gold tasks.

Sources: [466]

Evidence: Study result - Learning from Crowds with Crowd-Kit. For the D_Product categorical benchmark, Table 3 reports accuracy 0.940 for Dawid-Skene versus 0.897 for majority vote in the authors’ Crowd-Kit implementations.

Sources: [466]

Evidence: Study limitations - Learning from Crowds with Crowd-Kit. The library lacks distributed-cluster execution; M-MSR did not process D_Product in reasonable time and no implementation handled the large binary 3/binary 4 datasets. Its categorical aggregation methods assume one objective true label per task.

Sources: [466]

Analysis: Study assessment and change conditions - Learning from Crowds with Crowd-Kit. A common implementation helps compare aggregation under a common interface, but the target assumptions and execution limits still determine fit. Require held-out gold, fixed coverage and resource measurements on the buyer’s own label graph. Confidence is moderate for using the library as a baseline within this setting, without a universal algorithm winner. A representative throughput test and adjudication of systematic disagreements would change both the method and infrastructure choice.

Sources: [466]

Evidence: Reviewed-work atlas - Annotation engineering and labor: IMDB-WIKI-SbS: An Evaluation Dataset for Crowdsourced Pairwise Comparisons. IMDB-WIKI-SbS sampled 75 photos per age 10-70 and binary gender from IMDB-WIKI, yielding 9,150 photos, and constructed 250,249 comparison edges in a connected random graph. On Toloka, 4,091 workers each chose the older of two cropped faces; each pair had one unique annotator, with three regular and one control task per page. QC used top-half English-speaking workers, timing/skipping checks and 200 gold comparisons with over 20-year age gaps.

Sources: [467]

Evidence: Study result - IMDB-WIKI-SbS: An Evaluation Dataset for Crowdsourced Pairwise Comparisons. Whole-dataset NDCG@100 was 0.859 for Bradley-Terry, 0.803 for PageRank and 0.490 for the random baseline; these are age-ranking results on the connected comparison graph, not personal identity or fairness outcomes.

Sources: [467]

Evidence: Study limitations - IMDB-WIKI-SbS: An Evaluation Dataset for Crowdsourced Pairwise Comparisons. The authors noticed incorrect gender metadata in the source and allowed a person to appear in multiple age groups; within-age-group comparisons were harder and graph disconnection prevented meaningful Bradley-Terry ranking in some subgroup evaluations.

Sources: [467]

Analysis: Study assessment and change conditions - IMDB-WIKI-SbS: An Evaluation Dataset for Crowdsourced Pairwise Comparisons. Pairwise labels do not automatically produce a well-connected or demographically valid ranking problem. For a similar comparison service, inspect graph connectivity within intended subgroups, repeated identities and source-metadata errors before fitting a global rank. Confidence is high that these are separate failure channels and low that the reported age-ranking result transfers to other judgments. Independent subgroup metadata review and a matched connectivity experiment would change the assessment.

Sources: [467]

Evidence: Reviewed-work atlas - Human and model supervision: WebGPT: Browser-assisted question-answering with human feedback. WebGPT trained GPT-3-family models using human browser demonstrations, comparison-trained reward models, PPO or rejection sampling to generate answers with collected web references. The study collected about 6,000 demonstrations and 21,500 answer comparisons, mostly ELI5; around 25% of data came from 10 Upwork contractors and 75% from 46 Surge AI contractors. Paid trials, manual checks and reference-support annotations were used for QC.

Sources: [468]

Evidence: Study result - WebGPT: Browser-assisted question-answering with human feedback. On ELI5 evaluation, the 175B best-of-64 model was preferred to human demonstrator answers 56% of the time and to ELI5 reference answers 69%, counting ties as 50%; the reference-answer comparison stripped citations and used new contractors with minimal instructions.

Sources: [468]

Evidence: Study limitations - WebGPT: Browser-assisted question-answering with human feedback. Reference-answer comparisons have blinding, answer-intent and factual-checking limitations; citation-bearing answers can seem more authoritative and training can incentivize cherry-picked convincing references rather than a fair assessment of evidence.

Sources: [468]

Analysis: Study assessment and change conditions - WebGPT: Browser-assisted question-answering with human feedback. Citation collection, answer preference and evidence support are separate outputs. A procurement pilot should blind style where possible and independently check whether each material claim is supported, including omitted counterevidence. Confidence is moderate in this separation and unestablished for present-day product accuracy. Matched evidence-grounded evaluations with complete support audits would change the assessment; preference against a dated answer set does not establish a universal factuality advantage.

Sources: [468]

Evidence: Reviewed-work atlas - Weak supervision and data selection: Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision. Liger combines weak-source votes with fixed foundation-model embeddings to estimate local source accuracies and extend votes to nearby abstaining points; threshold radii and local smoothness control the extension tradeoff. Evaluation used Spam, Weather, Spouse and three video tasks (Commercial, Tennis, Basketball), with unlabeled train data, smaller hand-labeled development sets and held-out tests; GPT-3 Ada supplied text embeddings and CLIP supplied video-frame embeddings.

Sources: [469]

Evidence: Study result - Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision. Across the six tasks, the authors report that Liger improved coverage by 33.2 percentage points on average versus the vanilla weak-supervision label model; hand-labeled development data were used for hyperparameter tuning.

Sources: [469]

Evidence: Study limitations - Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision. The label model assumes conditionally independent weak sources for simplicity, and source extension can lose accuracy when radius is too large or embeddings lack smoothness; embedding-smoothness guarantees therefore do not apply without their stated assumptions.

Sources: [469]

Analysis: Study assessment and change conditions - Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision. Extending coverage is valuable only if errors introduced by the extension remain acceptable. A buyer should request accuracy-versus-coverage curves, source correlations, abstention behavior and an independently labeled development/evaluation split. Confidence is moderate in the documented local-extension idea under its assumptions, with no universal gain claimed. Shifted-domain tests and correlated-source stress cases would change the decision to automate or escalate labels.

Sources: [469]

Evidence: Reviewed-work atlas - Annotation engineering and labor: Clustering Without Knowing How To: Application and Evaluation. The system asks three workers per task to group small sets of clothing images by style using palette labels, then aggregates sparse judgments with a reimplemented probabilistic Crowdclustering model whose cluster count is estimated dynamically. Experiments used 2,000 FEIDEGGER dress images and 87 Toloka Shoes images; workers completed progressive five-page training and a qualification test, requiring at least 80 out of 100 to advance.

Sources: [470]

Evidence: Study result - Clustering Without Knowing How To: Application and Evaluation. For an intruder task that inserted an object from another cluster, evaluators selected the intruder with fractions 0.83 for the 2,000-dress experiment and 0.88 for the 87-shoe experiment.

Sources: [470]

Evidence: Study limitations - Clustering Without Knowing How To: Application and Evaluation. Style clustering lacks one uniquely correct gold grouping, so the evaluation measures consistency with other workers’ subjective judgments using intruders rather than recovery of an objective class truth.

Sources: [470]

Analysis: Study assessment and change conditions - Clustering Without Knowing How To: Application and Evaluation. For subjective style, consistency with a reference panel is the measured construct. An intruder task can identify obvious incoherence without establishing a unique correct taxonomy. Confidence is high in that measurement boundary and moderate in using it for a similar grouping workflow. A buyer should preserve competing clusterings and test downstream retrieval or merchandising outcomes; independently measured improvement on that intended use would change the value assessment.

Sources: [470]

Evidence: Reviewed-work atlas - Human and model supervision: Best Prompts for Text-to-Image Models and How to Find Them. A genetic algorithm proposed keyword sets; Stable Diffusion generated four images per description/set, humans compared image sets blind to keywords, and Bradley-Terry ranks averaged over descriptions supplied the optimization metric. The search pool used 100 popular keywords from Stable Diffusion Discord and 72 manually cleaned descriptions from Reddit/Lexica (60 train, 12 validation); Toloka workers below 80% on controls comparing to DALL-E Mini images were suspended.

Sources: [471]

Evidence: Study result - Best Prompts for Text-to-Image Models and How to Find Them. After 56 search iterations, Table 1 reports the best-training keyword set with mean rank 43.60 on 60 training descriptions and 38.00 on 12 validation descriptions, versus 14.25 and 12.50 for the top 15 popular keywords; higher rank is better among 56 sets.

Sources: [471]

Evidence: Study limitations - Best Prompts for Text-to-Image Models and How to Find Them. The authors cannot conclude the discovered set is optimal because genetic search can converge locally; rank-based metrics are relative and insensitive to convergence, and QC assumed DALL-E Mini images were less appealing.

Sources: [471]

Analysis: Study assessment and change conditions - Best Prompts for Text-to-Image Models and How to Find Them. Human preference can become an optimization signal, but control items can embed the researchers' own aesthetic premise. A buyer should hold out descriptions, generator versions and rater pools, and inspect whether optimized prompts merely exploit known controls. Confidence is moderate for the demonstrated search workflow and low for global optimality. An independent panel on new image subjects and fixed generation budgets would change the generalization assessment.

Sources: [471]

Evidence: Reviewed-work atlas - Process supervision and reasoning: Learning to Reason With Relational Abstractions. The study added abstract relational plans to arithmetic solutions and contrasted numeric-only, relational-first, interleaved and multitask training formats; in multitask training the model produced relational or numeric sequences on different batches and numeric sequences at test time. GSM8K-R extends GSM8K with human annotations pairing solution steps with short abstract relational explanations; annotators were instructed to keep equation structure while consistently naming the quantities. The published instructions include the actual labeling interface.

Sources: [472]

Evidence: Study result - Learning to Reason With Relational Abstractions. With GPT-2-XL and greedy decoding on GSM8K-R, multitask relational training achieved 28.05% top-1 test solve accuracy versus 22.97% for equation-only training; the much higher 66.26% ground-truth-relational-prompt result uses privileged test-time auxiliary information.

Sources: [472]

Evidence: Study limitations - Learning to Reason With Relational Abstractions. The authors note that human relational annotations can be ambiguous compared with formal variables; sequence length and ordering affect accuracy, and gains from supplying correct relational steps do not show the model can infer those steps unaided.

Sources: [472]

Analysis: Study assessment and change conditions - Learning to Reason With Relational Abstractions. The annotation specification describes relationships, not simply longer reasoning text. A pilot should independently verify relational labels and evaluate the trained model without privileged test-time steps. Confidence is moderate that the reported formats warrant a controlled comparison and high that auxiliary-information conditions must remain separate. Repeated matched-budget training with independently authored test problems would change the estimate of practical gain.

Sources: [472]

Evidence: Reviewed-work atlas - Human oversight and evaluation: Measuring Progress on Scalable Oversight for Large Language Models. The experiment compared unaided humans, a 52B helpful RLHF dialog model and human-model teams on MMLU and 5-minute-limit QuALITY questions; human conditions used 100 sampled validation questions per experiment. Surge AI recruited/trained/paid two groups of five participants; one unusually strong unaided participant per task was excluded, leaving four per task in reported results. Participants received answer feedback and at least$20 per active-work hour.

Sources: [473]

Evidence: Study result - Measuring Progress on Scalable Oversight for Large Language Models. Reported individual accuracy increased from 57.2% unaided to 75.4% with model help on MMLU and from 48.6% to 76.8% on timed QuALITY; assisted confidence-weighted group accuracy was 78.0% and 86.0% respectively.

Sources: [473]

Evidence: Study limitations - Measuring Progress on Scalable Oversight for Large Language Models. The authors say the controlled multiple-choice setting, answer revelation and removal of high-performing participants limit applicability, and the observed confident errors do not validate the protocol for high-stakes oversight of more capable future systems.

Sources: [473]

Analysis: Study assessment and change conditions - Measuring Progress on Scalable Oversight for Large Language Models. Assistance quality must be tested on human decisions after assistance, including confidently accepted errors. The excluded participants, feedback and time limit materially shape the reported team comparison. Confidence is moderate that this protocol identifies an oversight research question, with no high-stakes validation implied. A larger representative panel, retained exclusions and adversarially misleading assistance would change confidence in operational use.

Sources: [473]

Evidence: Reviewed-work atlas - Human and model supervision: Discovering Language Model Behaviors with Model-Written Evaluations. The method samples evaluation inputs conditional on intended labels and ranks/filters them with a discriminator; persona evaluation used a helpful RLHF generator and its preference model to form label-balanced agree/disagree datasets of at most 1,000 statements per behavior. Surge AI workers checked generated inputs and labels; three workers judged each example for 11 full datasets and 100 examples in each of 122 additional datasets, assessing relevance, label correctness and ambiguity.

Sources: [474]

Evidence: Study result - Discovering Language Model Behaviors with Model-Written Evaluations. In §3.3, the paper reports mean relevance 4.4 ±0.9 on a 1-5 scale, Fleiss κ0.875 between three workers and majority agreement with 95.5% of generated labels; these are validation judgments, not certified correctness of every released example.

Sources: [474]

Evidence: Study limitations - Discovering Language Model Behaviors with Model-Written Evaluations. Generator/discriminator biases and filtering can narrow diversity, models struggle with unfamiliar concepts or many constraints, and prompt wording or misunderstood instructions can change what an evaluation actually tests.

Sources: [474]

Analysis: Study assessment and change conditions - Discovering Language Model Behaviors with Model-Written Evaluations. The human audit checks a sampled generator/discriminator pipeline; it does not convert all generated labels into independent human labels. A buyer should retain selection scores and rejected items, then audit failure clusters and prompt sensitivity outside the generation rubric. Confidence is moderate in sampled validation as a useful control and low in interpreting its agreement as corpus-wide truth. Independently authored held-out behaviors and exhaustive review of a risk-focused subset would change the assessment.

Sources: [474]

Evidence: Reviewed-work atlas - Human and model supervision: DOC: Improving Long Story Coherence With Detailed Outline Control. DOC combines hierarchical detailed outlines with an OPT-350 m FUDGE-style controller trained contrastively on passage summaries; hard negatives splice off-topic continuations into initially relevant passages to maintain outline relevance. Controller training summarized WritingPrompts passages using InstructGPT3-13B; evaluation used 1,000-1,500-word corresponding passages from 20 stories, with three Surge AI annotators comparing each pair on coherence, relevance and interestingness.

Sources: [475]

Research gap: Study result boundary - DOC: Improving Long Story Coherence With Detailed Outline Control. No quantitative result assertion is adopted for this work in the approved synthesis input. The report uses its approved method, provenance and limitation observations and does not fill the result gap with a discovery snippet or a superseded candidate.

Evidence: Study limitations - DOC: Improving Long Story Coherence With Detailed Outline Control. DOC is specialized for English stories, expensive human evaluation limits detailed ablations, and factual consistency was not a focus; inconsistent names or genders can arise during outlining and cascade into drafting.

Sources: [475]

Analysis: Study assessment and change conditions - DOC: Improving Long Story Coherence With Detailed Outline Control. Long-form writing quality decomposes into coherence, relevance, interestingness and factual consistency. The inspected approved corpus supplies no adopted quantitative DOC result, so this entry does not invent one from the abstract or reputation. Confidence is moderate in the documented outline-control mechanism and unestablished in a buyer-specific gain. A blinded, complete-story evaluation with factual checks and a matched outline-only baseline would change the purchasing judgment.

Sources: [475]

Evidence: Reviewed-work atlas - Curation and provenance: StarCoder: may the source be with you!. StarCoder curation combined extension-level community inspection and source-code filters with a PII detector: human-labeled examples trained StarEncoder token classification for names, emails, keys, passwords, IP addresses and usernames; IDs were excluded because of poor annotation quality. Toloka engaged 1,399 workers from 35 countries to label 12,000 roughly 50-line files in 31 languages; 7,100 files were prefiltered to enrich rare PII and 5,100 were random. Annotators found 22,950 entities, and authors manually inspected 300 files for QC.

Sources: [476]

Evidence: Study result - StarCoder: may the source be with you!. For the 15.5B StarCoder model, Table 12 reports pass@1 of 33.6 on HumanEval and 52.7 on MBPP Python; the evaluation used sampling temperature 0.2 and 200 samples for open-access models. The fine-tuned PII baseline reports F1 above 90% for names, emails and IP addresses, 73.39% for passwords, 56.66% for keys and 59.39% for usernames; the split was 7,878 training and 4,000 test examples, supplemented in training by 400 earlier annotated files.

Sources: [476]

Evidence: Study limitations - StarCoder: may the source be with you!. The authors warn that license detection can misclassify repositories, repository-level opt-out can miss copied code, and PII detection has false positives/negatives; PII annotations are gated to approved individuals and published model evaluations are English-only.

Sources: [476]

Analysis: Study assessment and change conditions - StarCoder: may the source be with you!. The human annotation contribution concerns PII detector training and review; it does not make every code token human-verified or eliminate personal-data risk. A buyer should separately sample random files and rare-sensitive-type enriched files, retain per-type precision/recall and inspect opt-out propagation. Confidence is high in separating those denominators and moderate in the documented workflow’s relevance. An independently labeled target-code sample and end-to-end removal audit would change the risk assessment.

Sources: [476]

Evidence: Reviewed-work atlas - Evidence-grounded evaluation: Enabling Large Language Models to Generate Text with Citations. ALCE evaluates fluency, dataset-specific correctness, and citation quality separately; citation recall tests whether the concatenated cited passages entail a statement, while citation precision penalizes irrelevant citations. ALCE samples 1,000 development examples from each of ASQA, QAMPARI and ELI 5; its human evaluation samples 100 examples from ASQA and ELI 5, with workers paid an average of $20 per hour.

Sources: [549]

Evidence: Study result - Enabling Large Language Models to Generate Text with Citations. On ELI 5, Table 6 reports citation recall of 51.1 for ChatGPT VANILLA using five passages and 69.3 with reranking; the corresponding claim-correctness recalls are 12.0 and 11.4.

Sources: [549]

Evidence: Study limitations - Enabling Large Language Models to Generate Text with Citations. The authors acknowledge that NLI errors limit citation evaluation and cannot reliably detect partial support; automatically generated ELI 5 claims may omit valid answers, and the evaluated tasks omit math, coding and multi-hop reasoning.

Sources: [549]

Analysis: Study assessment and change conditions - Enabling Large Language Models to Generate Text with Citations. Citation recall can rise while claim correctness does not. The appropriate deliverable therefore has separate correctness, support, completeness and irrelevance judgments, with partial-support cases adjudicated. Confidence is high in the metric distinction and moderate in applying this evaluation structure. Independent expert scoring on the buyer’s evidence corpus, including unsupported compound statements and valid alternate answers, would change confidence in an automated citation judge.

Sources: [549]

Evidence: Reviewed-work atlas - Task construction and evaluation: Inverse Scaling: When Bigger Isn't Better. The Inverse Scaling Prize accepted datasets with at least 300 examples, evaluated submitted GPT-3 scaling patterns on held-out model series, and used classification or sequence-completion loss for winning tasks. Task labels were checked by contractors on random samples of at least 50 examples; contractors could use the internet, and multipart tasks were sampled by part or grouped parts rather than treating all labels as exhaustively verified.

Sources: [550]

Evidence: Study result - Inverse Scaling: When Bigger Isn't Better. Table 1 reports 11 winning tasks; for example, Sig Figs contains 20,897 examples with 99.5% sampled human agreement, while NeQA contains 300 examples with 98.0% agreement.

Sources: [550]

Evidence: Study limitations - Inverse Scaling: When Bigger Isn't Better. Inverse scaling can reverse into U-shaped scaling, and some tasks show inverted-U trends or multiple reversals through training; the authors cannot predict the scale needed for improvement on a given task.

Sources: [550]

Analysis: Study assessment and change conditions - Inverse Scaling: When Bigger Isn't Better. Sampled label agreement establishes only the inspected task sample under its labeling rule. A buyer should separate task novelty, label validity and behavior across model sizes rather than extrapolate an inverse trend indefinitely. Confidence is high that the observed shapes are conditional and low in predicting future model rankings. Matched versions and repeated evaluations across held-out model families would change the scaling interpretation.

Sources: [550]

Evidence: Reviewed-work atlas - Safety, feedback and verification: Universal and Transferable Adversarial Attacks on Aligned Language Models. The study optimizes one adversarial suffix across multiple training prompts and models, then evaluates whether the resulting suffix transfers to held-out behaviors and models. AdvBench distinguishes 500 target harmful strings from 500 harmful-behavior instructions; exact target-string matches and attempts to comply with a behavior use different success definitions.

Sources: [551]

Evidence: Study result - Universal and Transferable Adversarial Attacks on Aligned Language Models. Across 388 harmful behaviors, Table 2 reports ensemble transfer attack success rates of 86.6% for GPT-3.5 and 46.9% for GPT-4, compared with 1.8% and 8.0% for the behavior-only baseline; ensemble success means at least one suffix succeeds.

Sources: [551]

Evidence: Study limitations - Universal and Transferable Adversarial Attacks on Aligned Language Models. Transfer results depend on historical model versions, sampling and interface conditions; the authors discuss shared training/distillation as a potential transfer confound and disclosed findings before publication.

Sources: [551]

Analysis: Study assessment and change conditions - Universal and Transferable Adversarial Attacks on Aligned Language Models. Transfer attack success is conditional on an attack budget, model version and success rule. A defense pilot should report per-behavior and per-attempt outcomes and include newly authored harmful-intent tests without publishing operational attack recipes in its acceptance report. Confidence is high in the need for those denominators and unestablished for any current defense. Matched contemporary evaluation with independently validated harm and intent preservation would change the risk conclusion.

Sources: [551]

Evidence: Reviewed-work atlas - Human editing and evaluation: Beemo: Benchmark of Expert-edited Machine-generated Outputs. Beemo uses expert editors to refine machine outputs for factual accuracy, style, coherence and fluency; two lead editors manually validate submissions and annotators may skip already-suitable or substantially unsuitable examples.

Sources: [693]

Evidence: Study result - Beemo: Benchmark of Expert-edited Machine-generated Outputs. Table 4 reports Binoculars AUROC of 83.90% for human-written versus machine-generated text and 61.24% for human-written versus expert-edited machine text, on the paper’s out-of-domain benchmark.

Sources: [693]

Evidence: Study limitations - Beemo: Benchmark of Expert-edited Machine-generated Outputs. The study has no human detection baseline and does not evaluate commercial/API detectors; results can depend on editing prompts and may not transfer from expert editing to non-expert users. AUROC compares two labeled text populations; it is not accuracy at a chosen operating threshold. Expert-edited model text remains in the machine-origin class. Selection excludes skipped examples, so retained edits are not a random sample of all attempted editing work.

Sources: [693]

Analysis: Study assessment and change conditions - Beemo: Benchmark of Expert-edited Machine-generated Outputs. Expert editing is a distinct provenance stage that can change detector behavior without making the text wholly human-authored. A buyer evaluating authorship detection should include retained editing histories and task-matched human references, and compare detection errors by editing intensity. Confidence is moderate in this benchmark-specific sensitivity and low for unseen detectors. Independent evaluation of current tools, non-expert editing and different genres would change the assessment.

Sources: [693]

Evidence: Reviewed-work atlas - Task construction and evaluation: U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs. U-MATH sources expert-crafted problems from ongoing university courses, filters short-solution, multiple-choice and calculator-use problems, and uses small-model solution rates followed by expert review to select challenging problems. U-MATH has 1,100 problems across six subjects, including 900 text-only and 200 visual problems; in-house mathematical experts, Gradarius content staff and teaching academics review the selected material.

Sources: [552]

Evidence: Study result - U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs. Table 4 reports o1 accuracy of 86.8% overall, 93.1% on 900 text-only problems and 58.5% on 200 visual problems; text-only and visual scores measure distinct subsets.

Sources: [552]

Evidence: Study limitations - U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs. The authors note incomplete coverage of advanced mathematics, saturation of text-only questions, limited visual coverage and LLM-judge errors; visual problems are excluded from the meta-evaluation.

Sources: [552]

Analysis: Study assessment and change conditions - U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs. Overall mathematical accuracy masks a large distinction between text-only and visual subsets. A buyer should specify problem source, discipline, visual content and judge validation separately, with a held-out expert panel for the chosen deployment. Confidence is high that the subset distinction matters within this benchmark and unestablished for professional mathematics. Independent visual meta-evaluation and new expert-authored tasks would change confidence in the judge and model comparison.

Sources: [552]

Evidence: Reviewed-work atlas - Weak supervision and data selection: Weak-to-Strong Generalization Through the Data-Centric Lens. The proposed data-source selector treats sources as bandit arms and uses an upper-confidence bound on detected overlap density, targeting points where reliable weak-model signals coexist with patterns learnable by the stronger model. The language-model experiments use Qwen1.5 0.5 B as weak model, Llama 3 8 B as strong model and linear probing on 19 datasets; the weak-supervision experiments separately use nine WRENCH datasets.

Sources: [553]

Evidence: Study result - Weak-to-Strong Generalization Through the Data-Centric Lens. In the Amazon Polarity and DREAM examples, Figure 3 reports better weak-to-strong accuracy with UCB-based data-source selection than random source sampling, averaged over 20 repeated experiments with 25 selection rounds.

Sources: [553]

Evidence: Study limitations - Weak-to-Strong Generalization Through the Data-Centric Lens. The authors caution that few overlap points or noisy overlap detection can eliminate the gains; their mechanism assumes a simplified easy/hard feature structure rather than modeling every realistic weak-to-strong setting.

Sources: [553]

Analysis: Study assessment and change conditions - Weak-to-Strong Generalization Through the Data-Centric Lens. Source selection is part of the supervision recipe rather than a neutral sampling step. A pilot should record overlap estimates, chosen-source proportions and gains against random selection at the same annotation and compute budget. Confidence is moderate under the modeled feature assumptions and low where overlap detection is noisy. Shifted-domain experiments with independently measured overlap would change whether this approach is preferable to simpler selection.

Sources: [553]

Evidence: Reviewed-work atlas - Multilingual and multimodal supervision: JEEM: Vision-Language Understanding in Four Arabic Dialects. JEEM separates dialect-first captioning from question writing without access to the image, then answers questions with the image and captions; team leaders review submissions and may edit, reject or reassign them. JEEM contains 2,178 images and 10,890 question-answer pairs across Jordanian, Emirati, Egyptian and Moroccan Arabic; native-dialect recruitment yields teams of 10, 8, 10 and 9 annotators respectively.

Sources: [554]

Evidence: Study result - JEEM: Vision-Language Understanding in Four Arabic Dialects. On the sampled Jordanian image-caption evaluation, Table 3 reports human-rated dialect authenticity of 3.56 for GPT-4o captions and 4.82 for reference human captions on a 1-5 scale; human evaluation uses 50 images per dialect and one blinded rater per caption.

Sources: [554]

Evidence: Study limitations - JEEM: Vision-Language Understanding in Four Arabic Dialects. JEEM covers four dialects with an uneven, relatively small image set; automatic metrics may miss cultural/dialect nuance and human evaluation covers only a subset.

Sources: [554]

Analysis: Study assessment and change conditions - JEEM: Vision-Language Understanding in Four Arabic Dialects. Dialect authenticity cannot be established by a language code or a machine metric alone. A pilot should obtain native judgments in the target dialect, preserve the question-writing and image-viewing stages and report coverage per dialect. Confidence is moderate in this design lesson and low for a market-wide comparison from one sampled rater per caption. Multiple independent local raters and broader image/domain coverage would change the generalization assessment.

Sources: [554]

Evidence: Reviewed-work atlas - Annotation engineering and labor: Voices of Freelance Professional Writers on AI: Limitations, Expectations, and Fears. The study combines a 64-question questionnaire with a task in which selected multilingual writers prompt an AI chatbot for a local-event advertisement, revise it and submit the interaction and versions; two analysts inductively code free-text responses and resolve disagreements by discussion. The questionnaire includes 301 Toloka freelance writers from 49 countries; the interactive task includes 36 participants selected from 120 invitations using language proficiency, low self-ratings and/or detailed comments.

Sources: [555]

Evidence: Study result - Voices of Freelance Professional Writers on AI: Limitations, Expectations, and Fears. Among the surveyed professional writers, 73% report using AI chatbots; the paper reports that 75% improve outputs through iterative prompting and audience specification, as a sample-specific self-reported practice.

Sources: [555]

Evidence: Study limitations - Voices of Freelance Professional Writers on AI: Limitations, Expectations, and Fears. The authors caution that a single-platform, geographically uneven sample and only 36 behavioral participants limit generalization; language ratings with fewer than three respondents are individual observations, and model changes can age these ratings.

Sources: [555]

Analysis: Study assessment and change conditions - Voices of Freelance Professional Writers on AI: Limitations, Expectations, and Fears. The source gives writers' situated reports and observed editing artifacts, rather than an employment forecast. A buyer designing writer-assisted data work should pay for prompting, revision and explanation as well as the final text and test whether feedback improves later tasks. Confidence is moderate in the sampled workflow observations and low outside the platform. A longitudinal, multilingual sample with actual paid-time logs and independently scored outputs would change the labor interpretation.

Sources: [555]

Evidence: Reviewed-work atlas - Weak supervision and data selection: Shrinking the Generation-Verification Gap with Weak Verifiers. Weaver uses weak supervision to weight multiple binary verifier signals for response selection; its setup allows a small labeled development subset, roughly 1% of test queries, to estimate global statistics. The main evaluation generates 100 responses per query with Llama 3.3 70 B Instruct and combines a pool of 33 reward models and language-model judges; the paper compares selected-response accuracy with an oracle Pass@100 bound.

Sources: [556]

Evidence: Study result - Shrinking the Generation-Verification Gap with Weak Verifiers. Table 1 reports average success rates across four tasks of 87.7% for Weaver, 72.2% for majority voting and 91.9% for oracle Pass@100; these are response-selection results under 100-generation conditions.

Sources: [556]

Evidence: Study limitations - Shrinking the Generation-Verification Gap with Weak Verifiers. The aggregation model assumes conditional independence of verifier votes, and inference can require 30-plus verifier calls; on harder datasets with few correct candidate responses, selection remains difficult and is bounded by candidate generation.

Sources: [556]

Analysis: Study assessment and change conditions - Shrinking the Generation-Verification Gap with Weak Verifiers. The selection gain consumes a large generation and verifier budget, and depends on candidate quality. A buyer should compare total cost per accepted answer with simpler reranking at the same generation budget and preserve a gold holdout outside development calibration. Confidence is moderate for the reported setting and low for a blanket weak-verifier replacement of experts. Correlated-judge stress tests and matched-budget target-task results would change the economic conclusion.

Sources: [556]

Evidence: Reviewed-work atlas - Rubrics and executable evaluation: AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following. AdvancedIF retains prompts that trigger final-turn instruction-following failures; the same human annotator writes the prompt and decomposed rubric, and evaluation requires every applicable final-turn criterion to pass. AdvancedIF contains 1,645 dialogs across complex instructions, carried context and system steerability; RIFL separately uses human rubric/verification data to train a rubric generator and a verifier through supervised and reinforcement learning stages.

Sources: [557]

Evidence: Study result - AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following. Table 5 reports AdvancedIF average scores of 51.4 for Llama 4 Maverick and 58.1 after RIFL, with corresponding MultiChallenge scores 37.4 and 40.3; IFEval changes from 89.9 to 90.0.

Sources: [557]

Evidence: Study limitations - AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following. Rubric verifiers can be reward-hacked and can misjudge individual criteria; the authors leave a comprehensive reward-design study to future work, while their ablation compares all-or-nothing, fractional and hybrid rewards on one base model.

Sources: [557]

Analysis: Study assessment and change conditions - AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following. An all-criteria final-turn score differs from a fractional training reward and can depend on the same annotator’s prompt/rubric interpretation. A pilot should introduce independent rubric review and criterion-flipping examples, retaining both strict and fractional outcomes. Confidence is moderate in the method’s relevance to instruction following and unestablished for all instruction classes. Independent prompt/rubric authorship and a no-overlap held-out suite would change the transfer assessment.

Sources: [557]

Evidence: Reviewed-work atlas - Multilingual and multimodal supervision: Multimodal RewardBench 2: Evaluating Omni Reward Models for Interleaved Text and Image. MMRB2 filters near-unanimous model-judge pairs before collecting three blinded human preferences per generative pair; it discards large rater spreads, near-midpoint averages and tied majorities, then evaluates model judges in both response orders. MMRB2 retains 1,000 preference pairs for each of four tasks: text-to-image, image editing, interleaved generation and multimodal reasoning. For reasoning pairs, preferred responses must have correct answers and unanimous human agreement that their reasoning has no major errors.

Sources: [558]

Evidence: Study result - Multimodal RewardBench 2: Evaluating Omni Reward Models for Interleaved Text and Image. Table 2 reports Gemini 3 Pro judge-human agreement averaged across the four tasks at 76.3%, versus 72.2% for GPT-5 and 64.6% for Qwen3-VL-32 B; agreement evaluates the paper’s selected preference pairs rather than general model quality.

Sources: [558]

Evidence: Study limitations - Multimodal RewardBench 2: Evaluating Omni Reward Models for Interleaved Text and Image. The selected data exclude ambiguous or high-disagreement pairs and prioritize hard comparisons; the benchmark does not yet comprehensively cover multilingual, safety/bias, multi-turn, video or audio preferences.

Sources: [558]

Analysis: Study assessment and change conditions - Multimodal RewardBench 2: Evaluating Omni Reward Models for Interleaved Text and Image. The benchmark measures judge agreement on a selected, low-ambiguity but difficult pair population. Its filter can remove precisely the disagreements relevant to a buyer’s policy. Confidence is high in this selection boundary and moderate in using response-order swaps as a calibration check. A pilot should retain a separate ambiguous-pair stratum; broader modalities, languages and independently validated disagreement distributions would change the recommendation.

Sources: [558]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: APEX-SWE. APEX-SWE uses a persistent ReAct harness with a one-hour timeout; Integration is graded by service-facing pytest tests, Observability by fail-to-pass/pass-to-pass checks, and Pass@1 averages three independent runs. APEX-SWE has 100 Integration and 100 Observability held-out tasks authored by engineers with at least three years of experience; Integration includes prompt alignment, test validation and a gold solution, while Observability adapts complex real GitHub issue-PR pairs.

Sources: [559]

Evidence: Study result - APEX-SWE. Claude Opus 4.6 with High thinking scores 49.3% Pass@1 on Integration in Table 1 and 31.7% on Observability in Table 3; each task type contains 100 tasks and the percentages average three runs.

Sources: [559]

Evidence: Study limitations - APEX-SWE. The leaderboard’s binary test success differs from LM-judged engineering-quality rubrics, which are excluded from leaderboard scoring; the paper reports models that score well on rubrics while still failing task tests.

Sources: [559]

Analysis: Study assessment and change conditions - APEX-SWE. Executable integration success and engineering quality are complementary outputs with different acceptance meanings. A buyer should require both regression-safe function and an independently reviewed maintainability rubric, retaining per-task repeated runs. Confidence is high in separating the two measures and low in extrapolating their scores to an organization’s software backlog. Matched private repositories and prospective deployment results would change the capability assessment.

Sources: [559]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: The Hierarchy of Agentic Capabilities: Evaluating Frontier Models on Realistic RL Environments. The hierarchy study evaluates autonomous customer-support tasks in a stateful fictional PC-retail environment using identical system prompts, MCP tools and an LLM judge applying detailed human-written rubrics. Domain experts with customer-support experience design tasks; the December 2025 evaluation figure covers 150 workplace tasks, with trajectories logged and failed trajectories categorized for qualitative capability analysis.

Sources: [560]

Evidence: Study result - The Hierarchy of Agentic Capabilities: Evaluating Frontier Models on Realistic RL Environments. The paper reports GPT-5.2 completion of 61% compared with 57% for GPT-5 in its December 2025 update; even the strongest evaluated system fails approximately 40% of the study’s tasks.

Sources: [560]

Evidence: Study limitations - The Hierarchy of Agentic Capabilities: Evaluating Frontier Models on Realistic RL Environments. The authors identify systematic quantitative failure-distribution analysis, task-category breakdowns, cross-domain evaluation and benchmark-to-production validation as future work; the proposed hierarchy is therefore not a quantified causal model of deployment performance.

Sources: [560]

Analysis: Study assessment and change conditions - The Hierarchy of Agentic Capabilities: Evaluating Frontier Models on Realistic RL Environments. The hierarchy is a useful vocabulary for inspecting failure traces, with its causal interpretation still untested. A buyer should define reproducible categories and check coding agreement before using the taxonomy to prioritize remediation. Confidence is moderate for descriptive organization and low for predicting production performance. Independently categorized trajectories across different businesses, task mixes and model families would change its explanatory value.

Sources: [560]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: APEX-Agents. APEX-Agents grades each human-authored rubric criterion independently with Gemini 3 Flash and defines task success as satisfying all criteria; Pass@1 averages eight runs per task, while Pass@8 and Pass^8 measure any-attempt success and all-attempt consistency. The benchmark includes 480 tasks split equally among investment banking, law and management consulting across 33 worlds; 256 experts from Mercor’s marketplace create and review worlds, tasks, gold outputs and rubrics, and independent experts baseline 96 tasks.

Sources: [561]

Evidence: Study result - APEX-Agents. Table 3 reports Pass@1 of 24.0% for Gemini 3 Flash and 23.0% for GPT-5.2, with 95% task-bootstrap intervals 20.7-27.3 and 19.8-26.2; the paper’s corrected paired tests do not find their difference significant.

Sources: [561]

Evidence: Study limitations - APEX-Agents. Worlds are controlled and web search is disabled; the judge is also a leaderboard model, creating a self-preference risk, and aggregate successful-trajectory tool-use associations do not control for task difficulty or agent approach.

Sources: [561]

Analysis: Study assessment and change conditions - APEX-Agents. Repeated attempts expose a reliability question hidden by best-attempt success. The close leading point estimates and corrected paired tests also caution against treating displayed ranks as distinct tiers. Confidence is high in preserving Pass@1, Pass@8 and Pass^8 separately and moderate in the sample’s task relevance. An independent judge, representative held-out work and replicated expert baselines would change the deployment interpretation.

Sources: [561]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: Benchmarking Agents in Insurance Underwriting Environments. UNDERWRITE co-designs fictional insurance rules, data and tasks with Chartered Property Casualty Underwriters; a separate GPT-4.1 user simulator interacts with the agent, and outcome correctness, tool exceptions and uncertain user replies are distinct evaluation criteria. Six seed task types and 3,000 GPT-4.1-generated fictional applicant profiles inform task creation; experts revise and accept tasks, and the reported evaluation uses 300 accepted tasks with up to 50 user-interaction turns.

Sources: [562]

Evidence: Study result - Benchmarking Agents in Insurance Underwriting Environments. Table 1 reports Claude Sonnet 4.5 answer correctness of 90.30%, with 0.520 tool exceptions and 0.28 uncertain user replies on average per trace; correctness and interaction-error measures have different units.

Sources: [562]

Evidence: Study limitations - Benchmarking Agents in Insurance Underwriting Environments. The correctness judge is calibrated against one expert’s annotations on held-out 100-trace samples, and hallucination detection conservatively targets nonexistent insurance products; neither metric covers all business-domain reasoning failures.

Sources: [562]

Analysis: Study assessment and change conditions - Benchmarking Agents in Insurance Underwriting Environments. A correct final underwriting answer can coexist with interaction errors, and a narrowly defined hallucination check can miss other business mistakes. A buyer should predeclare separate outcome, policy and interaction gates and calibrate judges with multiple experts. Confidence is moderate in the benchmark-specific process lesson and unestablished for regulated production. Prospective domain validation, disagreement adjudication and a broader error taxonomy would change confidence in use.

Sources: [562]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: EnterpriseBench Corecraft: Training Generalizable Agents on High-Fidelity RL Environments. The Corecraft training study fine-tunes GLM 4.6 with GRPO using 16 rollouts per prompt and LM-judged rubric fractions as dense rewards; strict task pass rate separately requires all criteria to be satisfied. Training uses 1,000 tasks and 150 randomly held-out evaluation tasks in one fictional retailer world with more than 2,500 entities, 14 entity types and 23 MCP tools; the base model is a 357 B mixture of experts with 32 B active parameters.

Sources: [6]

Evidence: Study result - EnterpriseBench Corecraft: Training Generalizable Agents on High-Fidelity RL Environments. Table 3 reports the trained GLM 4.6 task pass rate increasing from 25.37% to 36.76% after one epoch, an 11.39 percentage-point gain on the held-out split available during training. Table 4 reports Toolathlon transfer accuracy rising from 18.8% to 25.6% and tau2-Bench Retail from 68.7% to 76.1% after Corecraft training; these are separate external benchmarks and differ from Corecraft task pass rate.

Sources: [6]

Evidence: Study limitations - EnterpriseBench Corecraft: Training Generalizable Agents on High-Fidelity RL Environments. The current frontier benchmark uses tasks added after training and is distinct from the training-time evaluation split; testing other base model families, more epochs and degraded-environment ablations remains future work.

Sources: [6]

Analysis: Study assessment and change conditions - EnterpriseBench Corecraft: Training Generalizable Agents on High-Fidelity RL Environments. The training-time held-out split, later frontier benchmark and external transfer tasks must retain their own identities. A dense criterion reward can improve while strict pass remains a more demanding target. Confidence is moderate in the reported single-run gain and low for a universal training advantage. Repeated seeds, other base families and degraded-environment tests at matched budgets would change the transfer judgment.

Sources: [6]

Evidence: Reviewed-work atlas - Safety, feedback and verification: Intent Laundering: AI Safety Datasets Are Not What They Seem. Intent Laundering tests whether safety refusal depends on overt wording while holding the underlying requested harmful intent as the evaluation target; revised responses receive independent safety and practicality judgments, and success requires both. The study evaluates 207 deduplicated AdvBench cases, 200 standard HarmBench cases and 288 StrongREJECT cases after removing 25 overlaps; judge validation compares 100 sampled responses with majority labels from three human experts.

Sources: [703]

Evidence: Study result - Intent Laundering: AI Safety Datasets Are Not What They Seem. On the 207-case AdvBench subset across ten evaluated models, Table 1 reports mean attack success rate of 3.86% for original requests and 87.54% for first revisions; the original and revised settings use the paper’s explicitly different success definitions.

Sources: [703]

Evidence: Study limitations - Intent Laundering: AI Safety Datasets Are Not What They Seem. The revised-request metric requires unsafe and practical content, while original-request ASR uses safety alone; judgments come from GPT-5.1, with human validation limited to100 responses and binary majority agreement rather than exhaustive intent-preservation review. The automated judge is GPT-5.1; the three-expert comparison covers 100 responses and does not establish intent preservation or correct harm judgments for every transformed request. The result retains the explicitly different original and revised outcome definitions.

Sources: [703]

Analysis: Study assessment and change conditions - Intent Laundering: AI Safety Datasets Are Not What They Seem. The two conditions use different success definitions, so their numerical gap cannot be read as a clean causal effect of wording alone. The practical value is a testable concern that safety evaluation can depend on overt cues. Confidence is moderate in that concern and low in a precise deployment-wide incidence. A matched-definition study with independent intent-preservation, practicality and harm adjudication would change the quantitative conclusion.

Sources: [703]

Evidence: Reviewed-work atlas - Evidence-grounded evaluation: τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge. Tau-Knowledge builds a structured fictional banking knowledge base, converts it to unstructured documents and co-constructs tasks/tools with human refinement; two reviewers independent of task creation check final states, minimal gold documents and one valid trajectory per task. The tau-Banking domain contains 97 tasks, 698 documents, 51 discoverable tools and 14 permanently available agent tools; its knowledge base spans 21 categories and 194,562 tokens under the stated tokenizer.

Sources: [563]

Evidence: Study result - τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge. Table 2 reports GPT-5.2 High pass^1 of 25.52% with terminal search and 32.73% with gold documents directly in context; Claude 4.5 Opus High reaches 39.69% in the gold setting. Here pass^k requires success in every one of k trials.

Sources: [563]

Evidence: Study limitations - τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge. The user simulator omits many characteristics of real human variation; latency depends on API serving, and unrestricted repeated search may differ from deployments allowing only one or a few searches.

Sources: [563]

Analysis: Study assessment and change conditions - τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge. Gold documents in context and terminal search change the information-access problem, while all-trials pass^k measures reliability rather than any-attempt discovery. A buyer should freeze search-call budgets, document versions and simulator assumptions and test retrieval failures separately from reasoning errors. Confidence is high in those distinctions and moderate in their buyer relevance. Real-user variation and a private, independently constructed knowledge base would change the transfer assessment.

Sources: [563]

Evidence: Reviewed-work atlas - Process supervision and reasoning: PivotRL: High Accuracy Agentic Post-Training at Low Compute Cost. PivotRL profiles assistant-turn states from demonstration trajectories, retains mixed-outcome difficult turns, and applies short on-policy rollouts with domain-specific functional rewards instead of requiring exact reproduction of a demonstrated string. Same-data SFT comparisons start from Qwen3-30 B-A 3 B-Thinking-2507 in four separately trained agentic domains; demonstration sources include synthetic conversational trajectories and model-generated coding, terminal and browsing traces, rather than an established human-annotator cohort.

Sources: [564]

Evidence: Study result - PivotRL: High Accuracy Agentic Post-Training at Low Compute Cost. Table 1 reports PivotRL versus same-data SFT scores of 63.81 versus 58.44 on tau2-Bench, 20.00 versus 13.75 on Terminal-Bench and 11.30 versus 1.50 on BrowseComp; on SWE-Bench Verified PivotRL scores 32.67 versus 37.40 for SFT. Table 2 reports average change on eight out-of-domain benchmarks across the four training runs of +0.21 percentage points for PivotRL and −9.83 for SFT, relative to the common base model.

Sources: [564]

Evidence: Study limitations - PivotRL: High Accuracy Agentic Post-Training at Low Compute Cost. The theoretical analysis assumes finite action spaces and positive reference support, and local verifier acceptance does not certify whole-task success; in SWE training the local verifier checks tool-call names without checking arguments or patch quality.

Sources: [564]

Analysis: Study assessment and change conditions - PivotRL: High Accuracy Agentic Post-Training at Low Compute Cost. Local rewards can improve selected decisions without validating arguments or whole-task patch quality. The SWE result also prevents a claim that the method dominates SFT across every domain. Confidence is moderate in testing short-horizon functional rewards and low for a universal efficiency advantage. A matched-budget study with complete local-check definitions, independent task-success scoring and repeated training runs would change the judgment.

Sources: [564]

Evidence: Reviewed-work atlas - Rubrics and executable evaluation: The Hitchhikers Guide to Rubric Quality Understanding and Enrichment. The rubric-quality framework combines content validity, domain mapping, cross-lens judging and weight calibration; its scoring panels are five GPT-5.5-generated expert personas per task, distinct from the human experts who developed the taxonomy. RIFT taxonomy development uses 85 rubrics and 255 expert annotations across five sources with four refinement rounds; the intervention experiment separately uses 60 human-authored rubrics, 30 each from GDPval and Terminal-Bench.

Sources: [565]

Evidence: Study result - The Hitchhikers Guide to Rubric Quality Understanding and Enrichment. Among 48 rubrics with nonconstant panel-rated necessity, 10 have negative weight-necessity rank correlation; Table 1 reports mean ordering correlation+0.195 for human-authored rubrics, scored by the model-persona panel.

Sources: [565]

Evidence: Study limitations - The Hitchhikers Guide to Rubric Quality Understanding and Enrichment. The controlled experiment injects one failure mode at a time and assesses absolute weight magnitudes; overlapping real-world failures, non-additive scoring/gates and downstream policy-learning consequences remain outside the experiment.

Sources: [565]

Analysis: Study assessment and change conditions - The Hitchhikers Guide to Rubric Quality Understanding and Enrichment. Human involvement in taxonomy design is separate from model-persona scoring of necessity. A buyer should inspect decisive conditions, weight sensitivity and cross-lens disagreement without treating generated personas as independent human experts. Confidence is moderate for this audit framework and low for inferring downstream policy gains. A blinded real-expert panel, overlapping failure cases and an end-to-end acceptance experiment would change its procurement value.

Sources: [565]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: ClawsBench: Evaluating Capability and Safety of LLM Productivity Agents in Simulated Workspaces. ClawsBench uses 44 structured tasks across five stateful SQLite-backed mock services and deterministic before/after state evaluation. Task Success Rate is measured on nonsafety tasks at score≥0.8, while Unsafe Action Rate is measured separately on safety tasks at score<0; domain skills and a meta prompt are varied with an intentionally incomplete factorial design.

Sources: [701]

Evidence: Study result - ClawsBench: Evaluating Capability and Safety of LLM Productivity Agents in Simulated Workspaces. In the six-model OpenClaw comparison, the paper reports that full skills-plus-meta scaffolding raises Task Success Rate from 0-8% without scaffolding to 39-63%. These are task/condition-level reported results with cluster-bootstrap uncertainty, not current-product safety rankings or proof of transfer to unseen tasks.

Sources: [701]

Evidence: Study limitations - ClawsBench: Evaluating Capability and Safety of LLM Productivity Agents in Simulated Workspaces. Scaffolding was iteratively refined on the evaluated task set, so the authors treat its measured lift as an upper bound on unseen-task transfer. The mocks omit rate limits, latency and concurrent access; the 44-task study lacks a human baseline and has only five of eleven harness/model pairs with full 2 × 2 coverage. The study separates 20 non-safety tasks from 24 safety tasks. Agents operate the mock REST services through terminal commands, not a browser. Figure 3 reports Flash-Lite Unsafe Action Rate as 22%, whereas Table 4 and adjacent prose report 23%; the discrepancy remains unresolved. Low unsafe-action rate can also reflect over-refusal.

Sources: [701]

Analysis: Study assessment and change conditions - ClawsBench: Evaluating Capability and Safety of LLM Productivity Agents in Simulated Workspaces. Scaffolding tuned on evaluated tasks can be useful diagnosis while overstating unseen-task lift. A pilot should freeze the scaffold before the holdout, complete the intended factorial comparisons and retain safety failures alongside task scores. Confidence is high in this validation requirement and unestablished for current product safety rankings. Prospective tasks with real service constraints and a suitable human baseline would change the generalization assessment.

Sources: [701]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: Riemann-Bench: A Benchmark for Moonshot Mathematics. Riemann-Bench describes 25 expert-authored research-level problems with unique closed-form answers, programmatic answer checking and two independent domain experts solving each from scratch before acceptance. It evaluates agents with code/search over 100 independent runs per problem and reports pass@k estimators.

Sources: [21]

Evidence: Study result - Riemann-Bench: A Benchmark for Moonshot Mathematics. Table 1 reports pass@1 between 2% and 6% for the seven listed API models on this private 25-problem evaluation with 100 runs per problem. This is a reported selected-benchmark snapshot, not a measurement of all mathematical reasoning or current model capability.

Sources: [21]

Evidence: Study limitations - Riemann-Bench: A Benchmark for Moonshot Mathematics. The problems have known solutions and the benchmark is private. The paper distinguishes this from genuinely open-ended mathematical research, where an answer’s existence can be unknown; benchmark privacy is an anticontamination design choice, not an independent demonstration that leakage or benchmark-selection bias is absent. Active advanced-mathematics researchers author the selected problems; two other domain experts independently solve them from scratch before acceptance. Private problems, credentials, check code and run traces were not independently inspected by this report author.

Sources: [21]

Analysis: Study assessment and change conditions - Riemann-Bench: A Benchmark for Moonshot Mathematics. Known-answer private problems isolate a demanding but bounded mathematical task. They do not test whether an agent can decide which open research question has a solution. Confidence is high in that distinction and moderate in the benchmark’s task-specific challenge. Independent reconstruction of answer checks, leakage controls and new expert problems would change confidence in measured capability; broader research usefulness would need a different outcome study.

Sources: [21]

Evidence: Reviewed-work atlas - Process supervision and reasoning: Learning from Less: Measuring the Effectiveness of RLVR in Low Data and Compute Regimes. The study procedurally generates counting, graph and spatial problems, calibrates difficulty using ten models, and compares 100/200/500-example easy versus mixed subsets while fine-tuning Qwen3-4B with LoRA and GRPO under fixed task budgets. Ground-truth checking is programmatic, but graph-answer normalization and spatial fallback parsing use GPT-4o.

Sources: [696]

Evidence: Study result - Learning from Less: Measuring the Effectiveness of RLVR in Low Data and Compute Regimes. Within Table 2’s counting-problem mean-test-accuracy metric, 100 mixed-difficulty examples and 500 easy examples each report 44.2%. The prose separately reports different solving-accuracy values; this table-specific observation is retained without harmonizing the two presentations or generalizing a fivefold gain to other tasks/models.

Sources: [696]

Evidence: Study limitations - Learning from Less: Measuring the Effectiveness of RLVR in Low Data and Compute Regimes. The authors did not repeat training over multiple seeds, used one 4B model with LoRA and fixed low-compute budgets, and did not test transfer to natural-language benchmarks. They explicitly frame results as empirical observations under interacting composition, size, token and optimization limits, rather than universal scaling laws. Counting reward also contains format and verbosity terms. Difficulty dimensions co-vary, and extraction failures and token limits constrain performance. Counting prose gives Mixed100/200/500 solving accuracies of 50.0/50.5/40.0, while Table 2 labels 44.2/43.4/35.5 as mean test accuracy; the two presentations are not reconciled.

Sources: [696]

Analysis: Study assessment and change conditions - Learning from Less: Measuring the Effectiveness of RLVR in Low Data and Compute Regimes. The equal table-specific scores are a useful counterexample to counting examples alone, but they do not establish a portable fivefold data-efficiency factor. A buyer should compare composition and compute jointly and preserve parsing-model dependence. Confidence is moderate in the reported low-budget observation and low in a general scaling claim. Multiple seeds, fixed definitions and matched tests across model sizes and task types would change the efficiency estimate.

Sources: [696]

Evidence: Reviewed-work atlas - Multilingual and multimodal supervision: StereoTales: A Multilingual Framework for Open-Ended Stereotype Discovery in LLMs. StereoTales generates stories under one imposed demographic attribute, extracts profiles with a three-LLM ensemble, and identifies overrepresented attribute associations before a separate harmfulness judgment. 247 UK English-language participants recruited through Prolific and the same 23 generating models rate associations, not the original stories.

Sources: [704]

Evidence: Study result - StereoTales: A Multilingual Framework for Open-Ended Stereotype Discovery in LLMs. For 1580 selected associations, the paper reports Spearman correlation 0.62 between per-association mean human and mean model harmfulness ratings. This is a correlation against one UK normative reference panel, not accuracy against universal moral ground truth or independent validation of all extracted stories.

Sources: [704]

Evidence: Study limitations - StereoTales: A Multilingual Framework for Open-Ended Stereotype Discovery in LLMs. The human reference is a single-country English-language panel, 89.9% White, while generated stories span ten languages. LLM profile extraction was manually checked only on 30 English/French stories; pairwise co-occurrence does not decompose causal mediation, and both extraction errors and nonuniform abstention can affect associations. There is an unresolved instrument conflict: section 3 and Figure 10 ask harmful-stereotype agreement, while Appendix F.2 describes acceptability with a different direction. Section 3 describes median aggregation, whereas section 4.3 and Figure 4 explicitly compare per-association means over 1,580 associations. The reported correlation is retained at that stated mean-comparison scope, without claiming the instruments are equivalent.

Sources: [704]

Analysis: Study assessment and change conditions - StereoTales: A Multilingual Framework for Open-Ended Stereotype Discovery in LLMs. Multilingual generation and the normative reference panel have different coverage. The measured correlation cannot certify each extracted association or stand in for local harmfulness judgments. Confidence is high in this boundary and moderate in using the framework to generate review candidates. A buyer should commission independent local panels and extraction audits; cross-country agreement and sensitivity to abstention would change the validity assessment. Before using the harmfulness correlation operationally, reconcile the source’s agreement-versus-acceptability instrument wording and mean-versus-median descriptions; a local-panel study cannot repair an unexamined measurement direction.

Sources: [704]

Evidence: Reviewed-work atlas - Human simulation and measurement: AttuneBench: A Conversation-Based Benchmark for LLM Emotional Intelligence. AttuneBench uses 200 multi-turn human/model conversations from 11 US native English participants. Participants supply emotional-state, observed/preferred behavior and response-ranking annotations, with GPT-4o-mini suggesting question subsets and a preference-informed model response helping the reference-editing workflow. Default evaluation predicts annotations without future turns; draft-judge scores are supplementary, not Composite inputs.

Sources: [699]

Evidence: Study result - AttuneBench: A Conversation-Based Benchmark for LLM Emotional Intelligence. Within the default-mode eleven-model panel, Emotion F1 did not reach omnibus significance($p=0.099$), while Pairwise Accuracy differentiated model behavior more strongly. The paper’s conversation-level correlations and model-level correlations are separately defined; these findings do not validate a universal one-dimensional emotional-intelligence ranking.

Sources: [699]

Evidence: Study limitations - AttuneBench: A Conversation-Based Benchmark for LLM Emotional Intelligence. Only 11 participants contribute 200 imbalanced conversations, with self-reported demographic/diagnostic data and cultural/language restriction. Pooled tests treat conversations as independent despite repeated participants. The authors explicitly exclude clinical mental-health validation and safety-critical deployment decisions from the benchmark's intended use. The three-rater human baseline covers seven conversations. Golden responses are participant edits after viewing original and preference-informed model outputs. Default, future-informed Omniscient and supplementary draft-judge settings differ; conversation-level Pearson r=-0.097 and model-level Spearman rho=-0.36 have different units.

Sources: [699]

Analysis: Study assessment and change conditions - AttuneBench: A Conversation-Based Benchmark for LLM Emotional Intelligence. The eleven participants, repeated conversations and task-specific annotations are the relevant denominators. A buyer should evaluate the intended conversational behavior with participant-level splits and preserve preferred behavior separately from inferred emotion. Confidence is moderate in the benchmark as a bounded research instrument and unestablished for clinical or safety-critical decisions. Larger culturally varied panels and participant-clustered inference would change its external-validity assessment.

Sources: [699]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: Agents' Last Exam. ALE separates a task workflow from runnable task instances and implements load/start/evaluate interfaces over expert-provided artifacts and references. Expert/engineering QC and code-based checks are the default, but selected visual/free-text workflows still use targeted LLM probes; static grader-type percentages refer to the open reference workflow tree, not all 1490 inventory instances.

Sources: [683]

Evidence: Study result - Agents' Last Exam. In Table 1's 67-instance Near-Term tier, the Codex/GPT5.5 row reports 38.1% full-pass and 64.7% mean score. Full-pass and fractional-score metrics are distinct, and repeated runs cover only a subset of configurations; these reported values do not establish universal professional-work success or economic output.

Sources: [683]

Evidence: Study limitations - Agents’ Last Exam. The release figure reports 150 public, 1017 private and 323 unverified instances out of 1490; later text describes a 152-task public set while the executed-inventory appendix says 150. This unresolved version/count discrepancy and workflow-versus-instance distinction must remain explicit rather than silently treated as one complete reviewed denominator. The thirteen-domain/fifty-five-subdomain taxonomy is SOC/O*NET-grounded and uses LLM-assisted mapping. Table 1 labels tiers of 67, 55 and 38 instances, which sum to 160, while other passages describe 150 or 152 public tasks. These contexts do not resolve whether the tiers overlap. The conclusion also calls 960 units task workflows, while the figure calls them external-submission variants. No reconciled inventory or implied overlap is supplied.

Sources: [683]

Analysis: Study assessment and change conditions - Agents’ Last Exam. Runnable instance counts, workflow definitions and public availability are separate inventory questions. A buyer should obtain an executed task manifest and grader map before interpreting the broad suite total or any pass rate. Confidence is high that full-pass and fractional scores must remain distinct, and moderate in the value of the documented interface. Resolved release counts and independently reproduced grader behavior across instance strata would change the coverage judgment. The 160 tier-count sum must be reconciled with the 150/152 public inventory descriptions before treating tiers as a complete partition.

Sources: [683]

Evidence: Reviewed-work atlas - Rubrics and executable evaluation: ComplexConstraints and Beyond: Expert Rubrics for RLVR. ComplexConstraints describes a three-stage expert prompt/reference/rubric authoring and independent-review workflow, plus iterative LLM-judge calibration and response edits designed to flip verdicts. Qwen3-4B training uses about 900 of the 1000 companion prompts, with GPT5-mini judging required criteria; the 75-prompt public benchmark is separate.

Sources: [5]

Evidence: Study result - ComplexConstraints and Beyond: Expert Rubrics for RLVR. Table 3 reports Qwen3-4B mean per-criterion pass rate increasing from 57.9% to 73.4% on the held-out ComplexConstraints training split. This criterion-average metric differs from the all-criteria task pass rates on the separate 75-prompt public benchmark.

Sources: [5]

Evidence: Study limitations - ComplexConstraints and Beyond: Expert Rubrics for RLVR. The instruction-following result is one training seed, and the authors have not quantified seed-to-seed variance. They describe causal isolation of rubric-design choices and transfer mechanisms as future work; comparisons with other pipelines use different base models/training setups and are suggestive rather than controlled. The second expert reviews and a senior reviewer adjudicates disagreements; GPT-5-mini supplies training judgments. Its two public comparison tables use different-time, different-suite snapshots. Connected CoreCraft results are prior-work reports, not an independently replicated result of this rubric study.

Sources: [5]

Analysis: Study assessment and change conditions - ComplexConstraints and Beyond: Expert Rubrics for RLVR. A criterion-average training improvement does not establish all-criteria success on the separate public benchmark. The independent author/reviewer and calibration loop is a concrete service unit worth testing. Confidence is moderate for that design relevance and low for a general RLVR superiority claim. Repeated seeds and a controlled rubric-design ablation on unseen prompts would change the assessment of which human stages produce value.

Sources: [5]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: GDP.pdf: Benchmarking Grounded Multimodal Reasoning over Professional PDF Documents. GDP.pdf retains 100 expert-authored question/PDF tasks across ten professional domains after at least two frontier models make material errors. Experts write atomic rubric criteria; each criterion is judged separately by Gemini 3.5 Flash without the source PDF or reference answer. Strict pass requires every criterion, while mean rubric score gives partial credit.

Sources: [8]

Evidence: Study result - GDP.pdf: Benchmarking Grounded Multimodal Reasoning over Professional PDF Documents. The July 2026 snapshot reports 2-30.7% strict pass rates across 17 listed models on 100 items, averaging five runs per model with native PDF input and no tools/additional context. These are all-criteria item pass rates on this selected suite, not document-work success rates in general.

Sources: [8]

Evidence: Study limitations - GDP.pdf: Benchmarking Grounded Multimodal Reasoning over Professional PDF Documents. Every retained item was selected because at least two frontier models failed it at collection time. The authors explicitly caution that absolute pass rates therefore describe adversarially selected professional tasks rather than professional document work at large; the live leaderboard also evolves beyond this frozen July 2026 snapshot. The judge does not receive the source PDF or reference answer; the authors report calibration and hand-check discussed failures, but this is not independent calibration by this synthesis. Provider-native PDF processing is part of the measurement. Apache 2.0 benchmark wording does not displace third-party PDF rights.

Sources: [8]

Analysis: Study assessment and change conditions - GDP.pdf: Benchmarking Grounded Multimodal Reasoning over Professional PDF Documents. Adversarial selection changes the meaning of absolute scores, and judging without the source PDF creates a separate evidence-access boundary. A buyer should retain the selection rule, native document presentation and strict-versus-partial metrics, then validate the judge on source-dependent errors. Confidence is high in these conditions and unestablished for general document-work accuracy. A representative private sample with source-visible expert adjudication would change the deployment estimate.

Sources: [8]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: HANDBOOK.md: A Benchmark for Long-Context Agentic Instruction Following. HANDBOOK.md describes 65 tasks using mutated expert-written 20-124-page policies and 824 programmatic criteria over final workspace/service state. Required outputs and prohibited behavior are checked separately; strict pass@1 requires every criterion and averages four trials per task, while N−1 is a separately labeled relaxed metric.

Sources: [10]

Evidence: Study result - HANDBOOK.md: A Benchmark for Long-Context Agentic Instruction Following. The frozen July 2026 Table 2 reports a highest strict pass@1 of 36.2% among 30 evaluated model/configuration entries on 65 tasks with four trials each. This is the averaged all-criteria trial metric in one specified harness, not a fraction of unique tasks permanently unsolvable or a current vendor ranking.

Sources: [10]

Evidence: Study limitations - HANDBOOK.md: A Benchmark for Long-Context Agentic Instruction Following. The study uses fictional company environments, expert-mutated policies, mock services and one common OpenHands harness under explicit call/time caps. Policy mutation is intended to resist memorization; the protocol does not by itself prove uncontaminated pretraining or establish naturalistic corporate compliance and safety outside these tasks. The common harness caps trials at 200 tool calls and one hour; provider/transport errors are excluded and rerun. The July table lists thirty configurations, while the cost section refers to twenty release configurations. Deterministic criteria check the specified state, without proving all possible corporate consequences are covered.

Sources: [10]

Analysis: Study assessment and change conditions - HANDBOOK.md: A Benchmark for Long-Context Agentic Instruction Following. Required and prohibited actions should be explicit acceptance dimensions. Mutated policies help pose a changing-instruction challenge, but a single relaxed missed criterion could be a decisive control. Confidence is high that strict and N−1 scores must remain separate and moderate in this testing approach. A buyer should add domain-specific critical gates; independently reviewed realistic policies and new harnesses would change the generalization assessment.

Sources: [10]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: APEX-Accounting. The pinned v2 paper describes 160 private tasks across ten synthetic company worlds and a separate 10-task public dev world. Experts author/solve tasks and unweighted binary rubrics, with synthetic world-file population under expert review. The main metric averages rubric criteria met over three selected runs per task; model executions number eight per task before infrastructure exclusions.

Sources: [697]

Evidence: Study result - APEX-Accounting. For the pinned nine-model Loop-harness snapshot, the paper reports a highest Mean Criteria@3 of 56.4%, while no model exceeds 2.6% Pass^8. Mean criterion completion and all-eight-attempt complete-task reliability are different metrics and do not measure a fraction of accounting jobs automated.

Sources: [697]

Evidence: Study limitations - APEX-Accounting. World/task difficulty was selected using low scores from three model families; clarification/human-in-the-loop tasks were excluded. The scope omits tax, audit, consolidation, multi-entity/multi-currency work and external reporting. Judge agreement is a calibration/selection-sample result with residual error, not independent proof of general benchmark accuracy. Forty-two experts specify/review worlds and author/solve tasks; most world files are synthetically populated under review, and independent experts re-solve twenty tasks. The GEPA-prompted DeepSeek-v4-Flash judge is selected/calibrated on 1,687 majority-vote labels from 120 trajectories and three human raters; 97.1% agreement is a calibration result. Mean Criteria@3 selects three runs lexicographically by task ID; model-caused failures score zero, while infrastructure exclusions leave 156-160 tasks in paired comparisons. The authors’ 37.1% judge-as-solver stress test does not independently eliminate self-preference. This paper-specific entry uses the v2 160-task count; the separately captured leaderboard conflict is disclosed in the family entry rather than attributed to this paper-only repair.

Sources: [697]

Analysis: Study assessment and change conditions - APEX-Accounting. Partial criterion completion is commercially different from repeatedly complete accounting work. Excluded clarification tasks also remove an interaction mode many deployments require. Confidence is high in these measurement distinctions and unestablished for accounting-job automation. A buyer should audit criteria with qualified practitioners, introduce ambiguous cases requiring escalation and retain all execution attempts. Real-world validated tasks and independently calibrated judging would change the use assessment.

Sources: [697]

Evidence: Reviewed-work atlas - Transfer and generalization: Cross-Benchmark Generalization in Long-Horizon Agents. The paper post-trains Qwen3.5-122B-A10B on 363 MCP tasks using a selected 3000-trajectory Kimi K2.6 supervised warmup followed by one-epoch RL. Deterministic criterion rewards are combined with a low-score effort floor capped at 0.40; evaluation uses native benchmark scores rather than this shaping reward.

Sources: [702]

Evidence: Study result - Cross-Benchmark Generalization in Long-Horizon Agents. For the updated greedy software-engineering evaluations, Table 2 reports SWE-BenchPro 20.5→26.3% and Terminal-Bench 2 44.9→47.7% after the combined SFT+RL pipeline. The authors report descriptive point estimates from one training run and distinguish these from older four-sample diagnostics and retained tau2 results.

Sources: [702]

Evidence: Study limitations - Cross-Benchmark Generalization in Long-Horizon Agents. Toolathlon informed initial base-family/teacher choices and contributed reused MCP surfaces, although no external benchmark tasks/graders entered training according to the authors. Pretraining contamination cannot be audited, the no-effort-floor ablation was not run, and exploratory outcome-conditioned behavior patterns do not establish a causal transfer mechanism. Updated greedy evaluations and older four-sample diagnostics have different provenance. The latter are not identified as standard unweighted any-of-four pass@4. One training run, correlated benchmarks and unavailable paired units limit uncertainty and replication claims.

Sources: [702]

Analysis: Study assessment and change conditions - Cross-Benchmark Generalization in Long-Horizon Agents. The combined supervised and reinforcement stages, task selection and shaping reward are a single treatment here. The missing no-effort-floor ablation prevents attributing transfer to that component alone. Confidence is moderate in the descriptive checkpoint comparison and low in a causal mechanism. Repeated seeds and equal-budget stage/effort-floor ablations with held-out model families would change the transfer explanation.

Sources: [702]

Evidence: Reviewed-work atlas - Transfer and generalization: Post-Training on Office Work Improves Software Engineering: A Behavioral Account of Cross-Domain Transfer. The study describes a combined SFT-plus-GSPO post-training experiment on 363 LHMTA office/general-tool tasks, followed by analysis of 103 base-fail/trained-pass task pairs using Claude Opus 4.8-assisted reports manually checked against raw trajectories. This is an outcome-conditioned analysis, not a random sample of all agent behavior.

Sources: [685], [702]

Evidence: Study result - Post-Training on Office Work Improves Software Engineering: A Behavioral Account of Cross-Domain Transfer. For the same base/trained checkpoints, Table 3 reports greedy SWE-BenchPro pass@1 changing from 20.5% to 26.3%. Over 731 paired SWE tasks, Table 5 separately reports the share containing a formal-test command changing from 37.5% to 73.3%; test invocation is an indirect behavior measure and does not establish semantic correctness.

Sources: [685], [702]

Evidence: Study limitations - Post-Training on Office Work Improves Software Engineering: A Behavioral Account of Cross-Domain Transfer. All results come from one model/training run. No task-demand ablation or equal-budget software-training counterfactual was performed; outcome-conditioned examples and analyst-selected explanations cannot establish a causal mechanism, prevalence of capabilities or that office training is the most efficient path to SWE improvement. The paper describes reward as trajectory-level criterion fraction, while companion 2608.00181v1 additionally describes an effort floor and Toolathlon-informed family/teacher choice. Matching checkpoints, recipe and values suggest shared experimental provenance; that is an author inference, not proof of independent replication or permission to equate every protocol description.

Sources: [685], [702]

Analysis: Study assessment and change conditions - Post-Training on Office Work Improves Software Engineering: A Behavioral Account of Cross-Domain Transfer. The successful-change subset is selected by outcome, so its behavioral themes are hypotheses rather than population frequencies or causal mediators. Test-command invocation is observable but does not establish a correct diagnosis or patch. Confidence is moderate in using the trace analysis to generate follow-up tests and low in a mechanism claim. A randomized task-demand ablation and equal-budget software-training counterfactual would change the explanation. Apparent agreement with the companion paper is not additional independent experimental replication, and the two reward descriptions retain their recorded difference.

Sources: [685], [702]

Evidence: Reviewed-work atlas - Multilingual and multimodal supervision: Chartography: A Benchmark for Professional Chart Understanding. Chartography uses 100 professionally authored chart questions verified by three additional experts and screened for at least one meaningful frontier-model failure. Each configuration receives 20 no-tools trials per task; Gemini 3.5 Flash judges the final answer against expert targets/ranges without seeing the chart, and multipart answers require all parts.

Sources: [3]

Evidence: Study result - Chartography: A Benchmark for Professional Chart Understanding. The frozen July 2026 table reports 45.0% mean pass@1 for the leading configuration across 100 tasks × 20 scored trials, with a reported 95% CI half-width 1.1 percentage points. The metric is the fraction of all graded trials passed; displayed ranks and overlapping intervals should not be interpreted as statistically distinct capability tiers.

Sources: [3]

Evidence: Study limitations - Chartography: A Benchmark for Professional Chart Understanding. Items are adversarially screened and require specialist standing knowledge under a no-tools single-image protocol. The judge shares identity with an evaluated configuration, and online chart provenance does not establish redistribution rights. Results do not estimate general professional decision accuracy, deployed agent performance or causal superiority of a specific visual mechanism. The 100 tasks use 98 byte-distinct images, including 81 online-sourced and 19 author-created charts. Twenty trials per task produce 2,000 responses per configuration, not independent new tasks. The self-judging configuration and qualitative trajectory explanations do not establish independent judge validity or a controlled perception-versus-reasoning causal ablation.

Sources: [3]

Analysis: Study assessment and change conditions - Chartography: A Benchmark for Professional Chart Understanding. Expert verification, adversarial screening and repeated model trials each answer a different question. The reported uncertainty does not establish distinct capability tiers, and an answer-only judge can miss visual-grounding errors. Confidence is high in retaining those boundaries and moderate in the benchmark’s target relevance. Source-visible independent adjudication, chart-rights review and representative professional-chart sampling would change buyer confidence.

Sources: [3]

Evidence: Reviewed-work atlas - Transfer and generalization: Cross-Benchmark Transfer from RL on Agentic Coding Tasks. The experiment applies one GSPO epoch with a rank 32 LoRA adapter to Kimi K2.7 Code on 1000 expert-built repository tasks and 700 terminal tasks, without a supervised stage. Reward is target-check fraction, but any failed pass-to-pass regression test zeroes repository reward; terminal tasks have no such regression-test gate.

Sources: [698]

Evidence: Study result - Cross-Benchmark Transfer from RL on Agentic Coding Tasks. For SWE-BenchPro’s 731-task public split under mini-swe-agent, Table 2 reports pass@1 increasing 60.1→64.8%, or 35 additional solved tasks, in matched in-house base/trained runs. This is one trained checkpoint/run and does not establish generalization uncertainty over training seeds.

Sources: [698]

Evidence: Study limitations - Cross-Benchmark Transfer from RL on Agentic Coding Tasks. Headline DeepSWE and Terminal-Bench 2.1 baselines are public scores, whereas other baselines are in-house; DeepSWE’s in-house base differs from its public baseline. Terminal-Bench 4 revises 3 and is not an independent task family, and eight trained Terminal 3 runs were excluded for harness failures. These provenance/denominator distinctions must accompany cross-benchmark aggregation. Partial credit weights checks/verifier criteria equally, which need not equal one vote per semantic requirement. Eight excluded Terminal-Bench 3 runs leave eight passes among 66 scored runs; Terminal-Bench 4 revises 3 within one family. Terminal-Bench 2.1 has no in-house base run, and its public baseline averages three runs.

Sources: [698]

Analysis: Study assessment and change conditions - Cross-Benchmark Transfer from RL on Agentic Coding Tasks. Protected regression tests define a meaningful acceptance constraint, but public and in-house baselines cannot be pooled as if all were matched reruns. Confidence is moderate in the matched SWE-BenchPro change and low in a universal transfer advantage. A buyer should require complete harness-failure accounting, an independent task manifest and repeated training seeds; an equal-budget control with stable benchmark versions would change the causal interpretation.

Sources: [698]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: DAYJOB: A Benchmark for Long-Horizon Professional Work. DAYJOB describes 130 tasks: 50 healthcare and 80 finance, of which all 50 healthcare and 50 finance are publicly released. A Claude Opus 4.8 agentic judge inspects delivered files against unweighted binary expert criteria. Strict pass@1 requires every criterion, averages scored attempts within each task and weights tasks equally by domain; error attempts are excluded.

Sources: [695]

Evidence: Study result - DAYJOB: A Benchmark for Long-Horizon Professional Work. The frozen September 2026 launch table reports its leading configuration at 24.7% strict pass@1 on healthcare and 23.9% on finance, with at most five scored attempts per task. Relaxed 95%/90% or one-missed-criterion scores are distinct and can accept a missed decisive control.

Sources: [695]

Evidence: Study limitations - DAYJOB: A Benchmark for Long-Horizon Professional Work. Healthcare completion-time figures are creator estimates rather than observed completion trials: range midpoints are averaged and more-than-40 hours is capped at 40. The selected task workspaces/rubrics and model judge do not establish population-level professional accuracy, safe autonomous practice or actual labor-time savings. The midpoint/capped-time convention is established for healthcare; the finance mean is separately reported on its dataset card and is not assigned that convention here. Section 5.3 cost covers agent and condenser model calls while excluding the judge. The 50 released finance tasks and the 80 evaluated finance tasks are different denominators.

Sources: [695]

Analysis: Study assessment and change conditions - DAYJOB: A Benchmark for Long-Horizon Professional Work. The buyer’s first question is complete delivery under required controls, followed by its cost and escalation burden. Estimated professional completion time cannot be multiplied by model scores to claim observed labor savings. Confidence is high in that restriction and moderate in the selected-task evaluation design. Prospective practitioner timing, complete execution-error accounting and independent file-level judging would change the economic and deployment interpretation.

Sources: [695]

Evidence: Reviewed-work atlas - Professional and long-horizon evaluation: PLE: Physical-Engineering’s Last Exam. PLE describes 85 tasks: 44 engineer-authored and 41 human/agent-coauthored. Coauthored tasks use expert briefs, a Claude Opus 5 Creator-Solver loop and final engineer validation. Expert tasks receive LLM-judged weighted rubric scores; most coauthored tasks combine constraint fraction with a hard-gated objective term, with four bespoke grading exceptions.

Sources: [700]

Evidence: Study result - PLE: Physical-Engineering’s Last Exam. For the pinned evaluation, Claude Opus 5.5 in Claude Code reports 50.6% mean task score across 85 tasks and three runs per task. This combines rubric and executable/hybrid scores and is not a strict task pass rate or the percentage of engineering assignments ready for licensed signoff.

Sources: [700]

Evidence: Study limitations - PLE: Physical-Engineering’s Last Exam. The hybrid score is s = 0.4 p + 0.6 G Oprime: a failed hard-constraint gate removes objective credit but does not generally zero the constraint-fraction term. Single-task run variance and model-judge self-preference are reported; benchmark scores therefore cannot certify safety, feasibility or real engineering deployment readiness. Thirty-seven coauthored tasks use the stated constraint/objective formula; four site/MEP tasks have bespoke graders with separately described gates. Fourteen models have restricted harness coverage, rather than every model being crossed with all seven common harnesses; Grok Build is model-specific. The paper plans release of 29 tasks without establishing their present availability. Model regrading is not fresh human engineering validation.

Sources: [700]

Analysis: Study assessment and change conditions - PLE: Physical-Engineering’s Last Exam. A hybrid mean task score is not the share of engineering deliverables ready for signoff. A failed hard gate can still leave constraint-fraction credit, so buyer acceptance must inspect the gate and each decisive physical requirement directly. Confidence is high in this scoring distinction and unestablished for engineering safety. Independent solver validation, verified physical checks and professional review of new tasks would change the deployment assessment.

Sources: [700]

Analysis: Cross-study synthesis - aggregation, selection and legitimate disagreement. MACE and Crowd-Kit assume an objective categorical target in their aggregation settings; style clustering evaluates consistency with subjective groupings; MMRB2 discards high-disagreement and near-midpoint pairs. Those procedures can improve a chosen reference fit while changing which cases are represented. The analyst judgment is to predeclare whether disagreement is error, uncertainty or an intended target distribution. Confidence is high in this design distinction. An independently adjudicated target-use study would change which aggregation/filtering choice is appropriate; no cross-paper accuracy winner is inferred.

Sources: [462], [466], [470], [558]

Analysis: Cross-study synthesis - transfer evidence is a hierarchy of comparisons. Corecraft reports a training-time held-out result and external transfer; the office-to-software papers describe connected checkpoints and an outcome-conditioned behavioral analysis; the coding-task study mixes public and in-house baseline provenance outside its matched SWE-BenchPro comparison. These reports should not be counted as independent replications of one causal mechanism. Confidence is moderate in the documented point-estimate improvements and low in explaining them through a single human-data property. Repeated seeds, equal-budget controls and ablations that isolate task demands, supervised warmup and reward shaping would change the causal assessment.

Sources: [6], [685], [698], [702]

Analysis: Cross-study synthesis - the verifier is part of the system under test. Weaver's selection depends on its candidate pool and many verifier calls; PivotRL's SWE local check can accept a tool-call name without validating arguments or patch quality; ALCE's NLI support metric can miss partial support; PLE's failed hard gate does not zero every score component. A technically inspectable acceptance report therefore retains input visibility, check logic, false accepts, false rejects and final-task outcomes. Confidence is high in these distinctions and unmeasured in any buyer deployment. Independent falsification cases and representative, blinded adjudication would change the trust placed in each verifier.

Sources: [549], [556], [564], [700]

Evidence: Reviewed-work atlas - Multilingual and multimodal supervision: CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. The v1 paper reports a four-stage chart selection process: SigLIP similarity filtering of figures extracted from January 2020-September 2023 arXiv preprints, graduate-student selection for visual diversity, removal of near-duplicates, and exclusion of blurry or unclearly labeled charts. The v1 paper acknowledges Microsoft Accelerate Foundation Models Academic Research Initiative support, an Apple Scholars in AIML Fellowship for Mengzhou Xia, and a Gordon Wu Fellowship for Luxi He. The v1 paper uses GPT-4o to extract responses and assign binary correctness grades rather than exact string matching, and acknowledges that automated grading errors or human participant mistakes may affect its measurements. The v1 paper describes graduate-student reasoning-question annotation using ten GPT-4V example question-answer pairs per chart; annotators could choose a suggestion, modify it, or write their own question.

Sources: [634], [639]

Research gap: Study result scope - CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. The admitted CharXiv records establish chart selection, question origins and model-assisted grading, with source-count and known-answer limitations. They do not supply a comparative model-performance table for this atlas entry; no chart-understanding score is inferred from the construction procedure.

Sources: [634], [635], [639]

Evidence: Study limitations - CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. Question authors may select or modify GPT-4V suggestions or write their own; human curation does not mean all questions are wholly human-authored. The descriptive interface also supplies model-proposed answers. GPT-4o grading can err, and funding acknowledgments do not establish annotation employment, commercial supplier delivery or chart rights.

Sources: [634], [635], [639]

Analysis: Study assessment and change conditions - CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. The useful provenance unit is the chart, question origin and grading path. A broad human-authored label would hide whether a suggestion was accepted, edited or replaced. Confidence is high that those stages should be retained and moderate in the benchmark as an instrument for scientific-chart questions. Compare a sample with model-independent questions and source-visible expert grading; resolved image-count and known-answer issues would increase confidence. Evidence of differential errors by question origin would change the recommended curation mix.

Sources: [634], [635], [639]

Evidence: Reviewed-work atlas - Hardware design and verification: Comprehensive Verilog Design Problems: A Next-Generation Benchmark Dataset for Evaluating Large Language Models and Agents on RTL Design and Verification. The v1 paper uses BLEU for RTL/testbench correspondence tasks and an o4-mini model judge for technical question answering, while stating that further investigation is needed into the technical reliability of those question-answer scores. The v1 paper reports engaging approximately 35 hardware engineers with more than four years of Verilog and verification experience to author problems, with doctoral or engineering-management subject-matter experts reviewing each problem for accuracy, task fit and scope. The v1 paper reports pass-at-one estimated from five samples and states that its Agentic datapoints were converted to Non-Agentic single-turn format for the published model evaluation. The v1 appendix describes sanity checks requiring each reference solution to pass and its initial context to fail, followed by model-judge scoring of ambiguity, consistency, category match and behavioral match with an aggregate passing threshold of 8.0.

Sources: [643], [644]

Research gap: Study result scope - Comprehensive Verilog Design Problems: A Next-Generation Benchmark Dataset for Evaluating Large Language Models and Agents on RTL Design and Verification. The admitted CVDP records establish task formats, five-sample pass-at-one estimation and validation procedures. They do not supply a separately admitted numerical model-performance result. The single-turn conversion of Agentic data is retained without inventing an agent-execution score.

Sources: [640], [641], [642], [643], [644], [645]

Evidence: Study limitations - Comprehensive Verilog Design Problems: A Next-Generation Benchmark Dataset for Evaluating Large Language Models and Agents on RTL Design and Verification. Agentic datapoints were converted to single-turn form for the published evaluation. BLEU, model-judged technical answers and executable design checks are different outcomes. Approximately 35 engineers and their experience are author reports, without independent staffing verification. The appendix does not report the automated prompt-revision experiment; no published-data benefit is inferred from it.

Sources: [640], [641], [642], [643], [644], [645]

Analysis: Study assessment and change conditions - Comprehensive Verilog Design Problems: A Next-Generation Benchmark Dataset for Evaluating Large Language Models and Agents on RTL Design and Verification. A hardware procurement brief must specify the verification toolchain and what the task actually exposes. Oracle context and a single-turn evaluation do not test full design discovery or a multi-turn engineering agent. Confidence is high in this scope distinction and unmeasured for fabrication readiness. A pilot should compile and simulate representative designs, test alternative valid implementations, and inspect any proprietary-tool dependency and component license gap. Independently checked behavior beyond the published tests would change the acceptance boundary.

Sources: [640], [641], [642], [643], [644], [645]

Evidence: Reviewed-work atlas - Research reports and source-grounded rubrics: DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report. The captured DeepResearch Bench II README distinguishes Apache 2.0 evaluation code from per-task data licenses, reporting 129 CC BY 4.0 tasks, two CC BY-NC 4.0 tasks and one CC0 task; the captured code LICENSE is Apache 2.0. DeepResearch Bench II requires rubrics to check essential, indivisible facts or inferences and encode the target content itself, rather than merely naming a broad topic. The creator README announces a separate DeepResearch Bench II dataset release containing articles generated by a subset of the evaluated models. The DeepResearch Bench II paper describes 132 research tasks across 22 domains, grounded in selected expert-written reports and evaluated with 9,430 rubrics. At the captured creator commit, the DeepResearch Bench II README describes migration to GPT-5.5 evaluation and temporarily excludes tasks 112 and 114 from official leaderboard scoring while retaining all 132 dataset tasks. The pinned DeepResearch Bench II README describes a text-only evaluation path: DOCX text and tables are extracted, embedded images are ignored, and Markdown is read as plain text.

Sources: [610], [611], [612]

Evidence: Study construction and protocol - DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report. DeepResearch Bench II reports human inspection of all generated tasks and rubrics, followed by more than 400 human-hours of domain-expert review, discussion and iterative refinement. The DeepResearch Bench II pipeline first uses an LLM to extract tasks and rubrics from expert articles, then scores each source article against its generated rubrics and regenerates them if Information Recall or Analysis scores below 90%. The original DeepResearch Bench II experiment uses Gemini 2.5 Pro as judge, evaluates batches of 50 rubrics, and describes task scores as the fraction of rubric requirements passed. The DeepResearch Bench II paper limits its Presentation dimension mainly to report formatting and layout and says adaptation to a user's knowledge and preferences remains unrealized. DeepResearch Bench II adds source-article titles and URLs to a prompt block list and performs a secondary check for generated answers supported by citations to those articles, excluding leaked answers from scoring.

Sources: [611]

Evidence: Study construction and protocol - DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report. For DeepResearch Bench II, the authors manually select source reports requiring open-source information gathering and substantial synthesis, excluding findings that depend on physical experiments or complex data-mining and stochastic machine-learning pipelines.

Sources: [611]

Evidence: Study result - DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report. The DeepResearch Bench II batch-size ablation uses ten human-labeled model reports in PDF, DOCX, Markdown and HTML formats; with Gemini 2.5 Pro judging 50 rubrics per batch, the paper reports 91.75% accuracy and F1 of 89.57. The captured DeepResearch Bench II README reports a three-way-label comparison on ten reports with 738 matched rubric judgments: GPT-5.5 achieves 91.19% exact-match accuracy, Cohen's kappa of 0.7993 and task-level Pearson correlation of 0.9587. At batch size 50 on the same human-labeled report set, DeepResearch Bench II reports Gemini 2.5 Pro accuracy/F1 of 91.75/89.57, compared with 88.36/78.28 for GPT 5 and 89.45/79.79 for Gemini 2.5 Flash. In DeepResearch Bench II Table 3, OpenAI-GPT-o3 Deep Research has the highest reported total score, 45.40, and Information Recall score, 39.98; Gemini 2.5 Pro Deep Research has the highest Analysis score, 51.91.

Sources: [611], [612]

Evidence: Study limitations - DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report. The original Gemini 2.5 Pro ten-report ablation differs from the later GPT-5.5 ten-report/738-judgment study. The later text-only path ignores embedded images, retains 132 tasks but scores 130 after excluding two, and changes per-task licensing from earlier blanket language. The selected contexts do not establish how a -1 leakage flag enters the final aggregate. Source-self-consistency filtering at 90% is not judging accuracy for arbitrary agent reports.

Sources: [610], [611], [612]

Analysis: Study assessment and change conditions - DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report. Content-bearing rubrics make omissions inspectable, but their source report can become a shortcut and their evaluator can miss information carried only in a chart. Confidence is high that source access, file extraction and scoring version belong in any comparison, and moderate in this rubric pattern for research procurement. A buyer should test blocked-source paraphrases, missing-figure evidence and subtle fabricated support with independent experts. Resolved leakage aggregation and representative held-out rubric judgments would change confidence in scoring.

Sources: [610], [611], [612]

Evidence: Reviewed-work atlas - Repository engineering and executable verifiers: DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. DeepSWE contains 113 tasks across 91 active public repositories in TypeScript, Go, Python, JavaScript and Rust; its repository-selection rules require at least 500 GitHub stars and a permissive open-source license. For eight models with publicly reported SWE-Bench Pro results, the DeepSWE paper reports a 69.8-percentage-point pass-rate range on DeepSWE versus 29.7 points on SWE-Bench Pro. DeepSWE evaluates 16 model-and-reasoning-effort configurations under mini-swe-agent with a shared prompt and single bash tool, pinned to commit adfe20233d456104c38c3129161b54f0fd39f2c7. DeepSWE uses binary functional rewards without partial credit and does not score code readability, idiomatic style, maintainability, performance or documentation. Every DeepSWE task includes a prompt, executable functional verifier and reference solution; verifiers exercise public APIs and observable outputs rather than requiring the reference implementation's internal structure. DeepSWE accepts tasks only after LLM-assisted analysis and independent human review of the prompt, verifier, reference solution and diagnostic agent runs.

Sources: [607]

Evidence: Study construction and protocol - DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. The pinned DeepSWE README says that since v1.1, grading uses a separate verifier environment and requires Pier newer than 0.3.0: agent commits are collected as a patch and applied in a pristine grading container. DeepSWE's harness pilot compares mini-swe-agent with each model's native harness on the same ten SWE-Bench Pro tasks, for Claude Opus 4.7, GPT-5.5 and Gemini 3.1 Pro. DeepSWE reference solutions are authored from scratch rather than copied or adapted from public patches, and the paper says the tasks are never merged upstream; task containers contain a shallow clone at the base commit. The DeepSWE paper announces release of the benchmark, executable verifiers and trajectory records while withholding the exact LLM-reviewer prompt; the captured repository contains an Apache 2.0 LICENSE. DeepSWE reports 95% intervals as pass@1 plus or minus 1.96 times the standard deviation of the roughly four benchmark-run scores divided by the square root of the number of runs.

Sources: [607], [608], [609]

Evidence: Study construction and protocol - DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. The DeepSWE experiment reports 7,174 scored runs across 16 configurations, with 428-452 scored runs per configuration after starting from approximately four attempts per task. DeepSWE's Figure 2 reports mean prompts of 2,158 characters and reference solutions with 668 lines added and 7.4 files edited. DeepSWE defines pass@1 as the equally weighted average of each task's scored-rollout pass fraction, while pass@4 is the fraction of 113 tasks with at least one success among up to four attempts. In the published DeepSWE comparison, a GPT-5.5 judge disagrees with executable verifier decisions on 10 of 735 DeepSWE runs, 1.4%, versus 256 of 789 SWE-Bench Pro runs, 32.4%.

Sources: [607]

Evidence: Study result - DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. In the paper's DeepSWE experiment, GPT-5.5 at xhigh reasoning effort has the highest reported pass@1, 70.0%, with a run-to-run 95% interval of 67.2-72.9%.

Sources: [607]

Evidence: Study limitations - DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. The headline interval measures rerun noise on the same 113 tasks, not task-selection uncertainty. Infrastructure errors are excluded without resampling, while context exhaustion and wall-clock timeout are failures. The pass@4 indicator has a smaller effective budget when fewer than four attempts remain. Judge/verifier disagreement on 735 runs is not an independently measured verifier error rate. The native-harness pilot uses ten SWE-Bench Pro tasks, not the DeepSWE corpus.

Sources: [532], [533], [607], [608], [609]

Analysis: Study assessment and change conditions - DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. Original reference patches and implementation-tolerant functional verifiers make the accepted change explicit, while binary reward still omits maintainability and incomplete progress. Confidence is moderate in the fixed-system task result and high in keeping its uncertainty and exclusions visible. A buyer should separately inspect alternative valid patches, regression coverage and production-code review. Representative repository sampling, more reruns and independent adjudication of disagreements would change confidence in comparative capability.

Sources: [532], [533], [607], [608], [609]

Evidence: Reviewed-work atlas - Scientific workflows and recoverable targets: GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine. The paper states that AI systems assisted problem design and review and manuscript editing, while the authors reviewed and approved the benchmark design, problem specifications, validation targets, evaluation results and final manuscript. The reported evaluation attempted ten independent runs per model-problem pair for standard evaluations and five for GPT Pro (Extended) and Claude Opus; container, tooling, provider and response-format errors were excluded from per-problem pass-rate estimates. The paper attributes benchmark conception, problem implementation and review, external-review coordination and benchmark testing to Jeremy Li, and model evaluation runs to Andrew Ho; both authors drafted and revised the manuscript. The reported primary metric uses Python binary grading: every prespecified target field must satisfy its exact-match rule or absolute numeric tolerance for the run to pass, and the free-text reasoning field is collected for qualitative analysis but is not graded. The paper reports evaluation of the full 129-problem suite across 60 model configurations, including mainline GPT, GPT Pro and non-GPT baseline rows.

Sources: [673]

Evidence: Study construction and protocol - GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine. The paper defines overall pass rate as the unweighted mean of per-problem pass rates across 129 problems and reports 95% hierarchical bootstrap intervals from 20,000 resamples, resampling problems and repeated runs within problems. The authors' manual review of selected GPT-5.5 and GPT-5.6 Sol reasoning traces suggests a notice-act gap: failures may identify local diagnostics without sufficiently revising downstream statistical methods or QC. The reported harness used a Linux Docker environment with Python and R scientific and bioinformatics software and no internet access, restricting agents to the prompt, staged files, installed software and model-internal knowledge. The inspected preprint banner states that the preprint was not certified by peer review and is available under CC-BY-NC 4.0. The paper's construction description and example prompt tell agents that analytical reasoning quality is graded alongside numerical correctness, while its Methods section states that free-text reasoning is not graded. The paper reports no additional uniform wall-clock budget imposed by the evaluation harness, while runs remained subject to provider and platform behavior in the evaluation stack.

Sources: [673]

Evidence: Study result - GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine. Supplementary Table 1 reports Claude Opus 4.8 at max reasoning with a 16.0% overall pass rate and a 95% hierarchical bootstrap interval of 11.0% to 21.2%. Supplementary Table 1 reports GPT-5.6 Sol at max reasoning with a 28.7% overall pass rate and a 95% hierarchical bootstrap interval of 22.5% to 35.1%. Supplementary Table 1 reports GPT-5.6 Sol Pro (Extended) with a 31.5% overall pass rate and a 95% hierarchical bootstrap interval of 24.3% to 38.9%.

Sources: [673]

Evidence: Study limitations - GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine. The paper and public card differ on whether agents are told data come from real experiments or uncertain provenance. The prompt/construction wording says analytical reasoning is graded, while Methods says the free-text reasoning field is not graded. Public release/reference answers, Artificial Analysis tasks and internal tasks are different visibility strata; external-review and release strata were not randomized. The preprint CC-BY-NC 4.0 statement is distinct from the public package MIT declaration.

Sources: [446], [673]

Analysis: Study assessment and change conditions - GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine. Recoverable target design is central: the correct estimand is what staged observations support, rather than an inaccessible simulation parameter. Binary all-field scoring nevertheless collapses an early failure and a late statistical mistake. Confidence is high in those measurement distinctions and moderate in the curated scientific-workflow relevance. The notice-act explanation is selected-trace interpretation, not a measured chain-length law. Independent review of unreviewed tasks, resolved prompt/grader wording and representative real-data studies would change confidence in generalization.

Sources: [446], [673]

Evidence: Reviewed-work atlas - Health conversations and rubric measurement: HealthBench: Evaluating Large Language Models Towards Improved Human Health. The v1 main text reports that 262 physicians created HealthBench tasks over eleven months, with practice experience in 60 countries and 26 specialties, and that all physician cohort members were compensated. The v1 paper describes most conversations as model-synthetic from physician-designed situation types, with additional physician red-team conversations and transformed HealthSearchQA queries described as released by Google; o1-preview filters for relevance, realism, self-consistency and complete messages. The v1 paper validates model grading against physician grades on the 34 predefined consensus criteria and selects GPT-4.1 as its default grader; it acknowledges that most example-specific criteria were written by individual physicians and were not validated by other physicians. The v1 paper grades each criterion as met or not met, applies its signed nonzero point weight from minus ten to ten, divides the sum by the available positive points per example, and clips the mean example score to zero through one.

Sources: [646], [647]

Evidence: Study construction and protocol - HealthBench: Evaluating Large Language Models Towards Improved Human Health. The v1 paper states that HealthBench evaluates single model responses to conversations and does not measure health outcomes of specific deployed workflows, identifying workflow-specific real-world outcome studies as future work.

Sources: [646]

Research gap: Study result scope - HealthBench: Evaluating Large Language Models Towards Improved Human Health. The admitted HealthBench evidence supports provenance, physician roles, scoring and validation limits. It does not supply a numerical benchmark leaderboard for this entry or a measured patient-outcome effect; neither is inferred from physician involvement.

Sources: [646], [647], [648]

Evidence: Study limitations - HealthBench: Evaluating Large Language Models Towards Improved Human Health. The main-text cohort of 262 physicians is distinct from appendix onboarding/rotation/removal counts; no final population is inferred by subtraction. Compensation and credentials are reported rather than independently verified. The 34 predefined consensus criteria receive a validation procedure that does not extend to every individualized rubric. A signed weighted score clipped to 0-1 is not a fraction of tasks fully solved.

Sources: [646], [647], [648]

Analysis: Study assessment and change conditions - HealthBench: Evaluating Large Language Models Towards Improved Human Health. The source makes an especially consequential distinction between judged response quality and patient outcomes. A system can improve a rubric mean while missing a rare severe requirement. Confidence is high in requiring workflow-specific evaluation and moderate in using the benchmark to expose response dimensions. A buyer should separately gate critical contraindications or missing escalation, calibrate individualized rubrics with independent clinicians, and preserve the full signed criterion vector. Prospective patient-relevant workflow studies would be needed to change the outcome conclusion.

Sources: [646], [647], [648]

Evidence: Reviewed-work atlas - Human-time calibration and composite evaluation: Measuring AI Ability to Complete Long Software Tasks. The paper reports approximately eight attempts per agent/task pair and uses general agent scaffolds without task-specific prompting, except that SWAA uses a simple prompting scaffold. The authors state that task-completion horizons depend on the task distribution and on the human reference population’s skill and context. For HCAST, task durations are calculated using the geometric mean time of successful human baseline attempts, with manual estimates for tasks lacking successful baselines; RE-Bench instead receives a fixed eight-hour duration rating. A separate internal-repository study uses five original issues and has maintainers grade submitted solutions manually, rather than requiring every issue to have an automatic checker. The authors label HCAST and RE-Bench tasks for the presence of 16 defined messiness factors and sum those labels into a score from zero to 16. The paper says supplementary material will provide code and data for some core figures, while the full codebase is infeasible to publicize and anonymize; code for additional figures may be provided on reviewer request.

Sources: [617]

Evidence: Study construction and protocol - Measuring AI Ability to Complete Long Software Tasks. The analysis binarizes task scores using task-specific human-performance thresholds: HCAST uses its target score and RE-Bench uses the average score of human runs lasting seven to nine hours. Section 2.1 describes a composite evaluation suite containing 97 selected HCAST tasks, seven RE-Bench environments and 66 Software Atomic Actions tasks. The method fits an agent-specific logistic model against log human task duration; the fitted horizon is the duration with predicted 50% success, and each task receives weight equal to the inverse square root of its task-group size.

Sources: [617]

Evidence: Study result - Measuring AI Ability to Complete Long Software Tasks. The paper extrapolates historical trends toward a one-month horizon defined as approximately 167 human working hours, while treating that result as conditional on continued trends and transfer to real-world software work. The reported 80%-success horizons are four to six times shorter than the 50%-success horizons, while the fitted doubling times are similar at 204 and 207 days respectively. For the reported 2019-2025 model series, the paper estimates a 207-day doubling time with a 95% bootstrapped interval of 166-240 days, using 10,000 resamples over task groups, tasks and runs.

Sources: [617]

Evidence: Study limitations - Measuring AI Ability to Complete Long Software Tasks. The 97 HCAST, seven RE-Bench and 66 SWAA components sum to 170, while the baselining section uses 169; both descriptions remain. Successful-baseline geometric means, missing-baseline forecasts and the fixed eight-hour RE-Bench assignment have different provenance. Humans and agents have different context/qualification conditions. Historical trend intervals do not include future trend changes or most external-validity uncertainty.

Sources: [614], [615], [616], [617], [618]

Analysis: Study assessment and change conditions - Measuring AI Ability to Complete Long Software Tasks. A fitted 50% horizon is a conditional relationship between this task distribution, a human reference and a system, not a universal duration an agent can safely work. The shorter 80% horizon shows why the reliability target must be chosen before applying the number. Confidence is high in this interpretation and low in unconditional extrapolation to workplace timelines. Measured long-task human baselines, explicit failure costs and new-domain validation would change the deployment inference; a prospective stable-suite trend test would change confidence in extrapolation.

Sources: [614], [615], [616], [617], [618]

Evidence: Reviewed-work atlas - Medical question construction and validation: MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. The v1 paper assesses leakage risk before and after rewriting with GPT-4o outputs, using perplexity and similarity to collected explanations measured by ROUGE-L and edit distance. The v1 methods describe model rephrasing intended to preserve question information, removal of weak distractors and setting the final choice counts to ten for Text and five for MM, using GPT-4o-2024-11-20 and Claude-3.5-Sonnet-20241022. The v1 paper describes filtering an initial 37,543-question medical exam/textbook bank using model votes, difficulty thresholds based on human-response Brier scores and prior expert difficulty ratings, and semantic-similarity filtering.

Sources: [649], [650], [651]

Research gap: Study result scope - MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. The admitted MedXpertQA evidence supports question filtering, augmentation and leakage-proxy procedures. It supplies no comparative accuracy value for this entry, and the report does not infer clinical safety or contamination freedom from those procedures.

Sources: [649], [650], [651], [652]

Evidence: Study limitations - MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. Approximately 16.78% of source questions lack the human-response/difficulty annotations needed for one filtering stage; that does not imply they skipped final physician review. Methods name two augmentation models despite introduction wording about three. Final MM choice counts do not mean new distractors were generated for every image-dependent question. Leakage proxies are not training-corpus inspection or clinical validation.

Sources: [649], [650], [651], [652]

Analysis: Study assessment and change conditions - MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. Reviewing original validity before augmentation is a useful control because rewriting can preserve an invalid premise as easily as a correct one. Text/MM, full-suite and ten-percent model samples need separate denominators. Confidence is moderate in the construction lesson and unestablished for clinical safety. A task-specific pilot would retain the invalid-question log, licensed-physician review roles and extraction scripts, then independently assess a stratified sample. Verified corrections and prospective workflow outcomes would change a deployment assessment.

Sources: [649], [650], [651], [652]

Evidence: Reviewed-work atlas - Knowledge and benchmark construction: MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. The reported MMLU-Pro evaluation generally uses five-shot chain-of-thought; the model table specifies zero-shot for Gemini-1.5-Pro and Gemini-1.5-Flash. Answer extraction tries two regular expressions and otherwise selects a random answer option. The v1 paper describes filtering out MMLU questions answered correctly by more than four of eight specified models, then using GPT-4-Turbo to extract answers and generate additional distractors, with manual checking of extracted answers. The v1 paper states that its multiple-choice format may fail to capture comprehension depth and creative responses as effectively as open-ended answers, and that its assessment is language-only rather than multimodal.

Sources: [653]

Research gap: Study result scope - MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. The current MMLU-Pro entry has approved construction, evaluation-setting and limitation observations without a separately admitted model-score table. No score is imported from a later corrected release or reconstructed from the benchmark title.

Sources: [653], [654]

Evidence: Study limitations - MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. Model-assisted distractor expansion and expert checking are separate stages. The v1 filter removes 5,886 MMLU questions at one construction stage; it is not the final MMLU contribution or a reviewer count. Five-shot chain-of-thought has named zero-shot exceptions, and random answer fallback is a scoring choice. Later answer/format corrections cannot retroactively validate the v1 model table.

Sources: [653], [654]

Analysis: Study assessment and change conditions - MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. Difficulty filtering partly defines what the benchmark asks: a harder retained set need not be a more representative professional workload. The random fallback also creates a reproducibility requirement for answer extraction. Confidence is high in these design boundaries and unmeasured for workplace validity. A buyer should freeze question and extraction versions, inspect alternative valid answers, and compare failure types on an independently selected target task sample. Replayed extraction under fixed seeds and independent domain review would change confidence in the score.

Sources: [653], [654]

Evidence: Reviewed-work atlas - Program reconstruction and behavioral testing: ProgramBench: Can Language Models Rebuild Programs From Scratch?. ProgramBench construction uses mini-SWE-agent with Claude Sonnet 4.5 inside Ubuntu 22.04 containers to build reference executables, generate behavioral tests and remove implementation details for inference. The published construction pipeline flags weak assertions for agent revision, then discards tests that do not pass the reference executable deterministically or that pass a dummy binary. The paper evaluates nine models with a single mini-SWE-agent scaffold, specifying 20 CPUs, 60 GB RAM, a 1,000-step limit and six hours per run. In its internet-access ablation, ProgramBench uses nine model judges from three model families to classify source lookup or binary wrapping, flagging a task when at least five judges identify a violation. Appendix A.3.4 reports manual inspection of test suites for five repositories identified as plausible risks for implementation-dependent output, finding no such numerical-output assertions in those inspected suites.

Sources: [655]

Research gap: Study result scope - ProgramBench: Can Language Models Rebuild Programs From Scratch?. The approved ProgramBench observations document construction agents, baseline resource settings, tests and limited manual review, without an admitted numerical model-completion result for this entry. Resource limits and test counts are not converted into a performance claim.

Sources: [655], [656], [657], [658]

Evidence: Study limitations - ProgramBench: Can Language Models Rebuild Programs From Scratch?. The construction agent can inspect source that the reconstruction worker cannot. Tests combine generated and harvested material, reported as 79.5% and 20.5%. Manual review covers five at-risk repositories and no formal human feasibility study. Finite tests omit some behavior and do not measure speed, memory or disk use. The paper and current card have different difficulty populations; model-panel cheating judgments are not independently established human truth.

Sources: [655], [656], [657], [658]

Analysis: Study assessment and change conditions - ProgramBench: Can Language Models Rebuild Programs From Scratch?. The acceptance unit is full tested behavior under isolation, not a fraction of a program inferred from partial tests. The dummy-binary filter is useful falsification but can miss subtle weak assertions. Confidence is high in retaining this distinction and moderate in using the benchmark for reconstruction tasks. A buyer should add metamorphic tests, alternative implementations and shortcut attempts while logging excluded tests. Independent black-box behavior checks, human feasibility studies and adversarial isolation tests would change confidence in generality and completeness.

Sources: [655], [656], [657], [658]

Evidence: Reviewed-work atlas - Scientific coding and expert review: SciCode: A Research Coding Benchmark Curated by Scientists. The v1 paper distinguishes numerical input/output tests from domain-specific tests, for which scientists check whether solutions reproduce published results or match analytical solutions from theoretical models. The v1 paper describes SciCode as scientific coding tasks mainly drawn from scientists’ everyday research scripts, organized into main problems with smaller subproblems and explicit input/output docstrings. The v1 paper reports three revision stages: at least two in-domain scientists cross-check the question, solution and domain tests; an out-of-domain scientist reviews clarity and sufficient information; then GPT-4 solutions undergo scientist error analysis and possible test or workflow revision.

Sources: [637], [638], [659]

Research gap: Study result scope - SciCode: A Research Coding Benchmark Curated by Scientists. The admitted SciCode paper observations establish research-script origins, scientific tests and staged review, without a numerical model-success table for this entry. No independent human reproduction or model score is inferred from the review-stage count.

Sources: [636], [637], [638], [659], [660], [661]

Evidence: Study limitations - SciCode: A Research Coding Benchmark Curated by Scientists. Creator FAQ clarifications limit the three-stage validation claim: human reviewers do not all independently write and execute complete solutions, and three subproblems receive supplied code context. Public underlying scientific results can exist even when the exact task is new. Payment, institutional supply contracts, test completeness and contamination freedom are not established.

Sources: [636], [637], [638], [659], [660], [661]

Analysis: Study assessment and change conditions - SciCode: A Research Coding Benchmark Curated by Scientists. Cross-domain clarity review complements in-domain correctness review, but neither makes a finite test suite exhaustive. The standard main-problem rule carries generated earlier answers forward, so errors can propagate through the workflow. Confidence is moderate in this construction pattern and low in extrapolating its pass rule to arbitrary research code. A pilot should preserve subproblem dependencies, distinguish supplied from generated code and review scientifically plausible false accepts. Independent reruns and adversarial tests grounded in analytical solutions would change trust in the checks.

Sources: [636], [637], [638], [659], [660], [661]

Evidence: Reviewed-work atlas - Chemistry curation and reasoning evaluation: SUPERChem: A Multimodal Reasoning Benchmark in Chemistry. The SUPERChem acknowledgments name Center of Mass Educational Tech. Co. Ltd. among educational platforms that contributed original problems used to build the benchmark. The SUPERChem acknowledgments name Chemy Group among educational platforms that contributed original problems used to build the benchmark. The SUPERChem acknowledgments credit Peking University's High-Performance Computing Platform and Computer Center's LLM Working Group with computational resources including LLM access and server support. The SUPERChem manuscript describes creator, reviewer and approver stages: a question enters the bank only after final approval, with return for revision or rejection available at review and approval. The manuscript reports 77 Peking University students contributing to curation, including 3 IChO medalists and 64 CChO medalists. Table 1 labels its reported frontier-model pass@1 figures as means over eight trials, while other model figures use the first trial; input modalities also differ by model. Appendix E defines Reasoning Path Fidelity as weighted recall of expert-authored reasoning checkpoints, with Gemini 2.5 Pro at temperature 1.0 judging semantic matches in the generated reasoning.

Sources: [662]

Evidence: Study construction and protocol - SUPERChem: A Multimodal Reasoning Benchmark in Chemistry. The SUPERChem acknowledgments name TAL Education Group among educational platforms that contributed original problems used to build the benchmark.

Sources: [662]

Research gap: Study result scope - SUPERChem: A Multimodal Reasoning Benchmark in Chemistry. The admitted SUPERChem paper observations establish curation, contributor roles, trial comparability and the RPF method, without a numerical model-performance table for this entry. The report does not promote the later repository demo to a reconciled manuscript result.

Sources: [662], [663], [664], [665], [666]

Evidence: Study limitations - SUPERChem: A Multimodal Reasoning Benchmark in Chemistry. The 77-person curation cohort differs from the 174-student human examination. Medalist groups are not assumed disjoint. Platform problem contributions and compute/tool assistance have explicit roles, without inferred contracts, ownership or annotation employment. Eight-trial frontier results and first-trial other-model results, modalities and question subsets are not directly interchangeable. Weighted checkpoint recall is not verified access to internal model reasoning.

Sources: [662], [663], [664], [665], [666]

Analysis: Study assessment and change conditions - SUPERChem: A Multimodal Reasoning Benchmark in Chemistry. Creator-reviewer-approver roles provide a concrete accountability design, while model-assisted OCR, translation and preliminary review still need their own error checks. Confidence is moderate in this workflow lesson and low in interpreting checkpoint recall as causal reasoning fidelity. The later DAG protocol should be evaluated as its own scoring version. A bilingual chemistry pilot should test image/text parity, chemically valid alternative paths and judge agreement on unseen solutions. Resolved manuscript/repository scoring differences and independent expert audits would change the assessment.

Sources: [662], [663], [664], [665], [666]

Evidence: Reviewed-work atlas - Repository engineering and repeated reliability: SWE Atlas: Benchmarking Coding Agents Beyond Issue Resolution. SWE Atlas describes professional software engineers creating tasks and reference solutions, review by an independent quality-control team, and independent audits by three trusted experts of instructions, verification and rubrics. The paper compares human and three model judges on 200 rubric gradings of responses from Opus 4.5 and GPT-5, then reports selecting Opus 4.5 as the judge for its runs. Appendix D states that SWE Atlas task authors and reviewers were compensated under contractual terms, with no unpaid volunteer labor. Appendix J reports Harbor-based experiments in Modal sandboxes with 16 CPUs, 16 GB memory and a six-hour limit, and Fireworks AI API inference for open models. SWE Atlas v1 distinguishes Pass@3, meaning any successful trial, Pass@1, meaning average per-trial success, and Pass3, meaning all three trials succeed.

Sources: [670]

Research gap: Study result scope - SWE Atlas: Benchmarking Coding Agents Beyond Issue Resolution. The admitted SWE Atlas paper observations establish expert construction, sampled judge studies, runtime roles and three-trial metric definitions. They do not supply a separately admitted model leaderboard for this entry, and a metric definition is not a measured success result.

Sources: [667], [668], [669], [670], [671], [672]

Evidence: Study limitations - SWE Atlas: Benchmarking Coding Agents Beyond Issue Resolution. The human/model agreement comparison has 200 rubric gradings, not 200 people or every benchmark response; separate studies have other denominators. Author/reviewer compensation and wage compliance are source reports without inspected contracts or payroll. Native and minimal harnesses differ. The later v1.1 judge/network controls cannot be applied retroactively to v1 paper results.

Sources: [667], [668], [669], [670], [671], [672]

Analysis: Study assessment and change conditions - SWE Atlas: Benchmarking Coding Agents Beyond Issue Resolution. SWE Atlas makes the completion contract more explicit than a single patch test: Q&A, test writing and refactoring have different decisive conditions. It also exposes a maintenance problem when rubrics and judging change while task identity stays fixed. Confidence is high in keeping any-of-three, average-trial and all-three success separate, and moderate in task-specific procurement relevance. A buyer should pin the grading release and falsify its rubric polarity and no-op tests. Independent same-task reruns under both release protocols would change confidence in longitudinal comparisons.

Sources: [667], [668], [669], [670], [671], [672]

Evidence: Reviewed-work atlas - Professional tasks, judging and economic interpretation: The AI Productivity Index (APEX). APEX-v1.0 grades each rubric criterion with three judges: o3 with thinking Low, Gemini 2.5 Pro with thinking Off, and Sonnet 4 with thinking Off. It uses a two-of-three or unanimous majority for each binary criterion grade. The APEX-v1.0 paper compares its model scores with publicly reported HLE, MMLU Pro, MMMU Val and GPQA scores and reports a mean cross-benchmark correlation of 0.79. The captured APEX-v1.0 paper parses evidence PDFs and CSVs with Reducto and appends the parsed material to each task prompt, using one parsing strategy across models and avoiding the closed models' native file-handling APIs. The APEX-v1.0 paper explicitly cautions that rubric scores need not track real-world value linearly: a response scoring 60% might be effectively useless when value is stage-gated. The APEX-v1.0 judge panel agrees with expert Pass/Fail labels on 89% of 5,818 criterion ratings collected for responses from one model, Gemini 2.5 Pro with thinking On.

Sources: [605]

Evidence: Study construction and protocol - The AI Productivity Index (APEX). In APEX-v1.0, repeating each judge three times on responses from one arbitrarily selected model, Grok 4, yields reported internal agreement of 99.5% for o3 Low, 99.4% for Gemini 2.5 Pro Off and 99.6% for Sonnet 4 Off. APEX-v1.0 rubrics lack negative criteria that penalize incorrect or irrelevant statements, so passing many positive criteria does not guarantee that a response contains fewer errors. The APEX-v1.0 observational comparison reports mean scores of 55.8% for models offering thinking tokens versus 40.4% for the others, with 16 of 23 models in the thinking group. APEX-v1.0 collects three responses from every model for each prompt and uses the median of their three scored responses for the leaderboard and analysis.

Sources: [605]

Evidence: Study result - The AI Productivity Index (APEX). In the captured APEX-v1.0 experiment, GPT 5 with thinking High has the highest reported mean score, 64.2%, followed by Grok 4 at 61.3% and Gemini 2.5 Flash with thinking On at 60.4%.

Sources: [605]

Evidence: Study limitations - The AI Productivity Index (APEX). The original 200-case August 2025 experiment uses median-of-three responses and a three-judge criterion majority. Its 89% expert agreement covers 5,818 ratings from one response model, with expert labels not themselves reviewer-validated. Near-99.5% judge repeatability covers another selected model and is not correctness. Cross-benchmark correlations have missing model scores. Thinking-group comparisons confound recency and model type, without a causal ablation.

Sources: [605], [606]

Analysis: Study assessment and change conditions - The AI Productivity Index (APEX). Providing the needed evidence in context makes this a professional response test rather than a live information-retrieval workflow. Positive rubric completion can coexist with unsupported extra claims because negative criteria are absent. Confidence is high in those restrictions and unestablished for economic value. A buyer should test decisive negative conditions and compare the full response distribution before choosing a best-of or median protocol. Representative prospective work, independent judge calibration and matched thinking on/off experiments would change the operational interpretation.

Sources: [605], [606]

Evidence: Reviewed-work atlas - Annotation methodology and maintenance: Towards a Science of Corpus Annotation. The authors propose keyboard-bound choices and simple annotation tasks, varying option order to reduce priming and showing all alternatives on one screen to avoid scrolling bias. The paper distinguishes intra-annotator stability over time from inter-annotator reproducibility and discusses simple agreement, Kappa variants and Krippendorff alpha as evaluation measures. The authors recommend developing the codebook with annotators and monitoring agreement so that category definitions remain stable as the scheme evolves. For contrastive annotation, the paper describes a common core scheme with language-specific extensions and illustrates this approach with English and Spanish thematic-field annotation in CONTRANOT. The paper discusses reconciliation and adjudication of annotation disagreements, cautioning that reconciliation can raise agreement scores while risking overtraining or drift and recommending it mainly for initial training. The paper emphasizes ownership and distribution licensing when maintaining annotated corpora and describes distributing annotation layers separately when the underlying corpus cannot be distributed directly.

Sources: [613]

Evidence: Study construction and protocol - Towards a Science of Corpus Annotation. The 2010 paper describes the QDAP annotation center at the University of Pittsburgh as performing annotation for hire for NLP and Political Science clients. The paper proposes an iterative annotation pipeline that selects representative text, instantiates a theory in a codebook, pilots annotation and agreement checks, scales annotation with intermediate checks, and trains and evaluates automation on held-out annotated material. The paper distinguishes corpus representativeness from category balance and describes a tradeoff between preserving natural category frequencies and obtaining enough rare-category examples for learning or theoretical analysis.

Sources: [613]

Research gap: Study result scope - Towards a Science of Corpus Annotation. This entry synthesizes the 2010 paper’s methodological proposals and cautions. Its illustrative agreement gate is not an experimental validation of a universal threshold, and no new controlled performance result is manufactured for this atlas.

Sources: [613]

Evidence: Study limitations - Towards a Science of Corpus Annotation. This 2010 methodological paper proposes procedures and illustrates tradeoffs; it does not supply a current controlled estimate of the best interface or agreement threshold. Its Figure 1 90% gate is illustrative, and later discussion says acceptable agreement depends on use and can be lower for complex tasks. Historical QDAP annotation-for-hire is not evidence of a current qualifying supplier.

Sources: [613]

Analysis: Study assessment and change conditions - Towards a Science of Corpus Annotation. The strongest connection to modern feedback work is that codebooks, interfaces, training and downstream use jointly define label quality. Agreement can rise through undocumented shared instructions or reconciliation without making the labels more reproducible. Confidence is high in using the paper to pose diligence questions and unmeasured for a universal numeric rule. A pilot should archive each guideline change and compare stability on fresh annotators. Controlled interface studies and downstream validation would change which recommendations merit mandatory acceptance criteria.

Sources: [613]

Evidence: Reviewed-work atlas - Aggregation, transcription and competition measurement: VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions. ROVER aligns submitted word sequences and votes at each token, allowing an output different from every worker submission; RASA and HRRASA instead choose one submitted transcription. The shared-task dataset was constructed from normalized English Wikipedia and BookCorpus sentences, synthesized as audio with Yandex SpeechKit and transcribed on Toloka by at least five different workers per recording. Table 1 lists 9,700 training recordings with 67,900 transcriptions and 4,502 test recordings with 31,514 transcriptions; the test set is split into 1,500 public and 3,002 private recordings. The competition scores mean per-recording word accuracy, AWAcc = mean(max(0, 1−WER(r))) × 100%, using JiWER for each recording; live competition scores use the public test subset and final standings use the private subset. The authors report releasing the raw texts, baseline methods and scoring program, and link the Toloka/VLDB2021_Crowd_Science_Challenge repository. The report identifies Toloka as the platform used to collect the shared-task transcriptions and thanks the Toloka team for providing the crowdsourcing budget.

Sources: [619]

Evidence: Study result - VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions. The winning system fine-tuned a BART summarization model on the shared-task data, shuffled input transcriptions for augmentation and normalized British word forms to American forms; it achieved reported private-test AWAcc of 95.73%.

Sources: [619]

Evidence: Study limitations - VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions. Table 1 separates 9,700 training and 4,502 test recordings; section 2 instead describes 99,414 transcriptions for 9,700 recordings. The mismatch is retained, not repaired by inventing a denominator. Audio is synthesized from normalized texts, so this is not natural recorded conversational speech. AWAcc averages clipped per-recording accuracy rather than computing one corpus WER. Public and private standings use different test subsets.

Sources: [619]

Analysis: Study assessment and change conditions - VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions. Choosing an existing transcription and constructing a token-voted transcription are different deliverables. The latter can recover useful fragments while creating an output no worker submitted. Confidence is high in that distinction and moderate in applying the shared-task lesson to similarly synthesized audio. A buyer should compare per-recording errors, rare words and cross-worker dependence, preserving the raw submissions. Natural-speech replications and independently checked reference normalization would change confidence in transfer.

Sources: [619]

Evidence: Reviewed-work atlas - Visual grounding and crowd collection: WSDM Cup 2023 Challenge on Visual Question Answering. The dataset contains 45,199 question-image instances: 38,990 training, 1,705 public-test and 4,504 private-test instances. The evaluation averages intersection over union between each predicted answer box and its ground-truth box; each item IoU is the boxes intersection area divided by their union area. Toloka workers selected images containing objects they found interesting, composed questions about those objects and then marked the answering objects with bounding boxes to exclude unanswerable questions. The CodaLab competition froze its dataset before the challenge, used public-test prediction submissions during evaluation and then required container-image solutions for organizer execution on the private test set to determine winners. The authors report releasing the complete dataset and ground truth after the competition under the same CC BY license as the selected COCO image subset, with Zenodo, Hugging Face, Kaggle and GitHub release links. The Toloka Visual Question Answering Challenge takes an image and an English question and asks for a bounding box around the object answering that question.

Sources: [620]

Evidence: Study result - WSDM Cup 2023 Challenge on Visual Question Answering. A separate Toloka annotation round measured non-expert human performance without providing ground truth; the report gives mean IoU of 0.87 ± 0.01 across the three dataset subsets. The winning system used a Uni-Perceiver backbone, ViT-Adapter localization and DINO prediction head with an auxiliary loss and test-time augmentation; Table 3 reports its private-test reproduction score as 76.347 on the displayed IoU ×100 scale. The starter baseline uses YOLOR to propose candidate rectangles, applies CLIP to compare the question with each bounded image region and selects the highest-similarity candidate.

Sources: [620]

Evidence: Study limitations - WSDM Cup 2023 Challenge on Visual Question Answering. Question/image instances are not unique images or workers. About 20% of boxes overlap COCO rectangles and all such instances enter training; pretrained-COCO access was allowed. The non-expert baseline is a separate annotation round, and alpha uses 1−IoU as response distance. The dataset DOI is distinct from this CEUR paper identity. Release licensing is an author statement, without an independent per-image rights audit.

Sources: [524], [525], [620]

Analysis: Study assessment and change conditions - WSDM Cup 2023 Challenge on Visual Question Answering. The operational lesson is to specify spatial ground truth and its disagreement geometry. A box-overlap metric measures localization, not whether a free-form answer is useful, and the separate crowd round cannot be treated as task-author self-agreement. Confidence is high in these distinctions and moderate in the collection protocol as a design reference. A buyer should inspect ambiguous object boundaries, unanswerable questions and image reuse across splits. New-image heldouts with independent annotators and container replay would change external-validity confidence.

Sources: [524], [525], [620]

Analysis: Cross-study synthesis - output completion, statistical recovery and rubric quality are different targets. GeneBench-Pro requires every prespecified output field and does not grade its reasoning text; HealthBench aggregates signed criterion weights; APEX-v1 reports median criterion completion; ProgramBench requires all behavioral tests and no cheating flag. A shared percentage sign is not a shared unit. Confidence is high in these distinctions and unmeasured for a universal best metric. A buyer should choose the actual accepted deliverable, identify decisive failures and retain the diagnostic vector behind the aggregate. Matched evaluations with independently accepted outputs would change the relative usefulness of each score.

Sources: [605], [646], [647], [655], [656], [657], [658], [673]

Evidence: Inspected-code and documentation supplement - MMLU and SWE-bench. At MMLU creator commit 4450500f923c49f1fb1dd3d99108a0bd9717b660, evaluate.py requests five demonstrations by default, reduces their count under context cropping, selects the highest A-D token log probability and reports sample-weighted aggregate accuracy. This code inspection did not execute the benchmark or establish byte-level identity with the separately pinned HF release. SWE-bench collection documentation at commit 02e7a74ffd0b707aab73d203fe87bdc7c76afc8e describes PR/issue scraping, gold patch/test requirements, installation setup and execution-based validation; it does not independently prove that every original instance followed every step. Original proposal-stage notes remain historical; the current approved records and exact receipts control admission.

Sources: [566], [567], [569]

Analysis: Reasoning-data provenance has at least three separate accounting units: the original question, the generated solution trace and the verification decision. AM distinguishes reused material from its own DeepSeek-R1 responses; s1K-1.1 retains the same questions while changing traces; Light-R1 describes curriculum selection followed by generated correct/incorrect pairs. The implication is to preserve prompt-source, response-source and verification lineage separately. A new teacher response does not establish a new independent question, a human preference or a proved-correct solution. Confidence is high in these documented distinctions and unmeasured in their comparative training benefit. Item-level lineage and matched independent verification would change the uncertainty about actual coverage.

Sources: [441], [443], [444]

Analysis: A benchmark judge’s information boundary is part of the measured system. Chartography’s final-answer judge does not receive the chart; GDP.pdf’s criterion judge receives neither PDF nor reference answer; Harvey LAB scopes its calls to declared output files. Those observations support an audit question before interpreting scores: can the judge actually observe each fact needed for the criterion, and where is the required evidence encoded? They do not demonstrate that these judges are inaccurate. Confidence is high in treating context visibility as a protocol condition; judge error remains unmeasured here. A blinded expert study with missing-evidence, contradictory-evidence and valid alternative-output cases could validate or revise the proposed acceptance design.

Sources: [2], [3], [7], [8], [772]

Analysis: Strict task completion, partial-credit quality and repeated-attempt reliability answer different questions even within one release. ComplexConstraints separates all-criteria success from mean criterion satisfaction; Harvey LAB's two-judge average permits a half score while its strict all_pass requires both judges; APEX-SWE separates pass@1 and pass@3 and keeps supplementary rubric scores outside its leaderboard. This supports reporting a small metric vector rather than converting every percentage to a common success rate. Confidence is high in the mathematical distinction and unknown in any cross-benchmark performance ordering. A common task population with a frozen rubric, judge context, repeat policy and failure denominator would be necessary for a narrower comparison.

Sources: [4], [559], [763], [772]

Analysis: Environment packaging and completed validation are separate observations. Multi-SWE-bench documents test-transition filtering and a manual review stage; the initial Multi-SWE-RL construction omits that manual stage. Terminal-Bench documents oracle/dummy controls and exploit review while retaining residual flaws; PostTrain Arena explicitly distinguishes static source acceptance from planned dynamic checks that runs do not yet await. The analytical implication is to request evidence for each gate rather than inherit confidence from a Dockerfile or a shared construction pipeline. Confidence is high in the documented stage boundaries and unmeasured for actual task failure rates. Independent replay of the oracle, no-op controls, alternate valid solutions and exploit cases would change the buyer's operational confidence.

Sources: [13], [377], [629]

Analysis: Difficulty selection changes what an absolute score represents. BrowseComp screens for difficult retrieval; AdvancedIF retains final-turn instruction failures against a supplied model; GDP.pdf requires meaningful failures by frontier models; μ-MATH intentionally selects text-only problems for judging difficulty rather than representing all U-MATH tasks. These observations support treating each as a stress distribution with an explicit selection rule. They do not yield the frequency of failure in an ordinary buyer workload or the share of a profession that can be automated. Confidence is high in that sampling limit. A prospectively sampled buyer workload, including routine and difficult cases, could establish an appropriately scoped external rate.

Sources: [8], [338], [552], [557]

5. Provider dossiers: offerings, delivery, differentiation and limits

Method: Dossier format: offering and scope; disclosed delivery and operating model; buyer assessment and limits. Every qualifying supplier in the approved input is covered. First-party marketing is evidence of what is offered or reported, not an independently observed delivery. A named case is preserved at its actual attribution, task and date. A dossier with no named customer artifact established here is not evidence that the supplier has never delivered one. Unrecorded language, modality, security or commercial facets are unknown, not negative capabilities. Differentiation means a documented design or responsibility distinction, not a score for provider quality.

Evidence: A.Team - offering and scope. Its RL engineering offer includes graders, rubrics, harnesses and rollout infrastructure. A.Team lists golden trajectories, expert data and RLHF/RFT support within post-training pipeline work. A.Team offers senior engineers to build and validate training environments, tasks and evaluations.

Sources: [285]

Evidence: A.Team - disclosed delivery and operating model. A.Team offers hourly, contract-to-hire, ongoing and advisory builder engagements. Its RL page says builders work within customer access controls, reviews and governance.

Sources: [285]

Evidence: A.Team - source-specific boundaries. These are engagement models, not client prices, worker compensation amounts or evidence of completed named-artifact delivery. Provider disclosure; no security audit report, scope or project execution was independently inspected.

Sources: [285]

Analysis: A.Team - buyer assessment and unknowns. The useful comparison is engineering accountability: request a concrete environment, grader or harness milestone and clarify whether the buyer is purchasing individual time or a warranted deliverable. Customer access controls should appear in the project specification. Builder engagement models alone do not establish delivered data quality, a fixed completion price or current specialist availability. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with independently accepted engineering milestones and access-control records.

Sources: [285]

Evidence: Abundant - offering and scope. Abundant lists coding, sciences, engineering, RSI, alignment and robotics as focus areas. Foundry offers to turn Fireworks customers’ own product data and workflows into RL environments. Abundant advertises creation of environments and datasets for reinforcement learning.

Sources: [205], [206]

Evidence: Abundant - disclosed delivery and operating model. Abundant says every Foundry task version receives human sign-off after automated and adversarial QA. Abundant says Foundry starts from customer product data including user sessions, tools and golden test cases. Abundant describes Oddish as tooling for trials, trace capture, scheduling and environment revision before human review.

Sources: [206], [692]

Evidence: Abundant - source-specific boundaries. Listed focus areas do not establish named customer work, linguistic coverage or worker headcount. The article separately describes baseline, adversarial and multi-agent stress tests; the execution of those checks was not independently audited.

Sources: [205], [206], [692]

Analysis: Abundant - buyer assessment and unknowns. The differentiating disclosed workflow starts from customer product traces and golden cases and closes with task-version sign-off. A pilot should inspect trace permissions, reset behavior and the recorded transition from automated stress tests to human acceptance. A list of scientific and coding focus areas does not establish task coverage or operational throughput in each domain. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with trace-to-task lineage, reset tests and independent task-version sign-off.

Sources: [205], [206], [692]

Evidence: AfterQuery - offering and scope. AfterQuery offers human demonstrations across browser and desktop environments as computer-use trajectories. AfterQuery offers SFT prompt-response pairs and reasoning traces for complex tasks. AfterQuery offers expert-designed prompts and grading frameworks for reasoning and code generation.

Sources: [207]

Evidence: AfterQuery - disclosed delivery and operating model. AfterQuery attributes office-task input for NVIDIA Nemotron 3 Ultra to its own delivery account. The linked NVIDIA report was not independently retrieved; this does not establish GDPval authorship or a share of all training data. AfterQuery’s website privacy policy excludes personal-information processing performed as a processor or service provider for customers. AfterQuery’s terms place service fees in separate agreements or order forms.

Sources: [208], [209], [719]

Evidence: AfterQuery - source-specific boundaries. Evidence is AfterQuery’s first-party delivery account, including its quoted NVIDIA-report passage; the linked NVIDIA PDF was not retrieved or independently read. It attributes office-task input, not GDPval authorship, all training data or model performance. The policy’s general security statement does not establish customer-processing contractual safeguards, certifications or specific worker protections.

Sources: [207], [208], [209], [719]

Analysis: AfterQuery - buyer assessment and unknowns. The buyer can distinguish demonstrations, prompt-response data and grading frameworks as separate outputs. The named office-task account is useful for requesting an artifact-specific sample, but its first-party attribution cannot be extended to all NVIDIA data or GDPval authorship. Obtain the customer-processing agreement and project order because the general website policy and terms leave those details elsewhere. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with customer-confirmed office-task provenance and independently scored trajectories.

Sources: [207], [208], [209], [719]

Evidence: Alegion / SanctifAI data services - offering and scope. Alegion offers data collection, annotation and quality control through managed workforce experts.

Sources: [119]

Evidence: Alegion / SanctifAI data services - disclosed delivery and operating model. Alegion’s official page describes Alegion as a division of SanctifAI. Alegion describes training and qualifying project annotators using ground-truth annotations before enrollment. Alegion describes a workforce-partner network supported by its project managers, data engineers and QA teams. Alegion describes QA review of submitted annotations with client acceptance or feedback. Alegion reports that Sportlogiq used its end-to-end managed service to scale annotation for a player-tracking system.

Sources: [119], [120]

Evidence: Alegion / SanctifAI data services - source-specific boundaries. Provider business-structure disclosure; legal ownership and corporate registry have not been independently checked. The same page also uses rebrand wording, so no separate SanctifAI supplier is counted. Access-to-workforce language is not a current paid-active count or assignment total; the source’s large pool figure is not adopted.

Sources: [119], [120]

Analysis: Alegion / SanctifAI data services - buyer assessment and unknowns. The disclosed model joins workforce partners with project management and QA. For a comparable bid, freeze the project ground truth, qualification gate, escalation path and client acceptance rule. The Sportlogiq account supports a scoped annotation case, not an independently verified accuracy gain or a current contract. Treat SanctifAI identity wording as one supplier boundary rather than an additional market participant. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with blind annotation samples, qualification records and the applicable contracting entity.

Sources: [119], [120]

Evidence: Alinia AI - offering and scope. Alinia advertises adversarial evaluation against a customer’s policy red lines. Alinia offers policy judges trained on internal and external policies to check AI outputs.

Sources: [288]

Evidence: Alinia AI - disclosed delivery and operating model. Its engagement process compares judge decisions with customer experts on real data. Alinia describes each engagement as a forward deployed legal and engineering project inside the customer environment. The website privacy policy names Alinia AI, Inc. as data controller.

Sources: [288], [289]

Evidence: Alinia AI - source-specific boundaries. This describes a proposed validation stage; it is not independent evidence of accuracy, legal compliance or expert sign-off for a particular deployment. Website-controller identity does not prove equity ownership or recertify every historical alias. The policy’s website scope does not replace a customer processing agreement.

Sources: [288], [289]

Analysis: Alinia AI - buyer assessment and unknowns. This is a policy-judge and integrated validation proposition. Require a customer-specific policy-to-test mapping and a blinded expert comparison before using judge outputs for operational decisions. The website controller identity is not a customer-data protection agreement, and an engagement description involving legal expertise is not a finding that a deployment complies with law. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with held-out expert policy judgments and complete false-accept/false-reject counts.

Sources: [288], [289]

Evidence: Appen - offering and scope. Appen offers repository tasks with verified patches, debugging traces, code preference pairs, and test-generation data. Appen offers expert alignment data, agentic workflows, multimodal data, physical-AI data, model evaluation, and research services.

Sources: [735]

Evidence: Appen - disclosed delivery and operating model. Appen describes project-level agreement metrics, continuous gold questions, and adjudication of disagreement. Appen's FY2025 report describes a global crowd of more than one million experts and identifies CrowdGen as a platform. The 2025 annual report identifies Appen Limited as the parent entity. Appen says its model-evaluation work is run by teams separate from the training-data producers. Appen's privacy statement permits service-provider processing and describes access controls and confidentiality duties for contractors. Appen markets consent records, licensing terms, and data lineage as delivery documentation. Appen reports FY2025 consolidated revenue from customer contracts of USD 230.774 million. Appen's security page describes SOC 2 Type II attestation, a HIPAA-specific solution, and named secure-facility options. Appen reports FY2025 revenue of USD 127.874 million for Appen Global and USD 102.900 million for Appen China.

Sources: [710], [735], [736], [737]

Evidence: Appen - source-specific boundaries. Crowd/network scope; no activity window, paid-contributor denominator, or paid assignment count is established. Report footnotes on geography/language need retention if those counts are used. Published privacy terms; the actual project subcontractor list, consent evidence, and contract-specific IP ownership were not inspected.

Sources: [710], [735], [736], [737]

Analysis: Appen - buyer assessment and unknowns. The breadth of disclosed modalities makes scope discipline central. A software-patch pilot and a physical-AI collection pilot should have separate acceptance units. Ask for evidence that the promised training/evaluation team separation and adjudication process operate on the proposed project. Group and geographic-segment revenue cannot identify the size or performance of expert generative-AI delivery. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with project-specific staffing, adjudication and matched accepted-unit results.

Sources: [710], [735], [736], [737]

Evidence: APTO - offering and scope. APTO offers commissioned solutions from data collection through model and system development. APTO offers its harBest platform with crowd-worker and in-house collection and annotation. Its homepage lists image, video, LiDAR, natural-language and audio data formats. APTO sells ready-made datasets in image/video, voice and natural-language categories.

Sources: [290]

Evidence: APTO - disclosed delivery and operating model. APTO's official company-information page names APTO, Inc. and a January 20, 2020 founding date.

Sources: [677], [678]

Evidence: APTO - source-specific boundaries. The inspected corporate-information page establishes the page’s APTO, Inc. rendering and founding-date statement. It does not establish a current privacy controller, processor terms, a legal name change or ownership. Listed formats and applications do not establish delivered-language coverage or independently tested capability.

Sources: [290], [677], [678]

Analysis: APTO - buyer assessment and unknowns. The documented comparison has two purchase paths: ready-made data and commissioned collection-to-system work. Require separate version, rights and acceptance schedules for each. A broad modality menu supports asking for representative image, LiDAR or audio samples; it does not establish domain-specific expert qualifications or a named completed customer delivery in the inspected dossier. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with exact licensed payloads, project acceptance results and verified corporate contracting records.

Sources: [290], [677], [678]

Evidence: Aya Data - offering and scope. Aya Data offers expert-annotator feedback for conversational-model evaluation and domain-specific AI annotation. Aya Data describes using human preference rankings to build reward models.

Sources: [123]

Evidence: Aya Data - disclosed delivery and operating model. Aya Data’s homepage describes measuring inter-annotator agreement on every delivery. Aya Data reports labeling aorta images for Cydar Medical and expanding the engagement to thrombus and stent labeling. Aya Data’s RLHF page describes stable contracts and protection from harmful content as worker-wellbeing practices.

Sources: [122], [123], [716]

Evidence: Aya Data - source-specific boundaries. Process assertion only; no measured agreement value, universal achieved quality score or independent audit is established. Provider-authored named case with no engagement calendar window or public dataset release identity. Clinical-outcome and diagnostic-accuracy claims are omitted.

Sources: [122], [123], [716]

Analysis: Aya Data - buyer assessment and unknowns. The Cydar account offers a concrete medical-image annotation case, while the conversational-feedback offering is a different task type. Do not transfer medical-image experience into a claim of superior preference data. A pilot should inspect how agreement is calculated, how difficult cases are retained and how harmful-content protection is implemented for the actual conversational task. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with independently checked task-specific agreement and correction records.

Sources: [122], [123], [716]

Evidence: BasicAI - offering and scope. BasicAI offers human-powered training data to fine-tune foundational language models. BasicAI’s RLHF service offers human-preference datasets formed by ranking multiple SFT-model responses.

Sources: [125]

Evidence: BasicAI - disclosed delivery and operating model. BasicAI’s security page requires every employee to sign a confidentiality agreement that survives employment. BasicAI states each labeling project undergoes multi-level inspections and validations. BasicAI states its labeling service maintains specialist teams with industry-specific knowledge rather than using crowdsourcing.

Sources: [124], [126]

Evidence: BasicAI - source-specific boundaries. Employee policy disclosure, not proof of a particular contractor’s terms or independent certification. The adjacent 99%+ headline is not adopted as a measured quality rating or universal guarantee.

Sources: [124], [125], [126]

Analysis: BasicAI - buyer assessment and unknowns. The stated specialist-team model differs from open participant access, but the buyer still needs project-specific staffing and review evidence. Request independent raw judgments before ranking aggregation, the definition of each inspection level and an escalation sample. A confidentiality agreement is one control; it does not by itself establish access isolation, secure deletion or regulatory scope. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with project staffing records, independent raw labels and escalation outcomes.

Sources: [124], [125], [126]

Evidence: Bespoke Labs - offering and scope. Bespoke advertises automated prompt and policy optimization using GEPA. Bespoke Labs offers RL environments and infrastructure for long-horizon agent training. Its environment offering includes company-scale simulations with real codebases and microservices.

Sources: [291]

Evidence: Bespoke Labs - disclosed delivery and operating model. Bespoke invites domain experts to build RL environments and data pipelines. Its November 5, 2026 website terms reserve product, customer-service and data-acquisition engagements to separate agreements.

Sources: [291], [292]

Evidence: Bespoke Labs - source-specific boundaries. A product capability description; no comparative accuracy or speed claim is carried forward. An expert invitation does not establish paid project availability, participant counts, rates or contributor agreements.

Sources: [291], [292]

Analysis: Bespoke Labs - buyer assessment and unknowns. The comparison centers on long-horizon environment engineering and optimization infrastructure. Test a resettable multi-service workflow with known failure states and independently reviewed rewards. Expert intake is a recruitment route, not a count of paid contributors. The separately contracted customer and data-acquisition services need explicit rights, maintenance and delivery obligations. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with reproducible environment resets, reward falsification and accepted engineering outputs.

Sources: [291], [292]

Evidence: Centific - offering and scope. Centific’s OneForma page describes connecting model labs and enterprises with vetted contributors and domain experts for data, human feedback and model evaluation. Centific offers human preference signals suitable for training and refining reward models.

Sources: [128], [129]

Evidence: Centific - disclosed delivery and operating model. OneForma describes conflict-resolution and consensus review workflows supported by automated labeling-error checks. Centific’s OneForma description states contributors are screened, identity-verified and continuously evaluated through quality checks. Centific describes trained evaluators ranking model outputs against nuanced criteria in its preference-optimization service. OneForma describes access-controlled environments and encrypted workflows for sensitive data.

Sources: [128], [129]

Evidence: Centific - source-specific boundaries. Provider process claim; no paid-active population or independent screening audit is inferred. Self-reported controls, not an independently inspected certification report.

Sources: [128], [129]

Analysis: Centific - buyer assessment and unknowns. OneForma’s screening, consensus and automated checks should be evaluated as a sequence rather than one quality score. Request the domain-language qualification test, conflict-resolution rubric and accepted-unit audit trail. Access-controlled facilities are a disclosed option, not proof that the quoted delivery will use them. No public pool description establishes paid availability for a specific project. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with locale-domain qualification outcomes and auditable conflict-resolution records.

Sources: [128], [129]

Evidence: CloudFactory - offering and scope. CloudFactory’s Accelerated Annotation combines a skilled workforce and ML automation for computer-vision data labeling.

Sources: [49]

Evidence: CloudFactory - disclosed delivery and operating model. CloudFactory reports delivering a first training dataset for an unnamed global AgTech company’s computer-vision model. CloudFactory’s AgTech case reports quality checks on all labeled data in that engagement. CloudFactory states its optional Endpoint offering uses centrally managed company-provisioned workstations with disk encryption. CloudFactory describes Network Security as included in each client engagement, with Endpoint Security offered as an additional tier.

Sources: [49], [50]

Evidence: CloudFactory - source-specific boundaries. The client remains unnamed. Timing is not dated to a calendar window, and comparative cost/speed claims are omitted. Project-specific provider disclosure; it does not establish an accuracy rate or a universal all-project guarantee.

Sources: [49], [50]

Analysis: CloudFactory - buyer assessment and unknowns. The reviewed case concerns computer-vision annotation, so it is most useful as a diligence reference for a similar labeling workflow. Confirm which controls are included in the quoted tier: the network baseline and optional managed endpoint are different promises. The unnamed AgTech case cannot establish current capacity, a generic model-performance gain or capability on unrelated expert preference tasks. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with tier-specific control evidence and independently scored vision deliveries.

Sources: [49], [50]

Evidence: Cogito Tech - offering and scope. Cogito describes expert-verified prompt-response pairs and SME ratings of model responses for RLHF. Cogito Tech offers managed annotation across text, audio, image and video.

Sources: [339]

Evidence: Cogito Tech - disclosed delivery and operating model. Cogito offers customized annotation workflows with project-specific quality checks, labeling schemes and reviewer trees. Cogito’s data-labeling page describes a human workforce of subject-matter experts, quality controllers and annotators. Cogito’s official registration page identifies Cogito Tech LLC.

Sources: [339], [340]

Evidence: Cogito Tech - source-specific boundaries. Evidence is the exact saved official web.run extracted-page snapshot; origin HTML and live HTTP status were not exposed. No workforce size, pay, audit or project-level performance is inferred.

Sources: [339], [340]

Analysis: Cogito Tech - buyer assessment and unknowns. The disclosed distinction is a managed multimodal workforce with project-specific reviewer trees and expert response ratings. A pilot should inspect who authors guidelines, who reviews corrections and how a response becomes accepted. The pinned evidence is an official extracted-page snapshot, not an origin-HTML capture or live-status check. Website registration identifies an operator without resolving current ownership or project contracts. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with current contracting records and independently inspected reviewer-tree outputs.

Sources: [339], [340]

Evidence: Datacurve - offering and scope. Datacurve offers full traces of expert execution including tool calls, checks, pivots and recoveries. Datacurve offers prebuilt datasets described as curated, quality-reviewed and structured for customers’ training stacks. Datacurve offers durable RL environments with long horizons, realistic tools, domain-specific judgment and partial progress.

Sources: [217]

Evidence: Datacurve - disclosed delivery and operating model. Datacurve presents DeepSWE as its published software-engineering benchmark; this is an artifact-specific first-party attribution. DeepSWE states that its tasks are written from scratch rather than adapted from existing commits or pull requests. The DeepSWE site describes hand-written verifiers that test software behavior rather than implementation details.

Sources: [215], [216]

Evidence: Datacurve - source-specific boundaries. An owned evaluation artifact; no supplier-to-third-party-training edge is inferred. Task-creation method is self-reported. This does not prove model-training contamination is absent.

Sources: [215], [216], [217]

Analysis: Datacurve - buyer assessment and unknowns. Expert traces and durable environments should be compared using recovery behavior, verifier coverage and reproducibility. DeepSWE's from-scratch task origin is a documented construction choice, not proof of no training overlap. Ask for a task sample that exercises realistic tools and partial progress, plus a negative case showing how the verifier rejects a superficially plausible but incorrect completion. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with private task provenance, negative verifier tests and reproducible expert traces.

Sources: [215], [216], [217]

Evidence: DataForce by TransPerfect - offering and scope. DataForce offers RLHF services incorporating human evaluation feedback into LLM fine-tuning.

Sources: [53]

Evidence: DataForce by TransPerfect - disclosed delivery and operating model. DataForce reports generating scenarios and annotating safe/unsafe prompts for IBM's Granite Guardian 3.0 dataset work. In the IBM case, DataForce describes jointly testing dataset groups and refining annotation protocols through feedback as model development progressed. DataForce identifies itself as part of TransPerfect and its platform as a proprietary TransPerfect solution. DataForce describes matching projects to domain experts/native speakers and using automated checks, multi-annotator consensus and expert adjudication. DataForce's vendor privacy notice states vendors must sign the applicable data-processing agreement before handling EU personal data and process it only under TransPerfect instructions.

Sources: [52], [54], [55]

Evidence: DataForce by TransPerfect - source-specific boundaries. A named delivery contribution is explicitly claimed by the provider. No public dataset release identity, complete dataset size, or independent model-performance result is inferred. A provider process disclosure, not an audited quality rating or evidence of a specific active contributor count.

Sources: [52], [53], [54], [55]

Analysis: DataForce by TransPerfect - buyer assessment and unknowns. The IBM case supplies an artifact-specific safety-data reference and an iterative protocol-design account. A pilot should separate scenario creation, safe/unsafe labels and expert adjudication, preserving IBM-specific claims within that case. The vendor privacy notice addresses EU-personal-data processing obligations; it does not establish every downstream dataset's consent chain or a current operational audit. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with artifact-specific customer confirmation and blind safety-label adjudication.

Sources: [52], [53], [54], [55]

Evidence: Dataloop - offering and scope. Dataloop provides data curation, cleaning, versioning and management tools for unstructured data selection. Dataloop offers software that integrates human feedback into AI data pipelines. Its human-feedback page offers prebuilt reinforcement-learning workflows containing human and AI feedback nodes for RLHF/RLAIF.

Sources: [341]

Evidence: Dataloop - disclosed delivery and operating model. Dataloop lists AI Annotation Assistant, Consensus, Honeypot and JavaScript validations as annotation automation features. The July 2025 privacy notice names Dataloop Ltd. as controller for the processing activities described there.

Sources: [341], [342]

Evidence: Dataloop - source-specific boundaries. Available controls do not establish project-level annotation accuracy. This establishes platform functionality, not a managed human workforce.

Sources: [341], [342]

Analysis: Dataloop - buyer assessment and unknowns. Treat the offering as data-pipeline and human-feedback software rather than infer a managed annotator workforce. A buyer should test routing between human and AI feedback nodes, versioning and the relationship between consensus, honeypots and corrections. Software feature availability does not establish project-level accuracy. Clarify which party recruits experts and warrants the final accepted data. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with workspace routing/versioning tests and a clear final-delivery responsibility agreement.

Sources: [341], [342]

Evidence: Dataocean AI (Speechocean domain association) - offering and scope. The approved service observation documents speech-data collection and transcription using linguists and native speakers. The broader general-offer assertion remains withheld and is not used to establish an off-the-shelf catalogue, licensed release or legal alias.

Sources: [370]

Evidence: Dataocean AI (Speechocean domain association) - disclosed delivery and operating model. The DOTS platform description offers automated pre-labeling, interactive assisted annotation, automated quality control and assisted quality inspection. The current people.speechocean.com login page is branded DATAOCEAN AI. Dataocean AI’s March 31, 2024 website privacy policy reserves project and product personal-data rules to the corresponding services or products.

Sources: [370], [371], [372]

Evidence: Dataocean AI (Speechocean domain association) - source-specific boundaries. DOTS describes available annotation and checking controls, without establishing delivered accuracy or certification. The branded people.speechocean.com login page establishes a brand/domain association only; it does not prove a registered company name, ownership relationship or dated rename. The March 31, 2024 website policy leaves product and project rules to their corresponding services. No paid-active worker denominator, complete dataset catalogue or dataset license is established.

Sources: [370], [371], [372]

Analysis: Dataocean AI (Speechocean domain association) - buyer assessment and unknowns. The graph retains one conservative entity for coverage accounting; that editorial treatment is not a positive legal-alias finding. The practical diligence path is a speech-collection proposal with the contracting entity, native-language qualification, recording permissions, service-specific processing terms and DOTS correction logs identified. Require release-specific rights evidence before treating any offered material as licensed. Confidence is moderate for the documented service scope and low for unresolved corporate identity; verified corporate records and the applicable signed agreement would change that assessment.

Sources: [370], [371], [372]

Evidence: Datasaur - offering and scope. Datasaur’s current Data Studio offers automated NLP data annotation with configurable project interfaces and quality-control workflows.

Sources: [344]

Evidence: Datasaur - disclosed delivery and operating model. Datasaur’s terms limit customer-data processing to providing the service, with an additional deidentified or aggregate improvement-use permission for free-access subscriptions. Datasaur describes its NLP labeling interface as intended for subject-matter experts and data scientists. Data Studio describes automated pre-labeling combined with human-in-the-loop verification. Datasaur’s terms identify Datasaur, Inc. as the contracting service provider.

Sources: [343], [344]

Evidence: Datasaur - source-specific boundaries. This records the published terms and free-access exception, not a claim about every customer’s negotiated agreement. The main company homepage also advertises private AI deployments; Data Studio is retained as a product, not a separate supplier.

Sources: [343], [344]

Analysis: Datasaur - buyer assessment and unknowns. Data Studio's expert-facing interface and human verification are product functions, not evidence of an internally employed expert pool. Test automated pre-labels and the correction audit trail on the buyer's NLP task. Review the free-access exception for deidentified or aggregate improvement use separately from paid or negotiated terms. Private AI deployment marketing is a separate scope from Data Studio annotation. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with task-specific prelabel/correction audits and applicable negotiated data-use terms.

Sources: [343], [344]

Evidence: Deepen AI - offering and scope. Deepen offers managed Annotate as a Service separately from customer-operated Annotate software. Deepen's managed annotation scope includes 2D image/video, 3D LiDAR/radar, fused-sensor labels and scenario or behavior tags.

Sources: [345]

Evidence: Deepen AI - disclosed delivery and operating model. Deepen describes batch sampling with inter-annotator agreement, in-tool correction routing and geometry/taxonomy/consistency validation. The August 2025 data-protection addendum names Deepen AI Inc. and forms part of its customer service agreement. Deepen describes its annotation service as staffed by specialists hired, trained, managed and quality-checked directly rather than crowdsourced.

Sources: [345], [346]

Evidence: Deepen AI - source-specific boundaries. No accuracy percentage, certification or completed project outcome is adopted. Provider process disclosure; no active paid headcount, labor-term compliance or staffing audit is established.

Sources: [345], [346]

Analysis: Deepen AI - buyer assessment and unknowns. The documented service/software separation is useful for procurement: managed specialists and a customer-operated tool allocate work differently. A multisensor pilot should examine geometry, taxonomy and temporal consistency, then inspect the batch sample and correction routing. Directly managed staffing is a disclosed process, not an independently audited headcount or labor-compliance result. The DPA identifies a customer-service boundary to verify. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with multisensor geometry/temporal acceptance results and assigned-team records.

Sources: [345], [346]

Evidence: Defined.ai - offering and scope. Defined.ai offers custom AI-training data collection across audio, image, video, text and multimodal inputs.

Sources: [56]

Evidence: Defined.ai - disclosed delivery and operating model. The robotics case describes standardized calibration/collection documentation and weekly delivery reports as quality controls. Defined.ai's Ethical AI page commits to explicit contributor consent and provenance clearance for AI-training data. Defined.ai says its crowd members must accept a community agreement, terms and privacy policy before work, and that jobs have their own statements of work and applicable transparency sheets. Defined.ai's supplier terms distinguish marketplace or third-party proprietary datasets from new collection, crowd services, evaluation and RLHF services. Defined.ai reports delivering 225 hours of synchronized multimodal robotics sensor recordings for an unnamed AI research and deployment company.

Sources: [57], [58], [59]

Evidence: Defined.ai - source-specific boundaries. This is a provider policy commitment, not proof that every offered dataset has legally sufficient consent or a certification of individual projects. Linked agreements were not separately captured; exact pay, rejection, dispute and IP conditions remain unknown.

Sources: [56], [57], [58], [59]

Analysis: Defined.ai - buyer assessment and unknowns. Marketplace licensing and custom collection require different diligence. For a robotics pilot, request synchronization tolerances, calibration records, participant permissions and an acceptance sample rather than infer quality from delivered recording hours. The 225-hour case is an unnamed-customer historical delivery account, not a current throughput guarantee. Ethical commitments need record-level implementation evidence. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with recording-level permissions, calibration logs and independently sampled usable units.

Sources: [56], [57], [58], [59]

Evidence: Deploy / Hive - offering and scope. Deploy's current schedule separately offers data annotation and sourcing using Hive Workers.

Sources: [352]

Evidence: Deploy / Hive - disclosed delivery and operating model. The October 26, 2026 Deploy terms identify Castle Global, Inc. as the company providing the covered AI-related services. The official Hive landing page says Hive is now Deploy and existing customer logins are unchanged. Deploy's AI-service terms provide for manual Hive Worker review of customer content to confirm model predictions unless the parties agree otherwise in writing. The Labeling Service terms describe Hive Workers as independent globally distributed workers and disclaim a guarantee that all task performance is accurate.

Sources: [352], [353]

Evidence: Deploy / Hive - source-specific boundaries. Hive and Deploy are retained as one supplier identity, not two companies. The labeling service is distinct from model API access and the customer-operated Arbiter moderation platform.

Sources: [352], [353]

Analysis: Deploy / Hive - buyer assessment and unknowns. Deploy/Hive is one supplier cluster with Castle Global named in the inspected terms. Separate labeling and sourcing through Hive Workers from model API access and customer-operated moderation software. The default manual-review clause is consequential for data access: confirm any written exception and downstream worker exposure. Contractual worker characterization and an accuracy disclaimer do not establish actual employment status or realized task quality. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with the operative review-access agreement and independently checked worker-exposure controls.

Sources: [352], [353]

Evidence: Digital Divide Data - offering and scope. Digital Divide Data offers ML data operations and training-data solutions for language and computer-vision systems. DDD’s service offering includes human preference optimization and large-language-model fine-tuning. DDD offers human-in-the-loop sensor-data annotation for robots, autonomous vehicles, drones and smart machines.

Sources: [347], [349]

Evidence: Digital Divide Data - disclosed delivery and operating model. The published DDD Terms of Use name DDD Ventures LLC and include investment and real-estate language. DDD says it can integrate with the customer’s existing technology stack rather than require its own tools. DDD’s physical-AI annotation page says safety data receives enhanced quality assurance and subject-matter-expert review matched to the customer’s risk profile.

Sources: [347], [348], [349]

Evidence: Digital Divide Data - source-specific boundaries. An impact-sourcing mission does not make the explicit commercial offer an adjacent grant program. No legal nonprofit status is inferred. The apparent mismatch with the AI-data business is unresolved. This records the page’s contents and does not identify DDD Ventures LLC as the settled AI-data contracting entity.

Sources: [347], [348], [349]

Analysis: Digital Divide Data - buyer assessment and unknowns. The offering is commercial AI data work despite its impact-sourcing mission; the mission does not establish a nonprofit legal classification or measured worker outcomes. A pilot should specify physical-AI safety review and platform integration responsibilities. Resolve the published terms-page mismatch before using it to name the AI-data contracting party; the record preserves uncertainty rather than treating DDD Ventures LLC as settled. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with a resolved contracting identity and independently accepted physical-AI outputs.

Sources: [347], [348], [349]

Evidence: Encord - offering and scope. Encord offers managed annotation projects using vetted domain experts for physical AI, multimodal systems and LLMs.

Sources: [135]

Evidence: Encord - disclosed delivery and operating model. Encord describes role-based platform permissions restricting team members to the information needed for their role. Encord describes matching medical imagery to clinicians, robotics tasks to specialist operators and LLM evaluation to native speakers with subject-matter backgrounds. Encord describes human experts reviewing and correcting AI-generated first-pass annotations. Encord describes starting annotation projects with a small pilot to identify guideline gaps and document edge cases with visual examples. Encord states that annotation decisions are logged and visible in real time for project monitoring.

Sources: [135], [136]

Evidence: Encord - source-specific boundaries. Platform control disclosure; no independently inspected assurance report or project-specific deployment is established. This is a disclosed workflow, not a measured accuracy or cost advantage.

Sources: [135], [136]

Analysis: Encord - buyer assessment and unknowns. The relevant proposition combines expert matching, model-assisted first passes and visible annotation decisions. Inspect a pilot’s guideline revisions and who can override generated labels, then verify role permissions on the actual workspace. A match between clinician, operator or native speaker and task is a starting condition, not an independently measured accuracy result. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with specialist qualification, guideline histories and blind correction samples.

Sources: [135], [136]

Evidence: G2i - offering and scope. G2i advertises senior engineering teams for evaluations, RL environments, agent workflows and training-data programs.

Sources: [218]

Evidence: G2i - disclosed delivery and operating model. G2i human-data terms state that services are limited to talent acquisition and disclaim responsibility for candidate work. G2i describes independent review by two developers per task and selective follow-up quality checks by senior engineers. G2i advertises SOC 2 Type II controls and documented encryption, monitoring, access reviews and incident response. G2i describes its bench as vetted senior and staff-level engineers with 10+ years building production systems. The inspected G2i account attributes TypeScript task verification for a Multi-SWE workflow, without establishing authorship of the entire benchmark or a measured contribution to model performance.

Sources: [218], [219], [717]

Evidence: G2i - source-specific boundaries. The same terms include contractor engagement, statements of work and work-product provisions. This unresolved legal/marketing scope tension prevents assuming managed-delivery responsibility. Selective senior checks are described separately in the same annotation-pipeline section; this is not a universal double-adjudication guarantee.

Sources: [218], [219], [717]

Analysis: G2i - buyer assessment and unknowns. The contract boundary is unusually important here: marketing describes data programs, while the inspected human-data terms describe talent acquisition. Establish who owns instructions, review, rework and acceptance before comparing the offer with a managed service. The TypeScript verification account remains a scoped workflow contribution; dual-developer review and advertised security controls are not independently audited outcomes. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with a signed allocation of acceptance/rework responsibility and task-level independent review logs.

Sources: [218], [219], [717]

Evidence: Giskard - offering and scope. Giskard’s red-team product generates adaptive multi-turn attacks using customer business context. Giskard offers software to build and maintain dynamic evaluation datasets for AI agents. The evaluation product lets business experts set ground truths, qualify failures and collaborate on tests. Giskard advertises structured assessment reports and a managed remediation and retesting process.

Sources: [301], [302], [303]

Evidence: Giskard - disclosed delivery and operating model. Its evaluation software can add detected vulnerabilities to a golden dataset.

Sources: [302]

Evidence: Giskard - source-specific boundaries. Service description does not certify a customer’s safety or show every vulnerability was found. No certification/audit report was accessed.

Sources: [301], [302], [303]

Analysis: Giskard - buyer assessment and unknowns. The output can be evaluation software, an evolving golden set or a managed assessment/remediation cycle. Specify which of these is being bought and who owns expert ground truths. For an adaptive attack pilot, hold out business scenarios and measure both discovered failures and false alarms. Tool-generated attacks are not evidence of exhaustive safety coverage or completed remediation. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with held-out business-scenario error counts and independently verified remediation.

Sources: [301], [302], [303]

Evidence: HackerRank Bench - offering and scope. HackerRank Bench offers private benchmarking and long-horizon task creation tuned to client objectives. Bench offers executable repository-scale RL environments with resettable workspaces, hidden verifiers and per-step evidence.

Sources: [305]

Evidence: HackerRank Bench - disclosed delivery and operating model. Its off-the-shelf data section says the offering concerns community submissions and excludes customer data. Bench says its benchmark verifiers are hand-written to test software behavior. HackerRank presents Bench as its software-engineering data, evaluation and environment service.

Sources: [305]

Evidence: HackerRank Bench - source-specific boundaries. The source commitment is not a dataset-license review, individual consent audit, proof of ownership of each submission or permission for every reuse. Benchmark-method disclosure; no benchmark ranking, lack of contamination or verifier soundness is independently established.

Sources: [305]

Analysis: HackerRank Bench - buyer assessment and unknowns. Bench offers a repository-scale environment package rather than only a problem list. Test resets, hidden verifiers, per-step evidence and behavior-based rejection on a small private repository. The stated exclusion of customer data from community submissions is a source boundary, not a record-level rights audit. Benchmark soundness and contamination require their own checks. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with executable private-task manifests, reset tests and hidden-verifier falsification.

Sources: [305]

Evidence: Handshake AI - offering and scope. Handshake operates a contractor program for annotated data and related Human Data Services.

Sources: [221]

Evidence: Handshake AI - disclosed delivery and operating model. The Handshake AI Contractor Agreement names Handshake AI Solutions, LLC as the contracting entity. Successful onboarding does not guarantee a Handshake AI project; contractor selection depends on expertise and project demand. Handshake sets payment amount, type and timing in each project invitation, including task, time or milestone payment bases. The contractor agreement permits technical and operational quality, integrity and compliance monitoring, including remote monitoring and automated tools. The global privacy policy defines Handshake as Stryder Corp. d/b/a Handshake and its affiliates and subsidiaries.

Sources: [221], [222]

Evidence: Handshake AI - source-specific boundaries. The July 2026 privacy policy separately defines Handshake as Stryder Corp. d/b/a Handshake and affiliates; corporate ownership shares are not established. This contractor-program scope does not identify a named customer or downstream model contribution.

Sources: [221], [222]

Analysis: Handshake AI - buyer assessment and unknowns. The dossier documents a contractor program and project invitation terms. For buyers, confirm delivery responsibility and acceptance separately from recruiting access. For contributors, onboarding and project availability are different states, and payment basis varies by invitation. Contractual monitoring rights do not establish actual project audit coverage, current paid headcount or the quality of accepted outputs. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with a specific project agreement, accepted deliverables and observed payment records.

Sources: [221], [222]

Evidence: Humans in the Loop - offering and scope. Humans in the Loop offers managed data-labeling teams, including specialized healthcare and real-time monitoring projects.

Sources: [140]

Evidence: Humans in the Loop - disclosed delivery and operating model. Humans in the Loop describes a two-hour annotation trial on sample data and instructions before final guideline and setup adjustments. Humans in the Loop describes delivering annotated data followed by a client post-mortem meeting to inform subsequent batches. Humans in the Loop’s Fair Work Policy applies to contractors performing paid work who are not full-time or part-time employees. The Fair Work Policy says annotators receive adequate training for every project. The Fair Work Policy calculates payments from the project unit price and units delivered on time and approved by the project supervisor.

Sources: [139], [141]

Evidence: Humans in the Loop - source-specific boundaries. The page calls the annotation-guideline stage “Fine Tuning”; it is not evidence of model fine-tuning services. Policy scope, not a current worker-count, legal employment classification or independently verified compliance finding.

Sources: [139], [140], [141]

Analysis: Humans in the Loop - buyer assessment and unknowns. The two-hour sample trial and post-mortem batch loop are useful mechanisms to inspect rather than performance guarantees. Require written guideline changes, rejected-unit treatment and the supervisor approval rule before scaling. The page’s annotation setup label Fine Tuning should not be read as evidence of model-training services. Worker-policy commitments require separate execution evidence. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with paid trial outcomes, written calibration changes and rejected-work treatment.

Sources: [139], [140], [141]

Evidence: iMerit - offering and scope. iMerit offers domain-expert feedback for model fine-tuning through RLHF services.

Sources: [62]

Evidence: iMerit - disclosed delivery and operating model. The same case study describes manual data correction involving iMerit experts and reviewers during persona-based prompt generation. iMerit states that employees, vendors and contractors are screened and sign NDAs, with project data restricted to the assigned team. iMerit’s case study reports that its annotators manually generated persona-based multilingual prompts for an unnamed Asian technology company’s LLM chatbot corpus. iMerit describes secondary-expert task review followed by sampled client audits and return of rejected tasks to review.

Sources: [61], [62], [63]

Evidence: iMerit - source-specific boundaries. Provider security disclosure is not a current independent certification report. The source’s nine-centre description is not adopted as current footprint. The case PDF is undated. Its later ranking exercises are a roadmap, not established delivered preference rankings. No client/model identity is inferred.

Sources: [61], [62], [63]

Analysis: iMerit - buyer assessment and unknowns. The multilingual prompt case establishes an authoring and expert-correction workflow for an unnamed customer. Keep the stated later ranking roadmap separate from delivered work. A comparable pilot should trace rejected items through secondary review and client audit, and inspect the assigned-team access rule. Domain-expert RLHF marketing alone does not establish measured downstream model gains. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with traceable secondary-review decisions and client acceptance on matched task types.

Sources: [61], [62], [63]

Evidence: Innodata - offering and scope. Innodata offers expert human-in-the-loop feedback for RLHF and direct preference optimization.

Sources: [142]

Evidence: Innodata - disclosed delivery and operating model. Innodata describes explaining response scores with details such as factual errors or logical inconsistencies. Innodata describes feedback types including thumbs-up/down, rating scales and nominal classifications in its RLHF service. Innodata describes a multi-pass assessment process for vetting model responses. Innodata states it provides documented annotation guidelines covering challenging edge cases.

Sources: [142]

Evidence: Innodata - source-specific boundaries. The provider’s absolute quality language is not adopted as an independently measured quality guarantee.

Sources: [142]

Analysis: Innodata - buyer assessment and unknowns. The disclosed signal menu distinguishes binary reactions, scales, categories and explanatory feedback. Select the label type that matches the intended learning or evaluation objective, then inspect multi-pass review on edge cases. A multi-pass process can add review stages without making judgments independent; request raw labels, reviewer roles and resolution reasons. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with raw multi-pass labels, reviewer independence and edge-case adjudication.

Sources: [142]

Evidence: Invisible Technologies - offering and scope. Invisible offers model-ready off-the-shelf datasets spanning professional work, code and customer interactions.

Sources: [64]

Evidence: Invisible Technologies - disclosed delivery and operating model. Invisible describes access to vetted domain experts across technical and specialized fields. Invisible reports hiring, training and coordinating expert teams for an unnamed AI developer’s adversarial NLI and RLHF processes. Invisible’s delivery case describes comprehensive onboarding and cross-training operators to move between processes. Invisible’s privacy policy states its software can collect keystrokes, mouse clicks, screenshots and webcam pictures from agents, with collection notices and possible loss of project access if declined.

Sources: [65], [66], [67]

Evidence: Invisible Technologies - source-specific boundaries. The larger “access to global experts” pool is not a count of paid or active workers. The client and model remain unnamed. The headline launch-time and accuracy comparisons are omitted rather than treated as independent measurements.

Sources: [64], [65], [66], [67]

Analysis: Invisible Technologies - buyer assessment and unknowns. The catalog and the expert-team case represent different purchase units. Ask whether a proposed program includes recruitment, onboarding, cross-training and accepted outputs, or a fixed licensed artifact. Agent monitoring disclosures create a concrete question about worker notices, proportionality and project access. They do not establish worker consent quality or production acceptance rates. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with the actual managed-program obligations and independently checked accepted-output logs.

Sources: [64], [65], [66], [67]

Evidence: isahit - offering and scope. isahit offers a managed workforce for annotation and LLM fine-tuning projects.

Sources: [144]

Evidence: isahit - disclosed delivery and operating model. isahit states each project has a dedicated delivery manager supporting the workforce daily. isahit’s contributor page requires proof of self-employed status and describes collecting identity, residence and self-employment documents during onboarding. isahit states it provides continuous coaching and at least four hours of training per person per project. isahit’s contributor page advertises remuneration above the SMIC of the contributor’s country plus bonuses.

Sources: [143], [144]

Evidence: isahit - source-specific boundaries. Public onboarding requirements, not a determination of employment status under any jurisdiction or proof of an individual’s enrollment. Provider-described training minimum; completion by any individual contributor has not been audited.

Sources: [143], [144]

Analysis: isahit - buyer assessment and unknowns. The disclosed model combines a delivery manager, project training and self-employed contributor onboarding. A pilot should measure training completion, revision handling and accepted output under the promised coaching structure. The remuneration statement is a provider policy, not measured earnings in a defined currency, country and time window; request a project-specific compensation and unpaid-time accounting. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with project-specific paid-time records, training outcomes and accepted-unit samples.

Sources: [143], [144]

Evidence: iSoftStone - offering and scope. iSoftStone lists search and advertising relevance as a managed service alongside in-person, remote and crowdsourced collection or labeling modes. iSoftStone offers voice-data collection, transcription, segmentation, ambient-noise processing, virtual-assistant tuning and text-to-speech services. iSoftStone’s computer-vision offer includes image/video labeling, boxes, categorization, facial-recognition data collection and human-gesture annotation.

Sources: [356]

Evidence: iSoftStone - disclosed delivery and operating model. iSoftStone describes quality assurance through error avoidance, early detection and correction, and continuous improvement during the project lifecycle. The current website privacy policy identifies iSoftStone, Inc. as its operator.

Sources: [356], [357]

Evidence: iSoftStone - source-specific boundaries. A process description, not independent validation of accuracy, annual volumes or labor conditions. This identifies the website-scoped US entity. Current ultimate ownership and group-wide contracting responsibility remain unresolved.

Sources: [356], [357]

Analysis: iSoftStone - buyer assessment and unknowns. The inspected US website supports speech, vision and relevance services with several workforce modes. Confirm the contracting entity, ultimate group responsibility and the project’s delivery mode rather than extend the website operator to every affiliate. A pilot should test error correction and task-specific quality on the relevant voice or vision data. No annual volume or labor-condition inference is warranted. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with verified contracting/group responsibility and mode-specific correction results.

Sources: [356], [357]

Evidence: Keymakr - offering and scope. Keymakr offers human-in-the-loop review for specialized AI through verified industry professionals.

Sources: [146]

Evidence: Keymakr - disclosed delivery and operating model. The SeeChange case identifies Keylabs as Keymakr's proprietary annotation platform used for that project. Keymakr's SeeChange case states that standardized annotation guidelines were drafted and followed to maintain consistency. Keymakr reports a core team of 30 annotators that scaled to 50 at stages of the SeeChange project. Keymakr reports delivering structured annotated video datasets for SeeChange's retail behavior-analysis model development.

Sources: [145]

Evidence: Keymakr - source-specific boundaries. Historical project-team disclosure without a calendar-defined paid-active window. It is not company headcount, recruitable pool, or paid assignment total. Provider-authored named case. No public dataset release, model identity, independent performance result or current contract is inferred.

Sources: [145], [146]

Analysis: Keymakr - buyer assessment and unknowns. The named retail-video case is useful for inspecting guidelines, video units and team scaling on a similar project. Its reported 30-to-50-person team belongs to that case, not a current active workforce. Request a reproducible clip-level acceptance sample and distinguish the Keylabs software role from Keymakr's managed annotation responsibility. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with clip-level acceptance data and a current project staffing/role record.

Sources: [145], [146]

Evidence: Kili Technology - offering and scope. Kili offers software for customer specialists to write post-training examples and judge AI evaluation data. Kili's image annotation tool offers boxes, polygons and masks with SAM drafting segmentations for experts to refine.

Sources: [359]

Evidence: Kili Technology - disclosed delivery and operating model. Kili's quality product supports specialist agreement checks, honeypot scoring and reviewer corrections. Kili's EULA distinguishes SaaS, customer-storage hybrid and customer-managed on-premise software, with selected services governed by the order form. Kili offers introductions to external specialist partners who work inside the customer's project and review workflow. Kili's May 29, 2025 legal notice names MYAI SAS as website editor with Paris registration 843 210 014.

Sources: [358], [359], [360], [361]

Evidence: Kili Technology - source-specific boundaries. Product controls are not audited project outcomes. Customer experts are not automatically Kili employees or a Kili managed annotation workforce.

Sources: [358], [359], [360], [361]

Analysis: Kili Technology - buyer assessment and unknowns. Kili’s software lets customer specialists author and judge data, while partner introductions remain distinct from direct staffing. A pilot should test consensus, honeypots, reviewer authority and the selected SaaS, hybrid or on-premise deployment. Resolve the legal-notice/EULA name inconsistency under the same registration number in the actual order. The discrepancy is not evidence of two independent suppliers. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with the executed order, deployment-specific permission tests and blind specialist outputs.

Sources: [358], [359], [360], [361]

Evidence: Label Your Data - offering and scope. Label Your Data offers data labeling by a trained workforce managed by experienced QA specialists.

Sources: [148]

Evidence: Label Your Data - disclosed delivery and operating model. Label Your Data describes tailored QA workflows in which specialists review outputs for accuracy, consistency and alignment with client requirements. Label Your Data’s notice effective 16 July 2024 identifies SupportYourApp, Inc. as the entity using the Label Your Data name. The same notice says Label Your Data processes client-submitted personal data for labeling services under client instructions or contractual requirements. Label Your Data reports providing polygon annotations to support Searidge Technologies’ computer-vision development. In the Searidge engagement, Label Your Data reports handling video-frame selection, polygon annotation and annotation-tool management using the client’s FTP for exchange.

Sources: [147], [148], [709]

Evidence: Label Your Data - source-specific boundaries. This is the notice’s identification, not a fresh corporate-registry or current ownership finding. Limited to the notice’s processor scope; no independently verified privacy compliance or certification status is asserted.

Sources: [147], [148], [709]

Analysis: Label Your Data - buyer assessment and unknowns. The Searidge case supplies a specific video-frame and polygon-annotation workflow. For a comparable project, define frame selection, polygon geometry, tool administration and transfer responsibilities before pricing. The processor notice identifies a contractual instruction boundary; it does not establish complete source rights, current secure transfer practice or downstream model quality. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with frame/polygon acceptance results and the actual processing/transfer agreement.

Sources: [147], [148], [709]

Evidence: Labelbox - offering and scope. Labelbox’s Alignerr offering delivers structured expert signal as preference pairs, ranked trajectories and rubric-graded assessments.

Sources: [70]

Evidence: Labelbox - disclosed delivery and operating model. Labelbox describes routing Alignerr tasks according to contributor credentials, experience and domain background. Labelbox's 14 October 2026 announcement reports more than $100 million in Series E financing, comprising equity and a credit facility, at a $1.25 billion pre-money valuation. Labelbox's privacy notice permits sharing personal information with service providers, consultants and Workforce Partners needing access to provide services. Labelbox describes continuously calibrating contributor performance against ground truth and validating outputs before pipeline delivery. Labelbox reports that Alignerr produced expert-graded speech signal for an unnamed text-to-speech company's post-training work.

Sources: [70], [71], [72], [680]

Evidence: Labelbox - source-specific boundaries. The source uses dollar symbols without explicitly naming an ISO currency. Financing and valuation are not revenue; the credit facility is not equity funding. This notice covers the stated services scope and explicitly excludes employee and job-applicant personal information. It is not a contributor pay or IP agreement.

Sources: [70], [71], [72], [680]

Analysis: Labelbox - buyer assessment and unknowns. The disclosed deliverables are structured expert signals, including preference pairs, trajectories and rubric assessments, with a separate speech case. Request domain-specific calibration and raw-to-accepted counts. Financing is not evidence of delivered capacity or revenue, and a pre-money valuation does not measure data quality. Workforce-partner privacy provisions need project-specific access and subprocessing detail. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with domain-calibration results, raw-to-accepted denominators and project subprocessing records.

Sources: [70], [71], [72], [680]

Evidence: LXT - offering and scope. LXT offers structured human-feedback workflows including pairwise output ranking, scaled ratings and open-ended feedback.

Sources: [75]

Evidence: LXT - disclosed delivery and operating model. LXT describes gold-task calibration, reviewer overlap scoring, audit sampling and performance tracking in its RLHF process. LXT reports contributors ranked, rated, categorized and rewrote model responses for an unnamed technology company's foundational-LLM fine-tuning program. LXT's privacy policy states client-submitted files are processed under client instructions and offers a data-processing agreement. LXT's case reports qualifying and onboarding 4,000 contributors across 50 languages in one week. LXT states that RLHF data is handled under NDAs and role-based access, with secure-facility execution available for sensitive projects.

Sources: [73], [74], [75]

Evidence: LXT - source-specific boundaries. Client/model identity and public dataset release remain unspecified. These tasks do not establish that LXT trained the client's model itself. The policy's effective date is 19 April 2020; a newer update date is not stated. This is a captured current page, not proof of current legal compliance.

Sources: [73], [74], [75]

Analysis: LXT - buyer assessment and unknowns. The reviewed case combines multilingual onboarding with ranking, rating, categorization and rewriting. The 4,000-contributor, 50-language, one-week account belongs to that engagement and cannot be a current hiring or throughput guarantee. A pilot should test each required language-domain intersection and inspect calibration, overlap scoring and rejected-unit correction under the proposed security setting. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with locale-domain sample results and complete onboarding/calibration/rework logs.

Sources: [73], [74], [75]

Evidence: Mercor - offering and scope. Mercor markets ready-to-license datasets across professional work, coding, search, research, and multimodal reasoning. Mercor offers custom datasets and research partnerships including model evaluation, failure-mode analysis, and post-training.

Sources: [690]

Evidence: Mercor - disclosed delivery and operating model. Mercor documents the APEX-Agents 1.1 release; the attribution is to that named release and its stated construction scope. Mercor's website privacy notice identifies Mercor.io Corporation as the operator of the website. Mercor's catalog advertises a network of more than five million domain experts. Mercor discloses hand-labeling rubric items for a revised APEX judge and an increased false-negative rate versus its naive grader. Mercor disclosed being affected by compromised LiteLLM versions in March 2026. Mercor says catalog tasks are created, vetted, and peer-reviewed by domain experts with automated quality controls. Mercor's website notice explicitly excludes the work.mercor.com service and directs experts to a separate notice. Mercor describes expert-written prompts, rubrics, golden solutions, workplace environments, and long-horizon workflows.

Sources: [690], [738], [739], [740]

Evidence: Mercor - source-specific boundaries. Scoped source statement: Mercor reports releasing APEX-Agents 1.1 with 80 tasks per professional domain, refined through three expert audits. No denominator definition, active-period window, currently paid contributors, or paid assignment count is given in this catalog. Do not equate pool size with production capacity.

Sources: [690], [738], [739], [740]

Analysis: Mercor - buyer assessment and unknowns. The disclosed catalog and custom programs center on expert-written tasks, rubrics, solutions and workplace environments. A pilot should inspect grader false negatives and long-horizon edge cases rather than treat a published benchmark as universal validation. The expert-network figure is not a paid-active count. Security diligence should use the actual work-service notice and scoped incident/remediation evidence. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with independently tested expert tasks, grader error counts and scoped service controls.

Sources: [690], [738], [739], [740]

Evidence: micro1 - offering and scope. Flow is described as the environment where experts create, review and deliver datasets in healthcare, legal and finance. micro1 Realm advertises human-in-the-loop data production for AI labs.

Sources: [228]

Evidence: micro1 - disclosed delivery and operating model. micro1 states that experts see a project pay rate before acceptance and are paid twice monthly. micro1 presents MissionReport as its named benchmark artifact, without an independently reproduced performance finding. Realm describes real-time tracking of data-production velocity, error rates, cost per task and quality. micro1 describes Zara as its AI recruiter that sources and vets domain experts.

Sources: [228], [229], [230]

Evidence: micro1 - source-specific boundaries. Public support-program commitments do not independently establish payment execution; active paid expert counts are not audited. Self-published benchmark construction; no government customer contract or downstream model-training contribution is inferred.

Sources: [228], [229], [230]

Analysis: micro1 - buyer assessment and unknowns. Separate recruitment through Zara, production within Realm/Flow and the named MissionReport artifact. The operational dashboard metrics are a disclosed tracking design, not audited velocity or error rates. A pilot should make each metric’s denominator visible and test review authority. The twice-monthly pay statement is a schedule disclosure, not proof of realized earnings or project continuity. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with metric definitions, reviewer decisions and observed accepted-output/payment logs.

Sources: [228], [229], [230]

Evidence: Mindy Support - offering and scope. Mindy’s FAQ offers image, text, speech-recognition, video and 3D point-cloud annotation services.

Sources: [362]

Evidence: Mindy Support - disclosed delivery and operating model. Mindy’s FAQ says client KPIs and measurement are discussed before project start, with the client ultimately determining the KPIs. Mindy says its R&D department selects and adapts annotation tooling to project needs. Mindy’s FAQ describes phone screening, a recruiter interview, written tests and an interview with the project manager. Mindy’s official FAQ offers a trial to assess deliverables before a longer engagement.

Sources: [362]

Evidence: Mindy Support - source-specific boundaries. Exact official web.run extraction is retained; direct site GET returned403. The FAQ is undated, so workforce and operating-location numbers are not treated as current. Evidence is the exact saved official web.run extracted-page snapshot; origin HTML and live HTTP status were not exposed.

Sources: [362]

Analysis: Mindy Support - buyer assessment and unknowns. The undated FAQ supplies an annotation, screening and trial-process disclosure, not current staffing, office footprint or active openings. Ask for the client-defined KPIs, adapted tool setup and paid trial terms in a current proposal. The exact saved official extraction is the reviewed evidence; it does not establish direct origin-HTML retrieval or a live availability result. The footer outside the frozen read range is not adopted. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with a dated proposal, paid-trial results and current KPI acceptance records.

Sources: [362]

Evidence: Objectways - offering and scope. Objectways offers expert human-in-the-loop labeling for text, image, audio, video and LiDAR training data.

Sources: [151]

Evidence: Objectways - disclosed delivery and operating model. Objectways describes client review of sample outputs and feedback before finalizing annotation guidelines. Objectways states employees are onboarded after an NDA and background checks. Objectways describes a test annotation job to estimate effort, timelines and output quality before scaling. Objectways offers expert annotation capacity from pilot to production for outsourcing or team augmentation.

Sources: [150], [152]

Evidence: Objectways - source-specific boundaries. Provider-reported employee onboarding control; no independent implementation audit, contractor coverage or named-project compliance is established. Availability language is not a paid-active workforce count or guaranteed individual assignment.

Sources: [150], [151], [152]

Analysis: Objectways - buyer assessment and unknowns. The sample-run and client-feedback loop provide a concrete way to test guideline ambiguity before scale. Choose a pilot with hard examples and inspect how sample feedback changes production instructions. NDA and background checks are onboarding disclosures; they do not establish complete access controls or acceptance quality. Team augmentation and outsourced accepted delivery should be priced and warranted separately. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with hard-example sample results and traceable guideline-to-production revisions.

Sources: [150], [151], [152]

Evidence: Ooak Data - offering and scope. Ooak Data offers RL environments built from real enterprise data and workflows.

Sources: [311]

Evidence: Ooak Data - disclosed delivery and operating model. Ooak describes sourcing enterprise material, anonymizing it into digital twins and generating expert-level tasks. Its October 27, 2026 process essay describes human QA, raw-copy deletion and source-company approval before dataset release. Its privacy policy names Ooak Data, Inc. for US website processing and Ooak Data SAS for European operations. Ooak’s website terms reserve data acquisition partnerships to separately negotiated written agreements.

Sources: [311], [312], [313], [314]

Evidence: Ooak Data - source-specific boundaries. A vendor process disclosure, not independent proof that a dataset is anonymous or every release received valid authorization. Controller roles are not proof of group equity ownership or a separate supplier count for the European entity.

Sources: [311], [312], [313], [314]

Analysis: Ooak Data - buyer assessment and unknowns. The distinctive disclosed pipeline converts enterprise material into anonymized digital twins and tasks, with source-company approval and raw-copy deletion described before release. A pilot should inspect one transformation end to end, including residual identifiers and the approval record. Website terms reserve partnership detail to separate contracts; they do not establish transferable rights for every underlying business document. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with independent residual-identifier tests and source-company rights/approval records.

Sources: [311], [312], [313], [314]

Evidence: Pareto AI - offering and scope. Pareto advertises specialist projects built around judgment, reasoning and domain knowledge for training models.

Sources: [235]

Evidence: Pareto AI - disclosed delivery and operating model. Pareto documents the AttuneBench artifact and its human-label contribution, under the specific release account. The Pareto site terms identify Pareto Inc., doing business as Pareto. Pareto says strong expert performance leads to more projects and specialized work across its network. Pareto reviews expert applications on a rolling basis and says it will reach out when there is a strong background match. Pareto's website terms grant customers a nonexclusive, perpetual license to original work product created for them.

Sources: [234], [235], [236]

Evidence: Pareto AI - source-specific boundaries. This is an owned evaluation artifact. The paper, code and dataset license were not independently reviewed in this pass. Forte and Hello Pareto are historical alias hints only; this pass does not verify a current separate product, employer or subsidiary under those names.

Sources: [234], [235], [236]

Analysis: Pareto AI - buyer assessment and unknowns. The reviewed scope combines specialist projects, matching and an AttuneBench artifact. The customer work-product license is nonexclusive and perpetual in the inspected terms, so exclusivity and underlying-source rights require explicit treatment. Rolling applications and performance-based project allocation do not guarantee work or establish a paid-active pool. Compare accepted expert judgments on the intended domain. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with domain-matched expert judgment audits and exact work-product/source-rights terms.

Sources: [234], [235], [236]

Evidence: Polymath - offering and scope. Polymath describes building simulation environments for training and evaluating autonomous agents.

Sources: [238]

Evidence: Polymath - disclosed delivery and operating model. Polymath's terms identify Polymath AI Labs, Inc. and cover both polymathlabs.ai and anthologyresearch.ai. Applying to or joining Anthology Fellows does not create employment; a separate contractor agreement is required before work or payment. Polymath presents Horizon-SWE as its named benchmark artifact; that does not establish the effectiveness of every simulation engagement. Polymath's terms say interview voice recordings and transcripts will not be used to train AI models. Horizon-SWE describes assessing agents on verifiable outcomes with partial credit.

Sources: [237], [239]

Evidence: Polymath - source-specific boundaries. A stated operator relationship is not proof of share ownership, acquisition history or a separate supplier count for Anthology. Community membership is not an active paid-worker count; the separate contractor agreement was not available in this source set.

Sources: [237], [238], [239]

Analysis: Polymath - buyer assessment and unknowns. The simulation offer and Horizon-SWE outcome checks should be inspected as a particular environment product. Test the partial-credit logic and negative cases before using it for training or evaluation. Anthology Fellows membership is not employment or paid work; a separate agreement is required. The interview-data restriction should not be extended to unrelated customer processing without a contract. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with executed task manifests, independent PF/PC grading checks and operative service terms.

Sources: [237], [238], [239]

Evidence: Prime Intellect - offering and scope. Prime Intellect provides an open-source environment framework and an environment hub for RL. Prime Intellect advertises SFT fine-tuning, hosted RL on managed Prime-RL, inference and compute within an integrated platform.

Sources: [240], [691]

Evidence: Prime Intellect - disclosed delivery and operating model. Model Factory terms meter hosted training per GPU-hour allocated to a run, prorated to the second; rates come from the order or platform. The NVIDIA case attributes verified enterprise tasks to the specific Prime Intellect workflow. It does not establish all NVIDIA training input or a causal share of model gains. The NVIDIA case study says enterprise tasks are scored on changes to simulated applications and files, with objectives and guardrails. Model Factory terms distinguish platform services from any personnel-performed post-training work, which needs a separate professional-services order.

Sources: [241], [242]

Evidence: Prime Intellect - source-specific boundaries. Pricing basis is distinct from quoted public compute example prices, data-service pricing, company revenue and customer spend. No commercial amount is asserted. Platform capability alone is not evidence of human-data staffing or bespoke task generation.

Sources: [240], [241], [242], [691]

Analysis: Prime Intellect - buyer assessment and unknowns. Separate open framework access, hosted compute/training and personnel-performed professional services. GPU-hours allocated to a run are a billing unit, not accepted data units or actual productive compute. For an enterprise-task pilot, inspect objective and guardrail checks in addition to final file/application changes. The named case does not establish all downstream model input or causal performance gains. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with allocated-resource logs, independent task outcomes and an explicit service/billing scope.

Sources: [240], [241], [242], [691]

Evidence: Prolific - offering and scope. Prolific’s AI tasker groups cover image, video, audio, comparative reasoning, fact-checking, and structured writing. Prolific offers both researcher-managed participant access and managed complete-dataset collection with multi-stage quality assurance.

Sources: [712], [720]

Evidence: Prolific - disclosed delivery and operating model. Prolific’s pricing page reports more than 300,000 active participants and separately over one million on its waitlist. Prolific offers participants selected through verified experience or targeted assessments for AI training and evaluation tasks. The inspected research case establishes recruitment through Prolific into a researcher-owned labeling interface, rather than Prolific authorship of all study data or analysis. Prolific’s captured pricing FAQ states usual platform fees of 42.8% for corporate customers and 33.3% for academic or nonprofit customers. Prolific states minimum permitted participant rates of GBP 6 or USD 8 per hour and recommends GBP 9 or USD 12. Prolific forbids under-18 recruitment and ordinarily disallows collection of directly identifying participant information. Prolific describes identity checks, ongoing pool monitoring, in-study checks, researcher approval history, and human appeals. Prolific states ISO 27001 and Cyber Essentials certification while describing SOC 2 Type 2 as a future plan. The case reports L1 English participants in the UK or US with at least 100 completed experiments and a representative-sample option.

Sources: [712], [720], [741], [742], [743], [744]

Evidence: Prolific - source-specific boundaries. The page explicitly says active, but does not define its activity window. Waitlisted people are not active participants. No currently paid workforce or assignment count is established. Role is recruitment. Researchers used their own modified Potato annotation interface; no evidence here that Prolific authored the labels, adjudicated them, or supplied a production model training set.

Sources: [712], [720], [741], [742], [743], [744]

Analysis: Prolific - buyer assessment and unknowns. Participant access can be appropriate when the research team owns the interface and judgment protocol; managed collection changes that responsibility. Define expertise and sampling rather than treating a representative-sample option as universal coverage. Active-participant, waitlist, tariff and reward-floor disclosures use different denominators. The inspected fee FAQ leaves the percentage denominator unspecified, so obtain a complete project quote. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with task-specific recruitment feasibility, full price/reward denominators and held-out label quality.

Sources: [712], [720], [741], [742], [743], [744]

Evidence: Proximal - offering and scope. Proximal describes a data research and engineering organization rather than a recruiting firm or talent marketplace. Proximal says it turns real-world traces and human-workflow artifacts into evaluations and targeted post-training data.

Sources: [244], [246]

Evidence: Proximal - disclosed delivery and operating model. In its September 29, 2026 announcement, Proximal reports scaling to $200M in annualized revenue. Proximal’s privacy policy requires prior written consent and an executed agreement before using customer material to improve generalized models. Proximal’s MSA identifies Proximal AI, Inc. and requires executed order forms for customer services. Proximal documents its scoped FrontierSWE collaboration; no broader customer contribution or exclusive production role is inferred. Proximal’s MSA permits independent contractors and keeps Proximal responsible for subcontracted services.

Sources: [243], [245], [246], [674]

Evidence: Proximal - source-specific boundaries. Self-reported annualized pace, not an audited annual revenue total, recognized revenue, profit, bookings, valuation or contractor payout. The source does not specify the annualization window, currency code, gross-versus-net accounting, or revenue by business line. This applies to generalized model training; it does not prohibit customer-authorized training services or prove compliance in individual projects.

Sources: [243], [244], [245], [246], [674]

Analysis: Proximal - buyer assessment and unknowns. The disclosed organization takes responsibility for data research and engineering rather than presenting itself as a recruiting marketplace. The MSA’s subcontractor responsibility is relevant to acceptance accountability. Inspect one trace-to-task transformation and consent boundary. The $200M announcement is self-reported annualized pace with undisclosed annualization window, currency code and gross/net basis, not audited annual or segment revenue. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with trace-to-task acceptance evidence and period/scope/basis-resolved financial disclosure.

Sources: [243], [244], [245], [246], [674]

Evidence: Refresh - offering and scope. Refresh offers simulation environments that turn enterprise workflows into measurable and trainable tasks. Refresh offers simulated worlds for coding and computer use with verifiable rewards.

Sources: [251]

Evidence: Refresh - disclosed delivery and operating model. Refresh presents BlenderBench as its published benchmark artifact; the associated verifier description remains source-reported. Refresh says it sources real tasks by partnering with expert engineers, reproduces failures and codifies verifiable tasks. Refresh’s website terms state that benchmarks, datasets and research may have separate licensing terms. Refresh says BlenderBench uses versioned verifiers checking saved scenes, protected starting content and rendered outputs, with task-specific rubrics.

Sources: [250], [251], [252]

Evidence: Refresh - source-specific boundaries. An owned benchmark and evaluation offering; its stated longer-term RL aspiration is not evidence of completed RL-training delivery. No named enterprise customer, completed contract or specific downstream training artifact is established by the homepage.

Sources: [250], [251], [252]

Analysis: Refresh - buyer assessment and unknowns. Versioned task verifiers and saved-scene checks make BlenderBench useful for asking concrete reproduction questions. For a computer-use or visual-editing pilot, test protected starting content and rendered outputs alongside task success. Expert-sourced tasks and separate artifact licenses do not establish unrestricted rights or complete contamination resistance. Maintenance and version migration should be part of the contract. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with versioned scene resets, protected-content tests and independent visual verification.

Sources: [250], [251], [252]

Evidence: RWS / TrainAI - offering and scope. TrainAI offers response rating, transcription, image segmentation and object-tracking data services. TrainAI offers automated and human-in-the-loop validation of AI training data.

Sources: [321], [324]

Evidence: RWS / TrainAI - disclosed delivery and operating model. The FAQ describes expert review, repeated-annotation consensus and golden-answer comparisons. The February 28, 2023 RWS announcement introduces TrainAI as an RWS brand. Its community FAQ describes language, general ML and project-specific screening tests.

Sources: [322], [323]

Evidence: RWS / TrainAI - source-specific boundaries. This records described QA techniques, not a measured error rate or independent verification of their application on each project. Historical brand launch, not a current separate legal-entity or ownership registry check. RWS group/Generate financials are not TrainAI standalone revenue.

Sources: [321], [322], [323], [324]

Analysis: RWS / TrainAI - buyer assessment and unknowns. TrainAI’s disclosed service menu spans response rating and perception tasks, with language and project-specific screening. A pilot should test the relevant signal type and inspect how repeated annotation and gold answers interact. The RWS brand relationship prevents double counting; it does not establish uniform staffing, security or acceptance performance across all programs. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with language/project qualification results and separately observed gold/repeated-label performance.

Sources: [321], [322], [323], [324]

Evidence: Sama - offering and scope. Sama offers human-verified annotation, validation and evaluation services for computer vision, NLP and multimodal AI.

Sources: [81]

Evidence: Sama - disclosed delivery and operating model. Sama describes maintaining annotation consistency through written guides and continuing calibration of labeling instructions as data changes. Sama reports its in-house data team reviewed and rewrote image captions to produce fine-tuning data for an unnamed Fortune 500 company. Sama describes using a full-time in-house workforce in a secure facility and states it does not keep or share customer datasets for training or other purposes. Sama states that project rooms and data are restricted to the assigned employees and production staff undergo background checks. In the same engagement, Sama says its team evaluated prompt-response turns and supplied rewritten prompts and rationale for preferred responses.

Sources: [80], [81], [82]

Evidence: Sama - source-specific boundaries. The client is anonymous; no named company, model, or public dataset is attributed. Published workforce/data-use representation, not a guarantee for all contracts or a verified pay/welfare finding. No wage rate or current paid headcount is disclosed here.

Sources: [80], [81], [82]

Analysis: Sama - buyer assessment and unknowns. The in-house secure-facility model is a disclosed delivery arrangement, while the caption and prompt cases describe specific authoring/rewrite work for an unnamed customer. A comparable pilot should preserve originals, edits and rationales and examine how calibration changes over batches. A statement about not reusing customer datasets requires confirmation in the actual project agreement and execution evidence. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with batch-level originals/edits, calibration histories and project access/data-use evidence.

Sources: [80], [81], [82]

Evidence: Scale AI - offering and scope. Scale offers tailored generative-AI training datasets curated by vetted subject-matter experts. Scale publicly offers model evaluation including targeted red-teaming.

Sources: [746]

Evidence: Scale AI - disclosed delivery and operating model. The HelpSteer3 source names Scale as an annotation contributor, with the more specific reviewed subset claims defining its scope. The HelpSteer3 paper assigns Scale recruitment to the general, STEM and code preference workflow. The InstructGPT paper names Scale as a recruitment channel for that study, without making Scale the author of every prompt, comparison or model component. Scale reports nearly $1 billion paid cumulatively to hundreds of thousands of contributors across more than 150 countries. The FY2025 modern-slavery statement is made on behalf of Scale AI, Inc. and its wholly owned subsidiaries. HelpSteer3 explicitly attributes Scale annotation to its general, STEM and code subset, separately from the multilingual subset. Scale says contributor-project matching uses background, skills, profile, and work performed on Outlier. Scale describes automated task review with human experts setting standards and reviewing edge cases. Scale publicly states it has a SOC 2 Type 2 report and ISO/IEC 27001:2022 certification. Scale describes Data Infrastructure and AI Applications as two primary business segments.

Sources: [713], [715], [729], [731], [745], [747], [748]

Evidence: Scale AI - disclosed delivery and operating model. For sensitive-content AI safety projects, Scale describes advance notice, ability to decline, and wellness resources as standard practice. Scale states contributors can access work through third-party BPOs or other service providers and proprietary platforms including Outlier. The pinned HelpSteer2 release attributes annotation to Scale under that release’s stated workflow. The pinned HelpSteer3 release records Scale annotation for its specified subset.

Sources: [451], [570], [713]

Evidence: Scale AI - source-specific boundaries. NVIDIA dataset-card attribution, limited to General, STEM and Code domains of HelpSteer3-Preference. The dataset card does not establish contract value, payment terms, an exhaustive supplier list, or contributions to other datasets.

Sources: [451], [570], [713], [715], [729], [731], [745], [746], [747], [748]

Analysis: Scale AI - buyer assessment and unknowns. The strongest artifact examples are release- and subset-specific: HelpSteer annotation and recruitment claims do not make Scale the sole creator of all related data. A pilot should inspect expert matching, automated review and human escalation on the proposed domain. Cumulative payouts and broad platform/BPO channels cannot establish present paid capacity, wages or the revenue of a specific business segment. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with release/subset-specific attribution, project-qualified labels and complete escalation logs.

Sources: [451], [570], [713], [715], [729], [731], [745], [746], [747], [748]

Evidence: Shaip - offering and scope. Shaip offers end-to-end RLHF preference datasets, rewrites and rationales produced by its own experts.

Sources: [157]

Evidence: Shaip - disclosed delivery and operating model. Shaip says experts calibrate on 100-200 items and enter production only after meeting its agreement threshold. Shaip offers preference-data delivery in JSONL or CSV with a summary of rejection-driving failure modes. Shaip describes matching evaluators to domain and language, with healthcare, legal, finance and code work assigned to credentialed professionals. Shaip describes sampled blind double-labeling, seeded gold pairs and weekly agreement reports during preference collection.

Sources: [157]

Evidence: Shaip - source-specific boundaries. A described onboarding gate, not a measured inter-rater agreement result. Delivery-format offering, not evidence of a completed named engagement.

Sources: [157]

Analysis: Shaip - buyer assessment and unknowns. The disclosed offer specifies preference files, rationales and rejection summaries, with credential matching and a calibration gate. Ask for calibration denominators and raw blind-double-label samples before reading weekly agreement as quality across all attempts. The stated 100-200-item calibration is a process description, not proof of current performance or an independently validated threshold. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with calibration denominators, blinded double labels and independently adjudicated failures.

Sources: [157]

Evidence: Snorkel AI - offering and scope. Snorkel says it supports text, code, image, audio, video, long-context, and vision-language datasets.

Sources: [711]

Evidence: Snorkel AI - disclosed delivery and operating model. Snorkel Data Series is described as non-exclusive, documented dataset and environment products refreshed quarterly. Snorkel’s privacy policy identifies Snorkel AI, Inc. Snorkel’s expert community offers authoring, peer review, and comparative evaluation under project instructions and rubrics. Snorkel documents its finance-reasoning benchmark at the captured release; no other financial benchmark relationship is inferred. Snorkel describes screening before project matching and typically advanced-degree, geographic, and independent-contractor eligibility. Snorkel describes project-based payment through HireArt: accepted work from roughly the prior week is processed weekly, the aim is within two weeks after acceptance, and exact timelines may vary by project requirements; experts may opt into matching projects. Snorkel describes calibration, dual-expert review, consensus, sampling, provenance, and programmatic quality checks. Snorkel states SOC 2 Type 2 compliance and scopes data-access and security controls at engagement start. Snorkel’s privacy policy contemplates staffing and workforce partners and client disclosure needed for an expert engagement.

Sources: [682], [684], [749], [750]

Evidence: Snorkel AI - source-specific boundaries. Exact captured page version only; no immutable release identifier or publication date is shown. Its LLM-judge evaluation is separate from expert creation of the questions. Scoped source statement: Snorkel says its Finance Reasoning benchmark contains 79 question-answer pairs co-created with financial experts. Rates depend on expertise and project complexity, acceptance is required, and invitations depend on eligibility and matching; neither typical earnings nor an assignment guarantee is established.

Sources: [682], [684], [711], [749], [750]

Analysis: Snorkel AI - buyer assessment and unknowns. The comparison must distinguish non-exclusive refreshed products from project-specific expert work. Request the exact quarterly version, review rubric, consensus treatment and security scope at engagement start. Opt-in, acceptance-conditioned processing, the two-week aim and possible project variation describe participation and payment administration, not guaranteed assignments, an unconditional payday or realized pay. A finance benchmark attribution does not generalize to all financial data. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with exact product versions, task-matched review evidence and engagement-specific rights/security scope.

Sources: [682], [684], [711], [749], [750]

Evidence: SuperAnnotate - offering and scope. SuperAnnotate provides data-annotation project expertise and service-team support for customized RAG evaluation workflows.

Sources: [84]

Evidence: SuperAnnotate - disclosed delivery and operating model. SuperAnnotate’s Databricks case attributes creation of golden datasets for benchmarking LLM judges to their partnership. SuperAnnotate’s case describes testing agreement between human evaluators and the model as part of feedback collection. SuperAnnotate reports developing a customized grading rubric with Databricks, including examples and dynamic annotation rules for evaluator consistency. SuperAnnotate’s Trust Center lists third-party platform support services, including cloud infrastructure, business communication, analytics and customer support. SuperAnnotate’s Trust Center reports SOC 2 Type II assurance and an ISO 27001:2022-certified information-security management system.

Sources: [84], [85]

Evidence: SuperAnnotate - source-specific boundaries. The attribution is a provider-published customer quote and case narrative. It does not identify a public benchmark dataset release or a training contribution to a named foundation model. No numerical accuracy or cost/speed result is adopted as an independent performance finding.

Sources: [84], [85]

Analysis: SuperAnnotate - buyer assessment and unknowns. The Databricks case is a focused reference for human/model-judge calibration and golden dataset construction. A buyer should hold out examples when refining the rubric and measure error types rather than only aggregate agreement. Trust-center statements and platform-support subprocessors describe disclosed controls and dependencies; underlying assurance scope and customer deployment still require diligence. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with held-out judge-calibration errors and deployment-specific assurance evidence.

Sources: [84], [85]

Evidence: Surge AI - offering and scope. Surge offers RLHF data, SFT demonstrations, and human model evaluation. Surge offers data for text, images, audio, and video. Surge offers agentic reinforcement-learning environments and behavior verifiers.

Sources: [753]

Evidence: Surge AI - disclosed delivery and operating model. Surge’s privacy policy identifies the service operator as Surge Labs, Inc. The pinned GSM8K documentation names Surge as a data-collection contributor; re-solving and repair are separately documented stages. The GSM8K card describes re-solving by different workers and repairing or discarding disagreements. Surge markets prebuilt datasets covering RL environments, coding, and core model capabilities. Surge’s privacy policy permits personal-information sharing with service providers including data-labeling providers. Surge’s published 2021 terms grant it a perpetual internal-business and service-improvement license for labeling User Data.

Sources: [751], [752], [753], [754]

Evidence: Surge AI - source-specific boundaries. The card quotes Appendix A of the original paper. It supports expanded collection, not sole authorship of every problem or all initial recruitment. Full paper was not read in this supplier invocation. Scoped source statement: OpenAI’s pinned GSM8K card says Surge AI helped scale data collection after an initial Upwork-collected set.

Sources: [751], [752], [753], [754]

Analysis: Surge AI - buyer assessment and unknowns. The disclosed portfolio combines feedback, demonstrations, licensed data and agent environments. GSM8K is a specific attributed collection and checking case, not evidence that Surge produced every later math dataset. Inspect rights terms closely: the captured 2021 User Data license concerns internal business and service improvement, and its applicability to a new engagement needs contract confirmation. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with artifact-specific contribution records, independent acceptance tests and applicable data-use terms.

Sources: [751], [752], [753], [754]

Evidence: TaskUs - offering and scope. TaskUs offers domain-specific data sourcing, structuring and labeling for training sets. TaskUs offers human review that ranks model outputs and refines prompts within model lifecycles.

Sources: [88]

Evidence: TaskUs - disclosed delivery and operating model. TaskUs's Global AI Policy prohibits processing client data or personal information through an AI system without explicit client approval. TaskUs policy requires qualified human review of internal AI recommendations affecting employment and disallows automation as the sole basis for high-stakes employment decisions. TaskUs reports redesigning inconsistent RLHF response reviews for an unnamed European AI company through shared review guidelines and audit/feedback loops.

Sources: [86], [87]

Evidence: TaskUs - source-specific boundaries. A policy obligation does not certify every operation complies. The June 2026 update date is a source label, not a new audit date. This describes the company's AI-use policy; it does not establish contractor pay, assignment continuity or data-labeling labor outcomes.

Sources: [86], [87], [88]

Analysis: TaskUs - buyer assessment and unknowns. The scoped European-client case describes review-process redesign, while policy disclosures require client approval for AI processing and human review of employment recommendations. A pilot should test the actual review loop and approval log. An internal AI policy does not establish a customer dataset's rights, a model's safety or the performance of every delivery team. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with customer-specific approval logs and independently observed review/rework outcomes.

Sources: [86], [87], [88]

Evidence: TELUS Digital - offering and scope. The brochure offers crowd-sourced digital data collection, on-site moderated projects and complex field operations. TELUS Digital's AI Data Solutions brochure offers human-aligned datasets generated by domain experts sourced programmatically from Experts Engine, for use with Fine-Tune Studio.

Sources: [89]

Evidence: TELUS Digital - disclosed delivery and operating model. The brochure labels more than one million people as AI Community members across more than 450 global locales. TELUS Digital describes software/process checks, expert human reviews, and contributor qualification and verification as its AI-data quality-assurance approach. The brochure describes secure data handling, SOC 2/TISAX-compliant platforms and ISO 27001-certified labeling facilities. TELUS Digital describes flexible AI-data workforce models including managed crowd, remote, secure in-facility and hybrid delivery.

Sources: [89]

Evidence: TELUS Digital - source-specific boundaries. Community membership is not paid-active headcount, completed assignments, or a current reporting-window labor denominator. Provider-reported compliance/certification; actual certificates, scope and audit reports were not inspected.

Sources: [89]

Analysis: TELUS Digital - buyer assessment and unknowns. The brochure distinguishes crowd, remote, secure-facility and hybrid arrangements. Obtain a proposal specifying which applies, with locale-domain qualifications and review responsibilities. Community membership and locale counts are pool descriptions, not active paid assignments. Security and facility certifications in a brochure should be matched to the particular legal entity, location, platform and project scope. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with the contracted delivery mode, local qualifications and site/system-specific assurance evidence.

Sources: [89]

Evidence: Terac - offering and scope. Terac offers verified expert access to human-data companies and frontier labs for post-training, evaluations and red teaming. Terac offers business-workflow partnerships producing training and verification examples within agreed anonymization and approval boundaries.

Sources: [329], [330]

Evidence: Terac - disclosed delivery and operating model. Terac says pay-as-you-go buyers pay expert payout plus its platform fee, with negotiated enterprise packages available. Its partnership page says businesses receive monthly USD payments per verified completion and set worker/business allocation. Terac describes identity, credential and conversational domain checks with ongoing performance scoring.

Sources: [329], [330]

Evidence: Terac - source-specific boundaries. Billing basis only; no actual fee, revenue, gross-versus-net accounting or guaranteed expert income is established. This is business-partner payout policy, not a worker’s stated rate, actual payment proof or promised earnings.

Sources: [329], [330]

Analysis: Terac - buyer assessment and unknowns. The expert-access and workflow-partnership offers have different acceptance units. Clarify the meaning of verified completion, platform fee, worker/business allocation and monthly USD payment before calculating project economics. Screening is a disclosed process, not a credential audit or a guarantee of availability. Agreed anonymization and approval boundaries need record-level tests. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with verified-completion records, complete fee/allocation terms and observed approved payments.

Sources: [329], [330]

Evidence: Toloka - offering and scope. Toloka offers SFT demonstrations, preference data, evaluation datasets, and hybrid synthetic/human data workflows.

Sources: [756]

Evidence: Toloka - disclosed delivery and operating model. Toloka’s notice describes anti-fraud access restrictions and a right to human review of automated decisions. Toloka’s notice bases biometric identity matching on explicit consent and warns that declining verification can prevent onboarding. Toloka discloses that some client coding deliverables may include contributor authorship metadata used in model development, evaluation, or training. Toloka’s notice applies to Toloka AI B.V. and relevant affiliates and covers the Toloka and Mindrift platforms. Toloka says its vetted experts have advanced degrees and industry experience. Toloka’s task guidelines state different conditions for public-space images, third-party personal information, and personal information collected from Tolokers; the third-party rule lists both anonymization before upload and consent agreements. Toloka’s January 2026 self-serve launch describes an AI-assisted pipeline with client-calibrated quality rules and revision routing. Toloka’s task guidelines require advance sensitive-content disclosure and an acknowledgement checkbox before labeling. Toloka describes taxonomy design by domain experts and data architects, followed by human verification of generated data.

Sources: [718], [755], [756], [757]

Evidence: Toloka - disclosed delivery and operating model. The pinned U-MATH card says math experts at Toloka AI and Gradarius validated university-curriculum problems authored by professors. Toloka’s platform distinguishes general annotators, trained AI tutors, and domain experts.

Sources: [718], [758]

Evidence: Toloka - source-specific boundaries. Specific-project policy category; not proof that any named person or particular repository was delivered. No personal contact data is included in these dossiers. These are published platform requirements, not evidence that any particular dataset complied. Conditions are kept separate by task category; third-party anonymization and consent are conjunctive. The linked Guidelines on data collection were not separately captured, so their additional requirements remain unknown.

Sources: [718], [755], [756], [757], [758]

Analysis: Toloka - buyer assessment and unknowns. The disclosed workflow distinguishes general annotators, AI tutors and domain experts and combines synthetic generation with human verification. Keep U-MATH validation separate from professor authorship. A pilot should inspect category-specific personal-data conditions, sensitive-content notice and automated-decision appeals. Biometric consent with onboarding consequences is a documented condition, not evidence of workers' experience or universal compliance. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with task-specific qualification and acceptance logs, with consent/appeal execution evidence.

Sources: [718], [755], [756], [757], [758]

Evidence: Trajectory Labs, PBC - offering and scope. Trajectory Labs describes creating safety tasks, environments and adversarial examples exposing agent failures. It describes building tools and agent workflows for labs to train and evaluate models.

Sources: [331]

Evidence: Trajectory Labs, PBC - disclosed delivery and operating model. Trajectory describes connecting red teamers, researchers and engineers to frontier-lab projects. Its privacy policy names Trajectory Labs, Public Benefit Corporation and describes the website as informational. Its homepage distinguishes novel expert-discovered attacks from automated testing of known attacks.

Sources: [331], [332]

Evidence: Trajectory Labs, PBC - source-specific boundaries. No named customer contribution, active paid contractor count, pay rate or hiring guarantee is inferred. Website privacy provisions do not resolve customer safety-data contracts, contributor terms or operational security controls.

Sources: [331], [332]

Analysis: Trajectory Labs, PBC - buyer assessment and unknowns. The documented distinction between novel expert attacks and automated known attacks is useful for safety pilot design. Specify threat model, protected information, finding adjudication and retesting, then hold out attack families. Website informational status does not establish customer-processing safeguards. A connection to frontier-lab projects does not identify an unnamed customer or prove comprehensive adversary coverage. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with independently adjudicated novel attacks, held-out families and scoped customer safeguards.

Sources: [331], [332]

Evidence: TrustScale / e2f - offering and scope. Humanus advertises text, audio, image, video and 3D labeling to specification with batch quality review. Humanus offers supervised fine-tuning examples written and reviewed by domain experts in the customer's field and language. Humanus offers responses ranked and rated by qualified reviewers against customer criteria for RLHF and preference data.

Sources: [686], [687]

Evidence: TrustScale / e2f - disclosed delivery and operating model. The pinned official application bundle contains Lawrence Snapp's June 16, 2026 Why We Became TrustScale article, which explicitly says e2f becomes TrustScale. Humanus describes expert qualification against golden data, reviewer agreement tracking, versioned guidelines and second-reviewer sampling. TrustScale's published investor page describes it as management-owned and independent, without institutional investor backing. TrustScale's privacy summary assigns it a processor role for organizational Humanus use, governed by the organization's agreement including its DPA.

Sources: [686], [687], [688], [689]

Evidence: TrustScale / e2f - source-specific boundaries. The rebrand bridge comes from the dated article decoded from the exact official application bundle; it does not establish rendered-page execution or current HTTP redirects. Humanus describes second-reviewer sampling of each batch, not independent second review of every item or measured achieved accuracy. The organizational-use row of the published privacy summary is distinct from its individual/website controller roles and does not supply an executed customer agreement or DPA.

Sources: [686], [687], [688], [689]

Analysis: TrustScale / e2f - buyer assessment and unknowns. TrustScale, e2f and Humanus are retained within one supplier cluster: a disclosed rebrand, organization and product relationship, rather than three competing firms. A pilot should inspect expert-written examples, second-reviewer sampling and golden-data qualification. Management-owned and independent are provider disclosures, not a cap-table audit. The organizational processing summary requires the actual agreement and DPA for customer-specific diligence. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with independent sampling of expert examples, reviewer roles and operative processing agreements.

Sources: [686], [687], [688], [689]

Evidence: Turing - offering and scope. Turing Frontier is offered as a platform connecting AI labs with verified US-based scientific and professional experts.

Sources: [708]

Evidence: Turing - disclosed delivery and operating model. Turing describes both AI-lab data and RL-environment work and enterprise agent-development work. Turing's privacy policy identifies Turing Enterprises, Inc. and its affiliates and subsidiaries. Turing's experts page advertises more than five million experts and coverage across more than 250 domains. Turing's California incentive notice describes opt-in participation and withdrawal that applies to future incentive offerings. Turing's privacy policy lists developing, training, fine-tuning, and prompting AI systems among uses of personal information. The same Turing case reports over 120,000 RLHF preference pairs contrasting compliant actions with named failure modes. Turing describes human-driven sessions in isolated cloud sandboxes with step-level capture across over 150 tools and skills. Turing says Frontier applies identity, background, and work-history verification to experts.

Sources: [708], [759], [760], [761]

Evidence: Turing - source-specific boundaries. Two described lines of activity; not independently reported financial segments. A talent-pool claim, without the number currently active, hired, assigned, or paid. The US-specific Frontier program is not equated with the entire global pool.

Sources: [708], [759], [760], [761]

Analysis: Turing - buyer assessment and unknowns. Separate AI-lab data/environment work from enterprise agent development, and isolate the named case's sandbox capture and preference-pair deliverables. Step-level traces are useful only if acceptance checks the relevant failure modes and rights. The advertised expert pool and domain count do not establish current paid activity; personal-data uses and opt-in incentive terms require service-specific interpretation. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with case-specific trace/pair provenance and independently validated task acceptance.

Sources: [708], [759], [760], [761]

Evidence: V7 / Darwin - offering and scope. Darwin offers AI-assisted labeling with Auto-Annotate and SAM2, video tracking and external-model pre-labeling. Darwin advertises access to specialist annotators including radiologists, pathologists, language experts and native speakers.

Sources: [378]

Evidence: V7 / Darwin - disclosed delivery and operating model. V7's December 17, 2025 Darwin MSA names V7 Ltd. and defines Darwin as its SaaS platform for structuring, labeling, processing, sharing and analyzing customer input data. The Darwin MSA says V7 acts as a reseller for some platform add-on services and does not itself deliver those services. Darwin lets users design multistage review workflows with roles, task assignment and conditional routing.

Sources: [378], [379]

Evidence: V7 / Darwin - source-specific boundaries. Advertised features; comparative speed and accuracy numbers are omitted. Darwin is a V7 product and is not a separate organization.

Sources: [378], [379]

Analysis: V7 / Darwin - buyer assessment and unknowns. Darwin is a V7 product, with software workflow controls and advertised specialist access. The MSA’s reseller clause applies to some add-ons, so identify the actual party delivering any annotator service before assigning acceptance responsibility. A pilot should inspect AI drafting, role routing and specialist correction. Advertised radiologist or language access is not evidence that V7 employs that workforce or completed a named delivery. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with the actual add-on delivery party, permission tests and blind specialist correction outcomes.

Sources: [378], [379]

Evidence: Welocalize / Welo Data - offering and scope. Welo Data offers expert-authored instruction-response pairs for supervised fine-tuning and human pairwise preference ranking.

Sources: [94]

Evidence: Welocalize / Welo Data - disclosed delivery and operating model. Welocalize’s December 2023 case reports Welo Data rewrote unacceptable LLM outputs and graded model responses for an unnamed foundational-model developer. The 2023 case reports mobilizing more than 9,500 remote workers across more than 35 locales for LLM evaluation workflows. Welo Data describes continuous session monitoring through NIMO and tracking inter-annotator agreement in its managed programs. Welocalize’s information-security policy requires security/privacy training for relevant parties and makes adherence mandatory for employees and contractors. The 2023 case says workers were informed about sensitive content and their consent obtained before exposure.

Sources: [92], [93], [94]

Evidence: Welocalize / Welo Data - source-specific boundaries. Historical 2023 engagement, not a current customer or a named model/dataset attribution. Mobilization is not a current paid-active count, annual employment headcount or paid-assignment total. The case has no paid-active reporting window.

Sources: [92], [93], [94]

Analysis: Welocalize / Welo Data - buyer assessment and unknowns. The December 2023 case supports a dated multilingual evaluation and rewriting workflow, including informed sensitive-content exposure. Its mobilized-worker and locale figures are not current workforce totals. A pilot should inspect how NIMO monitoring, agreement metrics and expert-authored SFT interact on the intended locale. Policy training commitments do not substitute for execution or worker-wellbeing evidence. Confidence is moderate for the disclosed scope and unestablished for independently measured operational performance. The assessment would change with locale-specific raw labels, accepted rewrites and paid-time/wellbeing evidence.

Sources: [92], [93], [94]

Method: The supplier count applies the graph's offering-based qualification rule. It requires an approved, included, full_source supplier record with its own approved has_capability assertion; it does not count a supplier only because a paper names a contribution. This distinction explains why the following narrow contributor observation can be discussed without becoming another broadly qualified provider dossier.

Evidence: Translated - narrow artifact contribution. The reviewed HelpSteer3 records identify Translated's multilingual annotation and recruitment role. They do not establish the rest of a current general supplier profile, its security controls, commercial terms, staffing availability or responsibility for other HelpSteer subsets. That scoped contribution is retained without inferring a general service catalog or counting a second broad offering from the paper relationship alone.

Sources: [570], [715], [729]

Evidence: BenchFlow - reviewed adjacent software, commercial qualification pending. The pinned README at commit a61a26f4db4d566de16928565ea58abf3dbaf064 describes BenchFlow 0.10, lists task sandbox backends and documents run rewards, token usage, trajectories and a JSONL training-export interface. Its repository LICENSE is Apache-2.0, without licensing every task, dataset or model. The available software does not by itself establish a commercial data-production offer or a delivered customer corpus. Separate captured documentation retains its own earlier version scope; no old receipt is retargeted.

Sources: [335], [336]

Analysis: Dossier-reading qualification: the newly reviewed HelpSteer and HelpSteer2 papers explicitly describe recruitment through Scale AI in their respective pinned versions. They support those scoped reported recruitment/workflow roles, rather than sole artifact authorship, other releases or measured current service quality. Conversely, the surgical QA study keeps its four companies anonymous, so its results cannot be assigned to any named dossier. Software used inside a paper, including LabelAId's Snorkel label model, does not by itself establish a commercial data-service contribution.

Sources: [429], [430], [432], [435]

Evidence: DataPure - offering and scope. DataPure offers AI-plus-human validation of model-flagged edge cases, LLM output checking and custom data gathering. DataPure describes AI crawlers that extract and organize online data for customer use.

Sources: [482]

Evidence: DataPure - disclosed delivery and operating model. A hosted testimonial attributed to Podchaser co-founder Cole Raven describes DataPure support for its podcast database and ad detection. The resources listing describes DataPure working with a mystery-shopping firm on a beverage-brand shelf audit.

Sources: [482], [483]

Evidence: DataPure - source-specific boundaries. The Podchaser account is a supplier-hosted testimonial, without a separate customer artifact or inspected contract. The shelf-audit observation is only a resource-listing summary for an unnamed buyer. Marketing does not establish expert qualifications, paid task rates, achieved accuracy or clearance for every web source and training use.

Sources: [482], [483]

Analysis: DataPure - buyer assessment and unknowns. Edge-case validation and general web-data collection require different acceptance briefs. A buyer should freeze the model-flagging rule, sample both flagged and unflagged cases, and measure false accepts rather than count completed checks. Web gathering also needs an exact source/use schedule. Confidence is moderate for the advertised scopes and low for measured outcomes. Independent customer confirmation, raw corrections and a representative blinded evaluation would change the assessment.

Sources: [482], [483]

Evidence: Datatang / Nexdata group - offering and scope. The issuer describes crowdsourced multilingual, multidomain, multimodal collection, annotation, proofreading and QA through its data-factory SaaS. The issuer describes licensing training-data resources and offering secure data computation and technical evaluation services.

Sources: [484]

Evidence: Datatang / Nexdata group - disclosed delivery and operating model. The 2026 H1 subsidiary table lists Nexdata Technology Inc. 100% directly held, Japanese Datatang 51% directly held, and Nexdata HongKong and Singapore 100% indirectly held. The 2026 H1 issuer report identifies Datatang (Beijing) Technology Co., Ltd. and NEEQ 831428. The official Japanese company page identifies Nexdata as the brand of Datatang株式会社 and states AI training-data and customization business. The issuer reports operating revenue of RMB 180,323,485.50 for January 1-June 30, 2026.

Sources: [484], [485]

Evidence: Datatang / Nexdata group - source-specific boundaries. The 2026 H1 group ownership table supports one Datatang/Nexdata supplier group, retaining regional legal entities and the Nexdata operating-brand record. The RMB 180,323,485.50 figure is unaudited group operating revenue for January-June 2026; it is neither Nexdata segment revenue, an annualized pace nor gross customer spend. The source read covers the stated issuer pages and brand texts, not the full report. Service descriptions grant no license to any particular corpus.

Sources: [484], [485]

Analysis: Datatang / Nexdata group - buyer assessment and unknowns. A buyer should contract with the identified legal entity and distinguish licensed resources, customized human production and model-generated products. Group ownership provides a deduplication boundary without making all brands' offerings interchangeable. Confidence is moderate for the disclosed group and business scope and unestablished for independently audited quality or licensing. A versioned product manifest, accepted customization samples and operative rights/processing terms would change the assessment.

Sources: [484], [485]

Evidence: FutureBeeAI - offering and scope. FutureBeeAI offers custom collection, annotation, human-verified transcription and human-in-the-loop model evaluation. FutureBeeAI offers on-site, crowdsourced, device-specific and environment-specific collection.

Sources: [487], [488]

Evidence: FutureBeeAI - disclosed delivery and operating model. About describes NDAs, project anonymity, trained contributors and multilayer quality assurance. Its published collection workflow includes screening/training/consent, pilot collection, sample checking, client feedback, scaled collection and final validation/delivery. The dataset-seller page requires ownership or sale rights, consent for personal information and documentation; it describes compliance review before negotiated terms.

Sources: [486], [487], [489]

Evidence: FutureBeeAI - source-specific boundaries. The collection modes are offers, not named model/paper deliveries. Screening, consent, review and NDA language are stated procedures without inspected contracts, certificates, active-worker counts or measured error rates. Seller-intake rules do not establish that every asset has cleared rights and do not grant this report a dataset license.

Sources: [486], [487], [488], [489]

Analysis: FutureBeeAI - buyer assessment and unknowns. The disclosed workflow is most testable at its transition from pilot samples to scaled collection. Freeze device/environment strata, consent language, rejection rules and validation checks before requesting more volume. Confidence is moderate in the described service scope and unestablished for executed quality or rights compliance. A buyer should inspect raw sample/rework logs and a source-by-source rights schedule. Independently accepted pilot units and operative agreements would change that assessment.

Sources: [486], [487], [488], [489]

Evidence: Globose Technology Solutions - offering and scope. GTS describes labeling and production pipelines using human-in-the-loop methods. GTS offers image, video, speech and text data collection, image/video annotation and audio transcription.

Sources: [491]

Evidence: Globose Technology Solutions - disclosed delivery and operating model. GTS describes explaining collection purposes and data use, secure processing, access controls and encrypted transfer/storage. About describes GTS Dash as a global work platform with flexible tasks and a multilingual workforce. The gts.ai footer identifies Globose Technology Solutions Pvt Ltd (GTS) as an AI data-collection company.

Sources: [490]

Evidence: Globose Technology Solutions - source-specific boundaries. The current gts.ai footer supplies the company-name evidence; the historic seed domain does not. GTS Dash membership is not an employee or active-paid contributor count. Consent explanations, encryption and access controls are provider statements without independently validated execution, regulatory compliance or named delivery results.

Sources: [490], [491]

Analysis: Globose Technology Solutions - buyer assessment and unknowns. A multilingual collection brief should identify language-modality intersections and the exact contributor qualification for each. The disclosed platform and human-in-the-loop process make task routing and access logs useful inspection points. Confidence is moderate in the offered modalities and low in unmeasured workforce and quality claims. Verified qualification, permission and accepted-output samples under a signed project scope would change confidence.

Sources: [490], [491]

Evidence: HabileData - offering and scope. HabileData offers sourcing, cleansing, enrichment, annotation, validation and synthetic generation for AI training.

Sources: [493]

Evidence: HabileData - disclosed delivery and operating model. A case for an unnamed German construction-technology company describes manual annotation, validation of automated tags, client-portal delivery and a two-step quality check. Its AI-service FAQ describes annotation reviews, expert audits, automated checks, statistical sampling, inter-annotator agreement and QA loops. Its security page describes agreed retention timelines, encrypted delivery, data destruction on request and customer ownership/control. HabileData states it uses encryption, role-based access, MFA, logged access and engagement NDAs.

Sources: [492], [493], [494]

Evidence: HabileData - source-specific boundaries. The German construction-technology case is first-party and the buyer unnamed; savings and model gains are not validated. Agreement, expert-audit and sampling methods are procedures without an observed common quality statistic. Retention, customer ownership, NDAs, encryption and MFA language are not executed agreements, third-party IP transfers or inspected certificate/deployment scopes.

Sources: [492], [493], [494]

Analysis: HabileData - buyer assessment and unknowns. The case distinguishes manual annotation from validation of automated tags, which a buyer should retain as separate work units. Request the two-step review log and compare raw labels with final portal outputs. Confidence is moderate for the disclosed workflow and unestablished for savings, achieved quality or applicable rights. Independent source-tag validation, error/rework records and customer-specific retention/access agreements would change the assessment.

Sources: [492], [493], [494]

Evidence: SunTec.AI - offering and scope. SunTec.ai offers labeled image, video, text, audio and 3D-sensor data using domain-trained teams and ML-assisted tools.

Sources: [500]

Evidence: SunTec.AI - disclosed delivery and operating model. The service page describes human/automated tiered review, gold standards, inter-annotator agreement, consensus and project KPIs/SLAs. Its FAQ describes following client retention policies, data destruction, standard/custom annotation formats and metadata/quality/audit information. A SunTec India-authored 2025 release identifies SunTec.ai as its AI/ML division. A service-page case describes manual polygon annotation of historical waterbody maps with guidelines and digital restoration.

Sources: [500], [501]

Evidence: SunTec.AI - source-specific boundaries. The 2025 PRLog release is SunTec India-authored and identifies a division; SunTec India, SunTec Data and SunTec.ai are not counted as separate suppliers here. The historical-map buyer is unnamed. Stated QA, KPI/SLA and deletion procedures are not numerical agreement results, an executed SLA, a source license or inspected security certification.

Sources: [500], [501]

Analysis: SunTec.AI - buyer assessment and unknowns. The map example makes guideline interpretation and restoration choices inspectable, rather than proving a universal annotation rate. A buyer should require original/restored views, polygon conventions, disputed boundaries and blinded review by the intended domain. Confidence is moderate for the described capability and unestablished for model gains or error rates. Accepted-unit audits and the operative format/retention/SLA schedule would change that judgment.

Sources: [500], [501]

Evidence: TagX - offering and scope. TagX describes schema-mapped public data delivered as JSON, CSV, XML, API or warehouse feeds. TagX offers managed collection, cleaning, validation and delivery of public web data, including for AI training.

Sources: [503], [504]

Evidence: TagX - disclosed delivery and operating model. Its privacy/data-ethics page restricts collection to public non-personal signals and excludes authenticated, paywalled, private and sensitive personal content. The TagX About footer names Inferentia L.L.C-FZ.

Sources: [502], [503]

Evidence: TagX - source-specific boundaries. The demonstration feed is explicitly illustrative. The Inferentia L.L.C-FZ footer is a legal-identity lead without independently verified registration or control, and creates no second supplier. A public non-personal collection policy is not verified source-by-source compliance or a downstream training license. No specialized human-preference annotation or model-effectiveness claim is established.

Sources: [502], [503], [504]

Analysis: TagX - buyer assessment and unknowns. Schema-mapped feeds are a different deliverable from human preference labels. A buyer should test source coverage, provenance, change detection and exclusion enforcement before judging freshness or volume. Confidence is moderate for the stated web-data scope and low for unverified legal/control and operating outcomes. An auditable source manifest, matched refresh logs, executed contracting identity and applicable use rights would change the assessment.

Sources: [502], [503], [504]

Evidence: Trupp Global - offering and scope. TruppGlobal offers dedicated outsourced teams and describes AI/ML metadata work among its BPO services. Its official 2024 annotation article explicitly offers text-labeling and outsourced data-annotation services.

Sources: [506], [507]

Evidence: Trupp Global - disclosed delivery and operating model. A March 2026 company announcement reports a third-party VAPT of internal network infrastructure with remediation within the assessment scope. Its privacy policy describes website forms, cookies and access logs, website-data security and marketing/newsletter use.

Sources: [505], [508]

Evidence: Trupp Global - source-specific boundaries. The 2024 annotation article is a commercial offer, not completed delivery to a named model or dataset. The March 2026 VAPT statement concerns internal network infrastructure and is a company announcement without the agency/report inspected. Website form/cookie/log and marketing terms do not establish customer training-data ownership, custody or an output license.

Sources: [505], [506], [507], [508]

Analysis: Trupp Global - buyer assessment and unknowns. A dedicated outsourced team should be evaluated through task-specific supervision and acceptance, not by extending an internal-network assessment to annotation correctness. Confidence is moderate for the documented service offer and unestablished for per-project quality or security. A buyer should inspect staffing/qualification, raw labels, escalation and the customer data-processing scope. The actual VAPT scope and independently assessed project outputs would change confidence.

Sources: [505], [506], [507], [508]

Evidence: Nexdata - operating-brand offer within the Datatang group. The paid agent-trajectory listing describes OpenClaw-generated trajectories for deep search, data analysis and industry research in JSONL. Nexdata offers on-demand multimodal annotation, supervised fine-tuning data, client-rule output ranking or scoring, and red-teaming. Nexdata documents multimodal annotation tools, task distribution, audit/permission controls and cloud/on-premises deployment. The company describes data-science, linguistics and project-management expertise, including specialized programmers and doctors. Its quality page describes mutual checking, project-team sampled inspection and a separate final-inspection center.

Sources: [484], [485], [495], [496], [497], [498], [499]

Evidence: Nexdata - operating-brand offer within the Datatang group. Source boundaries. Generation is provider-described synthetic/agent output, not proof of human expert trajectories. Paid commercial-use wording is not a complete contract or unrestricted corpus license. Capability does not establish a contribution to any particular buyer model or evaluation effectiveness. Documented software functions were not executed, penetration-tested or checked against an actual customer deployment. A claimed expert annotation team is not verified employee credentials, a registered pool total or a count of people active on an assignment. First-party procedure only; percentages are inspection rates, not independently verified accuracy or annotator headcount.

Sources: [484], [485], [495], [496], [497], [498], [499]

Analysis: Nexdata - operating-brand offer within the Datatang group. The exact agent-trajectory product is model-generated, not a human expert demonstration set. A buyer should inspect generator/version, source/task permissions, verifier coverage and the executed product license before using trajectories for training. Confidence is high in that disclosed origin distinction and unestablished for general product quality. A payload sample with independently verified task outcomes and complete upstream terms would change the use assessment.

Sources: [484], [485], [495], [496], [497], [498], [499]

Evidence: Clickworker - operating-brand offer within the LXT group. The terms index distinguishes client/worker agreements and jurisdiction-specific Clickworker and LXT entities. Clickworker offers custom image, video, audio and text training datasets with human labeling, annotation and validation in cooperation with LXT. The crowd page describes registered freelance users paid a fixed fee for completed jobs, with profile filtering, general tests and project tests/training. Clickworker explicitly identifies itself as a subsidiary of LXT.ai and member of the LXT Group. A first-party case describes crowd-submitted photo/video sets for an unnamed international software buyer, checked by the quality team and transferred by API.

Sources: [73], [74], [75], [477], [478], [479], [480], [481]

Evidence: Clickworker - operating-brand offer within the LXT group. Source boundaries. Linked individual agreements were not inspected; no ownership, reuse, consent or downstream training rights grant is inferred from this index. Provider capability, not independently audited model effectiveness or a particular lab contribution. This is registered-pool and process evidence. Eight-million versus ten-million page figures are retained as a source discrepancy, not active people or employees. LXT already qualifies in the pinned supplier roster. Retain the Clickworker identity without adding an independent supplier-group count. Buyer is anonymous. Submission eligibility, actual contributors and delivered items are distinct; no current volume or face-recognition accuracy is independently validated.

Sources: [73], [74], [75], [477], [478], [479], [480], [481]

Analysis: Clickworker - operating-brand offer within the LXT group. The inspected identity bridge places Clickworker within LXT, with jurisdiction-specific contracting entities retained. Pool marketing figures have different reported totals and are neither employees nor active-paid capacity. The photo/video case is provider-reported and its buyer anonymous. Confidence is moderate in the disclosed routing/qualification workflow and unmeasured for accepted quality. A buyer should inspect tests, contributor consent/IP terms and independently reviewed deliverables; operative contracts and raw QA records would change the assessment.

Sources: [73], [74], [75], [477], [478], [479], [480], [481]

6. Dataset and benchmark landscape: provenance, versions and use limits

Method: The atlas below covers the qualifying dataset and benchmark families in the approved graph, retaining exact-release observations beneath each family. A family is an intellectual or curation identity, not a sum of files, splits, formats, subsets or every upstream component. Family roll-up supports coverage only; it does not propagate licenses, supplier roles or quality claims across versions. Construction, supervision origin, review process, task unit, access and rights are read together. A benchmark artifact can also be a dataset in ordinary usage, but this coverage calculation keeps its declared family type consistent.

Analysis: For human preference data, compare the actual judgment unit. HelpSteer3 combines task/language configurations and independent preference judgments; HelpSteer2 uses attribute ratings and a closest-rating aggregation process; HH-RLHF preserves chosen/rejected dialogue structure; WebGPT comparisons carry evidence-bearing response material and a documented label convention. A conversion to one chosen/rejected JSONL schema may be technically convenient while losing rating strength, disagreement, evidence or contributor-level independence. Preserve the source fields before transformation and document each derived training view.

Sources: [202], [451], [452], [570], [715]

Analysis: For broad text curation, the lineage unit differs from the label unit. FineWeb documents crawl lineage, filtering and per-crawl deduplication; its educational subset uses model-scored supervision. Dolma reports source-specific deduplication and residual-risk limits. Common Pile applies a distinct inclusion definition for public-domain and openly licensed sources and excludes NonCommercial and NoDerivatives material under that definition. These are contrasting curation choices, not evidence that all three corpora have the same rights or that any one eliminates every privacy or copyright risk.

Sources: [134], [382], [734]

Analysis: For task-rich and composite collections, preserve upstream identity and new contribution separately. DataComp distinguishes a candidate pool, a filtered dataset and training samples seen; FLAN distinguishes datasets, task categories and dataset/category tasks; MathVista combines existing datasets with new sources; MuSiQue’s human-composed questions constitute one family rather than a new family per seed source or answerability serialization. Counting all these units as independent datasets would distort both coverage and an apparent cost-per-example calculation.

Sources: [130], [137], [177], [273], [274]

Analysis: Benchmark protocol is part of the artifact identity. AlpacaEval changes annotator and reference models across versions and provides a length-controlled measure; MT-Bench supports different grading/comparison modes; RewardBench 1 and 2 use different scoring structures; HumanEval’s pass@k depends on sample counts. Freeze the exact task set, model outputs, judge, prompt, aggregation and budget when comparing systems. A family name and a headline score alone are insufficient for reproduction or a purchase decision.

Sources: [29], [40], [171], [449], [573]

Analysis: Access and declared license do not automatically establish full rights. GAIA asks against crawlable redistribution and preserves private test information; MathVista restricts training use while retaining underlying source rights; RewardBench requires component-license compliance; MMMU’s creator card and collection guidance address different layers of rights. An approved catalogue observation describes what the inspected source says. It is not permission from every underlying rights holder, a legal opinion, or authorization for this report to redistribute the original corpus.

Sources: [177], [184], [185], [449], [694]

Evidence: Dataset-family atlas - AgentInstruct (AgentTuning): construction, signal and protocol. AgentTuning selects six held-in tasks from AgentBench and warns current AgentBench can diverge from the old evaluation code used for paper results. AgentInstruct describes new GPT-4 interaction trajectories filtered with strict task rewards and ReAct-style thought explanations. AgentInstruct card schema stores an id and conversation entries with from, value and boolean loss fields. AgentInstruct reports 1,866 filtered trajectories across ALFWorld, WebShop, Mind2Web, knowledge graph, operating system and database tasks.

Sources: [258], [259]

Evidence: Dataset-family atlas - AgentInstruct (AgentTuning): access, version and interpretation boundaries. AgentInstruct card task table reports an overall 5.24 average turns and per-task averages from 1.00 to 13.52, differing from its broad 5-35 range prose.

Sources: [259]

Evidence: Dataset-family atlas - Align-Anything: construction, signal and protocol. Align-Anything root card describes mixed text, image, video and audio inputs and outputs. Align-Anything reports annotation by GPT-4o, Gemini-1.5-Pro and human crowds. Align-Anything responses include self-constructed methods, open/closed-source model outputs and human-generated answers. Align-Anything uses general and modality-specific fine-grained annotation constitutions. Align-Anything language feedback includes critique and refinement suggestions for responses. Align-Anything root card declares CC-BY-NC-4.0 in metadata.

Sources: [95]

Evidence: Dataset-family atlas - Align-Anything: inspected-source qualification. Reviewed scope: Partial read of the1527-line root creator card: license metadata and complete narrative lines1319-1527. The remaining root YAML schemas and eight captured modality READMEs are unread; no full-family schema or per-subset read is claimed.

Sources: [95]

Evidence: Dataset-family atlas - APPS: construction, signal and protocol. APPS creator-linked card says authors manually curated problems from sites including Codewars, AtCoder, Kattis and Codeforces. APPS card lists 10,000 problems divided into 5,000 training and 5,000 test samples. APPS records contain an English problem, Python solutions, input-output test cases, difficulty, source URL and optional starter code.

Sources: [260]

Evidence: Dataset-family atlas - APPS: access, version and interpretation boundaries. APPS card reports 195 training samples without test cases and 1,235 test samples without ground-truth solutions. The APPS repository MIT file licenses software and associated documentation.

Sources: [260], [261]

Evidence: Dataset-family atlas - Aya Dataset: construction, signal and protocol. Aya reports 138,844 original annotations and 65,270 re-annotations, totaling 204,114 prompt-completion instances. Aya Dataset contains original human-written prompts/completions and human edits of automatically generated prompts/completions, plus anonymized annotator demographics. Aya describes volunteer crowd annotations followed by quality assessment on samples collected during May-December 2023.

Sources: [160]

Evidence: Dataset-family atlas - Aya Dataset: access, version and interpretation boundaries. Aya reports lack of toxic-speech flags and relabeling capabilities, with possible offensive data, incorrect language assignments and instruction-format violations. The Aya Dataset card permits academic or commercial use under Apache 2.0.

Sources: [160]

Evidence: Dataset-family atlas - BeaverTails: construction, signal and protocol. The captured BeaverTails card defines 14 harm categories. BeaverTails is described as human-labeled question-answer data with harm categories, allowing multiple categories per QA pair. The creators describe a safer-AI research purpose and advise against using the dataset to train dialogue agents because of harmful behavior risk. The default distribution lists 330k_train, 330k_test, 30k_train and 30k_test files within round0.

Sources: [30]

Evidence: Dataset-family atlas - BeaverTails: access, version and interpretation boundaries. The card states that BeaverTails and its family are released under CC BY-NC 4.0.

Sources: [30]

Evidence: Dataset-family atlas - BeaverTails: inspected-source qualification. Category labels are scoped to this version and are not interchangeable with the 19-category PKU-SafeRLHF scheme.

Sources: [30]

Evidence: Dataset-family atlas - Capybara: construction, signal and protocol. Capybara uses Amplify-Instruct to extend single-turn seeds into multi-turn synthetic conversations. Capybara creator reports personally vetting many single-turn seed datasets. Capybara filters conversations containing specified AI boilerplate and moralizing phrases. Capybara reports minhash comparisons down to 97% similarity against listed benchmarks. Capybara describes domain-expert volunteer correctness review as a future plan. The pinned Capybara card declares Apache-2.0 in metadata.

Sources: [96]

Evidence: Dataset-family atlas - Capybara: inspected-source qualification. Reviewed scope: Pinned official LDJnr/Capybara card. The metadata pretty_name differs from its official Capybara heading, so identity follows the creator description and repository.

Sources: [96]

Evidence: Dataset-family atlas - Civil Comments / Jigsaw toxicity annotations: construction, signal and protocol. Civil Comments source documentation distinguishes archived comment text from Jigsaw-added toxicity, identity-mention and covert-offensiveness labels. Civil Comments basic seven scores are fractions of crowd annotators assigning each attribute, on a 0-1 scale. The Civil Comments HF basic-label card lists 1,804,874 train, 97,320 validation and 97,320 test examples.

Sources: [262], [263]

Evidence: Dataset-family atlas - Civil Comments / Jigsaw toxicity annotations: access, version and interpretation boundaries. Civil Comments warns parent_text can leak information because splits ignored parent links and original annotators did not see parent text. TFDS warns of a potential CivilCommentsCovert quality issue and says underlying data may change. Civil Comments documentation declares CC0 for both the released dataset and underlying comment text.

Sources: [262]

Evidence: Dataset-family atlas - Code-Feedback: construction, signal and protocol. The Code-Feedback creator card says part of its data is generated by GPT-4-0613 and GPT-3.5-turbo-0613. The OpenCodeInterpreter repository describes Code-Feedback as 68K multi-turn interactions for dynamic code refinement. CodeFeedback seed queries aggregate Magicoder-OSS-Instruct, Python ShareGPT, Magicoder-Evol-Instruct and Evol-Instruct-Code sources. CodeFeedback seed selection uses Qwen-72B-Chat complexity scores from 1-5 and retains 4 or 5.

Sources: [161], [162], [163]

Evidence: Dataset-family atlas - Code-Feedback: access, version and interpretation boundaries. The final Code-Feedback card declares Apache-2.0 and directs users to OpenAI usage policy for generated content.

Sources: [161]

Evidence: Dataset-family atlas - CodeContests: construction, signal and protocol. CodeContests draws programming problems from CodeNet, description2code and Codeforces, with correct and incorrect human solutions. CodeContests distributes ContestProblem protocol buffers in Riegeli files with training, validation and test splits. CodeContests schema defines generated tests as modifications of public/private inputs validated using known correct solutions. CodeContests marks public tests as acceptable model inputs and private tests as unavailable before submission.

Sources: [264], [265]

Evidence: Dataset-family atlas - CodeContests: access, version and interpretation boundaries. CodeContests declares Apache 2.0 for code and CC BY 4.0 for non-code materials, while acknowledging separate component terms. CodeContests warns stored correct/incorrect solutions may compile or execute differently from original contest sites because of compiler, flags or library differences.

Sources: [264]

Evidence: Dataset-family atlas - CoT Collection: construction, signal and protocol. CoT Collection reports 1.84 million chain-of-thought rationales augmented across 1,060 tasks from FLAN. CoT Collection card lists 1,837,928 training rows and trains rationale followed by RESULT and target, parsing predictions after RESULT. CoT Collection repository describes per-subset rationale augmentation scripts using OpenAI API keys and storing rationale outputs separately from merged source data. CoT Collection separates source inputs, ground-truth target answers, explanatory rationales and source-task identifiers.

Sources: [266], [267]

Evidence: Dataset-family atlas - CoT Collection: access, version and interpretation boundaries. The CoT Collection HF card metadata lists CC BY 4.0. The creator CoT Collection repository restricts use to noncommercial purposes and applies OpenAI terms to generated data.

Sources: [266], [267]

Evidence: Dataset-family atlas - Daring-Anteater: construction, signal and protocol. The majority of Daring-Anteater is generated using NVIDIA proprietary models and Mixtral-8x7B-Instruct-v0.1. Daring-Anteater also sources examples from FinQA, WikiTableQuestions and commercially-friendly Open-Platypus subsets. Daring-Anteater is used for SFT before further preference tuning in the HelpSteer2 work. Daring-Anteater exposes conversations, mask, system and dataset-source columns. Daring-Anteater licenses synthetic subsets under CC-BY-4.0 and lists multiple permissive component licenses for Open-Platypus.

Sources: [98]

Evidence: Dataset-family atlas - Daring-Anteater: inspected-source qualification. Reviewed scope: Pinned NVIDIA creator card, including source-by-source license table and schema.

Sources: [98]

Evidence: Dataset-family atlas - Databricks Dolly 15k: construction, signal and protocol. The Dolly card describes more than 15,000 records across eight instruction categories. Dolly contributors were Databricks employees instructed to write prompts and responses without generative AI. Dolly states its succinct guidelines may trade rigorous rubric compliance for high task completion. Dolly uses contributor-selected Wikipedia passages for information extraction, closed QA and summarization; no passage-selection guidance was given.

Sources: [164]

Evidence: Dataset-family atlas - Databricks Dolly 15k: access, version and interpretation boundaries. Dolly declares CC BY-SA 3.0 for the collection and credits Wikipedia editors/contributors for included material.

Sources: [164]

Evidence: Dataset-family atlas - FineWeb: construction, signal and protocol. The paper-described FineWeb release is produced from 96 Common Crawl snapshots and contains about 15 trillion GPT-2-tokenized tokens. The FineWeb datasheet states that the released dataset still includes significant numbers of documents that may be toxic or harmful. The final FineWeb pipeline uses individual per-crawl MinHash deduplication. The FineWeb-Edu classifier is trained on 410,000 model-scored pages, with 50,000 held out for validation against the synthetic scores. Llama-3-70B-Instruct assigns educational-quality scores from zero to five to 460,000 sampled webpages used to develop FineWeb-Edu filtering.

Sources: [727], [734]

Evidence: Dataset-family atlas - FineWeb: inspected-source qualification. This is the release described by the pinned 2024 paper, not a current mutable Hub snapshot. Common Crawl sourcing does not establish a commercial supplier contract.

Sources: [734]

Evidence: Dataset-family atlas - GSM8K: construction, signal and protocol. The initial thousand GSM8K problems and solutions were collected through freelance contractors on Upwork; the authors then worked with Surge AI to scale data collection. The paper describes GSM8K as math problems and natural-language solutions created by human writers, with optional GPT-3-generated seed questions available to those writers. After collection, workers re-solved GSM8K problems they had not originally written; answer disagreements were repaired or discarded, followed by another agreement check on a smaller subset. Released problems use question and answer fields; calculation annotations occur in answers and the final numeric solution follows ####. The README describes 8.5K grade-school mathematics problems created by human problem writers, segmented into 7.5K training and 1K test problems. The separate Socratic format inserts model-generated subquestions before contractor-provided solution steps; the creators say it was not used in paper experiments.

Sources: [35], [90]

Evidence: Dataset-family atlas - GSM8K: access, version and interpretation boundaries. The provided example model solutions retain occasional calculator errors because earlier implementation bugs were fixed without regenerating the samples. The pinned Grade School Math repository contains an MIT license.

Sources: [34], [35]

Evidence: Dataset-family atlas - GSM8K: inspected-source qualification. These exact collection roles do not establish a provider-specific fraction beyond the stated initial pool or a current commercial contract.

Sources: [34], [35], [90]

Evidence: Dataset-family atlas - HelpSteer2: construction, signal and protocol. Human ratings cover helpfulness, correctness, coherence, complexity and verbosity on a 0-4 scale; dedicated preference labels and justifications complement the ratings. The card describes 3-5 independent annotators, aggregation of the three closest ratings and subsequent vendor and creator quality assurance. Prompts are mostly user-contributed ShareGPT, with approximately 5% human generated by Scale AI; responses come from ten early in-house LLMs.

Sources: [451]

Evidence: Dataset-family atlas - HelpSteer2: inspected-source qualification. Agreement filtering removes some disagreement; released disagreement annotations are separately identified.

Sources: [451]

Evidence: Dataset-family atlas - HelpSteer3: construction, signal and protocol. Each HelpSteer3-Preference sample is assigned to 3 to 5 independent human annotators for preference feedback. Postprocessing retains the three most agreeing preference annotations and excludes samples when their remaining disagreement is greater than two preference-strength points. The pinned collection contains Preference, Feedback, Edit, Edit-Quality and Principle configurations. Preference contains 40,476 samples with pairwise preference scores and individual concise reasons; card specifies a 95%/5% train/validation split.

Sources: [570], [715]

Evidence: Dataset-family atlas - HelpSteer3: access, version and interpretation boundaries. HelpSteer3-Preference does not contain samples at intersections of its General, STEM, Code and Multilingual areas.

Sources: [715]

Evidence: Dataset-family atlas - HelpSteer3: inspected-source qualification. Humans choose preference direction/strength or invalidity and provide brief justifications. The paper permits internet fact-checking but prohibits LLM tools during annotation.

Sources: [570], [715]

Evidence: Dataset-family atlas - HH-RLHF: construction, signal and protocol. Preference modeling artifacts contain chosen/rejected response pairs; helpfulness includes base, rejection-sampled and online tranches while harmlessness uses base-model data. Creator intends preference data for reward-model RLHF training and distinguishes red-team transcripts from preference-model and dialogue-agent fine-tuning data. The study recruited US master-qualified MTurk workers and Upwork contractors, using writing spot-checks and lightweight vetting.

Sources: [452], [577]

Evidence: Dataset-family atlas - HH-RLHF: access, version and interpretation boundaries. The study did not select workers by direct label-agreement metrics and reports approximately 63% researcher-worker agreement retrospectively.

Sources: [577]

Evidence: Dataset-family atlas - HH-RLHF: inspected-source qualification. Historical study-specific measurement, not a general platform quality claim.

Sources: [452], [577]

Evidence: Dataset-family atlas - HotpotQA: construction, signal and protocol. HotpotQA creators ask MTurk workers to write questions requiring reasoning over introductions of two Wikipedia articles. HotpotQA fullwiki dev context replaces gold paragraphs with the baseline retriever top ten, which may omit gold evidence. HotpotQA JSON defines question, answer, supporting title/sentence IDs and paragraph contexts, with answers/support omitted from test. HotpotQA creators separately collect supporting facts so QA models can learn from explicit evidence annotations.

Sources: [268], [269]

Evidence: Dataset-family atlas - HotpotQA: access, version and interpretation boundaries. HotpotQA declares CC BY-SA 4.0 for the dataset and Apache 2.0 for code.

Sources: [269]

Evidence: Dataset-family atlas - Human Preference Dataset (HPD/HPS): construction, signal and protocol. HPD v2 reports about 798k pairwise comparisons over 430k images/107k prompts, including 645k training pairs and 153k test pairs. HPD v2 asks human annotators to rank images generated by different models for the same prompt. HPD v2 sources prompts from DiffusionDB and MSCOCO Captions and cleans DiffusionDB prompts with ChatGPT. HPD v2 test schema includes averaged image ranks and per-annotator raw rank/user_hash records; training schema has human_preference, prompt and file paths.

Sources: [270]

Evidence: Dataset-family atlas - Human Preference Dataset (HPD/HPS): access, version and interpretation boundaries. The official HPDv2 card declares Apache 2.0 for the dataset. HPS v2 documentation says image-score comparisons are meaningful only for the same prompt and v2.0/v2.1 scores cannot be directly compared.

Sources: [270], [271]

Evidence: Dataset-family atlas - ImageRewardDB: construction, signal and protocol. All ImageRewardDB images are collected from DiffusionDB and grouped by their shared prompt. ImageRewardDB describes 137k expert comparison pairs over DiffusionDB prompts and model outputs, with its own assessment, annotator-training and quality-validation pipeline. ImageRewardDB metadata rows describe one image with prompt linkage, relative rank, overall rating, image-text alignment rating and fidelity rating. ImageRewardDB provides four train scales while keeping validation and test fixed across them.

Sources: [272]

Evidence: Dataset-family atlas - ImageRewardDB: access, version and interpretation boundaries. The ImageRewardDB card declares Apache 2.0 for the dataset and separately links MIT for Python code.

Sources: [272]

Evidence: Dataset-family atlas - Intel Orca DPO Pairs: construction, signal and protocol. Intel assigns OpenOrca completions generated by GPT-4 or GPT-3.5 as chosen responses. Intel generates corresponding rejected completions automatically with Llama-2-13b-chat. Intel documents selecting 12K examples from OpenOrca for the preference dataset. Intel documents DPO as a second stage after supervised fine-tuning for Neural Chat v3-1.

Sources: [172], [173]

Evidence: Dataset-family atlas - Intel Orca DPO Pairs: access, version and interpretation boundaries. The Intel Orca DPO Pairs card declares Apache-2.0.

Sources: [173]

Evidence: Dataset-family atlas - LMSYS-Chat-1M: construction, signal and protocol. LMSYS-Chat-1M logs conversations from the Vicuna demo and Chatbot Arena between April and August 2023. LMSYS-Chat-1M reports obtaining user consent through the collection website terms of use. LMSYS-Chat-1M reports PII removal and person-name redaction that can introduce incorrect redactions. LMSYS-Chat-1M explicitly states that benchmark decontamination was not run. LMSYS-Chat-1M has a custom agreement with model-specific terms and restrictions on third-party dataset transfer.

Sources: [99]

Evidence: Dataset-family atlas - LMSYS-Chat-1M: inspected-source qualification. Reviewed scope: Pinned lmsys/lmsys-chat-1m card, with full custom agreement and explicit non-decontamination statement.

Sources: [99]

Evidence: Dataset-family atlas - Magpie: construction, signal and protocol. Magpie generates queries and responses from aligned LLM pre-query templates without seed questions. The reviewed Magpie-Pro-300K-Filtered release is generated by Llama 3 70B Instruct. Its card describes quality, instruction-reward, repetition and incomplete-instruction filtering followed by selection of long responses, and warns that first turns largely overlap with the Pro multi-turn release. At the pinned versions, the release card declares llama3, the official repository README displays a CC BY NC 4.0 data badge and MIT code badge, and the repository LICENSE declares MIT for software.

Sources: [100], [101], [706]

Analysis: Dataset-family atlas - Magpie: version, rights and buyer interpretation. The Pro-300K-Filtered card revision 523df96eb7474e97bca6f378b3baa372a4735fcc and repository commit b734a36818e4ba7ef7f0f582fc55f7e860407d26 make separate declarations about different materials. This report does not harmonize llama3, a noncommercial data badge and a software license into one governing dataset license. Confidence is high that the inspected labels differ and unestablished for the legal permission governing a buyer’s exact payload. Obtain the applicable dataset instrument and upstream conditions for the intended use; neither a code license nor a top-level badge establishes all corpus rights.

Sources: [100], [101], [706]

Evidence: Dataset-family atlas - Math-Shepherd: construction, signal and protocol. Math-Shepherd defines label as the problem and solution with automatic step labels. Math-Shepherd uses plus for a good step that can lead to the correct answer and minus for a bad step. Math-Shepherd instances have input, label and task fields, with input comprising problem plus step-by-step solution. Math-Shepherd PRM loss is computed only at the special ки step-score positions.

Sources: [176]

Evidence: Dataset-family atlas - Math-Shepherd: access, version and interpretation boundaries. The Math-Shepherd card says its internal step-wise PPO training code cannot be open sourced, while SFT/PRM/RL checkpoints are provided.

Sources: [176]

Evidence: Dataset-family atlas - MedInstruct: construction, signal and protocol. MedInstruct describes 52,000 synthetic instruction/response pairs and a separate set of 217 clinical free-form evaluation tasks. The MedInstruct creator card says GPT-3.5-turbo synthesizes a response for each instruction. MedInstruct creator card describes 167 clinician-crafted seed tasks for GPT-4 instruction generation. AlpaCare documents clinician-evaluated 1-5 medical difficulty only for MedInstruct-test.

Sources: [182], [183]

Evidence: Dataset-family atlas - MedInstruct: access, version and interpretation boundaries. The MedInstruct creator card declares CC BY-NC 4.0. The AlpaCare repository schema names text-davinci-003 for output, while its response-generation section names GPT-3.5-turbo.

Sources: [182], [183]

Evidence: Dataset-family atlas - MuSiQue: construction, signal and protocol. MuSiQue README says its questions are composed from SQuAD, T-REx, Natural Questions, MLQA and Zero Shot RE seed questions. MuSiQue uses MTurk workers to write coherent new multi-hop questions from DAGs of source single-hop questions after checking bridge-entity identity. MuSiQue downloads contain train/dev/test for answerable and full configurations plus dev/test source single-hop IDs. MuSiQue authors selected 17 workers after a 100-worker, 20-example qualification round and sampled 10-20 examples per worker across nine annotation batches. MuSiQue says its model-development raw format differs from official released data only in field names/format, not qualitative content.

Sources: [273], [274]

Evidence: Dataset-family atlas - MuSiQue: access, version and interpretation boundaries. MuSiQue warns dev/test seed single-hop questions may already appear in seed training sets and releases exclusion IDs to mitigate leakage. MuSiQue README declares CC BY 4.0 for the dataset.

Sources: [273]

Evidence: Dataset-family atlas - MuSiQue: inspected-source qualification. Creator original methods statement, pinned by exact PDF body SHA256; method-page partial read only. TACL section5 stepS7, printedp545/PDFpage7. New human wording and entity checks support distinct annotation/content construction; single-hop text/paragraphs remain inherited.

Sources: [273], [274]

Evidence: Dataset-family atlas - Nectar: construction, signal and protocol. The card describes Nectar response rankings as GPT-4-generated seven-way comparisons. The good_natured prompt label is parsed from GPT-4 classification and is described as an approximation. The ranking rubric prioritizes harmlessness and also helpfulness for prompts classified as good natured. The card reports 182,954 prompts with seven responses per prompt.

Sources: [41]

Evidence: Dataset-family atlas - Nectar: access, version and interpretation boundaries. The card combines an Apache-2.0 declaration conditioned against competing with OpenAI with a later non-commercial research-preview statement and upstream model/data terms.

Sources: [41]

Evidence: Dataset-family atlas - Nectar: inspected-source qualification. GPT-4 ranking is generated supervision; human-origin source prompts do not make these preference labels human annotations.

Sources: [41]

Evidence: Dataset-family atlas - No Robots: construction, signal and protocol. No Robots is described as 10,000 instructions and demonstrations created by skilled human annotators, mostly single-turn. No Robots exposes prompt, prompt_id, role/content messages and category. The No Robots card lists 9,500 training and 500 test examples.

Sources: [187]

Evidence: Dataset-family atlas - No Robots: access, version and interpretation boundaries. No Robots leaves the annotation-process and annotator-identity sections as More Information Needed. No Robots declares CC BY-NC 4.0.

Sources: [187]

Evidence: Dataset-family atlas - NuminaMath: construction, signal and protocol. NuminaMath-CoT collects problems and reference solutions from exam PDFs and mathematics forums before OCR, translation and CoT processing. NuminaMath uses GPT-4o to translate and realign extracted solutions into CoT format with final-answer formatting. Some NuminaMath World Olympiad PDFs were manually segmented before OCR. NuminaMath reports exact-match and embedding decontamination with exceptions for MATH training and synthetic data. NuminaMath acknowledges challenges validating synthetic problem correctness and incomplete CoT detail on hard problems. The NuminaMath-CoT card declares Apache License 2.0.

Sources: [102], [103]

Evidence: Dataset-family atlas - NuminaMath: inspected-source qualification. Reviewed scope: Full pinned CoT card; partial textual read of pinned 16-page creator report (pages1-8 and12). Remaining experiment pages, references and visuals are not claimed fully reviewed.

Sources: [102], [103]

Evidence: Dataset-family atlas - OpenAI Summarize from Feedback: construction, signal and protocol. The OpenAI repository reports 64,832 TL;DR summary comparisons plus TL;DR comparison/Likert evaluation and CNN/DM Likert evaluations. The repository reserves valid2 labels for final evaluation because valid1 posts/articles were used for model selection. Summary comparisons are JSONL records of two summaries with choice, worker, batch and split metadata. The comparison note fields contain worker-written naive interpretations before seeing the post, with possible later edits.

Sources: [195]

Evidence: Dataset-family atlas - OpenAI Summarize from Feedback: access, version and interpretation boundaries. The repository attributes the original TL;DR source dataset to Syed et al. and declares it CC BY 4.0.

Sources: [195]

Evidence: Dataset-family atlas - OpenAssistant Conversations (OASST1 reviewed release): release observations. The ready-for-export subset has 88,838 messages in 10,364 trees and excludes spam/deleted messages; the all export contains 161,443 messages in 66,497 trees. The card describes 161,443 messages in 35 languages and 461,292 quality ratings from more than 13,500 volunteers in the full April 2023 corpus. These are card-specific export and collection observations.

Sources: [572]

Analysis: Dataset-family atlas - OpenAssistant Conversations (OASST1 reviewed release): workflow separation. A separately inspected OpenAssistant project-guidelines document at commit f1e6ed9526f5817531f3ab85441a40b3671ddccb distinguishes writing assistant replies and user prompts/replies, classifying replies and ranking replies, with factual accuracy and helpfulness prioritized in classification/ranking. This is a normative document workflow without a dataset-family, collection-release or parent identity edge. It is not established as the operating protocol of the April 2023 OASST1 snapshot and does not establish annotator compliance. A buyer needing that historical provenance would require a dated release-to-protocol record and collection logs.

Sources: [572], [707]

Evidence: Dataset-family atlas - OpenCodeInstruct: construction, signal and protocol. The OpenCodeInstruct card characterizes both data collection and labeling as automated/synthetic. OpenCodeInstruct reports five million coding question-answer pairs for supervised fine-tuning. OpenCodeInstruct defines generic/algorithmic domains and self-instruct/evol-instruct generation_algorithm values. OpenCodeInstruct stores LLM response judgments, assertion-list unit tests, pass/fail execution status and fraction of tests passed.

Sources: [188]

Evidence: Dataset-family atlas - OpenCodeInstruct: access, version and interpretation boundaries. NVIDIA declares OpenCodeInstruct under CC BY 4.0 and assigns users responsibility to check fitness of each dataset license for intended use.

Sources: [188]

Evidence: Dataset-family atlas - OpenHermes: construction, signal and protocol. The listed mixture sources include Airoboros 2.2, GPT-4-only ChatBot Arena and ShareGPT, Evol Instruct, MetaMath, and SlimOrca. The dataset is documented in ShareGPT-style conversations with from and value fields, with source metadata present in many examples. The card describes OpenHermes 2.5 as a continuation of OpenHermes 1 reaching about one million primarily synthetically generated instruction and chat samples.

Sources: [42]

Evidence: Dataset-family atlas - OpenHermes: access, version and interpretation boundaries. For the CoT Alpaca GPT4 component, the card says the source page was lost.

Sources: [42]

Evidence: Dataset-family atlas - OpenHermes: inspected-source qualification. This is creator-reported component provenance; source inclusion does not establish procurement or supplier contribution.

Sources: [42]

Evidence: Dataset-family atlas - OpenMathInstruct: construction, signal and protocol. OpenMathInstruct-1 problems come from the GSM8K and MATH training subsets. OpenMathInstruct-1 solutions are generated by Mixtral using text reasoning and Python code blocks. OpenMathInstruct-1 is_correct records final-answer correctness according to a grading script. OpenMathInstruct-1 train and validation subsets jointly cover the GSM8K and MATH training sets. The NVIDIA license requires its license text and existing notices to accompany redistribution of the work.

Sources: [104], [105]

Evidence: Dataset-family atlas - OpenMathInstruct: inspected-source qualification. Reviewed scope: Pinned NVIDIA OpenMathInstruct-1 card plus full NVIDIA LICENSE. No OpenMathInstruct-2 corpus or documentation is claimed reviewed.

Sources: [104], [105]

Evidence: Dataset-family atlas - OpenOrca: construction, signal and protocol. OpenOrca questions are sourced from the FLAN Collection. OpenOrca response fields contain GPT-3.5 or GPT-4 completions. The OpenOrca card reports approximately 1M GPT-4 completions and 3.2M GPT-3.5 completions. The pinned OpenOrca card states that the data are unsplit. The pinned OpenOrca card declares MIT in dataset metadata.

Sources: [106]

Evidence: Dataset-family atlas - OpenOrca: inspected-source qualification. Reviewed scope: Pinned Open-Orca/OpenOrca creator card; it calls the dataset a partial, ongoing completion of its intended distribution.

Sources: [106]

Evidence: Dataset-family atlas - OpenR1-Math: construction, signal and protocol. OpenR1-Math card metadata lists 225,129 all, 93,733 default and 131,396 extended training examples. OpenR1-Math describes prompting R1 on 400k source problems with two solutions per problem, sometimes four, and a 16k token generation limit. OpenR1-Math generates two-to-four DeepSeek-R1 reasoning traces for NuminaMath 1.5 problems. OpenR1-Math uses Math Verify for most answer checks and Llama-3.3-70B-Instruct as judge for 12% of samples.

Sources: [275]

Evidence: Dataset-family atlas - OpenR1-Math: access, version and interpretation boundaries. OpenR1-Math says its extended configuration adds sources such as cn_k12 and performed worse after SFT than default in creator experiments. OpenR1-Math declares Apache 2.0 for the dataset.

Sources: [275]

Evidence: Dataset-family atlas - OpenThoughts: construction, signal and protocol. The 114k card metadata lists 113,957 training examples in both default and metadata configurations. OpenThoughts metadata separates problem, ground_truth_solution, deepseek_reasoning, deepseek_solution, domain, source and code-only test_cases/starter_code. OpenThoughts-114k generates reasoning traces with DeepSeek-R1 from a curated source mixture and reports correctness verification. OpenThoughts-114k lists code collections, NuminaMath-CoT, CAMEL science and RiddleSense as sources.

Sources: [189]

Evidence: Dataset-family atlas - OpenThoughts: access, version and interpretation boundaries. The OpenThoughts-114k card declares Apache-2.0.

Sources: [189]

Evidence: Dataset-family atlas - Pick-a-Pic: construction, signal and protocol. Pick-a-Pic collects real users’ prompts and preferences over two generated images, allowing a tie. Pick-a-Pic paper identifies Stable Diffusion2.1, Dreamlike Photoreal2.0 and Stable Diffusion XL variants as image generators. Pick-a-Pic reports user authentication, activity monitoring, bans and NSFW phrase filters as data-quality controls. Pick-a-Pic paper forms validation/test data from unique-user prompts and excludes their prompts from training. Pick-a-Pic creator distinguishes original v1 data from v2 with more than one million examples. PickScore repository MIT terms apply to software and associated documentation.

Sources: [107], [108], [109]

Evidence: Dataset-family atlas - Pick-a-Pic: inspected-source qualification. Reviewed scope: Full pinned creator family card, PickScore README and code LICENSE; partial textual read of NeurIPS2023paper pages1-6and10. Dataset v1 API endpoint returned401; v1/v2 data access is not established.

Sources: [107], [108], [109]

Evidence: Dataset-family atlas - PKU-SafeRLHF: construction, signal and protocol. The card defines 19 harm categories and labels each question-answer pair with one or more categories. The card describes 83.4K preference entries with two responses per question, safety meta-labels, and separate human-preference dimensions of helpfulness and harmlessness. Helpfulness and harmlessness are ranked separately, with harmless responses always ranked above harmful responses in the harmlessness ranking. The documented responses come from Alpaca-7B, Alpaca2-7B and Alpaca3-8B; the latter two are described as Llama2-7B and Llama3-8B fine-tuned with Alpaca 52K.

Sources: [45]

Evidence: Dataset-family atlas - PKU-SafeRLHF: access, version and interpretation boundaries. The current card declares CC-BY-NC-4.0.

Sources: [45]

Evidence: Dataset-family atlas - PKU-SafeRLHF: inspected-source qualification. The 19-category current scheme must not be silently substituted for BeaverTails 14-category scheme.

Sources: [45]

Evidence: Dataset-family atlas - PRISM Alignment Dataset: construction, signal and protocol. PRISM participants complete a demographic/preference survey, then write prompts, rate LLM responses and give fine-grained conversational feedback. PRISM terms prohibit re-identification and warn that raw conversations can be unsafe or offensive. PRISM randomly selects four of 21 backend LLMs for the opening turn and asks participants to score responses from 1 to 100. PRISM utterance JSONL is a long-format presentation of the same conversations and repeats user inputs across candidate responses. PRISM reports 8,011 conversation trees and 68,371 rated human-prompt/model-response/score utterances.

Sources: [276]

Evidence: Dataset-family atlas - PRISM Alignment Dataset: access, version and interpretation boundaries. PRISM licenses human-written text under CC BY 4.0 and model responses under CC BY-NC 4.0, with original model-provider terms also applying.

Sources: [276]

Evidence: Dataset-family atlas - PRM800K: construction, signal and protocol. The paper describes the full unfiltered PRM800K release as 1,085,590 step-level labels over 101,599 solutions; the filtered training data contains about 800,000 labels over 75,000 solutions. Human labelers provide positive, negative or neutral judgments on steps in model-generated MATH solutions. The repository documents a nonstandard MATH split adding 4,500 original test problems to training and evaluating on the remaining 500 held-out problems. PRM800K is documented as 800,000 step-level correctness labels on model-generated solutions to MATH problems. The README describes human step ratings of -1, 0 or +1 and permits a human-written replacement step in phase 1 when all completions were rated -1. The documented annotation schema reserves common gold-labeled solutions for quality control and 30 initial screening questions for base task performance.

Sources: [44], [78]

Evidence: Dataset-family atlas - PRM800K: access, version and interpretation boundaries. The pinned PRM800K repository contains an MIT license.

Sources: [43]

Evidence: Dataset-family atlas - PRM800K: inspected-source qualification. These are paper-described raw versus filtered populations, not an independently counted downloaded release.

Sources: [43], [44], [78]

Evidence: Dataset-family atlas - ProsocialDialog: construction, signal and protocol. ProsocialDialog reports 58K dialogues, 331K utterances, 160K unique rules of thumb and 497K safety labels with free-form rationales. ProsocialDialog creator repo says HF dataframe and direct list-of-dialogues downloads differ only in format and contain the same content. ProsocialDialog creates new dialogues by having GPT-3 generate unsafe utterances and crowdworkers write prosocial responses. ProsocialDialog schema keeps final safety verdicts, three workers raw safety annotations and each workers free-form reasons. ProsocialDialog source field identifies socialchemistry, SBIC, ethics_amt and ethics_reddit seed text origins.

Sources: [277], [278]

Evidence: Dataset-family atlas - ProsocialDialog: access, version and interpretation boundaries. ProsocialDialog card declares CC BY 4.0 for the dataset.

Sources: [277]

Evidence: Dataset-family atlas - RichHF-18K: construction, signal and protocol. RichHF-18K reports 17,760 TensorFlow Examples: 15,810 training, 995 development and 955 test. RichHF-18K annotates generated images from Pick-a-Pic v1 and ships filenames rather than original image files. RichHF-18K releases newly collected subjective scores, human-labeled artifact heatmaps and prompt misalignment-token labels. RichHF examples contain filename, four scalar scores, artifact_map, misalignment_map and prompt_misalignment_label. RichHF defines all scores so higher is better; a higher artifact score means fewer artifacts.

Sources: [279]

Evidence: Dataset-family atlas - RLAIF-V: construction, signal and protocol. RLAIF-V metadata identifies a labeling model providing feedback for chosen and rejected answer pairs. RLAIF-V instructions draw from MSCOCO, ShareGPT-4V, MovieNet, Google Landmark v2, VQA v2, OKVQA and TextVQA. RLAIF-V also uses RLHF-V image-description prompts for long-form captioning. RLAIF-V card reports 83,132 preference pairs. The RLAIF-V card declares CC-BY-NC-4.0.

Sources: [110]

Evidence: Dataset-family atlas - RLAIF-V: inspected-source qualification. Reviewed scope: Pinned openbmb/RLAIF-V-Dataset card, including2024release and2025usage updates. Linked paper and source datasets were not read.

Sources: [110]

Evidence: Dataset-family atlas - RLHF-V: construction, signal and protocol. RLHF-V collects fine-grained segment-level human corrections for visual description and question-answering instructions. The RLHF-V card reports 5,733 preference pairs. RLHF-V January2024 expansion adds outputs from stronger models and image domains including WikiArt, landmarks and scene text. RLHF-V text data contains question, chosen and rejected entries. The RLHF-V card declares CC-BY-NC-4.0.

Sources: [111]

Evidence: Dataset-family atlas - RLHF-V: inspected-source qualification. Reviewed scope: Pinned openbmb/RLHF-V-Dataset creator card and its update history, including January2024 expanded correction release.

Sources: [111]

Evidence: Dataset-family atlas - Self-Instruct: construction, signal and protocol. The filtered GPT-3 davinci Self-Instruct corpus in this paper contains 52,445 instructions and 82,439 input/output instances. Self-Instruct starts with 175 author-written seed tasks and prompts the language model to generate further instructions and input/output instances. A single expert coauthor evaluates one randomly selected instance for each of 200 randomly sampled instructions; Table2 reports 54% with all fields valid.

Sources: [83]

Evidence: Dataset-family atlas - Self-Instruct: inspected-source qualification. These are distinct units and a paper-described corpus, not an independently recounted download.

Sources: [83]

Evidence: Dataset-family atlas - Social Bias Inference Corpus (SBIC): construction, signal and protocol. SBIC reports 82.4% pairwise agreement and average Krippendorff alpha of 0.45. SBIC adds structured categorical and free-text bias-implication annotations to social-media posts, with over 150k annotations and over 34k implications in its summary. SBIC splits posts together with their three annotation sets to keep a post from appearing across train/validation/test. SBIC card describes three MTurk annotations per post, with conditional target-group questions and two-to-four written stereotypes per selected group.

Sources: [280]

Evidence: Dataset-family atlas - Social Bias Inference Corpus (SBIC): access, version and interpretation boundaries. SBIC warns that predominantly White-aligned English and dialect/identity labeling bias should constrain deployment. The SBIC card declares Creative Commons Attribution 4.0.

Sources: [280]

Evidence: Dataset-family atlas - Stanford Alpaca: construction, signal and protocol. The Alpaca datasheet reports 52,002 instances and defines instruction, optional input, and text-davinci-003 output fields. Stanford Alpaca modifies Self-Instruct by using text-davinci-003 and generating one instance per instruction. The Alpaca datasheet warns that some generated instructions may be nonsensical and admit no good response. The Alpaca datasheet says all 52K examples were used for its models and recommends no dataset split.

Sources: [193], [194]

Evidence: Dataset-family atlas - Stanford Alpaca: access, version and interpretation boundaries. The same-revision Alpaca datasheet instead describes distribution under ODC-By. The pinned Alpaca README declares CC BY-NC 4.0 and research-only use for the dataset and models trained on it.

Sources: [193], [194]

Evidence: Dataset-family atlas - Stanford Human Preferences (SHP): construction, signal and protocol. SHP uses naturally occurring human-written Reddit posts/comments and collectively inferred preference labels rather than separately elicited individual pairwise ratings. SHP splits by post ID and includes 385,563 preference pairs across 18 subreddits, with 348,718 train, 18,436 validation and 18,409 test pairs in the card table.

Sources: [453]

Evidence: Dataset-family atlas - Stanford Human Preferences (SHP): access, version and interpretation boundaries. The card documents Reddit scraping terms and user ownership but does not grant a standard permissive dataset license.

Sources: [453]

Evidence: Dataset-family atlas - Stanford Human Preferences (SHP): inspected-source qualification. Do not assert the disputed timestamp-ordering direction.

Sources: [453]

Evidence: Dataset-family atlas - Super-NaturalInstructions: construction, signal and protocol. Natural Instructions documents a community effort to create tasks and natural-language definitions/instructions. The contribution instructions require JSON formatting tests and a maximum of 6.5K instances per task. Natural Instructions task JSON includes contributor/source fields, definitions, positive/negative examples and input-output instances. The official benchmark split trains on a subset of tasks and evaluates remaining unseen tasks for cross-task generalization.

Sources: [196]

Evidence: Dataset-family atlas - Super-NaturalInstructions: access, version and interpretation boundaries. Natural Instructions releases data other than task instances under Apache-2.0; instances retain original-dataset licenses in Instance License fields.

Sources: [196]

Evidence: Dataset-family atlas - TabMWP: construction, signal and protocol. TabMWP creators describe a new dataset containing 38,431 tabular math word problems. Each TabMWP question has a table represented as an image, semi-structured text and a structured table. TabMWP provides free-text/multiple-choice questions and gold solutions intended to reveal multistep reasoning.

Sources: [281]

Evidence: Dataset-family atlas - TabMWP: access, version and interpretation boundaries. TabMWP project documentation distinguishes 38,431 problems from 28,876 different questions and 35,442 different solutions. The TabMWP README explicitly licenses the dataset under CC BY-NC-SA 4.0.

Sources: [281], [282]

Evidence: Dataset-family atlas - TACO: construction, signal and protocol. TACO credits APPS and CodeContests for some source material while also describing manual curation from additional programming sites. TACO reports 26,443 problems with 25,443 training and 1,000 test samples. TACO schema describes tags as manually annotated algorithms needed to solve the problem and skill_types as mapped programming skills. TACO card says some test-set test cases were removed or modified and asks users to update to the newest version.

Sources: [283]

Evidence: Dataset-family atlas - TACO: access, version and interpretation boundaries. TACO declares Apache 2.0 for authored dataset contributions while acknowledging MIT/CC BY 4.0 material and explicitly unknown HackerRank legal rights/data license.

Sources: [283]

Evidence: Dataset-family atlas - ToxicChat: construction, signal and protocol. ToxicChat adds toxicity annotations to about 10K user prompts from the Vicuna demo through a human-AI collaborative annotation framework. ToxicChat 0124 reports 10,165 user prompts and 5,654 human annotations. For ToxicChat 0124, creator-reported error-analysis fixes were each checked by two annotators, and 20 additional examples received human annotation. ToxicChat schema separately stores user_input, model_output, human_annotation, toxicity, jailbreaking and OpenAI moderation output.

Sources: [284]

Evidence: Dataset-family atlas - ToxicChat: access, version and interpretation boundaries. ToxicChat warns that some personal information may remain despite creator removal attempts and prohibits identifying individuals. ToxicChat licenses both user prompts and model outputs under CC BY-NC 4.0.

Sources: [284]

Evidence: Dataset-family atlas - Tülu 3 SFT Mixture: construction, signal and protocol. The card prose reports 939,344 samples, while its dataset_info train metadata declares 939,343 examples. The card lists crowdsourced, expert-generated and machine-generated annotation creators and multiple source datasets. The card states that the Tülu 3 SFT mixture was used to train the Tülu 3 model series. Each documented mixture example contains an id, messages with user prompts and assistant responses, and a source dataset field.

Sources: [46]

Evidence: Dataset-family atlas - Tülu 3 SFT Mixture: access, version and interpretation boundaries. The collection declares ODC-BY-1.0, with different subset licenses, some non-commercial portions, and separate terms for third-party model outputs.

Sources: [46]

Evidence: Dataset-family atlas - Tülu 3 SFT Mixture: inspected-source qualification. The collection declaration must not be flattened into permission for every component; the card presents a research artifact.

Sources: [46]

Evidence: Dataset-family atlas - UltraChat: construction, signal and protocol. UltraChat uses separate ChatGPT Turbo APIs for generated user queries and assistant responses. UltraChat says Internet-available data are not directly used as prompts. UltraChat reports post-processing and filtering of generated dialogues. UltraChat covers world questions, writing and creation, and assistance on existing materials. The pinned UltraChat card declares MIT in dataset metadata.

Sources: [112]

Evidence: Dataset-family atlas - UltraChat: inspected-source qualification. Reviewed scope: Original creator card at stingning/ultrachat; multi-round conversations. No corpus files or linked repository implementation were read.

Sources: [112]

Evidence: Dataset-family atlas - UltraFeedback: construction, signal and protocol. GPT-4 annotates completions for instruction-following, truthfulness, honesty and helpfulness, providing numerical and textual feedback. The creator samples instructions from UltraChat, ShareGPT, Evol-Instruct, TruthfulQA, FalseQA and FLAN, then samples four models from a 17-model pool per instruction.

Sources: [454]

Evidence: Dataset-family atlas - UltraFeedback: access, version and interpretation boundaries. The repository documents a December 2023 correction of faulty overall-score values using fine-grained-score thresholds and re-annotation for remaining cases.

Sources: [580]

Evidence: Dataset-family atlas - UltraFeedback: inspected-source qualification. README-level evidence only; named implementation path returned 404 at this commit.

Sources: [454], [580]

Evidence: Dataset-family atlas - UltraInteract: construction, signal and protocol. The pinned pairwise distribution declares 219,522 training examples with trajectory, chosen, rejected, id and parent_id fields. The card reports approximately 86K instructions, 286K correct answers and 219K preference pairs. Correct actions and correct-ending trajectories are described as SFT material; paired correct and incorrect nodes or trajectories support preference learning. UltraInteract is documented as reasoning preference trees combining reasoning chains, environment interaction trajectories, critique and pairwise data.

Sources: [47]

Evidence: Dataset-family atlas - UltraInteract: access, version and interpretation boundaries. The pairwise card declares an MIT license.

Sources: [47]

Evidence: Dataset-family atlas - UltraInteract: inspected-source qualification. The declaration is scoped to this distribution; linked source-dataset licenses were not exhaustively reviewed.

Sources: [47]

Evidence: Dataset-family atlas - UltraMedical: construction, signal and protocol. UltraMedical combines adapted public sources with customized synthetic MedQA-Evol, TextBookQA and WikiInstruct data. UltraMedical card metadata lists 409,593 training examples, while narrative rounds the collection to 410,000 samples. UltraMedical documents manual or GPT-4 synthetic questions, GPT-4 responses, original answer keys and GPT-3.5-turbo instruction scores.

Sources: [201]

Evidence: Dataset-family atlas - UltraMedical: access, version and interpretation boundaries. UltraMedical authors acknowledge hallucinations/GPT-4 bias and describe verifying factual accuracy as future research. The UltraMedical card declares MIT.

Sources: [200], [201]

Evidence: Dataset-family atlas - VideoFeedback: construction, signal and protocol. VideoFeedback includes text-video pairs from eleven video-generative models plus real-world video augmentation. VideoFeedback creator describes human-provided multiaspect video scores. VideoFeedback ratings cover visual quality, temporal consistency, dynamic degree, text-video alignment and factual consistency on a 1-4 scale. VideoFeedback is used to train the VideoScore evaluator. The VideoFeedback dataset card declares Apache-2.0.

Sources: [113], [114], [676]

Evidence: Dataset-family atlas - VideoFeedback: inspected-source qualification. Reviewed scope: Pinned TIGER-Lab/VideoFeedback card and pinned official VideoScore README/LICENSE. Corrected the initial mistaken PKU-Alignment repository path; no source identity substitution in final claims.

Sources: [113], [114], [676]

Evidence: Dataset-family atlas - VLFeedback: construction, signal and protocol. VLFeedback preference annotations are produced by GPT-4V. VLFeedback samples have four responses selected from a pool of twelve LVLMs. VLFeedback annotations cover helpfulness, visual faithfulness and ethical considerations. VLFeedback card schema stores ratings and rationales under each completion annotation. The VLFeedback creator uses its dataset for DPO on a Qwen-VL-Chat backbone to obtain Silkie.

Sources: [115], [116]

Evidence: Dataset-family atlas - VLFeedback: inspected-source qualification. Reviewed scope: Pinned MMInstruction/VLFeedback card and pinned vlf-silkie/VLFeedback README. Embedded workflow/source diagrams and linked paper were not inspected.

Sources: [115], [116]

Evidence: Dataset-family atlas - WebGPT Comparisons: construction, signal and protocol. WebGPT quotes fields store model-found browsing extracts and source-page titles, while answers are composed using those quotes. WebGPT examples pair model answers to a question with human preference scores. WebGPT score_0 ranges from -1 to 1, sums to zero with score_1, uses zero as a soft 50% label and other scores by sign as hard labels. The WebGPT card describes all comparisons marked suitable for reward modeling by project end, totaling 19,578 comparisons. WebGPT token fields use GPT-2 tokenization of a truncated question-plus-quotes prefix and final answer completion for reward modeling.

Sources: [202]

Evidence: Dataset-family atlas - WildChat: construction, signal and protocol. WildChat collects human user conversations by offering free GPT-3.5 and GPT-4 access. The reviewed WildChat-1M variant removes conversations flagged toxic by OpenAI Moderation or Detoxify. WildChat de-identifies data using Microsoft Presidio and author-written rules. WildChat uses ODC-BY, whose license does not cover independent rights in individual database contents. WildChat documents empty user inputs that can cause hallucinated assistant responses.

Sources: [117], [118]

Evidence: Dataset-family atlas - WildChat: inspected-source qualification. Reviewed scope: Pinned allenai/WildChat-1M card and full LICENSE.md. Version changes include July 2024 toxic-content removal and October 2024 PII-related removals.

Sources: [117], [118]

Evidence: Dataset-family atlas - WildJailbreak: construction, signal and protocol. WildJailbreak transforms vanilla harmful prompts with 2-7 mined tactics using Mixtral-8×7B/GPT-4, filters low-risk/off-topic results and reuses counterpart vanilla refusals. WildJailbreak separately documents 210 adversarial benign and 2,000 adversarial harmful evaluation examples. WildJailbreak training schema exposes vanilla, adversarial, tactics, completion and four-way data_type fields. WildJailbreak describes 262K synthetic vanilla/adversarial prompt-response pairs with both harmful and superficially harmful-looking benign queries. WildJailbreak uses GPT-4 for vanilla harmful/benign prompts and GPT-3.5 for refusal or complying responses. The 262K figure is rounded card prose; the metadata category 100M<n<1B conflicts and is not used as an observed count. Only the pinned public documentation was inspected, without gated corpus bytes or established payload release identity.

Sources: [455], [705]

Evidence: Dataset-family atlas - WildJailbreak: access, version and interpretation boundaries. WildJailbreak public card declares ODC-BY and conditions dataset access on AI2 Responsible Use acceptance and completion of identity/intended-use fields. The license declaration and identity/intended-use access conditions are distinct; this synthesis accepted no gate and acquired no gated files.

Sources: [705]

Evidence: Benchmark-family atlas - AgentBench: construction, signal and protocol. For the Mind2Web browsing component, the reported primary metric is step success from element/action matching rather than whole-task success; this paper uses top-10 candidates and CoT few-shot prompts. The initial AgentBench paper defines eight environment families: operating systems, databases, knowledge graphs, digital card games, lateral-thinking puzzles, house-holding, web shopping and web browsing. For OS tasks, about half of instructions are human-created or collected; the other half are mainly GPT-4-generated QA tasks filtered by execution-based checks. Eight programming annotators select and construct human tasks.

Sources: [48]

Evidence: Benchmark-family atlas - AgentBench: inspected-source qualification. The setup differs from the antecedent top-50 Mind2Web setting.

Sources: [48]

Evidence: Benchmark-family atlas - AlpacaEval: construction, signal and protocol. The release documents about 20K human preference annotations, including 2.5K cross-annotations from four annotators over roughly 650 examples. AlpacaEval 2.0 uses weighted_alpaca_eval_gpt4_turbo as annotator with gpt4_turbo reference outputs; the original uses alpaca_eval_gpt4 with text_davinci_003. Length-controlled win rates fit a preference model including output-length difference and average counterfactual preferences at zero length difference.

Sources: [29]

Evidence: Benchmark-family atlas - AlpacaEval: access, version and interpretation boundaries. The README declares Apache-2.0 for code and CC BY-NC 4.0 for data. The creators caution that the instruction set may miss advanced usage, automatic evaluators can favor style over factuality, and AlpacaEval does not measure model harms.

Sources: [29]

Evidence: Benchmark-family atlas - AlpacaEval: inspected-source qualification. The code/data declarations are separate; this is not an audit of every upstream output provider term.

Sources: [29]

Evidence: Benchmark-family atlas - GAIA: construction, signal and protocol. The October 2025 GAIA format update replaces former JSONL-backed loading with Parquet splits and per-level views. GAIA describes more than 450 questions divided into three levels, with public validation/dev and tests having private answers and metadata. GAIA preserves task_id, Question, Level, Final answer, file_name, file_path and Annotator Metadata columns, with relative attachment paths. GAIA is documented as evaluating LLMs augmented with tools, prompting and search access. The pinned commit identifies documentation, not a verified corpus-byte release. Private answer values and payload row counts were not accessed or established.

Sources: [447], [694]

Evidence: Benchmark-family atlas - GAIA: access, version and interpretation boundaries. GAIA public metadata asks users not to reshare the dataset outside gated/private HF repositories and the card asks against crawlable validation/test redistribution. These are separately documented restrictions, not an unrestricted redistribution license or evidence that the researcher accepted access terms.

Sources: [694]

Evidence: Benchmark-family atlas - GPQA: construction, signal and protocol. The creator baseline code instructions distinguish closed-book prompting from Bing-based retrieval modes and expose a seed for shuffling answer choices. GPQA questions are written by domain experts in biology, physics and chemistry, with experts defined as people having or pursuing a PhD in the field. The original study describes 546 Extended questions, 448 main questions and 198 Diamond questions, with the latter two curated subsets. The collection pipeline uses first expert validation, writer revision, second expert validation and three non-expert validators from other domains with web access but no LM assistants.

Sources: [32], [33]

Evidence: Benchmark-family atlas - GPQA: access, version and interpretation boundaries. The pinned GPQA creator repository contains an MIT license.

Sources: [31]

Evidence: Benchmark-family atlas - GPQA: inspected-source qualification. Prompt type, subset, retrieval access and answer-order seed must be matched for comparison; no evaluation was run.

Sources: [31], [32], [33]

Evidence: Benchmark-family atlas - HarmBench: construction, signal and protocol. HarmBench evaluates whether each completion instantiates its corresponding behavior using separate text and multimodal classifier models. The HarmBench protocol uses max_new_tokens=512 for target completions and warns other limits can hinder comparisons. HarmBench accepts a behavior-dataset CSV and saves test cases under method/experiment directories with optional behavior-ID selection. HarmBench evaluates attack success rate through test-case generation, target-model completions and completion evaluation.

Sources: [167]

Evidence: Benchmark-family atlas - HarmBench: access, version and interpretation boundaries. The HarmBench repository contains an MIT license.

Sources: [168]

Evidence: Benchmark-family atlas - Helpful, Honest & Harmless alignment evaluation: construction, signal and protocol. HHH examples were written by contributors and each evaluation received discussion and approval from at least one other co-author. The pinned task evaluates binary response comparisons in helpfulness, honesty, harmlessness and other categories using multiple_choice_grade.

Sources: [448], [576]

Evidence: Benchmark-family atlas - Helpful, Honest & Harmless alignment evaluation: inspected-source qualification. Approval denotes the historical benchmark authors process, not independent ScienceToStartup review.

Sources: [448], [576]

Evidence: Benchmark-family atlas - HumanEval: construction, signal and protocol. HumanEval computes per-problem pass@k estimates from generated-sample counts and correct counts, then averages only when each problem has at least k samples. HumanEval predictions are JSONL records with task_id and completion-only text excluding the prompt. The pinned HumanEval evaluation function defaults to k=[1,10,100], four workers and a three-second correctness timeout. The OpenAI repository identifies HumanEval as a hand-written evaluation set with a harness for problem solving.

Sources: [170], [171]

Evidence: Benchmark-family atlas - HumanEval: access, version and interpretation boundaries. The pinned HumanEval repository contains an MIT license.

Sources: [169]

Evidence: Benchmark-family atlas - Humanity’s Last Exam: construction, signal and protocol. The documented simple evaluation defaults to temperature zero and advises at least 8,192 completion tokens for reasoning models. The pinned evaluator defaults to o3-mini-2025-01-31 and asks the judge to extract the final answer and compare it with the provided correct answer. The creator README describes HLE as a multimodal benchmark with 2,500 expert-developed multiple-choice and short-answer questions across dozens of subjects.

Sources: [36], [38]

Evidence: Benchmark-family atlas - Humanity’s Last Exam: access, version and interpretation boundaries. The pinned HLE repository contains an MIT license. The README overview describes 2,500 questions, while its sample evaluation output displays n = 2700.

Sources: [37], [38]

Evidence: Benchmark-family atlas - Humanity’s Last Exam: inspected-source qualification. This code/repository declaration does not establish the license of unretrieved HLE question content.

Sources: [36], [37], [38]

Evidence: Benchmark-family atlas - JailbreakBench: construction, signal and protocol. JailbreakBench provides a Llama3 70B jailbreak judge and a Llama3 8B refusal judge. JailbreakBench curates original behaviors and behaviors from TDC/HarmBench and AdvBench against a usage-policy taxonomy. JailbreakBench describes 300 judge-comparison examples with three human judge labels, majority labels and model-judge labels. JailbreakBench submissions record attack success rate and hyperparameters such as target, attacker, judge and token budgets. JailbreakBench documents 100 harmful behaviors and 100 thematically similar benign behaviors.

Sources: [174]

Evidence: Benchmark-family atlas - JailbreakBench: access, version and interpretation boundaries. The pinned JBB-Behaviors card says the dataset, like the code, is released under MIT.

Sources: [175]

Evidence: Benchmark-family atlas - MathVista: construction, signal and protocol. The official MathVista pipeline generates responses, extracts short answers using an LLM such as GPT-4, then calculates scores. MathVista reports 6,141 examples from 28 existing multimodal datasets plus newly created IQTest, FunctionQA and PaperQA. The pinned MathVista scorer normalizes extracted answers by answer type, compares to references, and computes correct/total accuracy. The MathVista card documents testmini with 1,000 development/validation examples and test with 5,141 standard-evaluation examples whose answers are not publicly released.

Sources: [177], [178], [179]

Evidence: Benchmark-family atlas - MathVista: access, version and interpretation boundaries. MathVista warns the third-party EvalScope workflow can differ in prompts, extraction and normalization and lacks a same-prediction parity report. MathVista declares CC BY-SA 4.0 for new contributions while original authors retain image/question copyright; it permits commercial test use and prohibits training use.

Sources: [177]

Evidence: Benchmark-family atlas - MBPP: construction, signal and protocol. MBPP describes around 1,000 crowdsourced entry-level Python problems, each with task description, solution and three automated tests. The MBPP study used three-shot prompts from task IDs 2, 3 and 4, with tests in the prompt and BEGIN/DONE solution delimiters. MBPP says a subset was hand verified by the creators and distributed as sanitized-mbpp.json. MBPP assigns task IDs 11-510 to testing, 1-10 to prompting, 511-600 to validation and 601-974 to training.

Sources: [180]

Evidence: Benchmark-family atlas - MBPP: access, version and interpretation boundaries. The pinned google-research monorepo LICENSE is Apache License 2.0.

Sources: [181]

Evidence: Benchmark-family atlas - MMLU: construction, signal and protocol. Graduate and undergraduate students manually collected the benchmark’s questions from freely available online sources. The paper designs MMLU as a text-only academic and professional question benchmark. The pinned HF distribution describes 57 multiple-choice subjects with a five-example development set per subject.

Sources: [571], [714], [722]

Evidence: Benchmark-family atlas - MMLU: inspected-source qualification. Sources include examination practice questions and educational material; the paper does not identify a commercial annotation supplier for this collection.

Sources: [571], [714]

Evidence: Benchmark-family atlas - MMMU: construction, signal and protocol. MMMU reports 11.5K multimodal questions collected from college exams, quizzes and textbooks across 30 subjects and 183 subfields. MMMU permits either final-answer files with user parsing or creator response parsing, and warns results can differ due to random selection for failed responses. MMMU documents 150 development samples for few-shot learning, 900 validation samples for model/debug tuning and 10,500 test samples. At the pinned 2026 repository revision, MMMU says test answers/explanations are released for local evaluation with answer_dict_test.json.

Sources: [184], [186]

Evidence: Benchmark-family atlas - MMMU: access, version and interpretation boundaries. MMMU says annotator guidelines required obeying source copyright/license rules and avoiding sites forbidding copying/redistribution, with removal after verified complaints. The MMMU creator HF card declares Apache-2.0.

Sources: [184], [185]

Evidence: Benchmark-family atlas - MT-Bench: construction, signal and protocol. The creators report releasing 3.3K human annotations for six models answering 80 MT-Bench questions. FastChat also supports pairwise comparisons against a baseline and across all model pairs. The recommended default is GPT-4 single-answer grading on a ten-point scale per turn, averaged over all turns. MT-Bench is documented as challenging multi-turn open-ended chat-assistant questions evaluated using an LLM judge.

Sources: [40]

Evidence: Benchmark-family atlas - MT-Bench: access, version and interpretation boundaries. The pinned FastChat repository LICENSE is Apache License 2.0.

Sources: [39]

Evidence: Benchmark-family atlas - MT-Bench: inspected-source qualification. This records the repository license; separate dataset-specific or upstream-content terms were not established.

Sources: [39], [40]

Evidence: Benchmark-family atlas - OSWorld: construction, signal and protocol. OSWorld evaluate() fetches result/expected state and invokes configured metric functions; multiple metrics use configured and/or combination. OSWorld warns missing Google-account/OAuth or proxy configuration can cause task execution failures and lower scores. OSWorld screenshot-baseline examples specify screenshot observations, a 15-step limit and a three-second post-action wait. OSWorld loads task id, instruction, setup config and an evaluator defining metric functions and result/expected getters.

Sources: [190], [191]

Evidence: Benchmark-family atlas - OSWorld: access, version and interpretation boundaries. The OSWorld repository declares Apache License 2.0. OSWorld announces OSWorld-Verified fixes and asks comparisons against new results when using the latest version.

Sources: [191], [192]

Evidence: Benchmark-family atlas - RewardBench (additional family): construction, signal and protocol. V1 succeeds when a chosen response receives a higher score than its rejected alternative; section aggregation weights prompts and balances math against code. V2 averages six categories; five compare one chosen response against three rejected responses while Ties uses a distinct correctness-and-margin metric.

Sources: [449], [573]

Evidence: Benchmark-family atlas - RewardBench (additional family): access, version and interpretation boundaries. RewardBench declares ODC-BY aggregation while requiring users to follow the licenses of its components, including listed noncommercial sources.

Sources: [449]

Evidence: Benchmark-family atlas - RewardBench (additional family): inspected-source qualification. Not comparable to v1 pairwise scores without a matched protocol.

Sources: [449], [573]

Evidence: Benchmark-family atlas - SWE-bench: construction, signal and protocol. Candidate task instances must include at least one test changing from fail before the reference solution to pass after it. The original full SWE-bench evaluation set contains 2,294 issue/PR-derived task instances from 12 Python repositories. The benchmark’s task instances in this paper are all in Python. A submitted patch counts as solved only when all reference FAIL TO PASS and PASS TO PASS tests pass. The original dataset tests repository issue resolution with 2,294 test issue/PR pairs from twelve Python repositories and uses test verification against post-PR behavior. The repository describes SWE-bench Verified as 500 engineer-confirmed solvable problems in collaboration with OpenAI Preparedness.

Sources: [574], [578], [726], [733]

Evidence: Benchmark-family atlas - SWE-bench: inspected-source qualification. The pipeline also removes installation/runtime failures; tests come from the corresponding merged PR. This is execution-based benchmark filtering, not fresh human preference labels.

Sources: [574], [578], [733]

Evidence: Benchmark-family atlas - TruthfulQA: construction, signal and protocol. TruthfulQA's 817 benchmark questions were written by the paper's authors. TruthfulQA combines 437 questions filtered against GPT-3-175B with 380 additional questions not tested on that target model. The creator repository describes a January 2025 binary multiple-choice setting using Best Answer and Best Incorrect Answer with randomized A/B order. The pinned HF distribution has generation and multiple_choice configurations of the same 817 questions. Generation evaluates truthfulness and informativeness separately; original MC1 accuracy and MC2 probability-mass metrics measure different behaviors.

Sources: [575], [579], [675], [723]

Evidence: Benchmark-family atlas - TruthfulQA: access, version and interpretation boundaries. TruthfulQA questions were authored rather than collected from a deployed system. Creator documentation says some questions were removed or revised and full validation of the remaining questions and sources was not completed.

Sources: [579], [675]

Evidence: Benchmark-family atlas - TruthfulQA: inspected-source qualification. Questions are designed to elicit imitative falsehoods across 38 categories; this is authored challenge material, not sampled deployment traffic.

Sources: [575], [579], [675]

Evidence: Benchmark-family atlas - WebArena: construction, signal and protocol. Authors contribute intents; authors and an external annotator curate information-seeking answers with two annotations and a third annotator resolving disagreement. Three authors contribute programmatic checks for the remaining tasks. The original paper-described benchmark contains 812 instantiated intents derived from 241 templates. WebArena uses standalone self-hosted website implementations with sampled real-world data to support reproducible evaluation without depending on live website changes.

Sources: [91]

Evidence: Benchmark-family atlas - WebArena: inspected-source qualification. Section3.2 also permits ChatGPT inspiration for intents; this is not an exclusively unaided-human task-generation claim.

Sources: [91]

Evidence: Benchmark-family atlas - WebShop: construction, signal and protocol. WebShop defaults to loading 1,000 products for preview and documents changing paths to load all products. WebShop reports 1.18 million real-world products and 12,087 crowdsourced text instructions in a simulated e-commerce task. WebShop setup downloads 50 randomly chosen MTurk-worker trajectories, and session logs store one instruction/web session with search/click steps. WebShop supports HTML and simple observation modes; simple removes extraneous metadata to ease training/evaluation.

Sources: [203]

Evidence: Benchmark-family atlas - WebShop: access, version and interpretation boundaries. The WebShop LICENSE.md is MIT for software and associated documentation.

Sources: [203], [679]

Evidence: Benchmark-family atlas - τ-bench: construction, signal and protocol. τ-bench also requires listed task outputs to appear case-insensitively in agent responses after removing commas, or sets reward to zero. The pinned τ-bench evaluator compares the final database hash with the hash produced by replaying reference task actions, assigning zero when they differ. τ-bench emulates conversations between a model-simulated user and a language agent with domain-specific APIs and policy guidelines. τ-bench defaults to gpt-4o as an llm-strategy user simulator and permits other models/strategies.

Sources: [198], [199]

Evidence: Benchmark-family atlas - τ-bench: access, version and interpretation boundaries. The pinned τ-bench repository contains an MIT license. The τ-bench README marks its airline/retail tasks outdated and directs users to successor repositories for fixed tasks.

Sources: [197], [199]

Method: Atlas interpretation: a source can document a dataset’s declared license, retained record count or human-checking procedure without this study inspecting every row or reproducing a training run. Unknown or restrictive licensing is a catalogue finding rather than an instruction to retrieve or redistribute the corpus. Empty taxonomy facets remain unknown. A release’s disclosed aggregate count can coexist with a README/example inconsistency; both are retained instead of silently manufacturing a reconciled number.

Analysis: Reference-label provenance should accompany benchmark protocol. The label-error study uses model screening followed by five-worker judgments and distinguishes ambiguous or unknown cases from correctable alternatives. Gilardi evaluates only texts on which two trained annotators agree; HelpSteer2 removes remaining high-spread responses. These are different decisions about which items enter a score. A buyer should retain the original denominator, review-selection rule and accepted reference rather than compare the resulting percentages as if they measured the same difficulty distribution.

Sources: [428], [430], [433]

Analysis: A paper’s dataset statistics are evidence about its described release, rather than independent confirmation of a current download. Aya separates original annotations from re-annotations and imposes a language inclusion threshold. WildChat v1 contains two unreconciled conversation totals, which this synthesis preserves. These additional paper reviews do not add another dataset family here. A future artifact inspection should bind the exact bytes, reconcile paper/card/repository counts and preserve component lineage before changing family or release accounting.

Sources: [422], [440]

Evidence: Benchmark-family atlas - Polymath Horizon-SWE: construction and protocol. The creator distinguishes Horizon-SWE-PF (pass/fail), requiring complete feature and deployment checks, from Horizon-SWE-PC (partial credit), weighted 60% feature correctness, 30% deployment/DevOps and 10% engineering quality. The creator protocol combines verifiable feature/deployment checks with model-based engineering-quality assessment and describes concrete failure patterns. Horizon-SWE describes 50 tasks in a running containerized application with synthetic traffic and an MCP tool surface.

Sources: [537]

Analysis: Benchmark-family atlas - Polymath Horizon-SWE: assessment and access boundaries. Fifty running-application tasks define one family, with strict PF and weighted PC views. A buyer should retain the 60/30/10 component breakdown and inspect feature/deployment checks separately from model-rated engineering quality. Confidence is high in the documented scoring distinction and unestablished for reproducibility of private payloads. Obtain the executed task/harness manifest, judge calibration and project rights before accepting the service as a matched benchmark. Source-code commit history is not a contributor headcount, and synthetic traffic is not naturally observed customer activity.

Sources: [537]

Evidence: Benchmark-family atlas - Toloka Visual Question Answering / WSDMCup2023: construction and protocol. The paper describes crowd workers choosing interesting objects/images, composing English questions and drawing the answer bounding boxes from scratch on a CC-BY COCO image subset. The challenge scores predicted bounding boxes with mean intersection-over-union; its paper separately reports a crowd baseline and annotation agreement using a 1-IoU distance. The paper explicitly checked overlap with COCO boxes and placed overlapping cases in the training set, then released the same complete dataset with ground truth through Zenodo, Hugging Face, Kaggle and GitHub. The paper says the complete question/box dataset was released under the same CC-BY license as its chosen COCO image subset. The creator challenge paper describes 45,199 image-question-box instances: 38,990 train, 1,705 public test and 4,504 private test.

Sources: [524], [525]

Analysis: Benchmark-family atlas - Toloka Visual Question Answering / WSDMCup2023: assessment and rights boundaries. The unit is an image-question-box instance: the 1,000-example train_sample remains part of the training split, and mirrors do not add cases or families. A buyer can compare box localization using the specified IoU protocol while independently checking ambiguous object references and source-image permissions. Confidence is high in the documented units and moderate in this measurement fit; a payload audit with per-image rights and adjudicated boxes would change the assessment. The creator’s CC-BY declaration is reported, without inferring blanket rights from the COCO brand or equating the crowd baseline with commercial annotation accuracy.

Sources: [524], [525]

Evidence: Benchmark-family atlas - Regulated Advice Bench: construction and protocol. The card reports tax nominal alpha 0.628 on 318 doubly read items including 58 ties excluded from release, and legal alpha 0.714 on 257 doubly judged items including ties. The pinned card states CC-BY-NC-SA-4.0 for this release. The full annotation policies are not included. The summarized tax policy follows UK, German and French reserved-advice rules with a protective default elsewhere; native speakers audited sample sets of automated translations. The card describes synthetic dialogue pairs followed by regulatory expert review of English replies: four tax reviewers and five legal reviewers, with two per item where possible, blind to the AI label. Regulated Advice Bench has 338 tax items expanded to 3,042 language rows and 332 legal items expanded to 2,988 rows; each item’s nine translations share a label.

Sources: [526]

Analysis: Benchmark-family atlas - Regulated Advice Bench: assessment and rights boundaries. The family contains tax/legal configurations and translated manifestations of 670 English-labeled items. Resample by item_id rather than treating 6,030 rows as independent judgments. The alpha denominators include excluded ties and differ from released-item counts. Confidence is high in these units and low in treating the labels as global professional advice standards. A buyer should obtain the full intended policy, qualified local review and translation-specific checks; private policies and inherited English verdicts prevent a universal legal-validity claim. CC-BY-NC-SA-4.0 is a release-specific declaration, not blanket upstream permission.

Sources: [526]

Analysis: Family counting and technical discussion answer different questions. This graph counts 85 dataset families and 87 benchmark families. A named derived collection can be distinct without disjoint source examples: the Flan Collection and xP3 disclose reuse and transformation of upstream tasks, while Chatbot Arena Conversations is defined by paired conversations and human winner votes rather than the single-model LMSYS-Chat-1M schema. Neither the 17 added paper works nor each release/configuration automatically adds a family. HCAST is a benchmark; SWAA is a named dataset task collection; the METR time-horizon lineage is a composite evaluation protocol with overlapping inputs. Versions, translated rows, submixtures, internal groups and public/private subsets remain components, not added families. Current source-backed identity dispositions do not erase earlier holds or establish example-disjointness.

Sources: [97], [165], [166], [204], [542], [545], [546], [547], [548], [617], [618]

Analysis: Benchmark selection and independent unit. GDP.pdf and Chartography select tasks after model failures; MMRB2 filters model-judge agreement and then human disagreement; Regulated Advice Bench translates one English-labeled item into nine rows. A representative denominator, an adversarial challenge denominator and a repeated-presentation denominator are all useful but answer different questions. Confidence is high that these selection rules must accompany scores and unestablished for their prevalence in a buyer's workload. Report item-level and participant-level grouping, not only row count, and use an independently drawn target-work sample before estimating operational success.

Sources: [3], [8], [526], [558]

Analysis: Corpus and documentation pins. Deduplication findings bind particular matching procedures and source corpora; StarCoder's PII project combines enriched and random files; Magpie's release card, repository README and code license make separate declarations. A buyer should require both an exact payload manifest and an applicable rights/provenance manifest. Confidence is high that a documentation revision alone does not establish corpus-byte identity or permission, and moderate that these manifests improve inspectability. Reconciled payload hashes, independent sampling and version-specific governing instruments would change the unresolved provenance/rights assessment.

Sources: [100], [101], [465], [476], [706]

Evidence: Dataset-family atlas - Chatbot Arena Conversations: construction, signal and protocol. Chatbot Arena Conversations contains cleaned conversations with pairwise human preferences. Chatbot Arena Conversations was collected on Chatbot Arena from April to June 2023. Chatbot Arena Conversations includes toxic tags generated by T5 and RoBERTa taggers trained on manually labeled data. Chatbot Arena Conversations intentionally retains unsafe conversations alongside moderation tags. Chatbot Arena user prompts are CC-BY-4.0, while model outputs are CC-BY-NC-4.0. The card describes terms-based consent and best-effort PII removal while deliberately retaining unsafe conversations with automated moderation and toxicity tags. Arena users converse with two anonymous models, vote their preferences and then see model identities; the platform was free, open without registration and did not pay users. The associated paper documents position, verbosity and reasoning limitations of LLM judging, with mitigation experiments and separate crowd/expert agreement samples. The pinned Chatbot Arena Conversations card describes 33,000 paired conversations from April-June 2023, with 20 model identities and a 13,383-user table.

Sources: [97], [542]

Evidence: Dataset-family atlas - Chatbot Arena Conversations: access and version observations. The card states CC-BY-4.0 for user prompts and CC-BY-NC-4.0 for model outputs, together with model-specific terms for using direct outputs.

Sources: [97]

Evidence: Dataset-family atlas - Chatbot Arena Conversations: interpretation boundaries. The 33,000 paired conversations, earlier 30K-vote paper snapshot, separate MT-bench expert set and single-model LMSYS-Chat-1M corpus are distinct units. Shared users/time/examples can overlap. A 13,383-user table or unique-IP description does not verify distinct people. Toxicity tags are model outputs, unsafe text is deliberately retained, and PII removal is best-effort. User prompts and model outputs have different CC terms plus model-specific conditions.

Sources: [97], [542]

Analysis: Dataset-family atlas - Chatbot Arena Conversations: buyer assessment and change conditions. Free, anonymous participation creates a selection process different from a paid calibrated panel. Confidence is high that prompts, winner votes, responses and toxicity tags need separate provenance, and low for population representativeness. A buyer should preserve paired context, duplicate/voter clustering and content safeguards rather than treat every vote as independent gold. Independent overlap and preference-validity audits would change confidence in training reuse.

Sources: [97], [542]

Evidence: Dataset-family atlas - ELOQUENT Preference Prediction: construction, signal and protocol. The protocol evaluates four-way preference classification with accuracy and explanations with ROUGE-L, BERTScore and an LLM judge, aggregating final metric rankings with Borda count. The pinned task protocol says human preferences and explanations were collected from scratch on Toloka for relevance, naturalness, truthfulness, safety and overall quality. ELOQUENT Preference Prediction’s pinned card lists 99 validation and 1,248 test examples in English, with two shared-task subtasks: preference prediction and preference-plus-explanation.

Sources: [543], [544]

Evidence: Dataset-family atlas - ELOQUENT Preference Prediction: access and version observations. The card names an MIT license and says the paper is TBA; its truthfulness-preference field description repeats the relevance label wording.

Sources: [543]

Evidence: Dataset-family atlas - ELOQUENT Preference Prediction: interpretation boundaries. The available card says Paper TBA and repeats relevance wording in the truthfulness field. The protocol does not establish annotator eligibility, multiplicity, adjudication or agreement. Development and final judges differ; labels were private during the competition and current payload-label availability was not inspected. MIT metadata does not establish upstream response rights.

Sources: [543], [544]

Analysis: Dataset-family atlas - ELOQUENT Preference Prediction: buyer assessment and change conditions. Preference prediction and explanation generation should be accepted as separate deliverables. Accuracy, lexical similarity, embedding similarity and a judge rank cannot substitute for one another. Confidence is moderate in using the task to compare these outputs and low in treating its labels as universal consensus. A buyer should resolve the field schema, collect repeated independent judgments and verify explanation support. Those records would change the QA assessment.

Sources: [543], [544]

Evidence: Dataset-family atlas - SWAA task collection: construction, signal and protocol. Software Atomic Actions (SWAA) consists of 66 short tasks representing software-engineering actions under one minute, using multiple-choice decisions and fill-in-the-blank completion questions. SWAA's authors wrote an initial 71 tasks without seeing AI attempts or running models on the developing tasks, then retained 66 after excluding tasks that more than one human baseliner failed. SWAA human baselines were collected from internal employees with relevant expertise using a custom web application, with four baselines per decision task and three per fill-in-the-blank task. The time-horizon paper credits Thomas Kwa with leading SWAA task development and Amy Deng with creating most SWAA tasks and collecting their human baselines and AI agent results.

Sources: [617]

Evidence: Dataset-family atlas - SWAA task collection: interpretation boundaries. Despite its retained benchmark-prefixed inventory ID, the semantic entity is a dataset task collection. Its five internal groups and 71 preliminary tasks are components; 66 survive the human-failure filter. Internal-employee timing excludes context acquisition, decision tasks permit one choice and fill-in tasks permit repeated attempts. Development was blind to AI task performance, with separate prompt-development tasks.

Sources: [617]

Analysis: Dataset-family atlas - SWAA task collection: buyer assessment and change conditions. Short atomic actions anchor a lower-duration regime without establishing a standalone operated benchmark or the time required for full software work. Confidence is high in this distinction and moderate in its role within the composite protocol. A buyer should test the same timing and retry rules on its own action population. Independent baselines including context acquisition would change operational relevance.

Sources: [617]

Evidence: Dataset-family atlas - The Flan Collection: construction, signal and protocol. FLAN v2 compiles Flan 2021, P3, Super-NaturalInstructions and additional datasets into mixed zero-shot, few-shot and chain-of-thought templates. FLAN downloadable submixtures can be filtered by template_type or task_name and combined at specified mixture rates. FLAN v2 names flan, t0, niv2, cot and dialog submixtures with prompt varieties controlling shots and answer options. The expanded Flan Collection combines prior instruction-task collections with prompt variants, input inversions and mixture weighting; the released example defines five principal submixtures. The paper studies mixed prompt formats, task scaling, input inversion and source balancing using T5-XL and held-in, CoT and held-out MMLU/BBH evaluations. The v2 README explicitly omits several tasks for legal constraints and excludes the Program Synthesis submixture, while the paper's described recipe includes program-synthesis data.

Sources: [165], [545], [546], [547]

Evidence: Dataset-family atlas - The Flan Collection: access and version observations. The FLAN root README licenses repository code through LICENSE. FLAN v2 says legal constraints excluded Dr Repair, DeepMind Code Contests and Task Master from generation scripts, leaving no program-synthesis submixture. The pinned release is a composite instruction-tuning recipe over upstream task datasets, with code under an Apache-2.0 header.

Sources: [165], [166], [547]

Evidence: Dataset-family atlas - The Flan Collection: interpretation boundaries. This is a derived weighted instruction mixture with heterogeneous upstream labels, templates and examples. A distinct collection identity does not establish example-disjointness from P3, Flan 2021 or Super-NaturalInstructions. Public v2 generation omits legally constrained tasks and Program Synthesis even though the experimental recipe includes such material. Apache code terms do not establish all component rights.

Sources: [165], [166], [545], [546], [547]

Analysis: Dataset-family atlas - The Flan Collection: buyer assessment and change conditions. Recipe reproducibility needs task filters and mixture weights, not just a collection name. Confidence is high in that engineering requirement and moderate in deriving task-specific experiments from the paper. A buyer should compare the exact obtainable recipe with the reported experimental mixture and quantify source overlap before combining collections. A reconstructed licensed manifest and held-out task ablations would change confidence in a proposed mixture.

Sources: [165], [166], [545], [546], [547]

Evidence: Dataset-family atlas - xP3: construction, signal and protocol. The xP3 card reports 78,883,588 samples in merged language JSONL files. xP3 defines inputs as the natural-language model input and targets as the natural-language generation target. The xP3mt variant uses prompts in 20 languages machine-translated from English. The xP3 family table identifies xP3 as 13 training tasks in 46 languages with English prompts. The pinned xP3 card distinguishes 13 training task types in 46 languages from the evaluation-inclusive 16-task variant, machine-translated-prompt variant and processed training format. The paper's rank-classification evaluations use five random prompts and median per-language scores; it notes unintentional exposure to supposedly unseen languages, limited language coverage and shorter-generation bias after multitask fine-tuning. The xP3 card's language table reports sample and kilobyte totals computed from merged language files. The paper extends P3 with multilingual supervised datasets and uses English prompts for xP3, while xP3mt machine-translates prompts for monolingual datasets.

Sources: [204], [548]

Evidence: Dataset-family atlas - xP3: access and version observations. The xP3 card declares Apache 2.0 for the released collection. The xP3 card states Apache-2.0 and identifies input/target fields; the paper lists xP3all, xP3mt and xP3megds as evaluation/translation/processing variants of xP3.

Sources: [204], [548]

Evidence: Dataset-family atlas - xP3: interpretation boundaries. Thirteen training task types and 46 languages differ from the evaluation-inclusive sixteen-task variant. The 78,883,588 sample total is a merged-file card count, not unique untemplated source examples or new human annotations. xP3mt, xP3all, processed xP3megds and expanded xP3x are related configurations, with explicit upstream overlap. Apache card metadata does not reconcile every component license.

Sources: [204], [548]

Analysis: Dataset-family atlas - xP3: buyer assessment and change conditions. Prompt language, target origin and byte/sample share answer different coverage questions. Confidence is high in keeping those units separate and low in inferring equal quality across languages from the aggregate. A pilot should stratify language-by-task and compare original versus translated prompts with native reviewers. A payload-level manifest, independent language validation and clarified exposure history would change transfer confidence.

Sources: [204], [548]

Evidence: Benchmark-family atlas - APEX-Accounting: construction, signal and protocol. The primary paper describes 42 accounting contributors, median 11 and mean 12.6 years of experience, constructing month-end-close work under U.S. GAAP accrual accounting. The paper grades binary rubric criteria with a DeepSeek-v4-Flash judge calibrated against 1,687 majority-vote expert criterion labels from 120 trajectories. The pinned Harbor edition renders the same ten public APEX-Accounting development tasks in one simulated company world; its card separately describes a closed 160-task, ten-world benchmark. The paper selected eleven difficult worlds from a 599-candidate-task pool using frontier-model scores; the ten hardest worlds became the closed benchmark and the eleventh supplied the public development set. It removed 24 incomplete-information tasks.

Sources: [509], [510], [511]

Evidence: Benchmark-family atlas - APEX-Accounting: access and version observations. The Harbor card states CC-BY-4.0, identifies three tightened rubric criteria, and supplies static accounting-system exports rather than the leaderboard accounting-software interface. The currently captured official leaderboard describes 100 tasks and 1,475 criteria, whereas the primary paper, creator post and Harbor card describe 160 closed tasks.

Sources: [509], [510], [511], [512]

Evidence: Benchmark-family atlas - APEX-Accounting: interpretation boundaries. The closed v2 paper has 160 tasks, while the separately captured leaderboard says 100 tasks and 1,475 criteria and calls k=4 canonical. The paper uses Mean Criteria@3. Neither counts nor metrics are harmonized here. The ten-task Harbor dev sample uses static exports and a different default judge/template; its tightened criteria and CC-BY sample terms do not make its scores comparable with the closed leaderboard.

Sources: [509], [510], [511], [512], [697]

Analysis: Benchmark-family atlas - APEX-Accounting: buyer assessment and change conditions. The selected month-end-close world is a bounded accounting slice with synthetic inputs and excluded clarification cases. Confidence is high in separating criterion completion, any-attempt success and all-attempt reliability, and unestablished for occupational automation. A buyer should require practitioner-reviewed critical gates and explicit incomplete-information escalation. A reconciled version ledger, independent judge test and representative real-work sample would change confidence.

Sources: [509], [510], [511], [512], [697]

Evidence: Benchmark-family atlas - APEX-Agents: construction, signal and protocol. The original paper describes 256 professional contributors and expert-built workplace worlds; each task has an expert-authored prompt, outcome rubric and golden output. The original paper evaluates criterion judgments against 747 human labels across 60 tasks and reports 98.5% judge agreement on that calibration. The original protocol distinguishes average single-attempt task success, any-of-eight success and all-eight consistency; release 1.1 introduces expert audits and judge changes intended to reduce scattergunning and ambiguity. APEX-Agents 1.1 has 240 tasks across 31 worlds and 955 binary criteria, with 80 tasks in each of investment banking, law and management consulting; the original paper describes 480 tasks across 33 worlds.

Sources: [513], [514], [515]

Evidence: Benchmark-family atlas - APEX-Agents: access and version observations. The pinned 1.1 card states CC-BY-4.0 and also restricts use to evaluation, forbids training/fine-tuning/parameter fitting and forbids crawling or scraping.

Sources: [514]

Evidence: Benchmark-family atlas - APEX-Agents: interpretation boundaries. The original paper’s 480 tasks and 33 worlds differ from release 1.1’s 240 tasks, 31 worlds and 955 criteria. Original judge calibration is not proof of release-1.1 accuracy. The card’s CC-BY-4.0 declaration coexists with evaluation-only, no-training and no-crawling restrictions; no payload rights reconciliation is performed.

Sources: [513], [514], [515]

Analysis: Benchmark-family atlas - APEX-Agents: buyer assessment and change conditions. For a buyer, this is an execution and repeated-reliability instrument, not the earlier in-context APEX response test. Confidence is high that version and any/all-attempt denominators are essential and unestablished for representative professional-work success. Freeze the world, judge and task manifest and test whether revised audits change false accepts. Independent reproduction and a matched target-work sample would change use confidence.

Sources: [513], [514], [515]

Evidence: Benchmark-family atlas - APEX-v1: construction, signal and protocol. For APEX-v1.0, Mercor sourced experienced professionals, interviewed them for 30-45 minutes, and paid selected experts to complete a 1-2 hour prompt-and-rubric assessment before successful applicants were contracted. APEX-v1.0 experts first described 3-5 common daily-work tasks and their required seniority, then used that workflow input to create prompts and find or create supporting evidence sources. The original APEX-v1.0 paper describes a closed heldout evaluation set and publishes four example prompts with rubrics that were created by the same expert group under the same instructions but are outside that heldout set. The paper's original APEX-v1.0 collection has 200 cases, 50 each for investment banking, management consulting, law and medicine, created and reviewed by industry experts. APEX-v1.0 reviews prompts before contributors write the corresponding rubrics, then reviews those rubrics through potentially multiple rounds; 200 of 300 started prompts were approved for the final collection.

Sources: [605]

Evidence: Benchmark-family atlas - APEX-v1: access and version observations. The pinned APEX-v1-extended card labels the release CC-BY 4.0 while also stating exclusive evaluation use and forbidding training, fine-tuning, parameter fitting, crawling and scraping. The captured APEX-v1-extended creator card describes 400 heldout tasks, 100 per job, plus 100 open-source development tasks; it replaces the original three-judge panel with Gemini 2.5 Pro thinking On as a single judge.

Sources: [605], [606]

Evidence: Benchmark-family atlas - APEX-v1: interpretation boundaries. The original 200-case, median-of-three, three-judge response experiment differs from the extended card’s 400 heldout plus 100 development tasks, eight-response mean and single Gemini judge. Expert recruitment and paid assessment do not establish hourly rates or employee status. Supporting sources can be created hypothetical materials. CC-BY labeling coexists with evaluation-only and programmatic-access restrictions.

Sources: [605], [606]

Analysis: Benchmark-family atlas - APEX-v1: buyer assessment and change conditions. This is an in-context response-quality family, distinct from APEX-Agents execution environments. Confidence is high in that boundary and unestablished for real professional value. A buyer should preserve the source packet, negative criteria and response aggregation rule and avoid converting criterion percentages to labor saved. Prospective accepted-deliverable studies and independent expert calibration would change confidence.

Sources: [605], [606]

Evidence: Benchmark-family atlas - BenchFlow SkillsBench: construction, signal and protocol. The paper acknowledges missing length-matched context controls, limited model/harness coverage, nondeterminism and contamination risks when interpreting skill gains. SkillsBench compares the same agent/model with and without reusable task skills; the original paper averages five binary-reward trials per task over a fixed 84-task denominator. The paper and pinned contributor protocol require outcome verifiers, passing reference solutions, human-written prompts, reusable skills and review for task realism, leakage and environment problems.

Sources: [369], [539], [540]

Evidence: Benchmark-family atlas - BenchFlow SkillsBench: access and version observations. The HF mirror calls its 87-active/14-excluded release v1.1 and points to commit 27738384b1df694ea2ae466e416f476e94d8fab9; README/CONTRIBUTING at that exact commit instead call it v1.2. The pinned native task.md release supports the BenchFlow 0.6.x CLI line and keeps credential-dependent/incompatible packages outside the default roster; its README states Apache-2.0.

Sources: [369], [539], [540], [541]

Evidence: Benchmark-family atlas - BenchFlow SkillsBench: interpretation boundaries. The paper’s 86 accepted and 84 evaluated tasks differ from a later 87-active/14-excluded roster. The HF mirror calls that pinned roster v1.1 while README/CONTRIBUTING call it v1.2. The paper’s Harbor setup differs from the native task.md/BenchFlow 0.6.x release. Valid-trajectory counts include timeouts and ordinary failures but exclude runtime/infrastructure errors.

Sources: [369], [539], [540], [541]

Analysis: Benchmark-family atlas - BenchFlow SkillsBench: buyer assessment and change conditions. Paired skill/no-skill tests are informative only after the context budget and task-selection rule are declared. Confidence is moderate in this task-specific comparison and low in a universal skill benefit without length-matched controls. Freeze the skill version, complete intended model/harness conditions and test held-out tasks. A controlled context-length study and reconciled release roster would change causal confidence.

Sources: [369], [539], [540], [541]

Evidence: Benchmark-family atlas - CharXiv: construction, signal and protocol. The v1 datasheet attributes all chart selection and question-answer annotation to Princeton University graduate students and states that the graduate students involved in collection were not compensated. The v1 paper reports 2,323 charts, with four descriptive questions and one reasoning question per chart: 9,292 descriptive and 2,323 reasoning questions; it splits charts into 1,000 validation and 1,323 test examples.

Sources: [634], [635]

Evidence: Benchmark-family atlas - CharXiv: access and version observations. The primary paper and creator releases distinguish CC BY-SA 4.0 questions and answers from Apache 2.0 code, while chart copyright remains with the original arXiv authors; the creator dataset card states that the dataset is for evaluation and must not be used for model training. The creator repository describes validation/test question files with test answers intentionally set to null; the pinned dataset card also records a known incorrectly annotated reasoning answer for 0.jpg.

Sources: [634], [635], [639]

Evidence: Benchmark-family atlas - CharXiv: interpretation boundaries. The paper reports 2,323 charts, while one README passage says 2,333. Test answers are intentionally null and the card names an incorrectly annotated reasoning answer; source Figure 4 percentage inconsistencies remain. Collection/annotation by uncompensated Princeton graduate students is author-reported, not a payroll audit. Question/answer CC BY-SA terms, Apache code, chart copyright and evaluation-only restrictions have separate scopes.

Sources: [634], [635], [639]

Analysis: Benchmark-family atlas - CharXiv: buyer assessment and change conditions. Four descriptive questions plus one reasoning question per chart create nested observations rather than independent new images. Confidence is high that chart grouping and known-label errors must remain visible and moderate in task relevance. A buyer should inspect unanswerable questions and source-chart-dependent mistakes with independent experts. Reconciled counts, corrected answers and a new-chart holdout would change measurement confidence.

Sources: [634], [635], [639]

Evidence: Benchmark-family atlas - ClawsBench: construction, signal and protocol. The paper evaluates a ragged set of model-harness and skills/meta-prompt conditions with five repeats, reporting 7,224 completed main trials after infrastructure losses. The paper explicitly limits interpretation because services omit latency, rate limits and concurrent access, there is no human baseline, evaluation is single-shot and the design has incomplete factorial coverage. ClawsBench uses 44 tasks across five mock productivity services, with 20 performance tasks and 24 safety tasks; grading compares service state before and after the attempt.

Sources: [516], [519]

Evidence: Benchmark-family atlas - ClawsBench: access and version observations. The pinned repository and card state CC-BY-NC-SA-4.0 and explicitly restrict commercial use of data, trajectories and environments. The trace card states 7,834 total traces while also listing 7,224 main trials plus 1,132 pilot traces, and says task definitions will be added later.

Sources: [517], [518], [519]

Evidence: Benchmark-family atlas - ClawsBench: interpretation boundaries. Twenty performance and 24 safety tasks have different outcome denominators. Services omit latency, rate limits and concurrency; same-task scaffold development and incomplete factorial coverage constrain transfer. The trace card’s 7,834 total conflicts with its listed 7,224 main and 1,132 pilot traces. Counts are not reconciled. CC-BY-NC-SA wording restricts commercial use of data, traces and environments.

Sources: [516], [517], [518], [519], [701]

Analysis: Benchmark-family atlas - ClawsBench: buyer assessment and change conditions. State changes expose side effects that answer-only scoring can miss, while deliberately unsafe tasks cannot estimate natural incident prevalence. Confidence is high in using both task success and unsafe actions, and moderate in the bounded mock-service diagnosis. A buyer should complete controls on unseen tasks and audit over-refusal separately. Real-service constraints and reconciled task/trace releases would change confidence.

Sources: [516], [517], [518], [519], [701]

Evidence: Benchmark-family atlas - Comprehensive Verilog Design Problems (CVDP): construction, signal and protocol. The v1 paper limits tasks to oracle contexts with minimal relevant information, and describes Docker-based harnesses using open-source EDA tools where possible while some design-verification tasks require Cadence Xcelium. The pinned creator README reports adoption of CVDP by the Silicon Integration Initiative’s LLM Benchmarking Coalition, describing extensions to categories and design domains, refinement of metrics, and maintenance of a leaderboard. The v1 paper reports 783 retained problems across thirteen task categories, partitioned into 617 Non-Agentic and 166 Agentic problems. The v1 paper explicitly thanks Turing for collaboration in developing the benchmark dataset and Cadence for providing EDA tool licenses that enabled collection of digital-verification datapoints.

Sources: [643], [644]

Evidence: Benchmark-family atlas - Comprehensive Verilog Design Problems (CVDP): access and version observations. The pinned dataset license distinguishes non-code material under CC BY 4.0 from original code under Apache 2.0 and says third-party-derived components may carry different upstream terms; its NOTICE lists the VGA-controller and Brent-Kung-adder components with no specified license. The pinned creator README says twenty datapoints were omitted from the initial public release because of harness issues and licensing restrictions, and that reference outputs and agentic patches were excluded from that initial release to mitigate contamination. The pinned dataset changelog describes v1.1.0 as adding context-heavy agentic tasks with public Git bundles, changing simulation tooling and recalibrating three datapoints; it also records earlier harness fixes and a split into commercial and non-commercial variants.

Sources: [640], [641], [642], [644], [645]

Evidence: Benchmark-family atlas - Comprehensive Verilog Design Problems (CVDP): interpretation boundaries. The paper’s 783 problems divide into 617 Non-Agentic and 166 Agentic tasks, while the published model evaluation converts Agentic material to single-turn form. Initial omissions are not subtracted to invent a current release size. Later v1.1.0 changes context and tools. Non-code CC BY, original-code Apache and third-party terms differ; two named NOTICE components have no specified license. Cadence tool licenses and Turing collaboration have narrow stated roles.

Sources: [640], [641], [642], [643], [644], [645]

Analysis: Benchmark-family atlas - Comprehensive Verilog Design Problems (CVDP): buyer assessment and change conditions. A buyer should select a reproducible design-verification environment before comparing model scores. Confidence is high that tools, oracle context and evaluation mode alter the task and unmeasured for complete chip-design competence. Test alternative valid RTL and check both executable and comprehension outcomes, with proprietary-tool and component-rights obligations explicit. Independent hardware validation would change acceptance confidence.

Sources: [640], [641], [642], [643], [644], [645]

Evidence: Benchmark-family atlas - Datacurve DeepSWE: construction, signal and protocol. The creator’s verifier comparison uses an LLM to assess 735 DeepSWE and 789 SWE-Bench Pro rollouts after excluding API errors, timeouts and transient harness failures. DeepSWE describes 113 original tasks across 91 active open-source repositories in five languages; tasks pin base commits and package prompts, verifiers and reference solutions. The creator protocol describes three verifier runs during authoring, regression checks, LLM-assisted analysis and independent human review of prompt-verifier alignment, implementation breadth, realism and environment cleanliness.

Sources: [532], [533]

Evidence: Benchmark-family atlas - Datacurve DeepSWE: access and version observations. The creator explicitly limits coverage to selected active repositories with at least 500 stars and five languages, with localization/refactoring and proprietary or long-tail code underrepresented. Since v1.1, DeepSWE’s README requires Pier newer than 0.3.0 and a separate verifier environment that extracts committed patches into a pristine grading container.

Sources: [532], [533]

Evidence: Benchmark-family atlas - Datacurve DeepSWE: interpretation boundaries. Repository selection favors active public projects with at least 500 stars and five languages, underrepresenting proprietary, long-tail and localization/refactoring work. Analyzer/verifier disagreement is not a ground-truth verifier error rate. The corpus has a no-training canary; pinned repository instructions do not establish full upstream rights clearance or task execution.

Sources: [532], [533], [607], [608], [609]

Analysis: Benchmark-family atlas - Datacurve DeepSWE: buyer assessment and change conditions. The pristine grading-container boundary is material because the evaluated patch should not control its own checks. Confidence is high in that separation as an acceptance requirement and moderate in relevance to substantial public-repository implementation. Inspect patch extraction, regression coverage and reasonable alternative solutions. Independent execution across less visible repositories would change the generalization judgment.

Sources: [532], [533], [607], [608], [609]

Evidence: Benchmark-family atlas - DAYJOB: construction, signal and protocol. The pinned cross-domain harness README instead documents a 0-1 trial reward, a task mean over five nonerrored attempts, and an unweighted mean across tasks in a domain. The finance and healthcare cards use 100% rubric satisfaction for a task pass, whereas the tax card states a 95% criterion threshold and published pass@1. The finance and tax cards describe professionals authoring assignments and three review layers for correctness, completeness and solvability; tax materials include synthetic documents modeled on practice. DAYJOB explicitly identifies itself as one benchmark family. Its finance card describes 80 full tasks and three samples, healthcare 50 full tasks and three samples, and tax 50 full tasks and two samples. The README specifies Harbor/OpenHands-based execution, an agent-phase allowlist for model API hosts and pinned OpenCode/rewardkit grading dependencies; cards reserve the tasks for held-out evaluation.

Sources: [520], [521], [522], [523]

Evidence: Benchmark-family atlas - DAYJOB: interpretation boundaries. The paper’s finance/healthcare snapshot and later cards/harness are distinct protocol contexts. The cards use full rubric satisfaction in finance/healthcare but 95% in tax; the cross-domain harness describes dense trial rewards and means over nonerrored attempts. Domain components and public samples do not add families. MIT metadata coexists with held-out-only restrictions against training, distillation and checkpoint selection.

Sources: [520], [521], [522], [523], [695]

Analysis: Benchmark-family atlas - DAYJOB: buyer assessment and change conditions. A domain name does not identify a common pass rule. Confidence is high that tax thresholds, strict completion and dense means must be displayed separately and unestablished for all-domain work readiness. A buyer should choose decisive criteria before evaluation, inspect synthetic versus natural inputs and retain errored-attempt counts. Reconciled card/harness manifests and independent professional grading would change confidence.

Sources: [520], [521], [522], [523], [695]

Evidence: Benchmark-family atlas - GeneBench-Pro: construction, signal and protocol. The reported task setup gives an agent an isolated workspace, staged files, a minimum viable prompt defining the scientific question and target estimand, and an output schema, while leaving the analysis workflow for the agent to recover. The paper reports 3 to 13 substantive inferential decision points per problem, with a median of 6. The paper's synthetic DRX1 carrier-screening example requires three linked stages: reportable carrier calling, calibration of assay performance and risk assembly using residual carrier risk and the full partner roster. The paper reports an external scientific reviewer pool of 11 domain experts spanning graduate students, postdoctoral researchers, industry scientists and professors. The paper reports external review of 84 candidate problems: 82 were included in the 129-problem evaluation, and two reviewed candidates were withdrawn for fatal issues. The reported external-review workflow gave reviewers private repositories for assigned problems, containing the prompt and data, a detailed design report, validation materials and an issue thread for feedback.

Sources: [673]

Evidence: Benchmark-family atlas - GeneBench-Pro: protocol continuation. Table 2 reports that external review identified a genetic-correlation reference implementation presented as LDSC that used incorrect M and OLS rather than WLS, and that the reference implementation was fixed to implement LDSC. The paper describes fully simulated problems and states that prompts tell agents the data came from a real experiment; the pinned public creator card instead quotes task provenance as uncertain and instructs scientific judgement. The pinned creator card states that reference answers and grading tolerances are intentionally public for case studies, reproducibility and model-analysis examples, rather than a hidden-answer leaderboard. The pinned creator card describes problems.csv as the one-row-per-problem inventory, manifest.json as containing per-problem paths, SHA-256 hashes and sizes, and checksums.sha256 as package-wide integrity information. The pinned creator card lists ten public case studies and describes each problem directory as containing an eval configuration, staged agent-visible data files and a public scientific report PDF.

Sources: [446], [673]

Evidence: Benchmark-family atlas - GeneBench-Pro: protocol continuation. The pinned creator card describes a Python 3.10-or-newer reference grader whose passed boolean is the authoritative grading decision and whose score is only a diagnostic summary of graded fields. The paper defines grading ground truth as the quantity recoverable from agent-visible data, rather than the hidden parameter used to generate those data; it gives the estimate produced by the correct analysis as an example. The paper says the Artificial Analysis subset prioritized relatively difficult tasks and that release-visibility and external-review strata were not randomized. The paper describes a tiered 129-problem release: ten public problem packages, 50 additional held-out problems provided to Artificial Analysis for third-party benchmarking, and 69 retained internally. The paper reports that external-review findings led to prompt clarifications or data-file changes for minor issues, and DGP or ground-truth changes, redesign or withdrawal for more substantive issues; corrected candidates returned through agent pilots and trace examination.

Sources: [446], [673]

Evidence: Benchmark-family atlas - GeneBench-Pro: protocol continuation. The GeneBench-Pro paper reports 129 evaluations across 10 primary domains and 21 terminal subdomains for multistage analysis in genomics, quantitative biology and translational biomedicine. The paper describes GeneBench-Pro problems as constructively simulated, with the full data-generating process simulated so that graded endpoints are recoverable from staged data. The paper describes checking plausible wrong decisions through ablations, reviewing scientific validity, methodology and target identifiability, and iterating frontier-model pilots and trace analyses to look for leakage, unintended solution paths, prompt-grader mismatch and robustness problems.

Sources: [673]

Evidence: Benchmark-family atlas - GeneBench-Pro: access and version observations. The pinned public creator card declares the public package available under the MIT License. The paper describes constructing GeneBench-Pro from GeneBench’s 103 problems by withdrawing three with fatal issues, significantly redesigning and hardening 54 of the remaining hundred, and adding 29 new problems.

Sources: [446], [673]

Evidence: Benchmark-family atlas - GeneBench-Pro: interpretation boundaries. One hundred twenty-nine problems span ten domains and 21 terminal subdomains; decision counts are qualitative metadata. External review covers 82 included tasks plus two withdrawn candidates, not all 129; 47 evaluated tasks lack that external-review stratum. Ten public case studies include answers/tolerances, while 50 further tasks go to Artificial Analysis and 69 remain internal. Visibility/review strata were not randomized. Constructive simulation does not recreate all real-data irregularities.

Sources: [446], [673]

Analysis: Benchmark-family atlas - GeneBench-Pro: buyer assessment and change conditions. Target identifiability and wrong-method ablations are strong questions for scientific-task QA. Public case studies are useful for explanation but should not be presented as a hidden-answer leaderboard. Confidence is high in that separation and moderate in the documented validation design. A buyer should independently review unreviewed tasks and reconcile provenance and reasoning-grading wording before reuse. Real-data workflow validation and audited target recovery would change confidence.

Sources: [446], [673]

Evidence: Benchmark-family atlas - HCAST: construction, signal and protocol. HCAST baseliners commit work to a repository and record screens and spoken work through Loom; the authors manually review most successful baselines, prioritizing unusual times and disagreement among attempts. HCAST recruited technical human baseliners mainly through METR employees' professional networks and required the large majority to pass domain-specific qualification tasks before main-suite work. Each HCAST task supplies instructions, a container where the agent runs as the agent user, and a scoring function running as root; continuously scored tasks are binarized at heuristically chosen success thresholds. HCAST's author-assigned contamination categories include 125 tasks with private solutions, 31 with public solutions judged hard to memorize, and 33 with public solutions judged easy to memorize. HCAST's contribution statement credits Chris Canal and Daniel O'Connell with helping manage task QA/testing and baselining, and its byline identifies both with Equistamp. HCAST v1 contains 189 individually created tasks: METR employees and contractors created 143, while 46 came through its independent-contributor task bounty.

Sources: [614]

Evidence: Benchmark-family atlas - HCAST: protocol continuation. HCAST combines fresh-human task testing with manually reviewed agent attempts: an unexposed person attempts a task in most task families, and reviewers read all 945 transcripts from five agent attempts on each of the 189 tasks. The HCAST v1 human reference comprises 563 attempts by 140 people, including 341 successful attempts; 110 tasks use the geometric mean of successful baseline times, and 79 use forecasts by a task creator or familiar author. HCAST's appendix task-family summaries were generated by claude-3-7-sonnet-20250219 from task code, tracked metadata truncated to 10,000 characters, and descriptive information supplied by task authors. HCAST reports paying baseliners $50-100 per hour plus $25-150 per hour in performance bonuses; Appendix E.2.1 defines a successful-completion bonus of (r/2)×t, where r is the baseliner's hourly rate and t is the average time spent on that task by all baseliners.

Sources: [614]

Evidence: Benchmark-family atlas - HCAST: access and version observations. The HCAST v1 paper releases 11 example task families and withholds the remainder to limit contamination and hill-climbing; the captured hcast-public README identifies its contents as a subset and marks the tasks deprecated.

Sources: [614], [615], [616]

Evidence: Benchmark-family atlas - HCAST: interpretation boundaries. The 78 internal setup/scoring groups are components of one 189-task benchmark. Of 563 human attempts by 140 people, 341 succeed; 110 task times use successful geometric means and 79 use forecasts, including 50 with no collected baseline. Humans receive some extra web access and exact thresholds. The public eleven-family subset is marked deprecated; its MIT terms do not expose the withheld whole. Affiliated authors' QA roles do not establish an organizational supplier contract.

Sources: [614], [615], [616], [617]

Analysis: Benchmark-family atlas - HCAST: buyer assessment and change conditions. Human calibration is a measurement process with incentives, corrections and missingness. Confidence is high in separating observed successful times from forecasts and moderate in task-specific horizon use. A buyer should retain unsuccessful attempts and the exact human/agent information difference, then measure new long tasks. Independently recorded baselines and representative workload sampling would change confidence.

Sources: [614], [615], [616], [617]

Evidence: Benchmark-family atlas - HealthBench: construction, signal and protocol. The v1 paper describes HealthBench Consensus as 3,671 base examples with positive consensus criteria, drawn from 34 predefined criteria that are assigned only when more than half of raters and at least two physicians agree that they apply. The v1 paper is authored under OpenAI and releases HealthBench through OpenAI's simple-evals repository; its acknowledgments name physician collaborators and contributors while explicitly saying their inclusion does not imply endorsement. The v1 paper reports 5,000 single- or multi-turn conversation examples with 48,562 unique rubric criteria, organized into seven conversation themes and five behavior axes. The v1 paper selects HealthBench Hard by scoring base examples across five model providers, excluding examples with no positive score from any model, and taking the 1,000 examples with the lowest average score.

Sources: [646]

Evidence: Benchmark-family atlas - HealthBench: access and version observations. The immutable simple-evals README lists HealthBench with an MIT license link and states that, from July 2025, the repository will retain reference implementations while no longer updating model or benchmark results; the primary paper asks that benchmark examples not be redistributed online to reduce leakage.

Sources: [646], [647], [648]

Evidence: Benchmark-family atlas - HealthBench: interpretation boundaries. The 5,000 conversations and 48,562 rubric criteria are source-reported units. Consensus and Hard are base-set variations, not new families. Consensus review of 34 predefined criteria does not validate all individualized base rubrics. The captured reference repository stopped updating model/benchmark results in July 2025. Creator MIT declarations coexist with a request against online example redistribution; upstream rights are not independently cleared.

Sources: [646], [647], [648]

Analysis: Benchmark-family atlas - HealthBench: buyer assessment and change conditions. Theme and behavior coverage is useful for selecting tests, but no rubric mean certifies clinical workflow outcomes. Confidence is high in that restriction and moderate in diagnostic use. A buyer should oversample severe misses and independently calibrate example-specific criteria without conflating them with consensus criteria. Prospective workflow studies and clinician-adjudicated false accepts would change confidence.

Sources: [646], [647], [648]

Evidence: Benchmark-family atlas - MedXpertQA: construction, signal and protocol. The pinned creator README reports that official model releases from Google DeepMind, Meta Superintelligence Labs, ByteDance Seed and Alibaba Qwen include MedXpertQA evaluations. The v1 evaluation generally uses zero-shot chain-of-thought with model-specific exceptions; o1 is evaluated on a ten-percent sample of Reasoning and Understanding questions in each subset with seed 42, rather than the full benchmark. The v1 paper uses GPT-4o to label medical task/subtask and Reasoning versus Understanding, and an LLM to assign a question’s most relevant body system. The v1 paper reports 4,460 questions across Text and MM, including 2,005 multimodal questions and 2,839 images; each subset includes five development questions, with comparison tables reporting 2,450 Text and 2,000 MM test questions.

Sources: [649], [650]

Evidence: Benchmark-family atlas - MedXpertQA: access and version observations. The v1 paper states that medical experts holding a physician's license review original and augmented questions and options for missing information, factual errors and contextual validity; its guidelines require reviewers to record and mark an invalid original question and take no further review steps for that question. The immutable creator dataset card declares MIT and exposes Text/MM development and test configurations with question, options, label and image-reference fields; the primary paper withholds per-question source provenance and asks readers not to share benchmark examples online.

Sources: [649], [650], [651], [652]

Evidence: Benchmark-family atlas - MedXpertQA: interpretation boundaries. Question, image and subset counts differ: 4,460 total questions include development, while 2,450 Text and 2,000 MM are test counts. Most MM tasks already have five options. Physician review does not establish reviewer headcount, independence or perfect labels; model-derived medical attribute labels are not physician annotations. The o1 sample and answer-cleaning exceptions prevent unqualified model-row comparisons. MIT metadata does not establish exam/image rights.

Sources: [649], [650], [651], [652]

Analysis: Benchmark-family atlas - MedXpertQA: buyer assessment and change conditions. Text and MM have different choices and filtering, so equal raw percentages need not mean equal difficulty. Confidence is high in that comparability boundary and unestablished for clinical deployment. A buyer should use a fixed subset, inspect invalid originals and validate answer cleaning before comparing systems. Independently reviewed questions and workflow-specific outcome studies would change the assessment.

Sources: [649], [650], [651], [652]

Evidence: Benchmark-family atlas - METR software-task time-horizon evaluation protocol: construction, signal and protocol. The TH1.1 release reports a post-2023 time-horizon doubling estimate of 130.8 days versus 165.3 days for TH1, using model sets that overlap substantially but are not identical. TH1.1 moves evaluation infrastructure from Vivaria to Inspect and re-estimates horizons for 14 of the 33 models with TH1 estimates. The TH1.1 creator compared Vivaria and Inspect using five models on the old TH1 tasks, and a separate three-model comparison on the new task suite.

Sources: [618]

Evidence: Benchmark-family atlas - METR software-task time-horizon evaluation protocol: access and version observations. The captured TH1.1 creator release changes the time-horizon suite from 170 to 228 tasks by adding 73 HCAST tasks and removing 15, while updating 53 existing tasks.

Sources: [618]

Evidence: Benchmark-family atlas - METR software-task time-horizon evaluation protocol: interpretation boundaries. The composite protocol reuses HCAST, RE-Bench and SWAA inputs. TH1.1 changes tasks, estimates and infrastructure, and re-estimates only 14 of 33 model horizons. The all-period fit is hybrid for earlier models; 130.8- versus 165.3-day post-2023 estimates have overlapping but unequal model cohorts. Of 31 tasks estimated at eight hours or more, only five have measured human times. No independent task-origin or replication credit follows.

Sources: [617], [618]

Analysis: Benchmark-family atlas - METR software-task time-horizon evaluation protocol: buyer assessment and change conditions. A versioned fitted horizon should travel with its task mixture, human-time provenance, scaffold and reliability target. Confidence is high in that requirement and low in interpreting a changed growth fit as pure model progress. A buyer should compare the same systems across old/new tasks and audit estimated long-task times. Prospective stable-suite data and measured long baselines would change trend confidence.

Sources: [617], [618]

Evidence: Benchmark-family atlas - micro1 MissionReport: construction, signal and protocol. Task reward is positive criterion points minus triggered penalties divided by possible positive points, clamped to 0-1; a missing/corrupt/empty report scores zero. The agentic judge sees the submitted Word document and expert report. MissionReport describes 20 reporting tasks across four report types and five organizational settings; its pinned public release contains five tasks and 42 source files. Each task combines a reporting-expert assignment, reader, working files, template and expert rubric; source files may be assembled, written or adapted by the expert.

Sources: [538]

Evidence: Benchmark-family atlas - micro1 MissionReport: access and version observations. The sample card states CC-BY-4.0, explicitly disclaims Department of Defense affiliation/endorsement and says reports are not official products for real reporting decisions. The public card documents current fixes to two expert reports and grading/source notes while retaining prompts, source files, criterion text and weights; published solver runs used OpenCode 1.18.31, three attempts and a 60-minute limit.

Sources: [538]

Evidence: Benchmark-family atlas - micro1 MissionReport: interpretation boundaries. Five public tasks and 42 files are a sample of the described 20-task family; 540 reports are attempts. The judge sees the submitted Word document and expert report but not source files, weights or grading guidance. Gold-report self-checks are not independent judge validation. Fictional tasks with real organization/doctrine names do not establish official government work or endorsement.

Sources: [538]

Analysis: Benchmark-family atlas - micro1 MissionReport: buyer assessment and change conditions. Positive points and penalties create a different acceptance surface from positive criteria alone, but an evidence-blind judge may miss unsupported factual synthesis. Confidence is moderate in the reported task design and unestablished for operational reporting. A buyer should validate penalties and source fidelity with source-visible experts, including corrupt/missing artifacts. Full-set access and independent false-accept analysis would change confidence.

Sources: [538]

Evidence: Benchmark-family atlas - MMLU-Pro: construction, signal and protocol. The paper describes expert answer/format review followed by Gemini-1.5-Pro flagging potentially valid distractors for human expert review; it reports 83% of questions retaining ten options and a mean of 9.47 options. The separately pinned creator card describes a panel of more than ten experts. The v1 source table reports12,032 questions across 14 discipline subsets:6,810 from MMLU,4,083 from the STEM Website,598 from TheoremQA and541 from SciBench.

Sources: [653], [654]

Evidence: Benchmark-family atlas - MMLU-Pro: access and version observations. The pinned creator card records correction of 15 medical-domain answers on 2025-04-06 based on medical-professional recommendations and a 2026-01-18 fix to leading spaces in answer options that could provide a shortcut. The immutable creator card declares MIT metadata and one default configuration with a 12,032-example test split and 70-example validation split; listed fields include question, options, answer/index, chain-of-thought content, category and source.

Sources: [653], [654]

Evidence: Benchmark-family atlas - MMLU-Pro: interpretation boundaries. The v1 table has 12,032 questions from MMLU, STEM Website, TheoremQA and SciBench; the release adds a separate 70-example validation split. More-than-ten experts is later card wording, not the v1 headcount. Medical answers and answer-option whitespace have later dated fixes. MIT declarations do not independently verify all upstream rights, and two review phases do not prove two independent reviewers per item.

Sources: [653], [654]

Analysis: Benchmark-family atlas - MMLU-Pro: buyer assessment and change conditions. A maintained benchmark can fix shortcuts while losing direct score comparability with its first paper. Confidence is high in preserving the corrected release identity and moderate in its intended challenge. A buyer should pin both answer extraction and question version, then inspect plausible distractors by domain. Independently replayed corrections and target-task validation would change confidence.

Sources: [653], [654]

Evidence: Benchmark-family atlas - Pareto AttuneBench: construction, signal and protocol. The primary scoring composite weights emotion tracking 24%, evaluation quality 49% and holistic comprehension 27%; Kendall tau and draft quality judging are supplementary rather than composite inputs. The paper documents informed consent, withdrawal, topic reassignment and manual PII redaction, while explicitly excluding clinical mental-health validation and safety-critical deployment conclusions. Participants had at least five exchanges with an anonymized original model, completed PANAS/binary labels, compared original and preference-informed model responses, and drafted human-edited golden responses. AttuneBench’s 200 conversations were collected from eleven US-based native-English-speaking adults, with each participant annotating their own emotional states, observed/preferred behavior and response preferences.

Sources: [527], [529], [530]

Evidence: Benchmark-family atlas - Pareto AttuneBench: access and version observations. The pinned repository separates MIT code from CC-BY-4.0 conversation data and asks that benchmark conversations/annotations/metadata not be used for training.

Sources: [527], [528]

Evidence: Benchmark-family atlas - Pareto AttuneBench: interpretation boundaries. Two hundred conversations and 50,000-plus annotations do not mean independent participants. The eleven adults are culturally and linguistically restricted, participation is imbalanced and diagnostic data self-reported. Consent/redaction are documented processes rather than proven de-identification. MIT code, CC-BY conversation data and requests against training use have separate scopes.

Sources: [527], [528], [529], [530], [699]

Analysis: Benchmark-family atlas - Pareto AttuneBench: buyer assessment and change conditions. The most useful acceptance target is the participant’s desired conversational behavior, separated from inferred emotion and model-written draft quality. Confidence is moderate in that bounded measurement design and unestablished for clinical use. Split by participant, audit the reference-editing stage and retain Default versus profile-informed contexts. Larger local panels and clustered inference would change external-validity confidence.

Sources: [527], [528], [529], [530], [699]

Evidence: Benchmark-family atlas - ProgramBench: construction, signal and protocol. The pinned usage guide specifies Linux amd64 containers, task_cleanroom_v6 inference images, task_v6 evaluation images and on-demand retrieval of test archives from Hugging Face. ProgramBench reconstruction tasks supply an executable and usage documentation while withholding original source and evaluation tests; the stated default blocks internet access and permits implementations in any language. ProgramBench defines resolved tasks as passing every associated test without a cheating flag; partial test pass rate records relative progress rather than a percentage of implemented functionality. ProgramBench v1 contains 200 reconstruction tasks sourced from open-source repositories, with a reported 248,853 behavioral test functions and a median of 770 per task.

Sources: [655], [656]

Evidence: Benchmark-family atlas - ProgramBench: access and version observations. ProgramBench's paper and current pinned dataset card report different difficulty populations: the paper lists 28 easy, 143 medium and 29 hard tasks, while the card lists 27 easy, 120 medium, 18 hard and 35 unrated. The pinned ProgramBench Generated Tests card declares MIT metadata while describing per-task attribution and original license files, including original terms for portions that may be derivative works under copyleft licenses.

Sources: [655], [657], [658]

Evidence: Benchmark-family atlas - ProgramBench: interpretation boundaries. The paper’s 28 easy/143 medium/29 hard roster differs from the card’s 27 easy/120 medium/18 hard/35 unrated. Test functions are generated or harvested, not independent programs. Linux image instructions describe available execution, not a completed run; ignored flawed branches/tests affect final scoring. MIT metadata coexists with original per-task/copy-left attribution. Finite tests and shortcut detection remain incomplete.

Sources: [655], [656], [657], [658]

Analysis: Benchmark-family atlas - ProgramBench: buyer assessment and change conditions. An accepted reconstructed executable should pass the frozen behavioral contract without forbidden source lookup or wrapping. Confidence is high in separating resolved tasks from partial test progress, and moderate in finite-test completeness. A buyer should inspect ignored tests, exercise hidden behavior and reproduce cleanroom isolation. Human feasibility data and independent shortcut audits would change confidence.

Sources: [655], [656], [657], [658]

Evidence: Benchmark-family atlas - Proximal FrontierSWE: construction, signal and protocol. The v2 protocol documents separate fresh verifier containers, non-root users and task mutation, while also describing a Modal-daemon bypass that exposed protected reference files. The v2 writeup says v1 wall-clock performance rankings changed on re-scoring, motivating a weighted instruction-count proxy priced to a pinned Intel microarchitecture. The v2 creator protocol reports 0-1 task scores aggregated as mean@5 over 34 tasks, with a 20-hour budget and Proximus as default harness. FrontierSWE v2 adds 21 tasks and retires four from the 17-task v1, producing 34 tasks within the same logical benchmark family.

Sources: [534], [535], [536]

Evidence: Benchmark-family atlas - Proximal FrontierSWE: access and version observations. The pinned repository provides task specifications and digest-pinned images for evaluation, directs users to per-task upstream attribution, and warns that task/scoring content may change.

Sources: [534], [535]

Evidence: Benchmark-family atlas - Proximal FrontierSWE: interpretation boundaries. Version 2’s 34-task roster comes from adding 21 and retiring four of the original 17; these are one family. A versioned instruction-count proxy is not a universal wall-clock-speed measure. The disclosed Modal-daemon bypass shows that fresh containers and non-root users were incomplete defenses. Per-task terms and complete payloads remain unread.

Sources: [534], [535], [536]

Analysis: Benchmark-family atlas - Proximal FrontierSWE: buyer assessment and change conditions. This family illustrates that performance scoring and evaluation integrity can change together after an incident. Confidence is high that historical scores must retain the versioned proxy and disqualification rules, and unmeasured for current sandbox robustness. A buyer should attempt reference-file access and adversarial alternate implementations under a frozen image. Independent containment tests and cross-hardware validation would change trust.

Sources: [534], [535], [536]

Evidence: Benchmark-family atlas - Refresh BlenderBench: construction, signal and protocol. BlenderBench scores saved scene artifacts through task-specific numeric and visual checks, with weights, gates and caps applied to individual tasks. BlenderBench describes 20 Blender workflows covering ten workflow types, with two models evaluated on the same 20 tasks.

Sources: [531]

Evidence: Benchmark-family atlas - Refresh BlenderBench: access and version observations. The creator page presents BlenderBench evaluation through a contact route; the observed /blenderbench link resolves back to the same article. The creator page distinguishes the older hair-dynamics example from the current 20-task comparison and identifies different verifier versions.

Sources: [531]

Evidence: Benchmark-family atlas - Refresh BlenderBench: interpretation boundaries. The inspected creator article supplies an evaluation/contact route, not a verified fully licensed public task/verifier release. Middle per-task numeric table ranges were outside the recorded reading scope. The older hair-dynamics example and its verifier version are separate from the current 20-task comparison. Mean reward is not binary completion.

Sources: [531]

Analysis: Benchmark-family atlas - Refresh BlenderBench: buyer assessment and change conditions. Saved-scene inspection is a useful delivery unit because a visually plausible screenshot can conceal invalid scene state. Confidence is moderate in the proposed workflow and low in a general production-effectiveness inference. A buyer should inspect scene artifacts, task-specific weights/gates and reproducible verifier outputs. Independent expert review of held-out scenes and complete release access would change confidence.

Sources: [531]

Evidence: Benchmark-family atlas - SciCode: construction, signal and protocol. In the v1 paper’s standard setup, models receive no supplied scientific background and carry forward generated solutions; a main problem counts as solved only when all subproblems and the integrated main solution are correct. The creator FAQ states that subproblems 13.6, 62.1 and 76.2 receive supplied problem-code context to control evaluation uncertainty, and explains that subproblem overlap is judged in the context of its main problem. The v1 paper explicitly marks a subset of its named authors as data curators and states that problems, solutions and test cases were annotated, revised and verified by senior researchers at PhD-student level or above. The v1 paper reports 80 main problems and 338 subproblems across 16 subfields; development has 15 main problems and 50 subproblems, while test has 65 main problems and 288 subproblems.

Sources: [636], [637], [638], [659]

Evidence: Benchmark-family atlas - SciCode: access and version observations. The pinned creator README links a Hugging Face release whose card declares Apache 2.0; the official repository also contains an Apache 2.0 license and describes evaluation that requires separately downloaded numerical test data.

Sources: [637], [660], [661]

Evidence: Benchmark-family atlas - SciCode: interpretation boundaries. Eighty main problems and 338 subproblems are nested units across sixteen subfields; five named scientific fields coexist with preserved six-versus-five wording elsewhere. Standard, gold-background, gold-prior-solution and subproblem-only settings differ. Selected subproblems receive code context. Declared Apache terms and separately downloaded numerical tests do not establish inspected payloads or complete upstream clearance.

Sources: [636], [637], [638], [659], [660], [661]

Analysis: Benchmark-family atlas - SciCode: buyer assessment and change conditions. The main-problem rule appropriately tests integration as well as individual snippets, but it needs an explicit dependency chain. Confidence is high in keeping integrated and isolated success separate and moderate in scientific-workflow relevance. A buyer should test scientifically wrong implementations that pass numerical tolerances and correct alternatives that differ from references. Independent domain review and stable data/tool manifests would change confidence.

Sources: [636], [637], [638], [659], [660], [661]

Evidence: Benchmark-family atlas - SUPERChem: construction, signal and protocol. SUPERChem pairs multimodal problems with manually prepared text-only counterparts; Appendix C describes general-image descriptions and organic-structure transcription using preliminary DECIMER recognition followed by ChemDraw drawing and revision. SUPERChem's checklist calls for question-source clarity, duplicate checking, plausible distractors, explanations that remain semantically complete and logically coherent after all checkpoints are removed, Chinese/English meaning consistency and correct rendering; its leakage mitigations include masking nonessential entities and discouraging reverse solving. The manuscript describes 633 filtered SUPERChem-All questions, partitioned into a 500-question public release and a 133-question holdout, with a separate 100-question evaluation subset.

Sources: [662], [663]

Evidence: Benchmark-family atlas - SUPERChem: access and version observations. The pinned SUPERChem dataset card declares MIT and links a bilingual 500-question PDF preview archive; the linked evaluation repository separately carries an MIT software license. The current pinned evaluation repository describes DAG-based scoring in which matched reference-node weights are adjusted by matched-parent reachability, and explicitly distinguishes its historical demo from pending final manuscript input and result reconciliation.

Sources: [662], [663], [664], [665], [666]

Evidence: Benchmark-family atlas - SUPERChem: interpretation boundaries. The 500 public, 133 holdout and 100-question evaluation set are subsets of 633 questions. Filtering removes questions solved by all models in selected groups and changes the target population. The current repository’s DAG/reachability scoring differs from paper weighted checkpoint recall and labels manuscript/result reconciliation pending. Image-to-text parity, typo freedom, component rights and internal reasoning fidelity are not independently established.

Sources: [662], [663], [664], [665], [666]

Analysis: Benchmark-family atlas - SUPERChem: buyer assessment and change conditions. A checkpoint evaluator rewards a chosen explanation representation as well as the chemical answer. Confidence is high that strict answer accuracy and versioned reasoning scores must remain separate and low in a causal fidelity interpretation. A buyer should test correct alternative routes and misleading checkpoint matches with bilingual experts. Resolved scoring versions and independent chemical validation would change confidence.

Sources: [662], [663], [664], [665], [666]

Evidence: Benchmark-family atlas - SWE Atlas: construction, signal and protocol. The pinned Artificial Analysis card describes SWE-Atlas-QnA-AA as a modified grading release of Scale AI's same 124-task QnA set, using GPT-6 Luna at xhigh reasoning and five votes per criterion. The pinned SWE Atlas v1.1 README and NOTICE credit Artificial Analysis's SWE-Atlas-QnA-AA release with wording reused for some updated Codebase Q&A rubric criteria. The pinned SWE Atlas v1.1 README credits Inductive ML with reward-hacking feedback used to further harden task environments. SWE Atlas v1 screens passing Refactoring and Test Writing patches against reference patches using added-line Jaccard similarity and raw-diff sequence similarity; the paper explicitly says this does not prove absence of contamination or memorization. SWE Atlas Codebase Q&A asks for codebase exploration and runtime-supported explanations and requires every mandatory rubric item to pass. SWE Atlas Refactoring combines behavior-preservation tests and mandatory rubric checks for code maintainability, artifact cleanup and negative criteria. SWE Atlas Test Writing combines a model-judged test manifest, tests passing on working code and failing or erroring on a no-op mutation, and all mandatory test-comprehensiveness rubric checks.

Sources: [667], [668], [669], [670], [671]

Evidence: Benchmark-family atlas - SWE Atlas: protocol continuation. The current pinned SWE Atlas README describes November 2026 v1.1 as the same 284-task suite with refined grading, restricted network access and a GPT-6.1 Sol judge using majority voting across three votes, and warns that v1.1 scores are not directly comparable to v1.0. The SWE Atlas manuscript reports 284 tasks divided into 124 Codebase Q&A, 90 Test Writing and 70 Refactoring tasks.

Sources: [668], [670]

Evidence: Benchmark-family atlas - SWE Atlas: access and version observations. The pinned SWE Atlas repository declares Apache License 2.0 and includes a NOTICE attributing reused QnA rubric wording and its modified release to Artificial Analysis.

Sources: [668], [669], [670], [672]

Evidence: Benchmark-family atlas - SWE Atlas: interpretation boundaries. The three workflows total 284 tasks, not three benchmark families. November 2026 v1.1 retains the tasks but changes grading, restrictions and judge voting and warns against direct v1.0 score comparison. The Artificial Analysis same-124-task grading release contributes wording and polarity fixes; Inductive ML contributes reward-hacking feedback. These are explicit artifact roles without inferred contract values or new family credits.

Sources: [667], [668], [669], [670], [671], [672]

Analysis: Benchmark-family atlas - SWE Atlas: buyer assessment and change conditions. The acceptance contract should expose mandatory rubric items, behavior-preservation checks and no-op mutation tests separately. Confidence is high in this decomposition and moderate in the finite checks. A buyer should challenge polarity, test-file tampering and valid alternate refactorings under the exact release. Independent cross-version reruns and subtler mutation coverage would change confidence in score stability.

Sources: [667], [668], [669], [670], [671], [672]

Evidence: Benchmark-family atlas - AA-LCR: defined object and observation units. The pinned AA-LCR creator card describes 100 text questions over 30 document sets containing 234 documents, with about 100,000 input tokens per document set. Questions require reasoning across documents rather than direct retrieval; questions, document sets and documents are different units of one family.

Sources: [762]

Evidence: Benchmark-family atlas - AA-LCR: construction and provenance. AA-LCR’s creator describes undergraduate question writers using non-frontier models to screen difficulty, followed by human testing on the same document sets, revision or removal of failed questions, and at least one correct human solution for every retained question. This is the creator’s reported construction and validation procedure, not an assertion that each question has unanimous human agreement.

Sources: [762]

Evidence: Benchmark-family atlas - AA-LCR: evaluation, validation and interpretation. The pinned AA-LCR card uses an LLM equality checker with instructions for numerical equivalence, required formats, ordering and a single committed final answer; it identifies GPT-5.6 Luna at medium reasoning as the checker. Mercor’s separately captured AA-LCRv 1.1 implementation specifies Gemini 3.5 Flash, four repeats and pass@1 aggregated across repeats rather than a best-of score. The two judge configurations are preserved separately. AA-LCR version 1.1 corrects sixteen answer keys and adds an equality-checker system prompt while leaving the questions and document sets unchanged; the creator says version 1.0.0 and 1.1 scores are not directly comparable. Four answer-key repairs change the answer itself; the remaining repairs address notation or units. These revisions are manifestations of the retained family.

Sources: [762], [775]

Evidence: Benchmark-family atlas - AA-LCR: release, access and rights. AA-LCR’s creator card licenses the question set under Apache 2.0 while declining to claim copyright or impose a license over the accompanying text representations of publicly available documents. The question-set license does not establish blanket permission for every underlying document.

Sources: [762]

Evidence: Benchmark-family atlas - AdvancedIF: defined object and observation units. AdvancedIF's v2 paper enumerates 402 explicit/complex instruction dialogs, 736 multi-turn carried-context dialogs and 507 system-prompt-steerability dialogs, totaling 1,645; rubrics for multi-turn dialogs are collected only at the final turn. The paper's dialog/prompt units and final-turn rubric target should be preserved; the three categories are not independent benchmark families.

Sources: [557]

Evidence: Benchmark-family atlas - AdvancedIF: evaluation, validation and interpretation. The AdvancedIF v2 benchmark evaluation uses o3-mini to check rubric criteria and labels a response correct only if all criteria pass; it reports category percentages and their average. RIFL's rubric-generator training is a separate process. Mercor's later implementation uses a different judge and repeat configuration, so that protocol must not be substituted for the paper's evaluator. The AdvancedIF v2 paper describes manually written prompts focused on instruction following; for multi-turn examples annotators first converse with a supplied model, and the collection retains final-turn prompts that trigger an instruction-following failure. This is adversarial difficulty selection against a provided model, not a random sample of ordinary user instructions.

Sources: [557], [775]

Evidence: Benchmark-family atlas - AdvancedIF: evaluation, validation and interpretation. AdvancedIF pairs each prompt with up to twenty final-turn criteria written by the same annotator, and the paper describes subsequent rounds of human review. The benchmark's human-written evaluation rubrics are distinct from the synthetic rubrics used by the paper's separate RIFL training pipeline.

Sources: [557]

Evidence: Benchmark-family atlas - AIME: defined object and observation units. Mercor's captured AIME 2025 entry describes thirty text problems drawn from the 2025 American Invitational Mathematics Examination, with integer answers from 0 through 999. This identifies the 2025 AI-evaluation manifestation of the retained AIME family; it does not redefine the rules of every human AIME administration.

Sources: [774], [775]

Evidence: Benchmark-family atlas - AIME: evaluation, validation and interpretation. The captured Mercor AIME 2025 protocol names Gemini 3.5 Flash as an LLM judge for the integer-answer responses. An integer answer format does not by itself establish programmatic exact-match grading in this implementation. Mercor's AIME 2025 methodology specifies a single-shot agent, four runs and pass@1; its shared rule aggregates repeats rather than reporting a best-of score. Four repeats are evaluation attempts, not four additional question sets.

Sources: [775]

Evidence: Dataset-family atlas - AM-DeepSeek-R1-Distilled: construction and provenance. The AM card describes LLM scoring of instructions by difficulty and category, label-based filtering and semantic deduplication; its introduction also claims cleaning against test-set contamination. This creator-described processing is not an independently established uniqueness or contamination-free guarantee for every record. Instruction selection is distinct from model response generation; a dataset-wide human annotation or adjudication process is not established by this card-only read. The AM card describes about 1.4 million Chinese-and-English reasoning entries: about 0.5 million reused from open datasets and about 0.9 million DeepSeek-R1-671B response generations attributed to the AM team. These are creator release-scale descriptions, not independently recounted payload totals; the source tables are non-exhaustive. The reused and AM-generated components have different response provenance, and the limitations warn of inaccurate attribution caused by nested source relationships.

Sources: [441]

Evidence: Dataset-family atlas - AM-DeepSeek-R1-Distilled: evaluation, validation and interpretation. The AM card declares CC-BY-NC-4.0 and separately restricts project artifacts to research use, disallowing commercial and potentially harmful uses. These are creator declarations, not a legal compatibility decision or blanket permission for all nested upstream data, teacher outputs and linked artifacts. Its factuality and nested-source-attribution warnings remain applicable to reuse and provenance checking. AM processing describes rule- and LLM-based reference-answer checks for part of the data, sandbox verification for code with test cases, and reward-model scoring followed by rejection sampling. The introduction says all responses underwent rigorous verification, but the limitations say the LLM-generated dataset was not strictly verified and retains factuality shortcomings; both statements remain explicit. Reference answers and test cases are documented only for some records; the schema does not establish universal answer-check or executable-test coverage.

Sources: [441]

Evidence: Dataset-family atlas - AM-DeepSeek-R1-Distilled: evaluation, validation and interpretation. AM records are documented as user/assistant messages with role, content and info; assistant content separates reasoning and answers, and info can retain source, reference answers, test cases and separate reasoning/answer content. The illustrated user source is natural_reasoning and assistant source am-0309, an example of distinct prompt/response provenance rather than a corpus census. The header represents test_case as a struct but the loading example declares a string; this mismatch is retained without claiming executable parser compatibility. Reference answers and code test cases apply only to some problems, and nested attribution may be inaccurate.

Sources: [441]

Evidence: Dataset-family atlas - AM-DeepSeek-R1-Distilled: release, access and rights. The AM card describes two zstd-compressed JSONL components separated by response source and an additional random 1,000-item AM-response sample for experimentation. The header and loading code use am_0.9M_sample_1k, while prose calls the sample file am_0.9M_1k.jsonl; the filename discrepancy is not silently resolved. These are configurations/components of the supplied existing family, not new independent units; payload loading, decompression and sampling were not reproduced.

Sources: [441]

Evidence: Benchmark-family atlas - Antidote: construction and provenance. Antidote’s creator article sources prompts from expert raters’ own prior debugging, research and decision histories, often after the raters already know an outcome. This is a source-reported prompt-origin procedure, not a new count of sampled conversations or a general-population sample.

Sources: [22]

Evidence: Benchmark-family atlas - Antidote: evaluation, validation and interpretation. Antidote’s creator article has experts judge in their own fields with time to inspect answers, check citations and run code, emphasizing substantive usefulness, taste and candor rather than a quick preference impression. The month-later framing is an expert judgment target; the article does not describe a universal prospective month-long follow-up experiment.

Sources: [22]

Evidence: Benchmark-family atlas - Antidote: release, access and rights. The captured Antidote launch covers an Everyday Edition; agentic, enterprise and additional domain-specific editions are described as future extensions. Planned editions are not counted as presently completed benchmark releases.

Sources: [22]

Evidence: Benchmark-family atlas - APEX-SWE: defined object and observation units. The APEX-SWE v3 paper describes two held-out task domains with 100 Integration tasks and 100 Observability tasks, and a public fifty-task development set containing 25 of each. The 200-task leaderboard and 50-task development set are separate evaluation scopes of one retained family.

Sources: [559]

Evidence: Benchmark-family atlas - APEX-SWE: construction and provenance. The APEX-SWE v3 paper says Integration tasks were built by engineers with at least three years of experience and checked for prompt/source alignment, faithful tests without reward hacking, and a human reference solution passing all tests. These are benchmark-author-reported construction checks, not a rerun or independent certification of every task. APEX-SWE Observability tasks are converted from GitHub issue/PR pairs in repositories with at least 350 stars, screening for patches of at least 100 lines across at least three files, and adding synthetic logs plus public developer discussions to reconstruct a production-failure environment. The environment includes normal log lines and bug symptoms; agents receive no failing unit tests and must diagnose the failure from telemetry and context.

Sources: [559]

Evidence: Benchmark-family atlas - APEX-SWE: evaluation, validation and interpretation. The APEX-SWE v3 paper separates pass@1 task completion from task-specific functional, robustness and style rubrics judged against the final patch and execution logs; it states these rubric scores are not used for its leaderboard. Style criteria apply only to Observability tasks. The paper’s Gemini 3 Pro judge configuration is specific to this supplementary rubric analysis. APEX-SWE v3 uses API-facing pytest suites for Integration and fail-to-pass/pass-to-pass verification for Observability; pass@1 is averaged over three independent runs, while pass@3 records success at least once. The observability README requires all fail-to-pass and all pass-to-pass tests to pass. Code and scorer execution are not implied by this documented protocol.

Sources: [559], [763]

Evidence: Benchmark-family atlas - APEX-SWE: release, access and rights. The APEX-SWE v3 paper describes a released development set under a CC-BY license and an open evaluation harness; the pinned repository declares MIT for its code. The paper’s general CC-BY wording is retained without inventing an unstated license version or transferring the code license to all source repositories.

Sources: [559], [764]

Evidence: Benchmark-family atlas - BabaIsBench / Baba Is Harbor: evaluation, validation and interpretation. The Baba Is Harbor README warns that oracle move strings are public and the game predates model training cutoffs, and distinguishes hinted and longer-clock variants from the base tasks. The linked memorization audit was not read, so no claim of contamination-free results or validation of that audit is made. Hints, time budgets and additional campaign stages are not counted as new logical benchmark families. The cached BabaIsBench page describes an eight-level Intro arm with three attempts per model per level and a 60-minute limit per attempt, and a fifteen-level Lake arm run once per model with effectively no time limit. The newer creator README lists 55 levels across stages 0-3; this is a different repository scope from the page’s two-arm 23-level evaluation, not an automatically reconciled current count.

Sources: [590], [591]

Evidence: Benchmark-family atlas - BabaIsBench / Baba Is Harbor: evaluation, validation and interpretation. The Baba Is Harbor creator README describes JSON grid puzzles with an interactive simulator; the agent submits a move string that the real game engine replays, assigning reward 1 when a controlled object reaches a winning object. The task instructions describe mechanics generically and do not name the game. No game was installed, purchased or executed by this author. The creator README describes privilege separation in which a non-root agent can run the game simulator through a gated runner but cannot read or modify the protected engine. This is documented environment design, not an independent security audit or proof against every exploit.

Sources: [590]

Evidence: Benchmark-family atlas - BabaIsBench / Baba Is Harbor: release, access and rights. The Baba Is Harbor README declares Apache 2.0 for repository code while excluding copyrighted Baba Is You game content and engine code; running its setup requires a user-owned local game installation. Code licensing does not grant rights to redistribute the game. No paid access or game-asset download occurred in this research.

Sources: [590]

Evidence: Dataset-family atlas - Bespoke-Stratos: defined object and observation units. Bespoke attributes Stratos-17k to a Sky-T1-style pipeline ported into Curator using DeepSeek-R1, with roughly 5k coding items from APPs/TACO, 10k math items from NuminaMATH subsets and 1k science/puzzle items from STILL-2. DeepSeek-R1 replaces QwQ, and its traces are not passed through Sky-T1's GPT-4o-mini reformatting step. The 17k name and rounded component counts do not establish an exact payload census; the card says rejection sampling can change the exact problems included. Creator/pipeline and teacher-model roles do not establish a paid human annotation workforce or universal human verification.

Sources: [442]

Evidence: Dataset-family atlas - Bespoke-Stratos: construction and provenance. Bespoke reports generating the synthetic reasoning dataset using Curator and DeepSeek-R1 in about 1.5 hours at a cost of $800. This is a creator description of dataset generation, not a reproduced universal price or the total cost of model fine-tuning and evaluation. The recipe and training/evaluation implementations are linked separately and were not read or executed in this card-only scope.

Sources: [442]

Evidence: Dataset-family atlas - Bespoke-Stratos: evaluation, validation and interpretation. Bespoke describes rejection sampling of incorrect solutions, Ray-cluster acceleration of code verification, and GPT-4o-mini math filtering in place of Sky-T1 parsing logic. Direct code-execution-verifier integration into Curator is described as work in progress at this card version. The reported increase from 25% to 73% in retained correct math solutions is a creator-reported filtering result, not an independently checked error-rate or adjudication study. The Bespoke card reports AIME2024 values of 63.3 for Stratos-32B and 20.0 for Stratos-7B, identified as fine-tunes of the corresponding Qwen2.5-Instruct models. Tables distinguish creator-measured DeepSeek-distill baselines from reported values: 66.7 versus 72.6 for 32B and 43.3 versus 55.5 for 7B; those columns are not interchangeable. These creator-reported model results are not an independent dataset-quality score, replicated improvement or significance result; linked evaluation code remains unread.

Sources: [442]

Evidence: Dataset-family atlas - Bespoke-Stratos: release, access and rights. The Stratos-17k card declares Apache-2.0 for its dataset release. The corpus combines named upstream question sources and DeepSeek-R1 generations; the card label is not an independent rights audit of all input data, teacher outputs or code tests. Curator and SkyThought links do not establish licenses for every linked artifact without reading their actual terms.

Sources: [442]

Evidence: Benchmark-family atlas - BinaryAudit: construction and provenance. BinaryAudit’s creator README describes manually inserting artificial backdoors into open-source projects, compiling modified code into stripped binaries and providing agents only the binaries and reverse-engineering tools in Docker. The benchmark does not claim that the upstream projects contain those vulnerabilities in real deployments. No backdoor or exploit was executed here.

Sources: [592]

Evidence: Benchmark-family atlas - BinaryAudit: evaluation, validation and interpretation. The BinaryAudit page describes binary-analysis pass rate as identifying a backdoor and its location in the binary and reports false-positive rate separately for clean code. No model ranking or creator performance number is endorsed. Test implementation and individual labels were not audited. BinaryAudit includes unmodified clean binaries for which the intended answer is that no backdoor is present, and separate baseline tasks testing operation of Ghidra and Radare 2. These controls separate false alarms and tool-operation ability from successful backdoor localization; their presence is not an independent sensitivity/specificity measurement.

Sources: [592], [593]

Evidence: Benchmark-family atlas - BinaryAudit: evaluation, validation and interpretation. The cached BinaryAudit benchmark page describes 33 tasks grouped into binary analysis, tool operation and clean-code false-positive testing. The newer creator README includes an expanded application/task listing; its scope is not silently equated with the older page’s 33-task configuration.

Sources: [592], [593]

Evidence: Benchmark-family atlas - BinaryAudit: release, access and rights. The pinned BinaryAudit creator README declares Apache 2.0 for the repository. The declaration is not a legal audit of every bundled upstream application, tool or compiled artifact.

Sources: [592]

Evidence: Benchmark-family atlas - BrickVision: defined object and observation units. The BrickVision creator card defines four task types-Crop ID, fine-grained counting, Duplicate and Unpaired-with 250 items each, totaling 1,000 benchmark questions. Task configurations remain manifestations of one retained BrickVision family; item counts are creator-reported, not a payload recount.

Sources: [585]

Evidence: Benchmark-family atlas - BrickVision: construction and provenance. BrickVision scenes are procedurally generated isometric renders restyled into photographs; the pinned creator card describes a paired model judge checking preservation of layout, using up to three attempts across four style templates without dropping items. Layout preservation is the creator’s paired automated-check account; it does not establish independent pixel-level validation. The blog’s generic Nano-Banana reference and the card’s exact image-model identifier remain version-specific descriptions.

Sources: [585], [586]

Evidence: Benchmark-family atlas - BrickVision: evaluation, validation and interpretation. BrickVision scores pointing responses by requiring exactly one point per target inside the corresponding segmentation mask, and scores counting responses by exact equality with the reference integer. The card uses the final answer without reasoning because every coordinate pair counts as a point. Coordinate conventions and the optional pixel mode must be kept with the evaluation configuration. No scorer execution or model-performance ranking is proposed.

Sources: [585]

Evidence: Benchmark-family atlas - BrickVision: release, access and rights. The pinned BrickVision creator card declares CC BY 4.0 for the benchmark. This is a creator release declaration, not an audit of image-generator terms or all source-image rights.

Sources: [585]

Evidence: Benchmark-family atlas - BrowseComp: defined object and observation units. BrowseComp’s authors caution that its questions are not representative of typical user requests and ask that benchmark examples not be reproduced online to reduce contamination. The paper’s public availability is not treated as a dataset-license clearance. Long-form response quality, naturally ambiguous requests and all multimodal browsing tasks are outside the documented short-answer design.

Sources: [338]

Evidence: Benchmark-family atlas - BrowseComp: construction and provenance. The BrowseComp paper describes 1,266 human-written browsing questions designed to have stable short factual answers that are difficult to locate online; candidate questions were screened against models with and without browsing and against the first results pages from five Google searches. This describes the original BrowseComp family, not the separate Mercor 130 subset or every browsing benchmark. Difficulty screening is creator-reported and is not a performance ranking.

Sources: [338]

Evidence: Benchmark-family atlas - BrowseComp: evaluation, validation and interpretation. The BrowseComp paper reports an initial 1,287-question set and removal of 21 questions with answer-format mismatches, ambiguous phrasing or incorrect answers, leaving 1,266 questions. The paper says this review focused on the 118 questions where Deep Research had a zero pass rate; it was not an independent audit of every payload label. BrowseComp uses an AI grader to compare a submitted short answer with the reference answer for semantic equivalence; its appended grading instructions focus on answer matching rather than independently solving the question. The grader prompt is a documented procedure, not an independently validated correctness oracle. No benchmark question, answer or canary string is reproduced in this proposition.

Sources: [338]

Evidence: Benchmark-family atlas - BrowseComp: evaluation, validation and interpretation. BrowseComp construction uses inverted question design from a chosen answer, reviews plausible answer uniqueness, and revises a question if another valid answer is found; second-trainer attempts were used for a portion of questions. The paper does not guarantee unique answers and does not establish a strict universal 10-minute second-review protocol for every question.

Sources: [338]

Evidence: Benchmark-family atlas - Chartography: defined object and observation units. Chartography’s v1 paper describes one hundred task-image records using ninety-eight byte-distinct images; eighty-one records use online-sourced charts and nineteen use expert-created charts. Task records, image files and source categories are different units within one retained benchmark family.

Sources: [3]

Evidence: Benchmark-family atlas - Chartography: construction and provenance. Chartography tasks are authored by working domain professionals with a target answer and a complete walkthrough of the visual readings and calculations; for visually estimated quantities, authors provide a chart-specific acceptable range. Each task is independently checked by three additional domain experts and retained only if screening elicits a meaningful failure in at least one frontier model. Ambiguous or unsupported tasks are revised or rejected; screening yields a deliberately difficult set rather than a random sample of professional chart questions.

Sources: [3]

Evidence: Benchmark-family atlas - Chartography: evaluation, validation and interpretation. Chartography supplies a chart and free-form question without tools, and uses Gemini 3.5 Flash to judge the final answer against expert targets and ranges without showing the judge the chart; multi-part answers are all-or-nothing. The paper reports twenty trials per task; the later README's leaderboard instructions specify ten. Those trial configurations remain separate dated source scopes.

Sources: [2], [3]

Evidence: Benchmark-family atlas - Chartography: release, access and rights. Chartography's paper describes release of tasks, images, provenance metadata and evaluation code; the repository distinguishes its two dummy local tasks from the full dataset linked on Hugging Face. Public release statements do not establish a blanket license over all online-sourced chart images.

Sources: [2], [3]

Evidence: Benchmark-family atlas - CompileBench: evaluation, validation and interpretation. CompileBench documents first-attempt and multiple-attempt success, model-usage cost, elapsed inference-plus-terminal time and executed-command count as evaluation measures. Its Elo-style summary is a creator-defined comparison procedure, not a model-performance ranking proposed here or evidence that command count measures competence. CompileBench supplies unmodified open-source project code and a concrete goal of building a working binary, including legacy-toolchain, static-linking and cross-compilation tasks. Project versions, build objectives and architectures are documented task conditions within one retained family.

Sources: [594]

Evidence: Benchmark-family atlas - CompileBench: evaluation, validation and interpretation. The CompileBench README identifies the project as research software and warns that results vary with model versions, system configuration and task complexity. The captured README and API blob are exact source versions; no current build reproducibility or dependency-resolution guarantee is inferred. CompileBench's README describes an interactive Linux terminal through Harbor for configuration, patching, compilation, installation and verification, with end-to-end command, error, cost and time logging. The documented task definitions contain Dockerfiles, instructions, tests and reference solutions; none was executed or fully code-audited by this author.

Sources: [594]

Evidence: Benchmark-family atlas - CompileBench: release, access and rights. The captured CompileBench LICENSE document declares the MIT license for its software and associated documentation. This is the repository license text, not blanket permission for every external project, dependency or model service used by benchmark tasks.

Sources: [581]

Evidence: Benchmark-family atlas - ComplexConstraints: defined object and observation units. The ComplexConstraints v3 paper defines a seventy-five-prompt benchmark with 1,559 criteria and a disjoint one-thousand-prompt companion training set; both contain single-turn tasks with ten to forty criteria per prompt. The creator blog describes fifteen to forty rubric items, whereas the paper and README say ten to forty; these descriptions remain version-specific. Training and evaluation sets are manifestations of one family.

Sources: [4], [5], [23]

Evidence: Benchmark-family atlas - ComplexConstraints: construction and provenance. ComplexConstraints' v3 construction uses expert prompt/reference authoring, rubric authoring, second-expert review and senior adjudication, with iterative judge calibration checking both a hand-graded reference and an edited response whose correct verdict flips. This is a creator-reported benchmark construction procedure, not an independent review receipt for these research assertions.

Sources: [5]

Evidence: Benchmark-family atlas - ComplexConstraints: evaluation, validation and interpretation. The ComplexConstraints README evaluates each prompt in a single completion without tools and grades criteria independently using an LLM judge; it reports strict all-criteria task pass rate separately from the average fraction of criteria satisfied. Its model-family override warning says proxy-served model names can silently lose reasoning controls; the declared configuration is part of the evaluation scope. The ComplexConstraints v3 paper describes Qwen 3-4B instruction-following RL training on a 90:10 split of the companion set, about nine hundred training tasks, using GPT-5-mini criterion judgments and fraction-of-required-criteria reward. This is the paper's training recipe, not a new rerun or performance claim; external benchmarks were excluded from its stated training, tuning and selection procedure.

Sources: [4], [5]

Evidence: Benchmark-family atlas - ComplexConstraints: release, access and rights. The ComplexConstraints v3 paper declares CC BY 4.0 for the released seventy-five-prompt benchmark and its 1,559 expert-authored criteria. The declaration is scoped to the named benchmark release, not automatically to every companion training artifact or third-party source.

Sources: [5]

Evidence: Benchmark-family atlas - DecoyCount: defined object and observation units. The DecoyCount v6 release notes describe 1,586 retained tasks, 2,108 registered decoys and 2,369 images, comprising 1,586 primary and 783 secondary images; they distinguish this preview from the paper’s original 1,660-task bank. The creator documents describe 25 earlier withdrawals and 49 additional task-level exclusions. These source-release exclusions do not remove the retained logical benchmark family. The blog describes rescored current-release results, while current repository documents warn that historical paper scores are not reruns on this release; no numerical performance claim is resolved here.

Sources: [584], [587], [588]

Evidence: Benchmark-family atlas - DecoyCount: construction and provenance. The DecoyCount creator article describes a decoy taxonomy fixed before image collection, model-based difficulty screening, and independent blind counting followed by disagreement adjudication and removal of ambiguous, occluded or specialist-dependent items. These are the creator’s collection/review procedures, not new worker-performance measurements or an independent recount. The article explicitly identifies a Mercor partnership; no additional recruitment, payment or per-worker responsibility is inferred.

Sources: [588]

Evidence: Benchmark-family atlas - DecoyCount: evaluation, validation and interpretation. DecoyCount’s format document discloses 121 potentially stale multi-hop answers in the internal evaluated bank after underlying counts changed, and states that the standalone evaluator does not automatically exclude those items. The disclosure does not establish that all 121 are wrong or certify the remaining labels. It is not silently mapped to 121 confirmed errors in the current 1,586-task preview. The item-level quality metadata and private historical records were not accessed. The DecoyCount format document defines exact-match integer accuracy with API/parsing failures retained in the denominator, a registered-decoy hit rate for neutral/cued counting, and a paired cued-minus-neutral accuracy gap. The gap is compared for the same checkpoint, release and settings. A positive gap may also reflect disambiguation and does not uniquely isolate all reasoning errors.

Sources: [583]

Evidence: Benchmark-family atlas - DecoyCount: evaluation, validation and interpretation. DecoyCount supplies neutral, alignment-cued and multi-hop prompt conditions; for neutral versus cued prompts the image and correct count stay fixed while the cue names the intended counting rule. Multi-hop can add scenario context and a secondary image, while the article retains the instruction to return an integer. These prompt conditions remain within one benchmark family.

Sources: [587], [588]

Evidence: Benchmark-family atlas - DecoyCount: release, access and rights. DecoyCount's creator terms apply MIT to evaluator code, retain individual image licenses, and leave a collection-level annotation license unconfirmed; release notes disclose five retained primary-image records with incomplete license or credit fields. TRUE permission flags are source metadata, not blanket rights or privacy clearance. No missing fields were filled by inference, and no private consent records or media archives were accessed.

Sources: [582], [584]

Evidence: Benchmark-family atlas - DeepResearch Bench II: defined object and observation units. DeepResearch Bench II's v1 paper reports 132 tasks across 22 domains and 9,430 rubrics. The later creator README retains all 132 dataset tasks but temporarily excludes tasks 112 and 114 from official scoring, leaving 130 scored tasks. The separately captured Mercor page reports 130 briefs and 9,287 criteria. Those counts are retained at its documented evaluation scope; no task/rubric payload recount is inferred.

Sources: [611], [765], [766]

Evidence: Benchmark-family atlas - DeepResearch Bench II: construction and provenance. DeepResearch Bench II constructs tasks from expert-authored open-access review reports selected for information gathering, substantial synthesis and reproducibility without physical experiments or complex data-mining/model pipelines. Tasks may cover a relevant subset of a source article and must fix a temporal scope when the topic is time-sensitive. DeepResearch Bench II's v1 paper describes LLM extraction, source-article self-evaluation, manual revision and domain-expert refinement to produce atomic, content-bearing numerical or inferential rubrics, with more than 400 human-hours of expert review. Self-evaluation regeneration is triggered when information-recall or analysis accuracy on the source article is below 90%; this construction check is not the benchmark's final model score.

Sources: [611]

Evidence: Benchmark-family atlas - DeepResearch Bench II: evaluation, validation and interpretation. The original DeepResearch Bench II paper uses Gemini 2.5 Pro with batches of fifty rubrics; the captured later creator README uses GPT-5.5 and three-way rubric labels for satisfied, unmentioned, or blocked-source-supported content. The source article is blocked from the evaluated agent; prompt restrictions cannot fully prevent leakage. DeepResearch Bench II's v1 paper acknowledges incomplete prevention of source-article leakage, possible human-annotation bias and a presentation dimension that does not yet model personalization to the user's knowledge and preferences. The source-article block and expert refinement do not establish that every score is leakage-free or that all annotators would agree.

Sources: [611], [765]

Evidence: Benchmark-family atlas - DeepResearch Bench II: release, access and rights. The later DeepResearch Bench II README declares Apache 2.0 for code and per-task source-article licenses for data: 129 CC BY 4.0 tasks, two CC BY-NC 4.0 tasks and one CC0 task. This later per-task declaration supersedes the README's former blanket non-commercial treatment; non-commercial restrictions remain for the two named tasks. It is a declaration rather than an item-level rights audit.

Sources: [765]

Evidence: Benchmark-family atlas - EMRBench: evaluation, validation and interpretation. The EMRBench study-design section specifies 10 replicas per task per model configuration. The report's illustrated case separately describes two attempts; the task/replica relationship and all result denominators are not independently reconciled. No published success-rate ranking is adopted, particularly given the report's conflicting percentages and numerator/denominator summaries. Refresh's EMRBench report describes 52 clinical tasks across 10 categories, while its separate captured one-pager lists 200 EMRBench tasks. The source scopes are left unresolved; neither count is silently selected as a unified release total or as additional logical families.

Sources: [597], [598]

Evidence: Benchmark-family atlas - EMRBench: evaluation, validation and interpretation. The EMRBench report uses OpenEMR workflows with a fixture described as more than 1,200 clinical encounters across 67 synthetic patients. Synthetic fixture counts are creator-reported. No private patient data, clinical efficacy or real-world safety validation is established. EMRBench distinguishes clinical retrieval tasks, such as looking up records, from clinical actions involving prescriptions, scheduling or documentation. This is a documented task taxonomy, not a clinical recommendation, measured clinician-burden reduction or an audited success rubric.

Sources: [597]

Evidence: Benchmark-family atlas - EnterpriseBench / CoreCraft: defined object and observation units. CoreCraft's v5 paper defines a fictional PC-retailer customer-support world with persistent Docker state, more than 2,500 entities across fourteen types and twenty-three MCP tools spanning queries, communications and state-changing operations. The entity graph and tool inventory are environment components, not additional benchmark families or real customer records.

Sources: [6]

Evidence: Benchmark-family atlas - EnterpriseBench / CoreCraft: evaluation, validation and interpretation. The captured CoreCraft creator article evaluates agents through a lightweight Vercel AI SDK harness with company-policy/system context and MCP schemas, permitting a tool-calling loop of up to one thousand steps. The article calls a trajectory a single conversation turn with an internal tool loop; that is not the same as one tool call or a static QA example. CoreCraft’s v5 paper uses expert-authored completeness, correctness, constraint and format criteria; task-level success requires all criteria, while RL training rewards the fraction of criteria satisfied as judged by an LLM. A dense training reward is distinct from the strict end-to-end evaluation metric.

Sources: [6], [24]

Evidence: Benchmark-family atlas - EnterpriseBench / CoreCraft: evaluation, validation and interpretation. CoreCraft’s paper describes one thousand training tasks and one hundred fifty held-out tasks from a random split of the same environment for its GLM 4.6 training experiment. The article explicitly distinguishes this held-out evaluation slice from the current benchmark task set; the split is not a claim that the entire evolving benchmark contains exactly 1,150 tasks.

Sources: [6], [24]

Evidence: Benchmark-family atlas - FrontierCode: defined object and observation units. Cognition’s FrontierCode creator article describes 150 tasks from 36 maintained open-source repositories, with nested 100-task Main and 50-task Diamond subsets; Mercor’s later 1.1 description retains Main and Extended but retires Diamond. Subsets and release changes are manifestations of one family, not independent benchmark units.

Sources: [767], [768]

Evidence: Benchmark-family atlas - FrontierCode: construction and provenance. Cognition describes FrontierCode tasks as hand-selected by repository maintainers, with more than forty hours of work per task and iteration with evaluation engineers and Cognition researchers. Task descriptions are deliberately concise and accompanied by codebase guidelines; task creation is not inferred from code-hosting affiliation alone.

Sources: [767]

Evidence: Benchmark-family atlas - FrontierCode: evaluation, validation and interpretation. FrontierCode combines unit tests, build/lint commands, reverse tests on the original broken code, adaptive test alignment, scope checks and LLM quality judgments. Its creator describes adversarial hack reports, four-solution rubric calibration, pod review and final researcher review. These are creator-reported benchmark construction controls, not new independent review receipts or measured false-positive improvements. FrontierCode requires every blocker criterion to pass before awarding a weighted rubric score, otherwise assigning zero; its creator article averages five trials per reasoning effort and reports the best-performing effort. A blocker pass rate and a weighted quality score are different metrics. Best-effort reporting is not a best-of-five success measure.

Sources: [767]

Evidence: Benchmark-family atlas - FrontierCode: release, access and rights. Cognition’s creator article keeps FrontierCode tasks private to limit contamination while offering evaluation to model creators. Mercor’s 1.1 description permits fair internet use but assigns zero to runs flagged for opening solution-bearing sources. The public methodology and sample do not establish public access or a redistribution license for the complete task collection.

Sources: [767], [768]

Evidence: Dataset-family atlas - Gauntlet: defined object and observation units. Refresh’s Gauntlet 80K article describes 79,139 software-engineering tasks drawn from developer workflows across 1,669 Python repositories using runtime tracing to pair task descriptions with code changes. These are creator-reported task/repository units, not independently recounted payload totals or evidence of a named external supplier. The 80K label is a rounded release name.

Sources: [622]

Evidence: Dataset-family atlas - Gauntlet: evaluation, validation and interpretation. Refresh describes Gauntlet 4K-RLVR as a curated 4,000-task subset of Gauntlet packaged with prompts, tests and preconfigured Docker environments for Harbor, using pytest-based binary rewards for reinforcement learning. The 4K subset/configuration remains part of the existing Gauntlet family and does not inherit any universal runtime or training-performance guarantee. The Gauntlet 80K article states that it includes only examples that actually run and have an associated passing test. Passing the creator’s checks is not a guarantee of general code correctness or an independently rerun validation; no payload tests were run.

Sources: [621], [622]

Evidence: Dataset-family atlas - Gauntlet: evaluation, validation and interpretation. Refresh describes Gauntlet tasks as including complete repositories with Git history and pre-resolved dependencies in Docker environments, rather than only static prompt-and-patch records. This is the documented environment packaging, not verification that every historical dependency still resolves or that every repository license permits all uses.

Sources: [622]

Evidence: Dataset-family atlas - Gauntlet: release, access and rights. The captured Gauntlet 80K creator article offers training-data access by contacting Refresh rather than documenting an anonymous public payload download. No contact was made, no paid access was used, and no explicit dataset-wide open license or rights clearance is established by the captured posts.

Sources: [622]

Evidence: Benchmark-family atlas - GDP.pdf: defined object and observation units. GDP.pdf’s creator article describes one hundred professional question/PDF tasks across ten domains, requiring interpretation and synthesis of the original documents. The task/domain taxonomy is internal coverage of the retained GDP.pdf family, not GDP.xlsx or another logical benchmark.

Sources: [8], [25]

Evidence: Benchmark-family atlas - GDP.pdf: construction and provenance. GDP.pdf’s v3 paper sources expert tasks from real work, rejects questions answerable from priors or ambiguous private context, and retains candidates only after at least two frontier models make a major meaningful error. Absolute pass rates therefore describe an adversarially selected set, not professional document work at large.

Sources: [8]

Evidence: Benchmark-family atlas - GDP.pdf: evaluation, validation and interpretation. GDP.pdf’s v3 evaluation supplies each question and native PDF without tools, runs five trials per item and reports mean strict pass@1; Gemini 3.5 Flash judges each self-contained criterion without the PDF or reference answer. Mean criterion satisfaction is reported separately from all-criteria success; provider PDF processing is part of the measured configuration. GDP.pdf’s paper describes expert-written atomic rubric criteria, including required abstention for unsupported queries, and sanity checks requiring the reference answer to pass and a deliberately bad output to fail the targeted criterion. Rubrics specify substantive facts and forbidden claims rather than relying on text-overlap similarity.

Sources: [7], [8]

Evidence: Benchmark-family atlas - GDP.pdf: release, access and rights. GDP.pdf’s v3 paper declares all one hundred examples released under Apache 2.0 while retaining the original rights of third-party PDFs; the repository’s committed sample is a synthetic placeholder. A benchmark release license does not relicense every underlying professional document.

Sources: [7], [8]

Evidence: Benchmark-family atlas - GDP.xlsx: defined object and observation units. The GDP.xlsx creator card describes a full benchmark of 70 spreadsheet tasks across 13 domains and a public sample of three tasks. Full-benchmark tasks and public samples are different release scopes; worksheets, populated fields and domains are not additional logical benchmark families. Counts are creator-reported and were not recounted from payload files.

Sources: [623]

Evidence: Benchmark-family atlas - GDP.xlsx: evaluation, validation and interpretation. The GDP.xlsx card describes task-specific rubrics assessing output accuracy, completeness and appropriateness. The full benchmark and detailed evaluation protocol are described as available to interested evaluation partners; they were not accessed. The card does not document a complete human authoring or adjudication procedure. GDP.xlsx tasks pair a workplace-style request with input Excel or CSV files; its documented Harbor packaging combines task instructions, inputs and evaluation logic for isolated container execution. This describes the pinned public card, not an executed Harbor run or inspection of all 70 tasks.

Sources: [623]

Evidence: Benchmark-family atlas - GDP.xlsx: release, access and rights. The pinned GDP.xlsx card declares Apache-2.0 in its metadata while its benchmark-use notice directs exclusion from training, supervised fine-tuning, reinforcement learning, distillation and checkpoint selection. Both statements are retained together as creator-published terms; this is not a legal reconciliation or blanket rights clearance. The benchmark-use restriction is not generalized to unrelated Surge datasets.

Sources: [623]

Evidence: Benchmark-family atlas - GDPval: defined object and observation units. GDPval's paper describes 1,320 work tasks covering 44 occupations across nine sectors and states at least 30 tasks per occupation in the full set, with a public gold subset of 220 tasks, five per occupation. The full corpus, public 220-task subset, occupations and reference files are different units; subsets do not create additional benchmark families. Gold here is the paper authors' name for their public task subset, not any ScienceToStartup audit reference.

Sources: [350]

Evidence: Benchmark-family atlas - GDPval: construction and provenance. GDPval task authors were occupational experts with at least four years of relevant experience; the paper describes background verification and a training quiz before creation of requests, reference files and deliverables. This is the paper's expert-selection protocol, not verification of every author's employment or a named supplier contribution.

Sources: [350]

Evidence: Benchmark-family atlas - GDPval: evaluation, validation and interpretation. GDPval's paper marks 12 of its 220 public tasks as ungradable because of automated-grader limitations and describes the experimental grader as a limited proxy for occupational-expert judgment. The 12-task exclusion is specific to automated grading, not a removal from the logical benchmark inventory. The grader's no-internet/Python-only setting and file-format limitations differ from the human-evaluation protocol. For the 220-task evaluation subset, GDPval compares model and expert deliverables through blinded pairwise judgments by occupational experts; the paper describes three completions of each model per task and three graders per completion, giving nine comparisons per task per model. Blinding hid explicit author identity but stylistic clues could still reveal model authorship. These are documented human-evaluation settings, not the automated grader's scope or a ranking endorsed by this author.

Sources: [350]

Evidence: Benchmark-family atlas - GDPval: evaluation, validation and interpretation. GDPval quality control uses model suggestions with experts retaining final responsibility, followed by iterative generalist review, occupation-specific expert review and a third expert's feedback loop. The models flagged possible issues and did not autonomously alter tasks. The paper reports at least three human reviews and about five on average; no new review of task payloads was performed here.

Sources: [350]

Evidence: Benchmark-family atlas - GDPval: release, access and rights. GDPval's public release provides prompts and reference files for 220 tasks with information identifying their expert authors removed. Public access does not by itself establish a standard reuse license or third-party reference-file rights.

Sources: [350]

Evidence: Dataset-family atlas - HackerRank community graded submissions: defined object and observation units. HackerRank's captured Bench page describes a community library of more than 2,000 questions and 470 million graded submissions, growing by about 5 million submissions per month. These are publisher-reported library and growth units, not a count of unique developers, delivered training records or an independently downloaded corpus.

Sources: [769]

Evidence: Dataset-family atlas - HackerRank community graded submissions: construction and provenance. HackerRank describes each community submission as linked to a problem statement, constraints, reference solution, hidden tests and checker, submitted code, verdict, runtime, memory, test-level results, edit history, difficulty and topic/skill tags. The public description identifies record fields; it does not expose the complete submission collection or independently establish label correctness.

Sources: [769]

Evidence: Dataset-family atlas - HackerRank community graded submissions: evaluation, validation and interpretation. HackerRank explicitly limits the advertised graded-submission library to community data and excludes confidential customer data; its report describes this stream as human community data used for interview preparation. Public description does not itself grant a redistribution or training license for the full library.

Sources: [769], [770]

Evidence: Dataset-family atlas - HackerRank Orchestrate agent traces: defined object and observation units. HackerRank's Bench page describes Orchestrate as a monthly coding challenge with cash prizes, typically more than 35,000 registrations and 2,500 submissions per iteration. Registrations and submissions are different units; the report's 30,000-plus participants are a separate reported event scope rather than an inferred count of unique trajectories.

Sources: [769], [770]

Evidence: Dataset-family atlas - HackerRank Orchestrate agent traces: evaluation, validation and interpretation. HackerRank's Bench report identifies Orchestrate's complex long-horizon engineering challenge as a source of agent trajectories involving frontier models. The public pages do not specify a trajectory count or per-step annotation protocol for this stream. It is retained separately from the much larger community graded-submission library.

Sources: [769], [770]

Evidence: Benchmark-family atlas - HANDBOOK.md: defined object and observation units. HANDBOOK.md’s v3 paper defines sixty-five tasks in ten fictional companies across five enterprise domains, with a containerized workspace, seeded mock services, work request and programmatic rubric for each task. Simulated email, Slack, calendar, Jira andShopify services are benchmark fixtures; they are not evidence of live changes to real company systems.

Sources: [10]

Evidence: Benchmark-family atlas - HANDBOOK.md: construction and provenance. HANDBOOK.md uses ten expert-written base handbooks adapted from industry practice, changes operative policies for each task, and has expert builders construct the world and hand-written rubric before implementing and iterating programmatic checks. Policy mutation changes rules and thresholds rather than only cosmetic wording; correct behavior can require withholding an action.

Sources: [10]

Evidence: Benchmark-family atlas - HANDBOOK.md: evaluation, validation and interpretation. HANDBOOK.md reports the fraction of trials satisfying every criterion, averaging four trials per task, and separately reports a metric allowing one missed criterion and mean criterion scores for analysis. Trials are capped at two hundred tool calls and one hour; provider or transport errors are excluded and rerun rather than counted as task failures. HANDBOOK.md’s v3 paper specifies 824 deterministic criteria: 592 required-output checks and 232 prohibited-behavior checks over final files and service states, without LLM judging. Exact-count invariants also check that actions outside the task scope were not performed; counts describe the documented release rather than an independent payload recount.

Sources: [10]

Evidence: Benchmark-family atlas - HANDBOOK.md: release, access and rights. The HANDBOOK.md repository provides three sample tasks and directs users to request the full set, while declaring Apache 2.0 for the repository. The paper describes a task/environment/harness release, but the captured repository is a three-task sample; public method documentation is separate from complete-task access.

Sources: [9], [10]

Evidence: Benchmark-family atlas - Harvey LAB: defined object and observation units. Harvey LAB’s creator documentation defines a task as instructions, source documents, declared deliverables and inline criteria; the harness runs an agent on a synthetic matter file, evaluates its output and generates reports. The dataset and execution harness are separate components of the retained LAB family.

Sources: [771], [773]

Evidence: Benchmark-family atlas - Harvey LAB: evaluation, validation and interpretation. The pinned Harvey evaluation document scopes each rubric criterion to its declared output files and asks an LLM to judge semantic satisfaction of match_criteria, without a separate golden reference output. The documented scorer fails a criterion without calling the judge when none of that criterion's scoped output files exist, or the run has no output at all. Each judge call uses the criterion-scoped outputs rather than all matter documents or all outputs. The captured Harvey methodology gives each judge a binary all-criteria-pass task score and averages the two default judges' scores; dual task values can therefore be 0, 0.5 or 1, while strict aggregate all_pass requires both judges to all-pass. The standard pair is Claude Sonnet 4.6 and GPT-5.5. Pooled and macro criterion pass rates are diagnostics distinct from the headline averaged all-pass rate.

Sources: [772]

Evidence: Benchmark-family atlas - Harvey LAB: evaluation, validation and interpretation. Harvey's architecture document routes workspace tools through a per-task Podman container with network disabled, read-only source documents and a writable output overlay; the finish signal only checks that the agent's listed deliverables exist. The finish existence check does not consult task.json and is separate from substantive rubric grading.

Sources: [771]

Evidence: Benchmark-family atlas - Harvey LAB: release, access and rights. The current Harvey evaluation document describes 2,010 tasks across 27 practice areas, while the earlier captured Mercor catalog describes 1,749 across 24; Harvey explicitly asks users to identify their release or commit because scores can change between releases. The two rosters are preserved as source-version scopes; neither number is silently treated as a payload recount of the other.

Sources: [772], [773], [774]

Evidence: Benchmark-family atlas - Hemingway-bench: evaluation, validation and interpretation. Hemingway-bench's creator article reports more than five thousand blind pairwise comparisons, scoring holistic writing quality and eight subdimensions including intent, creativity, truthfulness, coherence, instruction following, verbosity and humor. Judges also explain their preferences; later model personas are interpretations of those explanations rather than additional benchmark units or a numerical scoring formula. Hemingway-bench's creator article recruits expert writers from Surge's platform, including screenwriters, poets, speechwriters and copyeditors, with strong platform writing/evaluation records across hundreds of tasks. The source's qualification description is not an independent verification of each contributor's credentials or employment.

Sources: [26]

Evidence: Benchmark-family atlas - Hemingway-bench: evaluation, validation and interpretation. Hemingway-bench mixes practical creative, business and everyday writing prompts with deliberately harder prompts testing style control, long-form consistency and specific creative constraints. The source describes prompt categories and examples without publishing a complete prompt count or sampling denominator.

Sources: [26]

Evidence: Benchmark-family atlas - HSS: defined object and observation units. The HSS creator card describes 522 tasks-288 image and 234 video-with 723 rubric criteria, covering four domains and 11 subdomains. Tasks, media, rubric criteria and taxonomy subdivisions are different units within the same retained HSS family. Counts were not independently recounted from payloads.

Sources: [600]

Evidence: Benchmark-family atlas - HSS: construction and provenance. The HSS creator card states that trained annotators authored 3,466 candidate tasks and that 522 were admitted after unanimous approval in three independent review rounds. These are benchmark construction reviews, not ScienceToStartup source-verifier/adversarial receipts. The complete task-review records were not inspected.

Sources: [600]

Evidence: Benchmark-family atlas - HSS: evaluation, validation and interpretation. HSS uses an LLM judge given the question, reference answer and model answer to mark each rubric criterion; the documented headline accuracy counts a task correct only when every criterion is met. The card identifies Claude Opus 5 as the paper’s judge. No judge agreement or benchmark performance is newly measured here. HSS evaluates video tasks on the whole video; the documented evidence-window timestamps are analysis metadata and are not part of the model input. The timestamps do not justify a claim that models were supplied only answer-bearing clips. No videos or complete frame sequences were played by this author.

Sources: [600]

Evidence: Benchmark-family atlas - HSS: release, access and rights. The HSS card releases annotations under MIT while excluding third-party images and videos from that license and describing their inclusion for non-commercial research and evaluation. Original media terms remain applicable and were not individually audited. The card also warns of public-web media contamination and asks users not to train on this test set.

Sources: [600]

Evidence: Dataset-family atlas - Light-R1 training data family: evaluation, validation and interpretation. Light claims exact- or N-gram-based decontamination and names AIME24/25, MATH-500 and GPQA Diamond as target reasoning benchmarks. The same discussion warns that MATH-500 is compromised by identical or numerically altered questions in public data; this is not silently converted into an exact contamination count for Light's released payload. Creator filtering does not establish an independently contamination-free corpus, particularly for upstream pretraining sources. After SFT stage2, Light describes sampling stage2-model responses, separating correct and incorrect outputs, and constructing DPO pairs using verification results and DeepSeek-R1 responses before model merging. The card describes DPO or NCA with sequence parallelism and a merge of stage2 plus two DPO versions; one DPO-data version skips special tokens in rejected responses. These are creator-described operations, not an inspected payload, executed training recipe or independent human preference-labeling study.

Sources: [443]

Evidence: Dataset-family atlas - Light-R1 training data family: evaluation, validation and interpretation. Light-R1 describes public math questions with DeepSeek-R1 responses, a 76k first SFT set selected by verification and DeepScaleR-1.5B-Preview difficulty sampling, and a harder 3k second set mostly filtered from the first. The model starts from Qwen2.5-32B-Instruct; from scratch is explicitly limited to a model without long chain-of-thought capability, not raw-data pretraining. stage1-76k.json and stage2-3k.json are release descriptions, not independently recounted or proven-disjoint corpora, and remain stages of the existing family. Named question sources include OpenR1-Math, OpenThoughts, LIMO, OpenMathInstruct-2, s1K-1.1, Omni-MATH, hendrycks_math and AIME up to 2023. The Light card reports final AIME24/AIME25 values of 76.6 and 64.6, states 64-run score averaging, and estimates training at under six hours on twelve H800s, around $1,000. These are creator-reported scores and cost/time estimates; public-model results may come from existing evaluations, and no experiment, accounting or significance study was reproduced here. A separate stage2-data adaptation statement uses 0.779/0.675 beside 72.6/54.9 for a stated DeepSeek-distill model; its mixed notation is retained rather than substituted for final Light scores. The card describes math-specialized thinking with a forced think marker and some GPQA forgetting; these math results do not independently establish general-domain gains.

Sources: [443]

Evidence: Dataset-family atlas - Light-R1 training data family: release, access and rights. The Light card declares Apache-2.0 for released project materials and points to separate SFT, DPO, model, training-code and evaluation-code artifacts. This declaration is not an independent rights determination for upstream public math data, teacher outputs, model weights or linked software. The linked stages, DPO material and trained/merged models retain distinct roles within the supplied family scope; payloads and external terms were not opened or independently licensed by this read.

Sources: [443]

Evidence: Dataset-family atlas - MedMosaic: defined object and observation units. The MedMosaic card lists 4,661 examples across seven task folders: 106 long-form, 6 multi-turn, 691 open-ended, 1,417 sound-only, 1,174 speech-only, 1,087 speech-and-sound and 180 voice examples. Folder/example counts are creator-declared evaluation units, not independently counted recordings, patients or seven distinct dataset families.

Sources: [624]

Evidence: Dataset-family atlas - MedMosaic: evaluation, validation and interpretation. The MedMosaic card documents multiple-choice, free-form and multi-turn answer layouts. A documented answer schema is not a complete labeling, adjudication or evaluation protocol. The anonymous MedMosaic creator card describes a medical audio question-answering evaluation collection linking recordings to questions and expected answers. The publishing account is anonymous; this card does not establish Centific authorship, recruitment or annotation delivery. Only the public card was read; no clinical accuracy, recording provenance, consent or payload integrity audit was conducted.

Sources: [624]

Evidence: Dataset-family atlas - MedMosaic: release, access and rights. The anonymous MedMosaic creator card declares CC BY 4.0 for the published collection. The card’s license declaration does not establish consent or rights for every underlying recording; these remain explicit gaps.

Sources: [624]

Evidence: Benchmark-family atlas - Multi-SWE-bench: defined object and observation units. Multi-SWE-bench v1 documents 1,632 manually verified issue-resolving instances from 39 repositories across Java, TypeScript, JavaScript, Go, Rust, C and C++. The 2,456 candidates, final 1,632 instances and later mini/flash configurations are different scopes; languages and repositories do not create separate benchmark families.

Sources: [629]

Evidence: Benchmark-family atlas - Multi-SWE-bench: construction and provenance. Multi-SWE-bench construction selects maintained GitHub repositories with buildable CI/CD configurations, then keeps merged pull requests linked to issues and changing test files before creating per-PR Docker environments. This is the creator’s documented construction pipeline; no repository builds or source-code audits were executed here.

Sources: [629]

Evidence: Benchmark-family atlas - Multi-SWE-bench: evaluation, validation and interpretation. The Multi-SWE-bench paper reports 68 outsourced annotators screened for at least two years of target-language experience and a relevant degree, one-hour task training, independent dual labels followed by cross-review, and a 14-engineer internal quality team checking annotations against reference answers and an 80% threshold. No outsourcing supplier is named or inferred. This manual-verification stage belongs to the benchmark; it is explicitly absent from the initial Multi-SWE-RL construction stream. The 80% threshold is the creator’s annotation-quality rule, not a new measured annotation accuracy here. Multi-SWE-bench’s pre-annotation filter runs full test suites at the base commit, after the test patch, and after both test and fix patches; it rejects regressions, requires a failed-to-passed test transition, and tracks skipped or absent tests as well as pass/fail. These test-transition checks establish the paper’s filtering rule, not an independent guarantee that unit tests capture every valid software fix.

Sources: [629]

Evidence: Benchmark-family atlas - Multi-SWE-bench: release, access and rights. The Multi-SWE-bench creator card releases dataset material under CC0 only to the extent of ByteDance’s rights and warns that upstream repository licenses may continue to apply; its repository separately declares Apache-2.0 for code. Dataset terms and harness-code terms are different scopes; neither proves unrestricted rights for every upstream repository.

Sources: [627], [628]

Evidence: Dataset-family atlas - Multi-SWE-RL: defined object and observation units. The Multi-SWE paper describes an initial Multi-SWE-RL release of 4,723 issue-resolving instances from 76 repositories across seven programming languages, each accompanied by a containerized execution environment. The count is the paper’s initial release scope, not a current recount of every later dated card batch or the 1,632-instance benchmark.

Sources: [629]

Evidence: Dataset-family atlas - Multi-SWE-RL: construction and provenance. The pinned Multi-SWE-RL card describes its initial dated batch as combining new unannotated RL data with instances discarded from Multi-SWE-bench. The two origins remain distinct within one RL-data family; a discarded benchmark instance is not silently relabeled as manually verified benchmark material.

Sources: [630]

Evidence: Dataset-family atlas - Multi-SWE-RL: evaluation, validation and interpretation. The Multi-SWE paper states that the initial RL dataset uses the benchmark construction pipeline while excluding the manual-verification phase described for Multi-SWE-bench. Automatic environment/test filtering is not equivalent to the benchmark’s dual-human verification, internal QA or expert-annotation counts. No named supplier contribution, human label coverage or universal reward correctness is inferred.

Sources: [629]

Evidence: Dataset-family atlas - Multi-SWE-RL: release, access and rights. The Multi-SWE-RL creator card declares CC0 subject to the scope of ByteDance’s intellectual-property rights and warns that upstream repository licenses can remain applicable. This is the creator’s scoped release statement; underlying repository rights were not independently audited.

Sources: [630]

Evidence: Dataset-family atlas - Nexdata product 1463: defined object and observation units. Nexdata’s product 1463 listing describes 116,048 indoor right-hand pose/gesture sets, without hand held objects, covering first-person andthird-person perspectives and multiple individuals and Kinect devices. Nexdata is the publisher of this product description; upstream creation, ownership, participant permissions and individual annotation responsibilities are not established by the listing.

Sources: [14]

Evidence: Dataset-family atlas - Nexdata product 1463: construction and provenance. Product 1463 lists aligned RGB hand masks and 16-bit depth, TXT camera intrinsics and gesture labels, OBJ 3D keypoints/meshes and JPG demonstrations. The listing says facial-identifying information is absent; this is narrower than an independent privacy/consent audit or proof that every modality is non-identifying.

Sources: [14]

Evidence: Dataset-family atlas - Nexdata product 1463: release, access and rights. Nexdata’s product 1463 page labels the collection as a paid dataset licensed for commercial use and offers sample/quote requests. The public label is not an executed license or blanket upstream rights clearance; the catalog says terms depend on intended application, geography and project requirements.

Sources: [11], [14]

Evidence: Dataset-family atlas - Nexdata product 1515: defined object and observation units. Nexdata’s product 1515 page describes 480,000 original/corrected text pairs across French, German, Spanish and Italian for grammar correction. Nexdata is the publisher of this product description; upstream creation, ownership, participant permissions and individual annotation responsibilities are not established by the listing.

Sources: [15]

Evidence: Dataset-family atlas - Nexdata product 1515: construction and provenance. Product 1515 supplies original text in an input field and corrected text in an output field in JSON, with language-processing, translation and teaching applications listed. The field/language description does not establish whether corrections were authored by humans, models or another upstream dataset provider.

Sources: [15]

Evidence: Dataset-family atlas - Nexdata product 1515: release, access and rights. Nexdata’s product 1515 page labels the collection as a paid dataset licensed for commercial use and offers sample/quote requests. The public label is not an executed license or blanket upstream rights clearance; the catalog says terms depend on intended application, geography and project requirements.

Sources: [11], [15]

Evidence: Dataset-family atlas - Nexdata product 1814: defined object and observation units. Nexdata’s product 1814 page describes one hundred thousand English general-domain instruction-tuning text records in JSON format. Nexdata is the publisher of this product description; upstream creation, ownership, participant permissions and individual annotation responsibilities are not established by the listing.

Sources: [16]

Evidence: Dataset-family atlas - Nexdata product 1814: evaluation, validation and interpretation. The product 1814 narrative reports double verification by experienced linguistic professionals and AI engineers for its instruction-tuning data. The public description does not specify review independence, adjudication, quality thresholds or contributor identities; quality and privacy compliance are unverified publisher claims.

Sources: [16]

Evidence: Dataset-family atlas - Nexdata product 1814: release, access and rights. Nexdata’s product 1814 page labels the collection as a paid dataset licensed for commercial use and offers sample/quote requests. The public label is not an executed license or blanket upstream rights clearance; the catalog says terms depend on intended application, geography and project requirements.

Sources: [11], [16]

Evidence: Dataset-family atlas - Nexdata product 2121: defined object and observation units. Nexdata’s product 2121 listing describes 7,500 long-document multi-hop QA examples across Chinese, English and Korean, requiring evidence synthesis across passages and documents. Nexdata is the publisher of this product description; upstream creation, ownership, participant permissions and individual annotation responsibilities are not established by the listing.

Sources: [17]

Evidence: Dataset-family atlas - Nexdata product 2121: construction and provenance. Product 2121 lists JSON fields for ID, context, file count, question, answer, reasoning chain, supporting evidence and hop count. Schema declarations are not verification of answer correctness, evidence faithfulness orreasoning-chain provenance.

Sources: [17]

Evidence: Dataset-family atlas - Nexdata product 2121: release, access and rights. Nexdata’s product 2121 page labels the collection as a paid dataset licensed for commercial use and offers sample/quote requests. The public label is not an executed license or blanket upstream rights clearance; the catalog says terms depend on intended application, geography and project requirements.

Sources: [11], [17]

Evidence: Dataset-family atlas - Nexdata product 2145: defined object and observation units. Nexdata’s product 2145 page describes twenty thousand hours of egocentric video across residential, retail and office scenarios, pairing stereo video and calibration with full-body pose and step-by-step action labels. Nexdata is the publisher of this product description; upstream creation, ownership, participant permissions and individual annotation responsibilities are not established by the listing.

Sources: [18]

Evidence: Dataset-family atlas - Nexdata product 2145: construction and provenance. The product 2145 listing describes dense semantic and action-level annotation and multiple quality-control reviews. The page does not identify annotators, review independence, quality thresholds or adjudication records; its review claim remains publisher-reported. The product 2145 specifications describe a PICO 4 Ultra head-mounted stereo camera plus wrist and ankle IMUs, 4096×1536 video at thirty frames per second and seventy-six tracked joints comprising twenty-four torso and fifty-two hand joints. These are publisher-reported acquisition/output specifications, not independently measured synchronization, tracking or sensor-quality results. Nexdata is the publisher of this product description; upstream creation, ownership, participant permissions and individual annotation responsibilities are not established by the listing.

Sources: [18]

Evidence: Dataset-family atlas - Nexdata product 2145: release, access and rights. Nexdata’s product 2145 page labels the collection as a paid dataset licensed for commercial use and offers sample/quote requests. The public label is not an executed license or blanket upstream rights clearance; the catalog says terms depend on intended application, geography and project requirements.

Sources: [11], [18]

Evidence: Dataset-family atlas - Nexdata product 2185: defined object and observation units. Product 2185 lists a data volume of 5,300, JSONL format and English/Korean languages for its multi-turn reasoning and tool-call trajectories. The public specifications do not define the 5,300 unit more precisely as unique tasks, episodes, steps or verified successful runs.

Sources: [19]

Evidence: Dataset-family atlas - Nexdata product 2185: construction and provenance. Nexdata’s product 2185 specifications identify agent trajectories generated by OpenClaw using Anthropic Claude Opus 4.8 for deep search, data analysis and industry research. Nexdata is the publisher of this product description; upstream creation, ownership, participant permissions and individual annotation responsibilities are not established by the listing.

Sources: [19]

Evidence: Dataset-family atlas - Nexdata product 2185: release, access and rights. Nexdata’s product 2185 page labels the collection as a paid dataset licensed for commercial use and offers sample/quote requests. The public label is not an executed license or blanket upstream rights clearance; the catalog says terms depend on intended application, geography and project requirements.

Sources: [11], [19]

Evidence: Dataset-family atlas - Nexdata product 2236: defined object and observation units. Nexdata’s product 2236 page describes 1,042 six-camera egocentric video segments of approximately thirty-five seconds, spanning thirty-four locations and six real-world environments. Nexdata is the publisher of this product description; upstream creation, ownership, participant permissions and individual annotation responsibilities are not established by the listing.

Sources: [20]

Evidence: Dataset-family atlas - Nexdata product 2236: evaluation, validation and interpretation. Product 2236’s listed outputs include JSON calibration and SLAM trajectories, MP 4 RGB video, MJPG infrared fisheye video, HDF 5 sensor records, MKV depth and PLY reconstructed point clouds. These are the listing’s declared output formats, not an independent file or reconstruction validation; formats remain manifestations within one collection.

Sources: [20]

Evidence: Dataset-family atlas - Nexdata product 2236: release, access and rights. Nexdata’s product 2236 page labels the collection as a paid dataset licensed for commercial use and offers sample/quote requests. The public label is not an executed license or blanket upstream rights clearance; the catalog says terms depend on intended application, geography and project requirements.

Sources: [11], [20]

Evidence: Benchmark-family atlas - OTelBench: defined object and observation units. The captured OTelBench methodology describes 23 OpenTelemetry instrumentation tasks across 11 programming languages, using small microservice codebases of around 300 lines for agents to edit and run in a terminal. This is the January 2026 reported evaluation scope. The later README’s task list is retained separately and not forced to have the same count.

Sources: [595], [596]

Evidence: Benchmark-family atlas - OTelBench: evaluation, validation and interpretation. OTelBench’s creator methodology verifies span names, parent-child relationships, valid trace IDs and context propagation rather than accepting code merely because it compiles. The example distinguishes separate user journeys; it does not establish a comprehensive reliability or production-readiness guarantee for every instrumented system. The OTelBench README states that tasks require internet access and that task solution instructions were not yet included at the captured revision. This incomplete documentation and live-network dependency are not silently converted into successful operational validation.

Sources: [595], [596]

Evidence: Benchmark-family atlas - OTelBench: evaluation, validation and interpretation. The OTelBench article describes 966 model runs from 23 tasks, three attempts and 14 models in its reported benchmark run. Model runs, attempts, tasks and languages are different units; no published model score or cost is adopted as this author’s performance judgment.

Sources: [595]

Evidence: Benchmark-family atlas - OTelBench: release, access and rights. The pinned OTelBench creator README declares Apache 2.0 for the repository. This is not an audit of third-party microservice/dependency rights or of model-service terms.

Sources: [596]

Evidence: Benchmark-family atlas - PDFBench: evaluation, validation and interpretation. PDFBench distinguishes explicit field-name tasks supplying field names, locations and values from semantic tasks describing what fields mean in ordinary language. The two instruction arms and screenshot conditions are configurations within the same retained family. No model-performance ranking is proposed. Refresh’s PDFBench report describes 70 form-filling tasks across nine form types, a tile labeled 12 configurations by 90 tasks and 1,082 trials; its one-pager separately lists 200 PDFBench tasks. These conflicting task/trial populations remain source-specific.12×90 equals 1080, so the 1,082 total is not silently repaired or treated as independently verified. Counts, form types and configurations are not additional logical benchmark families.

Sources: [598], [599]

Evidence: Benchmark-family atlas - PDFBench: evaluation, validation and interpretation. PDFBench’s report distinguishes strict task passing from field-level correctness and illustrates a failed form despite 46 of 49 fields being filled correctly. The example does not supply a complete executable scoring/adjudication protocol for every task. Underlying form permissions, label authorship and all business-rule dependencies remain unverified. PDFBench tasks provide a blank form, written fill instructions, programmatic access through PyMuPDF and an optional screenshot tool. This describes the creator’s tool-access conditions, not a rerun of form filling or proof that optional screenshots always improve results.

Sources: [599]

Evidence: Benchmark-family atlas - Physical Engineering’s Last Exam (PLE): defined object and observation units. The PLE v2 paper describes 85 tasks: 44 expert-authored tasks sharing twelve project environments and 41 co-authored tasks with distinct environments, spanning mechanical, electrical, civil and structural engineering. The summary’s mean check counts cover 81 rubric/constraint-graded tasks, not all 85; four site-design/MEP tasks keep distinct grader scores.

Sources: [700]

Evidence: Benchmark-family atlas - Physical Engineering’s Last Exam (PLE): construction and provenance. PLE’s co-authored route starts from engineers’ structured seed briefs, uses Creator and Solver agents to construct and challenge environments and graders, and ends with expert validation and difficulty intensification. The solver is isolated from graders and references; the described loop tests mean performance across typically three attempts rather than assuming a task is hard after one failure. For PLE’s expert-authored route, practising engineers write a realistic work request, construct a multi-format project environment and specify weighted evidence-bound rubric checks on the deliverable; the paper reports 941 rubric items across 44 tasks. These are reported task-construction stages, not an assertion of independent inspection of every environment or engineer’s records.

Sources: [700]

Evidence: Benchmark-family atlas - Physical Engineering’s Last Exam (PLE): evaluation, validation and interpretation. PLE’s documented evaluation gives an agent the instruction and environment without rubric, grader or reference access, then grades collected deliverables outside the agent container; model/harness pairs run every task three times with a six-hour agent budget. Infrastructure failures are rerun, while runs ending without a gradable deliverable score zero. Results are scoped to the model plus harness and cannot be interpreted as real engineering sign-off. PLE’s v2 paper grades expert-authored tasks by weighted binary rubrics and most co-authored deliverables by 0.4 times constraint compliance plus 0.6 times a hard-constraint-gated normalized objective score. Twelve co-authored graders also use LLM judgments for narrative soft constraints; two site-design and two MEP-coordination tasks retain their own scores. The hybrid formula is not universal to all 85 tasks.

Sources: [700]

Evidence: Benchmark-family atlas - Physical Engineering’s Last Exam (PLE): release, access and rights. The PLE repository describes 29 public development tasks behind a Hugging Face terms gate and 56 hidden test tasks; it declares Apache 2.0 for the project, while the paper states that environments contain no client-confidential material and use generated or redistributable source files. The code declaration and paper’s source-rights account do not replace the gated dataset’s terms or an item-level license audit.

Sources: [700], [776]

Evidence: Benchmark-family atlas - PostTrain Arena: evaluation, validation and interpretation. The captured PostTrain landing page accepts contributions while challenge runs are paused pending evaluation repairs/base-model measurement. The later README separates public configuration and bounded historical pipeline checks from proof that a full held-out training run completed. Seen-task historical practice and a small same-domain diagnostic are not evidence of broad held-out generalization; illustrative landing-page scoreboards are not measured results. PostTrain Arena's organizer specification fixes the model, training recipe and held-out suite for a challenge and ranks contributed environment collections by the change in held-out pass@1 after versus before training in the same run. An organizer reviews a result before ranking; the landing page ranks collections by mean change over accepted runs and warns that a small change from one noisy run is not evidence of improvement.

Sources: [1], [12], [13]

Evidence: Benchmark-family atlas - PostTrain Arena: evaluation, validation and interpretation. PostTrain Arena documents static overlap/leakage checks and planned dynamic reference/no-op/difficulty controls; static acceptance reads source without executing it and is explicitly not proof that the oracle, runtime or verifier works. The planned dynamic checks rerun reference and no-op controls eight times and target one to three base-model successes in four attempts. The specification says runs do not yet wait for these verdicts. PostTrain Arena accepts collections of one to two hundred task packages containing a prompt/limits file, Docker environment, verifier and reference oracle; verifiers run only after the agent finishes and must score substantive output rather than file existence. The documented held-out attempt passes only if its verifier passes, and a time-limit attempt fails. This package contract does not establish that every submitted task's verifier works.

Sources: [13]

Evidence: Benchmark-family atlas - PostTrain Arena: release, access and rights. PostTrain Arena records declared task author, license and original/adapted/generated origin fields and pins submitted content to a repository commit; its README applies AGPL 3.0 to repository contents unless otherwise noted and describes draft CC BY 4.0 data/text and Apache 2.0 code submission rules. Contributor metadata is declared rather than independently verified; missing or invalid license/origin causes a warning rather than a structural block. Draft submission terms require checking the final rules.

Sources: [12], [13]

Evidence: Benchmark-family atlas - RedlineBench: defined object and observation units. RedlineBench's creator README defines 140 runnable tasks across three negotiation scenarios and four turns; tasks in the same input group share model-facing input but differ in the attorney rubric set used to grade it. A rubric variant is not a new underlying negotiation input or an independent benchmark family.

Sources: [778]

Evidence: Benchmark-family atlas - RedlineBench: evaluation, validation and interpretation. RedlineBench gives the agent a party-specific playbook, commercial context and a source contract: a clean template at Turn 1 and a previously redlined draft with tracked changes visible at Turn 2 and later. The agent produces a native Word redline with tracked edits and comments; counterparty playbooks and verifier-side rubrics are not visible to the agent. Two acceptance-only final-turn tasks intentionally lack an attorney golden redline; golden files support document-native behavioral metrics rather than being agent inputs. RedlineBench's source report describes independently branching attorney negotiation paths and rubrics authored around the most important redline decisions; the creator card says senior technology-transactions attorneys wrote weighted criteria in real time as they negotiated each turn. The five rubric dimensions cover legal correctness, commercial context, negotiation quality, counterparty acceptance and deal-closing orientation; they encode contextual judgments rather than a universal single correct redline.

Sources: [777], [778], [779]

Evidence: Benchmark-family atlas - RedlineBench: evaluation, validation and interpretation. The RedlineBench README describes a loadable-Word-output validity gate, weighted rubric scores with penalties clipped to zero through one, and strict-majority voting by a three-LLM panel; the card aggregates within input groups and then equally across twelve scenario-by-turn cells. The source report separately computes behavioral metrics directly from document edits without LLM judges. Reproduction is non-deterministic because of agent sampling and LLM judging.

Sources: [777], [778], [779]

Evidence: Benchmark-family atlas - RedlineBench: release, access and rights. RedlineBench's creator card declares CC BY 4.0 for its data and MIT for the accompanying code, and describes fictional parties, scrubbed contact/document metadata and no real client material. These are the release's stated license and anonymization properties, not a review of every file or blanket clearance of independently sourced legal materials.

Sources: [777]

Evidence: Benchmark-family atlas - Riemann-Bench: defined object and observation units. Riemann-Bench's v3 paper describes twenty-five research-level mathematics problems drawn by active mathematicians from their own work, with definite verifiable answers requiring extended theoretical reasoning. The problems have known solutions; they are not themselves open conjectures such as the Riemann hypothesis.

Sources: [21]

Evidence: Benchmark-family atlas - Riemann-Bench: evaluation, validation and interpretation. Riemann-Bench’s v3 paper permits coding and search tools with no token-budget/interaction-format constraints and describes one hundred independent attempts per model/problem for pass@k estimation. The blog body retains an earlier one-hour timeout while its update note removes that limit; those conflicting time-budget scopes are preserved without selecting a favorable performance number. Riemann-Bench’s paper requires two independent domain experts to solve each problem from scratch without seeing the author’s solution, checking ambiguity, underspecification and difficulty, with revision or removal after failed verification. The benchmark’s double-blind verification is its reported construction protocol, not an independent peer decision on this research claim.

Sources: [21], [27]

Evidence: Benchmark-family atlas - Riemann-Bench: release, access and rights. Riemann-Bench’s paper keeps the full dataset private and offers a controlled model-evaluation service, while publicly describing its construction and verification procedure. Privacy is an access policy, not independent proof that no model has ever encountered related material.

Sources: [21]

Evidence: Dataset-family atlas - s1K: construction and provenance. s1K-1.1 distinguishes Gemini Flash Thinking Experimental reasoning/Gemini responses from DeepSeek-r1 traces/responses and credits Ryan Marten with generating the r1 traces. Questions and other data are attributed to original s1K; source_type and metadata retain the originating dataset/category and original fields. The illustrated AI-MO/NuminaMath-CoT/aops_forum item supplies one example's lineage, not provenance proportions of every question. Trace-generation credit does not establish paid annotator status or a human correctness audit of all outputs. s1K documents original solutions/source metadata beside model traces and attempts, while the header declares separate Gemini and DeepSeek grade and grade-reason fields. The example has cot set to None and a generic attempt field, whereas prose names gemini_attempt and deepseek_attempt; header/prose/example differences remain unresolved without a payload parser. Grade-field names and illustrated traces are not evidence that this researcher executed a grader or verified every model solution; this card supplies no complete item-validation protocol.

Sources: [444]

Evidence: Dataset-family atlas - s1K: evaluation, validation and interpretation. The s1K-1.1 card says it retains the same 1,000 questions as s1K while replacing reasoning traces with DeepSeek-r1 generations. Its train split declares 1,000 examples, without an independent payload recount or a new independent question-family count. The summary says the new traces improve performance but gives no quantified comparable evaluation in this card, so no numerical or reproduced gain is inferred.

Sources: [444]

Evidence: Dataset-family atlas - s1K: release, access and rights. The pinned s1K-1.1 card declares an MIT license label and links the original s1 paper and repository. It also preserves upstream question metadata and multiple model-generated trace fields; the label is not an independent rights finding for every question or teacher output. Only the complete creator card was read; linked paper, code, payload and external rights terms remain outside scope.

Sources: [444]

Evidence: Benchmark-family atlas - ScienceVision: defined object and observation units. ScienceVision’s creator article describes 250 expert-level questions, 50 each in biology, chemistry, medical imaging, mechanical CAD and architecture, across 22 subdomains, paired with native scientific artifacts and task-specific reference answers. The article defines expert-level by domain knowledge and conventions, not a human-difficulty baseline. Domains and access settings do not create additional benchmark families.

Sources: [589]

Evidence: Benchmark-family atlas - ScienceVision: construction and provenance. The ScienceVision creator article attributes joint construction to Elorian AI and UniPat AI and reports a 66-person annotation pool covering question writing, revision, reference-answer annotation and review across the five disciplines. This is the explicit creator-reported contribution scope; recruitment, compensation and independent worker qualification were not established.

Sources: [589]

Evidence: Benchmark-family atlas - ScienceVision: evaluation, validation and interpretation. The ScienceVision article describes mean answer scores on a zero-to-one scale, counting ineligible trials as zero, averaging available repeats within each question and then across the 250 questions. The source discloses unequal repeat counts for some configurations and different scaffolds for GPT versus other models. Neither uniform three-repeat coverage nor model-only causation is inferred. ScienceVision’s creators caution that a correct answer under native-source access can result from direct computation over the artifact and is not by itself evidence of successful visual observation. The interactive webpage converts two demonstration artifacts, including CT resampling; those demonstrations are not treated as byte-identical native evaluation inputs. The full manuscript, native files, interactive sessions, task payloads and reuse rights were not independently audited.

Sources: [589]

Evidence: Benchmark-family atlas - ScienceVision: release, access and rights. ScienceVision compares native-file source access, source access plus MCP inspection tools, and a fixed six-rendered-view control; MCP tools permit changing views, isolating structures, navigating slices and measuring geometry. The settings use the same underlying artifacts but provide different information and interactions. Public webpage viewers are described as expert-facing demonstrations rather than the evaluation harness.

Sources: [589]

Evidence: Benchmark-family atlas - Search-PI: defined object and observation units. Search-PI Simple combines 10 benign SimpleQA tasks with 12 malicious tasks into 120 task pairs in the documented evaluation. Benign tasks, malicious goals, task pairs and repeated attacker evaluations are different units; Simple is a subset of the retained Search-PI family.

Sources: [625]

Evidence: Benchmark-family atlas - Search-PI: construction and provenance. Meta states that Search-PI Complex was developed with Trajectory Labs, which pre-generated injections using its proprietary attacker for multi-location attacks; the report includes 30 benign-malicious task pairs and reserves remaining tasks as held-out. Trajectory Labs' contribution is limited to the explicitly described Complex subset; it is not generalized to every Meta benchmark or the Simple subset. The private attacker, held-out tasks and full corpus were not accessed, and no open payload license is inferred.

Sources: [625]

Evidence: Benchmark-family atlas - Search-PI: evaluation, validation and interpretation. Meta's Search-PI method evaluates prompt-injection attacks on an information-seeking assistant using web search, open and find tools with simulated personal information; Simple uses live web search while Complex uses a mock search index built from cached web content. This describes the public report's bounded Search-PI method, not the entire SIREN suite, actual personal-data exfiltration or a performed attack here. For Search-PI Complex, the report's pass@5 procedure repeats the same static attack five times because only outputs of the proprietary attacker are available. This is not five newly adapted attack strategies; Simple's adaptive-attempt procedure is a different setting. No performance ordering or operational security guarantee is proposed.

Sources: [625]

Evidence: Benchmark-family atlas - Search-PI: evaluation, validation and interpretation. The Search-PI Simple method evaluates each task pair with four handcrafted static injection templates and with an LLM attacker making six adaptive attempts, yielding 480 static and 720 adaptive evaluations per model. These are documented evaluation budgets, not counts of unique logical benchmark families or an independent attack-success measurement.

Sources: [625]

Evidence: Dataset-family atlas - SmallThoughts: construction and provenance. SmallThoughts links a distillation pipeline and semantic-similarity visualization, but its README does not specify a teacher checkpoint or an item-level answer-verification procedure. The one-command generation description is a creator claim; external pipeline and Atlas content were not opened or executed here. Visualization labels and links do not establish correct answers, dataset-wide human review or a verified filtering protocol.

Sources: [445]

Evidence: Dataset-family atlas - SmallThoughts: evaluation, validation and interpretation. SmallThoughts declares 100,000 training examples and 2,000 test examples while assigning size category 10K<n<100K. The category conflicts with the declared training-count boundary and combined split counts; both statements remain explicit, without harmonization. These are card-declared counts, not a payload recount; split independence, overlap and deduplication were not checked. SmallThoughts describes synthetic math, science, code and puzzle reasoning intended to make DeepSeek-R1-distilled trajectories more precise and concise while retaining reflection. This is a creator design aim, not an independently checked brevity, correctness or retained-performance study. The exact source is SmallDoge/SmallThoughts; it is not automatically conflated with other similarly named datasets or upstream OpenThoughts.

Sources: [445]

Evidence: Dataset-family atlas - SmallThoughts: evaluation, validation and interpretation. SmallThoughts declares problem, solution, messages and system_prompt fields and suggests problem/solution for TRL GRPO and messages for SFT without additional preprocessing. These are usage instructions, not evidence that this reader executed either training recipe or measured an outcome. Payload/pipeline compatibility, actual prompts and generation implementation were not independently checked.

Sources: [445]

Evidence: Dataset-family atlas - SmallThoughts: release, access and rights. The SmallThoughts card declares the dataset released under Apache-2.0. This card declaration does not independently audit rights to upstream problems, teacher outputs or generation code. Header and license paragraph agree at the pinned version, but the linked software license was not opened or generalized to all artifacts.

Sources: [445]

Evidence: Benchmark-family atlas - Snorkel Finance Reasoning: construction and provenance. Snorkel describes co-creating the Finance Reasoning question-answer pairs with its financial-expert network for questions requiring reasoning and SQL tool use over relational tables converted from company 10-K filings. This is the creator’s explicit co-creation account; it does not establish recruitment, pay, per-question annotator counts or an independent accuracy audit.

Sources: [632]

Evidence: Benchmark-family atlas - Snorkel Finance Reasoning: evaluation, validation and interpretation. Snorkel’s Finance Reasoning method uses Claude 3.7 Sonnet as an LLM judge supplied with a reference answer to assess final-answer correctness and completeness. The judge configuration is creator-described; judge decisions were not independently reproduced or equated with human adjudication. Snorkel’s pinned creator card describes an initial release of 357 agent-trace rows based on 79 financial question-answer pairs. Trace rows and underlying QA pairs are different units, not 357 independent financial questions. This release is distinct from the earlier Snorkel Finance benchmark linked by the card.

Sources: [631], [632]

Evidence: Benchmark-family atlas - Snorkel Finance Reasoning: evaluation, validation and interpretation. The captured Finance Reasoning methodology describes an evaluation cap of 15 minutes or 100 turns and includes traces ending in API errors, recursion limits or premature termination when reporting accuracy. This limit belongs to the captured methodology version, not every historical or future trace release. Including failed traces describes the reported evaluation denominator; it is not evidence that every released trace completed successfully. Snorkel's Finance Reasoning creator card warns that the dataset is based on financial documents from specific years and may become stale without regular updates. This temporal limitation is not financial advice or evidence that every historical question is currently false.

Sources: [631], [632]

Evidence: Benchmark-family atlas - Snorkel Finance Reasoning: release, access and rights. The pinned Finance Reasoning creator card declares the released dataset to be Apache-2.0 licensed. This is the creator's release-license declaration, not an independent rights audit.

Sources: [632]

Evidence: Benchmark-family atlas - sudo L7: defined object and observation units. sudo L7's creator article describes sixty tasks across backend, frontend, security, reliability and architecture, grounded primarily in private production codebases alongside open-source projects. The task collection is one retained family; repositories, languages andbehavior dimensions are internal coverage.

Sources: [28]

Evidence: Benchmark-family atlas - sudo L7: construction and provenance. The sudo L7 article says every task is authored by a staff-level-or-above engineer and begins with a situation in which a frontier model meaningfully failed. The source describes contributor roles and affiliations; it does not establish a new supplier contribution by each cited employer.

Sources: [28]

Evidence: Benchmark-family atlas - sudo L7: evaluation, validation and interpretation. sudo L7 evaluates staff-level engineering behavior using expert-authored task-specific criteria across eight dimensions; grader models use unit tests, static analysis and other executable checks as evidence where appropriate. The documented target includes architectural judgment, verification and truthful communication rather than code-test success alone.

Sources: [28]

Evidence: Benchmark-family atlas - sudo L7: release, access and rights. The sudo L7 creator article directs readers to request access to its benchmark set; it does not expose the complete private-production task collection in the article. Public methodology does not establish redistribution rights for private repositories or access to their task payloads.

Sources: [28]

Evidence: Benchmark-family atlas - SWE-Marathon: defined object and observation units. The SWE-Marathon v1 paper describes 20 accepted project-scale tasks authored by 11 contributors across library reproduction, product cloning, ML engineering and algorithmic optimization. The captured Refresh fork’s literal Markdown task rows 19-37 enumerate 19 tasks, while the creator paper describes 20. These are different historical scopes of one retained family; the earlier inherited 18-task note is not reused as an observed count.

Sources: [601], [603]

Evidence: Benchmark-family atlas - SWE-Marathon: construction and provenance. SWE-Marathon’s paper describes proposal review, CI and LLM-assisted checks, pilot agent trials, adversarial exploit search and final human approval, revising and revalidating tasks with confirmed quality failures or exploits. These are the benchmark creators’ task-quality procedures, not ScienceToStartup peer receipts or proof that all future exploits are excluded.

Sources: [601]

Evidence: Benchmark-family atlas - SWE-Marathon: evaluation, validation and interpretation. SWE-Marathon tasks use Harbor instructions, Dockerized starter environments, development feedback, hidden verifiers, held-out human-written solution oracles and wall-clock limits, with final scoring based on the submitted container state. Visible feedback differs from stricter hidden final checks; some tasks omit visible tests. The SWE-Marathon paper’s reported protocol uses five trials for each of 13 agent-model configurations across 20 tasks, yielding 1,300 trajectories, and defines headline pass@1 as the fraction passing the task verifier. The paper discloses expensive runs, seed variance, a single execution backend and unmeasured exploit false negatives. No new trial, ranking or successful exploit audit is claimed.

Sources: [601]

Evidence: Benchmark-family atlas - SWE-Marathon: evaluation, validation and interpretation. SWE-Marathon combines task-specific tests, behavioral parity, performance gates, replay, integrity checks and-in selected product tasks-a browser-based rubric stage whose evidence is scored by a separate judge. The appendix makes the S3-console UX stage optional while other named product tasks require both stages; no universal all-tasks UI requirement is inferred. Verifier families are scoring mechanisms, not new logical benchmark families.

Sources: [601]

Evidence: Benchmark-family atlas - SWE-Marathon: release, access and rights. The SWE-Marathon creator README declares Apache 2.0 but describes its roughly 800 GB trajectory bucket as public while also asking users to request S3 credentials for downloading logs. The contradictory access description remains unresolved; no credentials, outreach or bulk log download was attempted. Code/repository licensing is not generalized to every upstream task artifact.

Sources: [602]

Evidence: Benchmark-family atlas - Terminal-Bench: construction and provenance. Terminal-Bench 2.0 was selected from 229 contributed tasks by 93 contributors; the paper describes difficulty curation and multiple experienced reviewers before accepting 89 tasks. Contribution, selection and review counts are the paper authors' reports, not an independent contributor census or a guarantee of error-free tasks.

Sources: [377]

Evidence: Benchmark-family atlas - Terminal-Bench: evaluation, validation and interpretation. The Terminal-Bench contributor checklist requires human-written task instructions and human-written reference solution scripts with minimal model assistance; the paper separately discloses Claude 4 Opus use to shorten author-provided instructions into appendix descriptions. A contributor checklist is a declared requirement, not an independent authorship audit of each file. The appendix-description use is not evidence that models authored every original task. Terminal-Bench's paper identifies internet access, public task/solution availability, mutable external dependencies and hardware or API variation as residual risks to contamination control and reproducibility. Containerization and a canary do not establish a contamination-free or deterministic benchmark. No agent runs, dependency-resolution audit, task-license audit or independent performance measurement was conducted.

Sources: [377]

Evidence: Benchmark-family atlas - Terminal-Bench: evaluation, validation and interpretation. The Terminal-Bench paper's 2.0 benchmark contains 89 tasks, each with a natural-language instruction, a containerized environment, tests, a reference solution and a time limit; outcome tests inspect the final container state rather than a prescribed command sequence. This proposition is scoped to Terminal-Bench 2.0 in the pinned v1 paper, not every earlier or later registry version. Harbor is evaluation infrastructure and Terminus 2 is an agent scaffold, not additional Terminal-Bench task families. Terminal-Bench's verification combines automated oracle-pass and dummy-fail checks, contributor checklists, LLM-assisted checks, expert manual review, agent-trajectory review and a separate adversarial exploit audit with human inspection. The adversarial procedure described by the benchmark authors is not a ScienceToStartup peer-review receipt and was not rerun here. The paper still acknowledges residual specification and evaluation flaws.

Sources: [377]

Evidence: Benchmark-family atlas - Trajectory 72-scenario indirect prompt-injection evaluation: evaluation, validation and interpretation. The commissioning article identifies Claude Code v2.1.205 and Codex v0.144.5 for the evaluation and warns that a later Auto-review release could change results. The methods refer to publicly available application versions as of July 17, 2026. A later article update is not a reset of that evaluation date or evidence for current-product performance. The 72-scenario manifest, payload, release license and proprietary attack-generation details were not publicly established by this source read. The commissioning article states that the 72-scenario attacks were optimized against Claude Opus 4.7, which was outside the reported model set, so the reported models encountered them as unseen attacks. Unseen is the publisher’s evaluation-design statement, not an independently verified contamination or robustness guarantee. No attack was performed by this author.

Sources: [604]

Evidence: Benchmark-family atlas - Trajectory 72-scenario indirect prompt-injection evaluation: evaluation, validation and interpretation. The Trajectory 72 evaluation used the same Trajectory Labs browser integration for Claude Code and Codex, exposing MCP tools as wrappers around Chrome APIs; first-party browser-extension safeguards were outside the test. The article distinguishes text tool-return and screenshot-browser tasks. Its scope is not the complete safeguard stack of every deployed application. Anthropic’s commissioning article describes a Trajectory Labs evaluation of 72 indirect prompt-injection scenarios held out from Anthropic, tested 10 times each. This establishes the documented 72-scenario evaluation scope and 720 attempts per evaluated configuration, not a public scenario manifest or additional families per model/mode. No overlap or identity merge with Search-PI or Apollo’s separate attack datasets is inferred.

Sources: [604]

Evidence: Benchmark-family atlas - U-MATH: defined object and observation units. U-MATH is documented as 1,100 university-mathematics problems across six subjects; the v4 paper table distinguishes 900 textual problems and 200 problems with visual elements. The paper rounds its visual share to about 20%; the exact tabulated counts are 900 and 200. U-MATH problem-solving tasks are distinct from the derived μ-MATH solution-judgment samples.

Sources: [552]

Evidence: Benchmark-family atlas - U-MATH: construction and provenance. The U-MATH v4 paper describes sourcing previously unpublished problems and solutions crafted by subject-matter experts for ongoing university courses through Gradarius, then filtering short-solution, multiple-choice and calculator-dependent problems. Previously unpublished and expert-authored are the paper authors' provenance statements; original course materials and permissions were not independently audited. Visual-problem selection additionally keeps only problems containing a single image. U-MATH construction uses several small language models to select difficult problems, followed by manual curation by in-house mathematical experts and the Gradarius team and review by university teaching experts. Model-assisted difficulty selection and subsequent human review are different stages. The precise selection statement concerns 150 challenging problems per subject; no uniform human-labeling multiplicity for every problem is inferred.

Sources: [552]

Evidence: Benchmark-family atlas - U-MATH: evaluation, validation and interpretation. The U-MATH paper defines accuracy using an LLM judge against reference answers and considers a problem solved only when all questions in its statement are answered correctly and fully. This is the documented all-questions scoring rule, not an independent validation of judge accuracy or a partial-credit measure.

Sources: [552]

Evidence: Benchmark-family atlas - U-MATH: release, access and rights. The pinned Toloka U-MATH creator card states that all U-MATH dataset contents are available under the MIT license. This is the creator card's license declaration, not a permissions audit of underlying educational materials. This U-MATH statement is not transferred to µ-MATH machine-generated model outputs.

Sources: [633]

Evidence: Benchmark-family atlas - µ-MATH: defined object and observation units. µ-MATH is a solution-judgment meta-evaluation dataset of 1,084 labeled samples drawn from 271 U-MATH problems, rather than another spelling of U-MATH's 1,100 problem-solving tasks. The retained family IDs remain distinct despite inconsistent umath/mumath link spellings inside the cards. These are documented release units; no payload recount was performed.

Sources: [552], [626]

Evidence: Benchmark-family atlas - µ-MATH: construction and provenance. For µ-MATH construction, the v4 paper describes four model-generated solutions per selected text-only problem, formal answer-equivalence checks through Gradarius-API, and in-house expert labeling of solutions receiving negative API responses. A negative API response is not itself treated as a wrong solution because extraction and expression conversion can fail. The paper names Llama-3.1 8B among the four generators; the pinned creator card names Llama-3.1 70B. That source discrepancy is unresolved and neither size is silently substituted.

Sources: [552], [626]

Evidence: Benchmark-family atlas - μ-MATH: evaluation, validation and interpretation. μ-MATH evaluates a judge given a problem statement, reference answer and candidate solution as a binary correctness classifier, with macro-averaged F1 as its primary metric and positive/negative predictive values and true-positive/true-negative rates as secondary metrics. These are method definitions, not performance rankings or new experimental measurements. The μ-MATH paper describes selecting 271 U-MATH problems using expert estimates of evaluation difficulty and explicitly states that the resulting set is not intended to reflect the overall U-MATH distribution. The meta-evaluation sample is text-only and selected for judging difficulty; it does not establish judge quality for U-MATH visual problems or all university mathematics.

Sources: [552], [626]

Evidence: Benchmark-family atlas - μ-MATH: release, access and rights. The pinned μ-MATH creator card places its machine-generated model_output field under the underlying models' license terms and places the other fields under MIT. The YAML MIT label is not treated as an unconditional license for every column; underlying model terms were not separately audited.

Sources: [626]

7. Human expertise, recruitment, disagreement and labor

Analysis: Expertise has at least four operational components: eligibility, task familiarity, calibrated judgment and permission to perform the work. The inspected sources use different proxies: Turing describes identity/background/work-history verification; Shaip describes credentials and a calibration gate; Encord describes task-to-domain matching; Prolific offers verified experience or targeted assessments. These are not one standardized credential. For each task, ask which evidence predicts successful completion and which screening merely establishes identity or administrative eligibility.

Sources: [135], [157], [708], [712]

Analysis: Recruitment and contribution remain separate stages. HelpSteer3 names different recruitment/annotation roles for Scale and Translated; Prolific's studied role is recruitment into a researcher-owned interface; U-MATH credits Toloka AI and Gradarius with validation of professor-authored problems. A contributor manifest should record authoring, recruitment, labeling, review, adjudication and final release authority separately. Affiliation, acknowledgments or a capability page alone should not be used to infer a contribution edge.

Sources: [715], [741], [758]

Analysis: The human experiment on instruction-tuning disagreement uses the same annotators and items for individual ratings and paired preferences, yet reports different ambiguity patterns and no evidence in that setup that preference annotation is easier. This cautions against assuming that a change from scales to pairs resolves judgment ambiguity. A pilot should test the intended formulation with representative difficult items, retain task-specific reasons and distinguish unclear instructions from legitimate preference diversity.

Sources: [133]

Analysis: DICES separates two rater pools and applies explicit quality exclusions, while its expert reference can disagree with crowd majority. The Minority Reports paper defines a minority relative to final majority consensus and warns that RLHF preference tasks may lack a single clear ground truth. Together these sources support keeping minority labels inspectable. They do not establish that every disagreement is valuable or that a majority is always wrong. The analyst recommendation is to adjudicate the reason for disagreement before filtering the person who expressed it.

Sources: [132], [410]

Evidence: The approved Minority Reports pruning observation is scoped to the §5.1 prose: unseen workers or tasks are retained, while otherwise the estimated minority-report probability is compared with a threshold. The reviewed Algorithm 1 line and surrounding prose have a documented consistency concern in the source review. This report does not certify that pseudocode and prose are equivalent. The simulation reference compares post-pruning majority votes with the counterfactual unpruned majority, rather than establishing an external truth label for every judgment.

Sources: [410]

Analysis: Calibration and adjudication should have separate logs. Appen describes gold questions and adjudication; DataForce describes automated checks, consensus and expert adjudication; Snorkel describes calibration and dual-expert review; Shaip describes sampled blind double-labeling and weekly agreement reports. Request the raw judgments before consensus, the rule for escalation, the adjudicator’s independence and the number of corrected accepted units. A final agreement percentage is a process output whose meaning depends on those rules, not a universal reliability score.

Sources: [54], [157], [684], [735]

Analysis: Training instructions are themselves a research input. The reviewed factorial study crosses instruction design with incentives and measures accuracy, helpfulness and difficulty separately. Data Cascades examines high-stakes data-work failures; CrowdWorkSheets asks why an annotator pool fits intended uses and impacted communities. These works motivate testing guideline clarity and the consequences of exclusion before scaling. They do not establish a universal wage or training-duration effect for all suppliers. Intervention results should retain task, sample, payment and outcome definitions.

Sources: [127], [389], [401]

Analysis: Multilingual coverage must be tested at the intersection of language, domain, rubric and modality. HelpSteer3 identifies missing domain intersections; M-IFEval removes language-inappropriate constraints and uses native-speaker leads; BenchMAX uses human post-editing of translations; MENLO separates fluency, tone and localized factuality. Translation of a single English benchmark is not evidence that every language variant measures the same construct. A buyer should sample required intersections and use native-domain review before claiming coverage.

Sources: [381], [390], [391], [715]

Analysis: Model-judge agreement is also conditional. The multilingual-judge study uses parallel ground-truth answers, pointwise criteria and a particular prompting setup, with cross-language consistency measured by treating language-specific judgments as raters. M-RewardBench translates general-task instances with human-reviewed filtering and separately uses human-rated translation pairs. These are different validation designs. Neither supports an unqualified conclusion that an English-calibrated judge is reliable in every locale or that translated evaluation is equivalent to native task authoring.

Sources: [394], [396]

Analysis: Pool statistics should not become labor statistics. Appen’s global crowd, Mercor’s advertised expert network, Turing’s advertised pool, TELUS Digital’s community membership and Prolific’s active/waitlist figures have different definitions. None should be converted into the number of people who earned money this week, the number available for a specific assignment or the number of accepted tasks. A useful labor ledger distinguishes registered profiles, screened people, matched people, active paid contributors, attempted units, accepted units and paid units under explicit windows.

Sources: [89], [690], [720], [736], [760]

Analysis: Workforce arrangement changes the questions to ask. Sama describes full-time in-house delivery; TELUS Digital lists several workforce modes; Alegion describes workforce partners; Scale describes BPOs and proprietary platforms; isahit describes self-employed onboarding. These descriptions cannot establish legal status or worker experience across jurisdictions. For a project, record the contracting entity, subcontracting chain, training obligation, supervision, payment authority and appeal channel rather than assume the same conditions from a shared annotation label.

Sources: [82], [89], [120], [143], [713]

Analysis: Compensation disclosures concern different units and conditions. Prolific publishes permitted minimum and recommended participant hourly rates; Handshake puts amount, type and timing in project invitations; Humans in the Loop uses supervisor-approved delivered units; Snorkel describes weekly HireArt processing of prior accepted work, aims for processing within two weeks after acceptance, and allows project-dependent variation; micro1 states a twice-monthly schedule. These are rules or schedules, not comparable realized wages. Any labor analysis should count qualification, waiting, revisions, rejections and administrative time separately and retain the project's currency and payment basis.

Sources: [139], [221], [229], [682], [720]

Analysis: Sensitive-content and monitoring rules deserve operational evidence. Scale describes notice, the ability to decline and wellness resources; Toloka requires disclosure and acknowledgement before labeling; the dated Welo case describes informed consent; Invisible discloses keystroke, screenshot and webcam collection with possible access consequences if declined. These sources describe different settings. They do not prove that every worker could safely refuse, obtained support or experienced proportionate monitoring. A buyer should request project notices, escalation handling and aggregate worker-feedback evidence without seeking unnecessary personal records.

Sources: [67], [92], [713], [755]

Analysis: Labor research has its own sampling limits. The Data-Production Dispositif combines interviews, observation and instruction-document analysis in specific Latin American settings; the Latin American labor article states that unknown platform-population boundaries limit statistical representativeness; the requester-proxy study limits its conclusions to US requester perspectives and self-reports. These studies illuminate mechanisms and research questions. They should not become global percentages, a census of all workers or evidence that a named supplier currently practices every behavior described.

Sources: [385], [409], [414]

Research gap: Important labor unknowns remain: independently measured effective pay, qualification rejection and appeal outcomes, subcontractor payment reconciliation, harmful-content exposure, contributor retention and the relation between monitoring and acceptance. The public-source corpus does not resolve these for the covered market. Research can progress through consented worker studies, project-level anonymized logs and comparable definitions, while avoiding inference from registered pools or recruitment advertisements.

Analysis: Disagreement routing should distinguish uncertainty from error. Experts Rise uses cross-model entropy to select expert attention, while Inherent Disagreements models multiple interpretations of the same short sentence pairs. CrowdTruth defines mutually dependent consistency scores but leaves broader comparative validation open. The practical choice is whether the task seeks one verifiable answer, a distribution of valid interpretations or a codebook revision. Record that choice before using disagreement to exclude workers, remove items or declare a quality improvement.

Sources: [424], [426], [431]

Analysis: Instruction design and assistance need separate acceptance checks. The surgical study's comparison is conditional on instruction richness and project-specific QA; LabelAId's intervention improves precision within a 34-person study without an equivalence finding on time. Neither supports a general staffing reduction. In a proposed pilot, log instruction revisions, assistance shown, human corrections and downstream acceptance, then inspect whether the intervention changes errors or merely which cases remain. Preserve any effect on contributor effort and disagreement as part of the operating record.

Sources: [432], [435]

Analysis: Qualification labor is part of the production process. The qualification-threshold study observes task-dependent effects and self-selection; CrowdSpeech/VoxDIY reports compensation for attempted exams and completed pre-block tasks; the writers study makes revision and interaction visible. The analyst judgment is to budget and record qualification, calibration, waiting, revision and rejected-but-payable work separately from accepted-unit production. Confidence is moderate in the relevance of that accounting and unestablished for any supplier's actual effective pay. Consented paid-time logs and task-specific performance comparisons would change the labor assessment.

Sources: [458], [464], [555]

Analysis: Human expert and model persona are different provenance categories. U-MATH describes mathematical experts reviewing tasks; APEX-Agents records expert creation and independent baselines; RIFT's taxonomy uses human expert annotations while its scoring panels are generated expert personas. A contract should state which decisions require a credentialed person, whether review is independent of authorship and how unresolved judgments are escalated. Confidence is high in those distinctions and unmeasured for outcome superiority. Blinded assessment of raw expert decisions, panel disagreement and final accepted artifacts would change the staffing design.

Sources: [552], [561], [565]

Analysis: A reference panel has a population boundary. StereoTales compares multilingual generated associations with one UK English-language panel; AttuneBench has repeated conversations from eleven participants; the personality study measures model questionnaire responses rather than human personalities. A buyer should identify whose behavior, values or preferences the labels are meant to represent before scaling collection. Confidence is high that these are distinct populations, with no universal normative standard inferred. Larger local panels, participant-level holdouts and transparent disagreement distributions would change confidence in transfer.

Sources: [461], [699], [704]

Analysis: Labor comparison - task authors, baseliners, annotators and expert reviewers are different populations. CharXiv reports uncompensated graduate-student collection/annotation; HealthBench reports a compensated physician cohort; SWE Atlas reports compensated task authors and reviewers without inspected payroll; HCAST reports hourly pay plus task-dependent bonuses for baseliners. These do not establish a common labor price or that one staffing arrangement causes better data. Confidence is high in keeping role, qualification, pay basis and period separate. A buyer should budget paid qualification, revision and rejected work explicitly. Verified contracts, complete time records and matched accepted outputs would change the economic comparison.

Sources: [614], [615], [616], [634], [635], [639], [646], [647], [670]

Analysis: Guidelines and agreement - a final reconciled label can conceal changes in the task itself. The 2010 annotation-methodology paper warns about undocumented instructions and reconciliation drift; SciCode adds out-of-domain clarity review; GeneBench-Pro reports changes to data-generating processes or ground truth after expert review. Confidence is high that these are distinct review interventions and moderate that separate logs improve reproducibility. Preserve original judgments, codebook revisions, scientific-target changes and adjudication reasons as separate artifacts. A new annotator team reproducing labels from those records, or failing to do so, would change confidence in the process.

Sources: [446], [613], [637], [638], [659], [673]

Evidence: InstructGPT - recruitment-channel boundary. The pinned paper identifies Upwork as a channel through which some contractors were hired for demonstrations, comparisons and main evaluations. This does not establish Upwork as their employer, a managed annotation vendor or the producer of the full dataset. The preserved original-stage qualification predates current exact-content approval; it is not a new current approval claim by this report author.

Sources: [731]

Analysis: Human contribution should remain stage-specific when a collection also uses models. Multi-SWE-bench reports outsourced annotation and internal engineering review without naming an outsourcing supplier; ScienceVision reports Elorian AI/UniPat AI joint construction and an annotation pool without establishing compensation or independent credential checks; s1K credits a person for generating model traces without establishing paid label work. The inference is to record actual task authoring, labeling, review, recruitment, trace generation and funding roles independently. Confidence is high in avoiding transitive supplier attribution and unknown for undisclosed contracts. Project-specific agreements and role records could resolve particular commercial relationships.

Sources: [444], [589], [629]

Analysis: Expert-review intensity is useful process evidence but does not supply a shared quality measure. GDPval describes iterative human reviews; HSS admits candidates after unanimous independent review rounds; Riemann-Bench asks two experts to solve each problem without the author’s solution. Their different task populations, acceptance rules and reviewer information prevent a ranking by the number of reviewers alone. A buyer should retain which reviewers saw earlier answers, how disagreement changed the task and which candidates were removed. Confidence is high that these fields are needed to interpret the process; comparative error reduction is unmeasured. Independent post-release error sampling could test whether the retained process predicts delivered correctness.

Sources: [21], [350], [600]

8. Commercial evidence and procurement economics

Evidence: Appen - reported revenue. The audited FY2025 statements report consolidated customer-contract revenue of USD 230.774 million for 1 January-31 December 2025, including Appen Global USD 127.874 million and Appen China USD 102.900 million. These geographic operating segments do not isolate expert generative-AI data or the market definition used in this report. The figures are full-year reported revenue, not annualized pace, gross transaction volume, contributor pay or a comparable human-feedback segment across suppliers.

Sources: [736]

Evidence: Proximal - annualized pace. Its 29 September 2026 announcement reports $200 million in annualized revenue. The inspected disclosure does not specify an ISO currency code, annualization window, gross-versus-net basis or revenue split by business line. It is self-reported pace, not an audited annual revenue total, profit, bookings, valuation or contractor payout. No annual recognized-revenue amount is calculated from it in this report.

Sources: [246]

Evidence: Scale - cumulative contributor payout. Its FY2025 statement, approved in June 2026, reports nearly $1 billion paid to hundreds of thousands of contributors across more than 150 countries to date. The statement does not define a start/end period or a dollar currency code for that aggregate, and payment reconciliation was not inspected. The number is not annual revenue, annual labor cost, a current paid-active workforce total or a wage estimate.

Sources: [713]

Evidence: Labelbox - financing and valuation. The 14 October 2026 company announcement reports more than $100 million in Series E financing comprising equity and a credit facility, at a $1.25 billion pre-money valuation. The inspected source uses dollar symbols without an explicit ISO currency. The facility is not equity, and financing and valuation are not revenue or evidence of accepted data volume. No implied annual growth or provider-performance score is calculated.

Sources: [72]

Evidence: Prolific - tariff and reward rules. The captured FAQ states usual platform fees of 42.8% for corporate customers and 33.3% for academic/nonprofit customers, but the bound passage does not explicitly state the percentage denominator. It separately sets minimum participant rewards of GBP 6 or USD 8 per hour and recommends GBP 9 or USD 12, with higher expert compensation recommended. Managed services are custom-priced. These are published rules and tariff disclosures rather than realized earnings, margins or a complete specialist-project quote.

Sources: [720]

Evidence: Other disclosed billing units differ. Prime Intellect’s Model Factory terms meter allocated GPU-hours prorated to the second, with rates in the order or platform; personnel post-training needs a separate professional-services order. A.Team lists hourly and other builder-engagement models. Terac describes expert payout plus a platform fee and separately monthly USD payment per verified completion for partnerships. AfterQuery reserves fees to order forms or separate agreements. These documents describe commercial structure, not a comparable price per accepted preference pair or environment.

Sources: [209], [242], [285], [329], [330]

Analysis: The analytical rule is to keep commercial denominators separate. A company-wide annualized rate cannot be divided by a benchmark count to estimate task price; cumulative payouts cannot identify a current take rate; financing cannot measure demand; a broad expert pool cannot be a capacity denominator. The selected disclosures support questions for diligence, not supplier market shares. A defensible future comparison would need the same reporting period, currency, gross/net basis, organizational segment and accepted-unit definition across every included provider.

Sources: [72], [246], [690], [713], [736]

Method: Reproducible buyer calculation: accepted-unit cost equals all attributable project cash costs divided by independently accepted units. The numerator should state whether it includes recruitment, qualification, paid training, tool licenses, compute, project management, review, rework and buyer-side acceptance. The denominator should identify the unit: independent judgment, resolved item, trajectory, verified environment, recorded hour or licensed release. This is a proposed accounting method; the report does not insert hypothetical numbers or claim observed costs for suppliers.

Method: Reproducible review calculation: first-pass acceptance equals accepted units on first submission divided by all submitted units; final acceptance equals accepted units after permitted rework divided by all submitted units. Keep attempt, exclusion and unresolved counts. For expert-time projects, report both accepted outputs per paid hour and buyer QA effort per accepted output, without interpreting either as worker wages. For environment projects, separate authoring, maintenance, execution and recurring license costs, and preserve the term of access.

Method: Procurement comparison procedure: issue the same task brief and acceptance rubric, obtain responsibility and rights schedules, run a paid calibration sample, then quote a frozen production scope. Compare totals only after normalizing included review/rework and access terms. Log scope changes as new versions rather than treating a renegotiated specification as supplier underperformance. Proposed service-level thresholds should come from the buyer's loss tolerance and pilot uncertainty; they are not reported market standards.

Research gap: The commercial research gap is task-matched economics. This corpus does not supply comparable current contracts, realized gross/net human-data revenue, project margins, cancellation/rework charges or accepted-unit costs across suppliers. Missing financial disclosure is unknown rather than zero. No financial forecast, market-size projection, provider-return model or inferred valuation multiple is warranted by this evidence.

Analysis: Cost comparisons must include the measured boundary. LMCrowd excludes model, prompt and worker selection from its tariff comparison; CrowdAgent holds human annotation proportions rather than all computation or operating cost equal. HelpSteer2 discards about half of collected annotations overall, so accepted responses and purchased ratings have different denominators. Request setup, failed/rejected work, QA, adjudication, model calls, compute and accepted-unit totals explicitly. These studies inform accounting questions; they do not provide current service prices or a cross-provider cost ranking.

Sources: [423], [430], [434]

Analysis: Worker-time accounting is also distinct from project billing. The historical MTurk log study produces different work-unit rates under accept-to-submit and temporal-cluster definitions; its task/cluster medians are not median worker wages. Selected plugin users, unknown active effort and noncausal associations constrain inference. For a new project, request consented, task-appropriate compensation and time evidence rather than convert historical platform figures or supplier revenue into a claim about present worker earnings.

Sources: [425]

Analysis: Cost comparison must preserve generation, verification and acceptance budgets. Weaver generates 100 candidates per query and can call many verifiers; APEX-SWE averages repeated one-hour task runs; APEX-Agents evaluates repeated attempts; DAYJOB excludes error attempts from its score and uses creator-estimated healthcare time and a separately reported finance mean. These choices affect measured performance and a buyer's cost ledger differently. Confidence is high that a single score or token bill is insufficient and unestablished for relative supplier unit costs. Record all attempted work, retries, judge calls, human escalation and accepted outputs before estimating a project-specific cost per usable unit. DAYJOB's reported agent/condenser call cost excludes the judge and is not an all-in service price.

Sources: [556], [559], [561], [695]

Analysis: Reported professional-task performance is not revenue or observed economic output. DAYJOB's healthcare times are creator estimates; its finance mean has a separate dataset-card basis; APEX-Accounting measures rubric completion and repeated reliability on selected synthetic worlds; PLE reports a mean hybrid score. Multiplying those scores by an occupation's wages would introduce unmeasured task representativeness, deployment overhead, review, failure cost and acceptance assumptions. Confidence is high in rejecting that conversion from this evidence. Prospective matched deployments with observed professional time, complete exception costs and verified accepted deliverables would be needed to change the economic conclusion. DAYJOB's reported agent/condenser call cost excludes the judge and is not an all-in service price.

Sources: [695], [697], [700]

Evidence: Datatang / Nexdata group - unaudited six-month revenue. The issuer reports operating revenue of RMB 180,323,485.50 for January 1-June 30, 2026, and explicitly labels the report unaudited. The reported group ownership table includes Nexdata entities. This is group revenue for that period, not a standalone Nexdata segment figure, a run rate, task price or gross customer spend.

Sources: [484], [485]

Analysis: Budget comparison - a timed run, a valid scored attempt and a billable production unit can diverge. DeepSWE excludes provider/verifier/network errors without resampling but counts timeouts as failures; GeneBench-Pro excludes execution/format errors and uses unequal attempted-run budgets for some configurations; HCAST's human baseline combines hourly rates and performance incentives. Confidence is high that these denominators must be retained and unestablished for a common accepted-unit cost. A buyer should reconcile all attempts, provider charges, human time, reruns and independently accepted outputs. Prospective matched project ledgers would change a unit-cost assessment.

Sources: [607], [608], [609], [614], [615], [616], [673]

Analysis: The new reasoning-data disclosures add unlike cost units to the commercial ledger. Bespoke describes about $800 and 1.5 hours for synthetic dataset generation, while Light-R1 estimates about $1,000 and under six hours on twelve H800s for training. Generation expense and training estimates are different cost boundaries, and neither is a current supplier quote or a fully reconciled accepted-item cost. HackerRank's submission library and challenge registrations describe yet other units, with no disclosed delivered-training-record price. Confidence is high that these figures must remain separate and unestablished for economic superiority. A buyer would need complete compute, human review, licensing, rejection, retry and acceptance accounting under the same deliverable definition before making a cost comparison.

Sources: [442], [443], [769], [770]

9. Rights, security, privacy and attribution governance

Analysis: Dataset rights should be modeled as a chain of permissions rather than a single badge. The Data Provenance Initiative distinguishes creator-attributable data licenses from code licenses and explicitly limits itself to best-effort self-reported categorization. Common Pile requires confidence that license information originates with a rights holder. These methods support asking who can grant which rights for each component, transformation and release. They do not establish enforceability or replace project-specific legal review.

Sources: [153], [382]

Analysis: Web availability and collection permission are distinct. Consent in Crisis reports inconsistencies between robots.txt signals and terms of service in selected corpora. Documenting C4 recommends documenting included and excluded text and leaves some PII/copyright auditing to future work. FineWeb and Dolma acknowledge residual harmful or sensitive content despite filtering. A buyer should retain capture-time permissions, exclusion rules and removal paths, while treating a crawlable page as a source observation rather than automatic authorization to train or redistribute.

Sources: [134], [383], [404], [734]

Analysis: License scope can change within a research artifact. AlpacaEval declares different code and data licenses; RewardBench aggregation remains subject to component terms; MathVista retains original image/question copyright and prohibits training use; GAIA asks against public redistribution. The correct unit for diligence is the exact component and use, not a repository’s top-level license string. A restrictive finding does not imply that this report possesses or can redistribute the underlying data.

Sources: [29], [177], [449], [694]

Analysis: Customer material and website visitor data have different policy boundaries. AfterQuery explicitly excludes processor/service-provider customer processing from its website policy; Mercor excludes its work platform; Label Your Data and LXT describe client-instruction processing; Proximal requires written consent and an executed agreement for generalized-model use of customer material. Obtain the project processing agreement, subprocessors, retention and secondary-use terms. A public policy’s existence does not establish that the proposed delivery is covered by every described safeguard.

Sources: [74], [208], [243], [709], [739]

Evidence: Entity names and terms require exact scope. Kili’s May 2025 legal notice names MYAI SAS, while its same-date EULA names KILI-TECHNOLOGY SAS under the same Paris registration number; the public inconsistency is unresolved. DDD’s published Terms of Use name DDD Ventures LLC and use investment/real-estate language, so that page does not settle the AI-data contracting identity. iSoftStone’s privacy policy identifies a website-scoped US operator without resolving ultimate ownership or group-wide contracting responsibility. These findings preserve uncertainty rather than invent corporate transactions or additional suppliers.

Sources: [348], [357], [361]

Analysis: Secondary use and human access can be explicit commercial conditions. Datasaur's terms contain a free-access deidentified/aggregate improvement exception; Deploy's terms default to manual Hive Worker review of customer content unless otherwise agreed in writing; Dataocean's website policy leaves product/project rules to the corresponding services. None of these clauses alone establishes what happened to a particular customer's records. A buyer should identify the applicable subscription, written exceptions, worker exposure, use permissions and service-specific privacy terms before choosing a delivery path.

Sources: [343], [352], [372]

Evidence: Toloka's published task guidelines distinguish public-space images, third-party personal information and personal information collected from Tolokers. For third-party information, the listed conditions include both anonymization before upload and consent agreements. Collection from Tolokers also carries lawful-collection, administering-entity, purpose and policy conditions. These are category-specific platform rules, not a finding that every dataset complies. The linked additional collection guidelines were not separately captured in the reviewed scope.

Sources: [755]

Analysis: Biometric and authorship metadata deserve separate consent and use records. Toloka's notice bases identity matching on explicit consent and says refusal can prevent onboarding; it also discloses authorship metadata in some coding deliverables. Turing's policy lists AI development/training uses of personal information, while Polymath excludes interview voice recordings and transcripts from model training. These statements concern different records and services. Preserve which person, purpose, data type and downstream use each notice covers rather than combine them into a general permission claim.

Sources: [237], [757], [761]

Analysis: Security disclosure must retain attestation type and scope. Scale states SOC 2 Type 2 and ISO/IEC 27001:2022; Appen describes SOC 2 Type II, a HIPAA-specific solution and facility options; Snorkel states SOC 2 Type 2 and engagement-start scoping; Prolific describes ISO 27001 and Cyber Essentials while treating SOC 2 Type 2 as a future plan. Underlying reports, certificate scopes and customer deployments were not inspected here. These statements should not be flattened into equivalent independently verified controls.

Sources: [684], [710], [743], [748]

Analysis: Operational controls can be tiered or project-specific. CloudFactory distinguishes included network controls from optional endpoint controls; Sama describes assigned-employee project rooms; Encord describes role permissions; DataForce's vendor notice requires an applicable DPA before EU-personal-data processing. A buyer should verify the selected tier, location, legal entity, access matrix and actual data flow. The report records disclosed controls, without certifying that all personnel or projects implement them.

Sources: [50], [55], [82], [136]

Analysis: Incidents and remediation should be compared on evidence, not on whether one company has a public disclosure. Mercor's company investigation account describes the March 2026 compromised LiteLLM event and is explicitly not an independently audited impact/remediation finding. This report has no common incident-discovery process or equal disclosure window across suppliers. It therefore cannot rank relative security from the presence of that one record or assume that an unrecorded incident never occurred elsewhere.

Sources: [740]

Analysis: Artifact attribution is nontransitive. Scale's HelpSteer annotation, Translated's multilingual role, Prolific's recruitment case, Toloka's U-MATH validation and Surge's GSM8K collection are specific relationships. A paper citing a dataset, an author's affiliation, a supplier's general service menu or a benchmark family's parent link does not establish another supplier contribution. The graph is sparse in explicit object relationships; many approved observations are typed text. A reader should not confuse catalogue richness with an exhaustive supplier-to-artifact map.

Sources: [570], [741], [745], [751], [758]

Method: Research governance: the publisher's public identity discloses applications for paid work at some covered companies, membership in A.Team's talent network and a separate proposed AI research lab. No profiled company has paid for coverage or is a customer, investor or affiliate of the publisher under that disclosure. External reviewer credit requires consent. This preparation claims actual separate AI source review and will require separate exact-report evidence-auditor and adversarial review; it does not claim human peer review or an external operational certification.

Research gap: Governance unknowns include project-specific contracts, current assurance-report scopes, removal-request outcomes, consent withdrawals, subprocessor execution and worker appeal effectiveness. These cannot be resolved by treating a top-level license, policy statement or source-content review as an operational audit. Buyer diligence should obtain the relevant documents and test the actual project boundary; future research should record both support and unresolved limitations.

Analysis: Disclosure completeness is an observable governance input, while actual conduct remains a separate question. GIGO audits publication reporting in a selected literature; Who Annotates extracts task-level disclosures with an automated model and a separately adjudicated reference. Both reject equating missing reporting with absent practice. Their results justify asking for training, language proficiency, instructions and reconciliation records. They do not justify ranking a supplier's ethics or data quality from how many public fields are filled.

Sources: [427], [439]

Analysis: A documented consent workflow does not finish privacy diligence. WildChat reports an agreement plus secondary confirmation and multilingual PII filtering, yet warns that hashed IPs and request headers can link conversations. The reported process was not independently tested here. For a proposed conversational-data use, examine the exact released fields, linkage routes, permissions and intended population; preserve the difference between a source's stated control and demonstrated residual risk reduction.

Sources: [440]

Analysis: Public availability and private evaluation are separate diligence questions. Riemann-Bench is private; APEX-Accounting distinguishes private tasks and a small public development world; ALE has unresolved public-instance count wording; DAYJOB evaluates more finance tasks than its public finance release. None of these facts alone invalidates the scientific use, but each changes independent reproducibility and a buyer's access obligation. Confidence is high in this distinction. Require a precise access manifest, permitted reproduction scope and complete task/attempt accounting; inspected payloads and executable grading checks would change reproducibility confidence.

Sources: [21], [683], [695], [697]

Analysis: Rights evidence should follow each component. A paper's benchmark declaration does not establish rights to every included chart, source image, copied repository or professional document. StarCoder explicitly retains license/opt-out detection limits; Chartography's online provenance does not establish chart redistribution rights; the new VQA family reports a creator CC-BY release statement without a per-image audit. Confidence is high in the need to keep those scopes separate and unestablished for any particular buyer's legal permission. Obtain governing release instruments and component-level provenance rather than convert source review into a rights certification.

Sources: [3], [476], [524], [525]

Analysis: Versioned rights comparison - permission statements attach to different objects. GeneBench-Pro's preprint carries CC-BY-NC 4.0 while its public case-study package declares MIT; CVDP separates non-code, original code and upstream components with two unspecified NOTICE entries; DeepResearch Bench II distinguishes evaluation code from task-specific licenses and later generated model reports. Confidence is high in keeping those declarations separate and unestablished for a buyer's exact downstream grant. Build a component-and-use schedule with source documents and obtain the applicable instruments. A verified exact payload and rights chain would change the assessment; the synthesis itself grants no source-corpus rights.

Sources: [446], [610], [611], [612], [640], [641], [642], [643], [644], [645], [673]

Analysis: A benchmark-use instruction can coexist with a permissive metadata license. GDP.xlsx preserves an Apache label beside directions excluding training and checkpoint selection. μ-MATH separates model-output license terms from other fields. AA-LCR licenses its questions while declining to claim rights over accompanying document text, and the Nexdata listings describe paid commercial access whose terms depend on the project. These observations support a component-and-use matrix covering questions, source media, traces, annotations, code and execution environments. They do not provide a legal reconciliation or blanket permission. Confidence is high in the disclosed scope distinctions; a rights judgment would require the actual applicable terms and permissions.

Sources: [11], [17], [623], [626], [762]

Analysis: Declared privacy and synthetic fixtures should be reported at the object actually described. EMRBench’s fixture is described as synthetic; RedlineBench describes fictional parties and scrubbed metadata; Nexdata’s hand-pose listing says facial-identifying information is absent. None of these observations is an independent consent, reidentification or downstream-use audit. The reasoning is to request a data-flow and provenance record for each delivered component rather than infer complete privacy protection from a category label. Confidence is high in preserving these public-source limits and unmeasured in any private payload’s actual properties. An authorized component-level inspection and project-specific provenance evidence could narrow those unknowns.

Sources: [14], [597], [777]

10. Buyer decision guides: use cases, pilot designs and acceptance

Method: The following are proposed buyer procedures derived from the reviewed evidence, not reports of completed supplier trials. They do not recommend a universal winner. Each use case states the required deliverable, a candidate responsibility model, a paid pilot and acceptance evidence. The buyer should pre-register critical errors, slices, exclusions, rework limits and decision rules before seeing outputs. Numeric targets should follow the buyer’s risk and uncertainty requirements; no universal accuracy threshold is invented here.

Analysis: Buyer use case 1 - Expert preference data for a specialized assistant: requirement and comparison. Specify prompts, candidate responses, domain credentials, language, independent judgments per item, preference strength and rationale fields. Compare a managed preference program with participant access only after assigning task design, recruitment, adjudication and acceptance. HelpSteer3, HelpSteer2 and InstructGPT show why final pair counts conceal different repetition and aggregation structures.

Sources: [451], [715], [731]

Analysis: Buyer use case 1 - Expert preference data for a specialized assistant: proposed paid pilot. Use a paid calibration batch spanning routine, ambiguous and deliberately difficult cases. Freeze the raw-label schema before consensus. Have a separate reviewer inspect disagreements and rejected items without knowing the supplier or initial majority. Keep duplicate prompts and ranked-set expansion identifiable, then hold out a second batch for acceptance.

Sources: [451], [715], [731]

Analysis: Buyer use case 1 - Expert preference data for a specialized assistant: proposed acceptance and limits. Accept only against the predeclared target: valid independent judgments, traceable exclusions, correct domain-language intersections and adjudication supported by reasons. Report first-pass and final acceptance with all submitted items in the denominator. Do not let post-filter agreement substitute for coverage of difficult cases. Reject an unsupported claim that more released pairs necessarily means more independent human evidence.

Sources: [451], [715], [731]

Analysis: Buyer use case 2 - Repository engineering and coding-agent environments: requirement and comparison. Define repository language/version, issue requirements, permitted tools, test suite, reset behavior, task horizon and artifact rights. Compare custom environment builders, engineering teams and licensed task packages on the same accepted unit. G2i’s contract boundary and A.Team’s engagement models require special attention when responsibility for the final environment could remain with the buyer.

Sources: [227], [285], [305], [568], [717], [733]

Analysis: Buyer use case 2 - Repository engineering and coding-agent environments: proposed paid pilot. Run known-correct, known-incorrect and superficially passing patches in isolated copies. Repeat resets and tests, record installation/network failures separately and sample passing changes for completeness, readability and unintended behavior. Freeze the environment image, dependencies, time limits and verifier version. Inspect a hidden test that catches a plausible shortcut rather than merely reproducing the reference implementation.

Sources: [227], [285], [305], [568], [717], [733]

Analysis: Buyer use case 2 - Repository engineering and coding-agent environments: proposed acceptance and limits. Accept reproducible environments and task-scoped outcomes under the agreed protocol, with infrastructure failures explicitly adjudicated. Require pass/fail consistency, regression protection and side-effect review. SWE-bench cautions that passing tests does not guarantee complete solutions, while MBPP shows hard-coded-input gaming. A supplier’s benchmark authorship or engineering credentials alone does not satisfy these tests.

Sources: [227], [285], [305], [568], [717], [733]

Analysis: Buyer use case 3 - Enterprise browser and desktop workflow training: requirement and comparison. Specify an initial state, user goal, authorized actions, protected data and prohibited side effects. Decide whether the deliverable is human demonstrations, a simulated environment, a verifier or all three. AfterQuery’s demonstrations, Abundant’s customer-workflow conversion and Refresh’s versioned visual checks represent different disclosed units. A recording without an outcome specification is not yet an accepted RL task.

Sources: [91], [191], [206], [207], [233], [250], [257]

Analysis: Buyer use case 3 - Enterprise browser and desktop workflow training: proposed paid pilot. Use synthetic or approved nonproduction accounts and fixed starting documents. Capture action traces, final state and constraint violations. Include recoverable mistakes, irreversible actions, ambiguous requests and tasks that should be refused. Test evaluator behavior on completed tasks with a deliberately harmful extra action, and on valid alternate solutions that differ from the reference sequence.

Sources: [91], [191], [206], [207], [233], [250], [257]

Analysis: Buyer use case 3 - Enterprise browser and desktop workflow training: proposed acceptance and limits. Require reproducible resets, authorized-state transitions and preservation of protected content, alongside end-goal success. OSWorld explicitly identifies side-effect detection limits and configuration failures; WebShop documents lexical reward errors; WebArena uses mixed semantic and state checks. Acceptance should distinguish environment, policy, model and judge failures so a low score does not silently become a provider-quality claim.

Sources: [91], [191], [206], [207], [233], [250], [257]

Analysis: Buyer use case 4 - Multilingual assistant and localization evaluation: requirement and comparison. State language variety, script, locale, domain, task type and whether native authoring or translation is required. Compare native task creation, post-edited translation and multilingual judging as different workflows. HelpSteer3’s missing intersections and M-IFEval’s language-specific constraint changes make a broad language count an inadequate specification.

Sources: [381], [390], [391], [396], [715]

Analysis: Buyer use case 4 - Multilingual assistant and localization evaluation: proposed paid pilot. Construct a matrix of required language-domain pairs and sample every promised intersection. Use native-domain reviewers, preserve original and translated prompts and test local factuality, tone and instructions separately. Include constraints whose English formulation is invalid in the target script. Blind model identity and candidate order, then compare judge decisions with held-out human decisions under the actual target-language protocol.

Sources: [381], [390], [391], [396], [715]

Analysis: Buyer use case 4 - Multilingual assistant and localization evaluation: proposed acceptance and limits. Accept documented coverage and calibrated errors within each required slice, with no unsupported extrapolation from an overall average. Report translation revisions and unresolved dialect disagreements. MENLO’s dimensions, BenchMAX’s post-editing and the multilingual-judge setup show why aggregate cross-language consistency cannot replace validity on the buyer’s real conversations.

Sources: [381], [390], [391], [396], [715]

Analysis: Buyer use case 5 - Medical, legal or financial expert data: requirement and comparison. Specify the permitted professional task, jurisdiction or clinical context, credential evidence, privacy boundary and intended model use. Shaip’s credential matching and Encord’s clinician/task matching are offering disclosures, while Aya Data’s Cydar case concerns specific image labels. Do not transfer evidence from image annotation to clinical advice quality or from general finance work to jurisdiction-specific legal accuracy.

Sources: [135], [157], [351], [716]

Analysis: Buyer use case 5 - Medical, legal or financial expert data: proposed paid pilot. Use de-identified or consented material under a project processing agreement. Have qualified reviewers build an independent reference and define cases requiring escalation rather than forced answers. Record source authority, uncertainty and conflicting professional opinions. Separate rubric authoring from acceptance where feasible, and hold out difficult or out-of-scope cases that test whether annotators recognize the boundary.

Sources: [135], [157], [351], [716]

Analysis: Buyer use case 5 - Medical, legal or financial expert data: proposed acceptance and limits. Accept only the defined artifact under the qualified review protocol, with provenance and privacy records and no unresolved critical error in the buyer’s predeclared category. Credentials establish eligibility, not correctness of every output. GPQA’s authors reject representativeness of scientific practice and flag selection-biased validator accuracies, illustrating why expert benchmark performance cannot certify a professional service.

Sources: [135], [157], [351], [716]

Analysis: Buyer use case 6 - Safety evaluation, red teaming and policy judges: requirement and comparison. Choose a threat model, target policies, allowed attack surfaces, severity taxonomy, benign near-neighbors and remediation owner. Compare novel expert attack discovery, automated attack generation and an integrated policy judge separately. Giskard, Alinia and Trajectory document different approaches; a single attack-success rate is not a common deliverable or a claim of exhaustive safety.

Sources: [154], [223], [225], [288], [301], [331]

Analysis: Buyer use case 6 - Safety evaluation, red teaming and policy judges: proposed paid pilot. Run attacks adapted to the actual defense, with fixed generation budgets and held-out attack families. Ask independent reviewers to adjudicate both successful attacks and benign refusals. Record severity and plausible impact separately from whether a classifier marks harmful content. Inspect worker notice, opt-out and support arrangements before exposure to sensitive material.

Sources: [154], [223], [225], [288], [301], [331]

Analysis: Buyer use case 6 - Safety evaluation, red teaming and policy judges: proposed acceptance and limits. Accept a reproducible finding set, false-positive accounting and verified remediation/retest evidence under the scoped threat model. HarmBench shows budget confounds; JailbreakBench distinguishes adaptive defense evaluation and paired benign behaviors; the human red-team study flags incomplete expertise and coverage. No finite pilot proves absence of all harmful behavior.

Sources: [154], [223], [225], [288], [301], [331]

Analysis: Buyer use case 7 - RAG evaluation and LLM-judge calibration: requirement and comparison. Define the evidence corpus, query distribution, answer criteria, citation validity and abstention behavior. A human golden set, a tuned judge and an operational dashboard are distinct deliverables. SuperAnnotate’s Databricks account documents rubric construction and human/model agreement, while Giskard’s tooling permits expert ground truths and qualification of failures. The buyer still owns the intended-use definition.

Sources: [29], [84], [302], [367]

Analysis: Buyer use case 7 - RAG evaluation and LLM-judge calibration: proposed paid pilot. Sample answerable, unanswerable, misleading and conflicting-evidence questions. Keep calibration examples separate from final acceptance. Blind source model and response order, test stylistic/length variation and examine errors by failure type. Have an independent reviewer inspect whether a plausible citation actually supports the answer, rather than treating a judge’s high scalar rating as evidence fidelity.

Sources: [29], [84], [302], [367]

Analysis: Buyer use case 7 - RAG evaluation and LLM-judge calibration: proposed acceptance and limits. Accept the frozen judge only within the measured scope, documenting thresholds, abstentions and known blind spots. RewardBench 2 warns that reward benchmark accuracy alone is insufficient for strong RLHF training; AlpacaEval warns style can outweigh factuality and that harms are not measured. A judge calibrated on one response population needs renewed testing when the generator or domain changes.

Sources: [29], [84], [302], [367]

Analysis: Buyer use case 8 - Licensed or off-the-shelf training data: requirement and comparison. Specify exact release, use rights, exclusivity, refresh entitlement, permitted transformations, deletion/removal obligations and overlap exclusions. Snorkel’s non-exclusive refreshed products, Mercor’s catalog and Defined.ai’s marketplace/service separation show why a dataset purchase needs its own license and provenance schedule. A quantity quote should identify rows, tokens, hours or tasks rather than use examples as an undefined unit.

Sources: [29], [59], [177], [449], [684], [690]

Analysis: Buyer use case 8 - Licensed or off-the-shelf training data: proposed paid pilot. Inspect a representative sample and the component/source manifest before full acceptance. Check format validity, duplicates, contamination against the buyer’s holdouts, sensitive information and the documented origin of labels. Review a random sample as well as hard cases so cherry-picked examples cannot define the result. Resolve license conflicts and restricted components before using the data for the intended training or redistribution.

Sources: [29], [59], [177], [449], [684], [690]

Analysis: Buyer use case 8 - Licensed or off-the-shelf training data: proposed acceptance and limits. Accept the versioned artifact only if source rights and technical fitness meet the agreed project scope. RewardBench component terms, AlpacaEval’s code/data split and MathVista’s training restriction demonstrate why a top-level license label is insufficient. A public source or approved catalogue entry is not a rights clearance or a warrant for downstream commercial use.

Sources: [29], [59], [177], [449], [684], [690]

Analysis: Buyer use case 9 - Multimodal and physical-AI data collection: requirement and comparison. Specify sensors, synchronization, environment, participant permissions, modality alignment, annotation ontology and the accepted recording unit. Defined.ai’s robotics case, TELUS Digital’s collection modes and APTO’s modality menu are starting points for scope, not a common measured service. A recorded hour can contain different usable-event density, missing streams or annotation depth.

Sources: [57], [89], [135], [290]

Analysis: Buyer use case 9 - Multimodal and physical-AI data collection: proposed paid pilot. Run a controlled collection with calibration records, known synchronization events and independent inspection of unusable segments. Test edge conditions, privacy redaction and the linkage between raw recording and derived labels. Keep human review of model-generated first passes traceable, as Encord describes, and include downstream consumers when deciding which errors invalidate a unit.

Sources: [57], [89], [135], [290]

Analysis: Buyer use case 9 - Multimodal and physical-AI data collection: proposed acceptance and limits. Accept usable synchronized units with documented rights, calibration and error handling, rather than raw storage volume or gross duration. Report discarded or repaired recordings and modality-specific failure rates. The reviewed robotics hours and collection brochure do not establish success rates for the buyer’s sensors, field conditions or safety requirements.

Sources: [57], [89], [135], [290]

Analysis: Buyer use case 10 - Research-agent and long-horizon professional-task evaluation: requirement and comparison. Define task horizon, starting resources, permitted external code, human/tool assistance, time budget and the difference between partial and complete success. PaperBench, RE-Bench, ScienceAgentBench and TheAgentCompany expose different research-work abstractions. Their results should inform test design, not be translated into a general claim about replacing a profession or performing an arbitrary research project.

Sources: [334], [365], [366], [368]

Analysis: Buyer use case 10 - Research-agent and long-horizon professional-task evaluation: proposed paid pilot. Freeze a small private task set with independently authored rubrics and executable checks. Record intermediate artifacts and final outputs; have reviewers inspect whether a claimed replication actually reproduces results under the permitted resources. Match human and agent tool access where the comparison requires it, and report wall time, compute and expert-review effort separately.

Sources: [334], [365], [366], [368]

Analysis: Buyer use case 10 - Research-agent and long-horizon professional-task evaluation: proposed acceptance and limits. Accept a reproducible task-scoped result with transparent leaf decisions, execution failures and uncertainty. PaperBench’s Code-Dev correlation limit, RE-Bench’s representativeness warning and ScienceAgentBench’s language selection boundary show that even a well-defined score can be narrow. A complete-success claim must not be substituted for partial credit or a rubric-weighted average.

Sources: [334], [365], [366], [368]

Method: Common pilot record: versioned brief; source and rights manifest; task/environment version; contributor qualification rule; raw judgments or trajectories; review and exclusion log; adjudication reasons; acceptance denominator; cost inclusions; and residual risks. Separate supplier process errors from buyer guideline defects, environment failures and model failures. The final purchase decision should state which evidence established suitability and which assumptions remain conditional. Keep unsuccessful pilots and rejected examples in the internal learning record so later comparisons do not inherit survivorship bias.

Analysis: Pilot refinement for expert preference and multilingual data: retain an unfiltered calibration sample alongside the accepted sample. Score raw agreement, exclusions, adjudication and acceptance separately; inspect whether difficult but relevant cases disappear. HelpSteer2's κ change is conditional on multistage filtering, Aya's language total does not guarantee uniform contributor diversity or safety review, and Inherent Disagreements warns against pooling batch-normalized scales. Predeclare whether the buyer needs consensus labels, calibrated scalar attributes or a distribution of valid preferences before requesting a price per example.

Sources: [422], [430], [431]

Analysis: Pilot refinement for research agents and feedback services: freeze the feedback visibility contract. RECODE-H's richest condition supplies correct code and its evaluated expert is simulated; AutoLibra collects one feedback element for a complete trajectory and does not demonstrate live step-level guidance. Compare these against the buyer's actual need: post-hoc diagnosis, interactive help or verifiable task scoring. Count feedback rounds, solution material disclosed, human effort and independent final success separately so richer assistance is not mistaken for better unaided agent capability.

Sources: [421], [436]

Analysis: Pilot refinement for label automation and judge calibration: establish reference labels independently before comparing the system under test. Who Annotates resolves two independent annotations through discussion and third-person adjudication; the benchmark-label study retains uncertainty even after five-worker review. Include easy, ambiguous and rejected cases, then state whether the metric measures agreement with a chosen reference or verified outcome correctness. This proposed acceptance design should document disagreements and confidence rather than silently discard cases that challenge either the reference or the evaluated system.

Sources: [428], [433], [439]

Analysis: Acceptance refinement for use cases 1, 4 and 9 - choose the independent unit before setting a tolerance. Preserve item, participant, conversation, source-file and translation-group identifiers. Require complete delivery of raw labels, exclusions and adjudication reasons, and evaluate a held-out sample before aggregation/filtering. Set the acceptable error and coverage targets in the signed brief by severity and stratum. Do not accept a high aggregate agreement score when a required language-domain intersection has no reviewed items. Confidence is high in these denominator requirements; target tolerances remain buyer choices rather than measured results of this report.

Sources: [526], [554], [558], [699]

Analysis: Acceptance refinement for use cases 2, 3 and 10 - retain both strict completion and repeated reliability. For each frozen task, store every launch, infrastructure failure, retry, terminal state and violated constraint. Predeclare which failures are infrastructure exclusions, report the exclusion denominator and rerun repaired infrastructure without silently replacing scientific failures. Require all decisive safety/authorization/regression conditions to pass for an accepted task; display fractional quality as a separate diagnostic. Test known-good alternate solutions, known-bad outputs and superficially compliant outputs. Confidence is high in this protocol requirement; an observed pilot result is still needed to select a supplier.

Sources: [10], [561], [698], [700]

Analysis: Acceptance refinement for use cases 5 and 7 - give the judge the evidence required by the intended decision. Have independent experts adjudicate unsupported claims, incomplete but plausible answers, conflicting documents and decisive missed criteria. Report false accepts and false rejects separately by severity, alongside any agreement statistic. A self-judge or answer-only judge requires an explicit calibration rationale; a benchmark’s calibration sample does not certify the buyer’s domain. Confidence is high in this separation and unestablished for any candidate judge. Choose the operating threshold only after a blinded held-out evaluation.

Sources: [3], [8], [549], [562]

Analysis: Acceptance refinement for use case 6 - compare safety conditions under the same outcome definition. Retain target harmful intent, attack family, attempt budget, current defense version, model-side refusals and practical harm judgments. Include both overt and revised wording while independently checking that the underlying request remains comparable. Report classifier-flagged candidates separately from adjudicated failures and preserve false-negative sampling. Confidence is high in those measurement requirements and unestablished for deployment-wide attack prevalence. Acceptance should be against the predeclared threat model and error tolerances, without interpreting a selected red-team set as a random incidence estimate.

Sources: [459], [551], [703]

Method: A proposed statistical acceptance rule should make its assumptions explicit. For n independent randomly sampled units with zero observed critical defects, a one-sided 95% binomial upper bound is $1 - 0.05^{(1/n)}$; excluding a true defect rate p^* at that confidence requires $n \geq \log(0.05)/\log(1 - p^*)$. This is a design calculation, not an observed quality result. Repeated attempts, translations, shared source items and repeated participants can violate independence; use the original sampling unit and cluster-aware inference or report descriptive results instead. A zero-defect pilot does not prove zero population risk, and safety-critical controls can require a separate exhaustive gate.

Method: Proposed commercial acceptance record. Freeze the deliverable version, coverage matrix, rights/processing instruments, permitted subcontracting, rejection/rework rules, escalation path and payment basis before production. Require a complete manifest and reproducible checks for every accepted unit, plus a stratified blinded sample for substantive correctness. Log paid qualification and revision time separately from production units and include failed attempts, review and infrastructure costs in the project ledger. These are analyst-proposed conditions, not a representation that any named supplier has accepted them or that a pilot has occurred.

Analysis: Pilot refinement for original use cases 2, 5, 7 and 10 - separate evaluation integrity from score quality. Require a pristine grading boundary, reject attempts to read hidden references and test correct alternative implementations as well as deliberately invalid ones. DeepSWE’s separated verifier environment, ProgramBench’s cleanroom and dummy-binary filtering, and SWE Atlas’s no-op mutation checks suggest different falsification cases; none guarantees complete coverage. Confidence is high in retaining these controls and unmeasured for a particular supplier. Predeclare the critical false-accept conditions and inspect all exceptions. Independent adversarial execution on a frozen pilot would change acceptance confidence.

Sources: [532], [533], [655], [667], [668], [669], [670], [671], [672]

Analysis: Pilot refinement for original use cases 1, 4, 8 and 9 - separate source examples from prompted rows and repeated judgments. Flan and xP3 transform/reweight upstream material; Chatbot Arena Conversations pairs human winner votes with two model responses; ELOQUENT separately predicts preference classes and explanations. Confidence is high that counting all rows as independent human labels would be misleading. Require source/group identifiers, derivation manifests, raw judgments and version-specific component terms before computing acceptance statistics. A stratified independent sample of the buyer’s actual use and a complete overlap audit would change confidence in reuse.

Sources: [97], [165], [166], [204], [542], [543], [544], [545], [546], [547], [548]

Analysis: Extension to the existing environment and software pilots. Freeze the exact task roster, container/dependency manifest, instructions, hidden references, judge version, tool access, time limit and repeat aggregation before comparing candidates. Terminal-Bench’s mutable dependencies, SWE-Marathon’s distinction between development feedback and hidden final checks, and APEX-SWE’s telemetry-based diagnosis without failing tests show why a single runnable-environment label is insufficient. The proposed acceptance exercise should separately record startup failure, infrastructure interruption, ungradable output, task failure and constraint violation, then challenge the scorer with a no-op and a valid alternate implementation. This is an extension of the retained pilot designs, not a newly executed trial. Confidence in any supplier remains unmeasured until the buyer’s own held-out acceptance results exist.

Sources: [377], [559], [601], [763]

Analysis: Extension to the existing document and professional-work pilots. Preserve native deliverables and evaluate validity separately from substantive utility: PDFBench distinguishes a form's strict pass from field-level correctness; RedlineBench requires a loadable Word redline and contextual attorney criteria; Harvey LAB separates deliverable existence from rubric satisfaction; PLE separates infrastructure reruns from runs with no gradable deliverable. A proposed pilot should include apparently polished artifacts with a critical missing field, prohibited edit, unsupported claim or invalid native structure, together with valid alternatives. These examples motivate acceptance design, not a forecast of legal, clinical or engineering job replacement. Confidence is high in the metric separation and unmeasured for a new deployment. Independent domain review of the buyer's artifacts would determine whether the proposed checks are adequate.

Sources: [599], [700], [771], [772], [777], [778], [779]

Analysis: Extension to the existing supervised-data and feedback pilots. Request item-level prompt and response provenance, the exact selection/filter decision, any reference or test availability, raw and accepted denominators, and applicable component terms. AM's verification-language conflict, s1K's unresolved schema differences, SmallThoughts' missing item-verification procedure and Multi-SWE-RL's omitted manual phase make these concrete inspection targets. The proposed pilot should sample rejected and accepted items and independently inspect plausible false positives and valid discarded alternatives. No training or supplier comparison was executed for this report. Confidence is high in the disclosed documentation gaps; measured error, utility and cost require the buyer's own matched experiment.

Sources: [441], [444], [445], [629]

11. Research outlook: testable judgments and open questions

Analysis: Hypothesis - stage-level provenance will become a more useful comparison than a binary human/synthetic label. The sample already contains human-authored preferences, model-generated feedback, synthetic educational scores and human correction of model responses. Confidence is moderate for the usefulness of this distinction and unestablished for its effect on procurement outcomes. The hypothesis would strengthen if provenance-complete manifests predict fewer acceptance surprises on matched tasks; it would weaken if those details fail to add information beyond ordinary samples and contracts. This is a research proposition, not a forecast of market adoption.

Sources: [296], [454], [715], [734]

Analysis: Hypothesis - environment and verifier maintenance will be a distinct recurring delivery burden. Refresh describes versioned verifiers, OSWorld announces a verified-version change, and SWE-bench Multimodal reports repeated-test filtering. Confidence is moderate that version and stability matter within these examples, but the corpus supplies no market-wide maintenance cost. A longitudinal study should record environment breakage, verifier corrections and acceptance changes under frozen agents. If stable tasks require little upkeep across relevant software changes, the hypothesized burden would be narrower than suggested.

Sources: [191], [250], [374]

Analysis: Hypothesis - multilingual validity will require native task design as well as translation quality. M-IFEval modifies constraints, MENLO isolates local dimensions and BenchMAX post-edits translated data. Confidence is moderate for the methodological distinction, without an estimated commercial premium. A useful experiment would match translated, post-edited and native-authored tasks on intended constructs and independently score local failures. Equal validity under those controlled conditions would weaken the need for a separate native-authoring stage in that particular use case.

Sources: [381], [390], [391]

Analysis: Hypothesis - reliability reporting will need both repeated execution and disagreement preservation. τ-bench distinguishes all-trials reliability from at-least-one success, while DICES and the disagreement study expose meaningful variation in human labels. Confidence is high that these are different denominators; the effect on supplier selection remains untested. A matched pilot should compare decisions made from aggregate success alone with decisions using repeated trials, raw labels and adjudication reasons. The hypothesis would weaken if added detail does not change acceptance or predict deployment errors.

Sources: [132], [133], [254]

Analysis: Hypothesis - reward-model and monitor evaluations will need adversarial separation from the optimization loop. Summarization overoptimization, verifier search exploitation and the reasoning-monitor study demonstrate specific mechanisms by which optimization can outpace a proxy. Confidence is moderate for the need to test this risk, not for its prevalence in current services. Hold out failure classes and reviewer inputs when optimizing the system, then measure independent outcome errors. A robust matched study showing stable external validity across optimization intensity would narrow this concern.

Sources: [90], [158], [393]

Analysis: Hypothesis - labor and documentation quality should be evaluated as operating processes, not only disclosure artifacts. CrowdWorkSheets asks for maintenance and pool-selection rationale; Datasheets asks about collection participants and compensation; the labor study warns about unknown population boundaries. Confidence is moderate for the research value of these fields, with no causal claim that publishing them improves wages or model quality. Follow-up research should compare documentation with consented worker evidence and project logs, retaining discrepancies and avoiding a disclosure-completeness score that masquerades as performance.

Sources: [401], [409], [730]

Research gap: Priority evidence gaps: comparable accepted-unit economics; contributor-level paid-active windows; independent project outcomes; version-specific rights clearance; non-English and non-text coverage intersections; benchmark side-effect detection; and a complete source-supported supplier contribution map. The current graph's richer catalogue does not resolve these automatically. Research should prioritize decision-changing evidence rather than optimize counts, and should document negative findings and unresolved cases alongside additions.

Method: The weekly outlook should be revised through falsifiable change conditions. Each stable executive finding records confidence, reasoning, limitations and what would change the assessment. New primary evidence can strengthen, narrow or supersede a finding; a changed source version should not silently overwrite its predecessor. A meaningful revision will identify which observed assumption changed and whether the effect is task-specific or supports a broader conclusion. No growth, revenue, employment or technology-adoption forecast is produced from this selected corpus.

Analysis: Hypothesis - a stronger documentation audit will improve diligence only if reporting signals are checked against project evidence. GIGO and Who Annotates observe disclosure, not undisclosed practice; the latter also exposes extraction uncertainty. Confidence is moderate in the need to distinguish those objects and unmeasured in the commercial value of an audit score. A follow-up could compare disclosed instructions, training and compensation fields with consented project records and independently sampled outputs. If completeness fails to predict agreement between disclosure and practice, its role should remain a transparency measure rather than a performance proxy.

Sources: [427], [439]

Analysis: Hypothesis - complete-task reliability will be more decision-relevant than fractional rubric completion in workflows with decisive controls. APEX-Accounting, HANDBOOK.md and DAYJOB distinguish strict/relaxed or mean-criterion/repeated-complete metrics; PLE's hybrid score exposes a different tradeoff. Confidence is moderate in the hypothesis and high that the metrics differ. A prospective workflow study should compare acceptance and downstream failure predicted by each metric under a fixed task population. If fractional scores predict accepted outcomes as well after severity and review costs are controlled, the proposed emphasis should narrow.

Sources: [10], [695], [697], [700]

Analysis: Hypothesis - purported transfer mechanisms will become clearer when data design is treated as an experimental treatment. The connected office-work studies generate descriptive transfer and trace hypotheses; Corecraft and the coding-task study add different recipes and baseline conditions. Confidence is low in a shared causal explanation from these reports. A multi-seed, equal-budget factorial study separating task domain, demonstration quality, reward shaping and base-model choice could strengthen a particular mechanism. Null or domain-specific findings would weaken broad claims that one data category generalizes across professional work.

Sources: [6], [685], [698], [702]

Analysis: Hypothesis - disclosure of excluded work and selected populations will improve procurement decisions more than larger raw row counts. Qualification self-selection, disagreement filtering, adversarial task screening and translated rows all change the effective denominator in the reviewed studies. Confidence is moderate in the decision relevance and unmeasured in commercial impact. A blinded buyer study could compare choices made from headline totals with choices made from full admission/exclusion manifests, then validate the selected service on independent work. No effect on selection accuracy or accepted cost would narrow this hypothesis.

Sources: [8], [458], [526], [558]

Analysis: Hypothesis - benchmark maintenance will become a separate recurring research and procurement cost. SWE Atlas changes grading while retaining tasks; MMLU-Pro corrects answers and whitespace; ProgramBench has paper/card difficulty differences; TH1.1 changes tasks and human-time estimates as well as infrastructure. Confidence is moderate in the maintenance-burden hypothesis and high that these are real version distinctions. A prospective cost ledger should compare revalidation, regrading and downstream decision changes under frozen versus updated protocols. If stable old protocols predict accepted outcomes equally well, the proposed maintenance premium should narrow.

Sources: [618], [653], [654], [655], [656], [657], [658], [667], [668], [669], [670], [671], [672]

Analysis: Testable judgment - release evolution can alter a score without a new logical family. AA-LCR repairs answer keys and warns against direct score comparison; DecoyCount distinguishes release withdrawals, stale derived answers and historical scores; Harvey LAB changes task rosters; DeepResearch Bench II changes the scored subset and judge protocol. The proposed hypothesis is that a meaningful portion of apparent score movement in an evolving evaluation could arise from these protocol or population changes. No fraction is estimated here. Confidence is high that matched-version replay is necessary and unknown about the contribution of each factor. Re-evaluating unchanged stored outputs under old/new scorers, then rerunning a fixed task intersection under fixed scaffolds, would separate these explanations.

Sources: [583], [584], [587], [588], [611], [762], [765], [766], [772], [773], [774]

Analysis: Testable judgment - downstream improvement needs a controlled training contrast rather than a promising environment specification. PostTrain Arena describes a fixed before/after held-out comparison, but its captured landing state pauses challenge runs and its later README distinguishes bounded pipeline checks from a completed held-out training run. The implication is methodological: freeze base model, compute, recipe, held-out suite and acceptance rules, then report repeated-run uncertainty before attributing a gain to contributed environments. This report asserts no completed PostTrain gain or supplier ranking. Confidence is high in the documented distinction and unknown for the intended program’s effectiveness. Actual accepted held-out runs with matched controls would supply the evidence that changes it.

Sources: [1], [12], [13]

12. Methods, coverage, bibliography and reproducible appendices

Method: Evidence boundary and source hierarchy. The factual boundary is the approved, validated merged graph selected by the explicit partition index. Discovery candidates, provisional reviews, search snippets and unpinned working files do not become report facts. Primary creator papers, official repository/card versions, provider policies, audited statements and first-party case accounts retain their source type. Provider marketing is reported as a disclosure. A source-content review checks whether an exact claim is supported; it does not transform self-reported operations into an independent audit.

Method: Assertion approval and review identity. Each approved assertion has exact-content author, source-verifier and adversarial receipts from separate actual AI invocations. Approval binds wording, subject, predicate, object or typed value, qualifications, source version, locator and excerpt digest. Report-level evidence-auditor and adversarial reviews are separate requirements bound to the exact report digest and complete graph digest. Prior report or graph reviews cannot approve this expanded synthesis. No fabricated human reviewer, honorific, institutional endorsement or review receipt is added.

Method: Original-stage notes and current status. Frozen source packets can retain author-only or proposal-stage wording from their original creation. Those notes are preserved as historical provenance; current assertion status and exact-digest three-role receipts determine admission. Superseded original assertions remain candidates, and new approved replacements retain their actual source and review dates. This synthesis incorporates qualified wording changes, not only ID substitutions. In particular, an admitted paper or family does not erase source conflicts, unread spans, experiment-reproduction limits or unresolved rights.

Method: Reading depth. full_source is a bounded editorial category: methods plus main results or worked examples and limitations, with separately recorded spans. It does not mean all appendices or proofs were read, every experiment reproduced or all corpus records checked. A truthful full_document_read false is compatible with adequate bounded main-section review. Abstract-only material cannot be promoted through a claim count. Dataset construction may be reviewed through creator methods and documentation without adding the original paper work as a fully reviewed paper or establishing corpus-byte identity.

Method: Capture-specific evidence. Cogito and Mindy observations bind the exact saved official extracted-page snapshots available to the reviewers; origin HTML and live HTTP status were not exposed by that extraction. Separate direct retrieval attempts had their own failures and do not turn the extraction into a verified live-origin response. The Mindy FAQ is undated, so current staffing/location figures are not adopted. TrustScale's rebrand article was decoded from the pinned public application bundle rather than established by rendered-page execution. These capture boundaries remain attached to the underlying source records and dossier limitations.

Method: Version and identity accounting. A paper work counts once through its reviewed manifestation; a dataset or benchmark family counts once through approved family/release evidence. Formats, configs, splits and subset releases do not add families. FineWeb-Edu is currently one FineWeb subset release, correcting the initial eight-family baseline's separate-family treatment without changing that archive's bytes. MuSiQue's new human-composed questions support one family; inherited seed corpora and answerability/serialization variants do not add families. Corporate aliases and product brands do not automatically add suppliers.

Method: Reproduction snapshot. This authoring report binds the complete partitioned load_graph result at canonical JSON SHA-256 2a7364f8adb6887fd8b005c7e8013d6ebaec9f10f29c320aa28fac91022d14ed, graph input release research-2026-11-11.candidate.1, and graph as_of 2026-11-12T17:50:15.040816Z. The report release research-2026-11-12.1 is independently versioned from that evidence input. Canonical dataset version 2.5.0 retains file SHA-256 948e69229fa24a4a1e6b5eed6d9c11570568a3c05bf9512e442d19611061a061. Recompute the graph digest after load_graph validates the approved-partition index and canonical bytes. The actual frozen-core identity digest is a413cd22ee5dceb633d2495e3d1bc592e75a15ca6ffd82f2f95beea8d119e5b4, frozen_on 2026-11-12T17:42:16.963477Z, with captured-source cutoff 2026-11-12T17:32:13.289438Z. These identifiers describe the research inputs and authoring provenance; archived-edition identity and public availability require the separately verified publication receipt. A graph, cutoff, release-identity or prose change requires fresh exact synthesis review bindings.

Method: Computed coverage appendix. Coverage method 1.3.0 counts 2,548 approved, 195 candidate and 39 historical assertions. There are 3,261 scoped source records, of which 3,260 are captured, representing 783 distinct captured document versions under the version/response-hash rule. The selected qualifying corpus has 73 supplier groups, 225 distinct reviewed paper works, 85 dataset families and 87 benchmark families. All 39 legacy organizations have bounded inclusion audits and 38 have conclusive dispositions; BenchFlow remains commercially pending. Operating brands such as Clickworker/LXT and Nexdata/Datatang are retained without extra group counts. The 416-unit frozen core has 389 deep dispositions and 27 outstanding deep-review gaps. The numerator and denominator are source-qualified inventory units, not all discoveries, all scientifically reviewed works, all potential supplier artifacts or the world market.

Method: Relationship appendix. Exactly 225 approved assertions use explicit object relationships. Other observations use source-bound typed values. Predicates are nontransitive: parent links support work/family counting only, and group identifiers deduplicate supplier coverage without extending every brand claim to every subsidiary. Software use, funding, contributor affiliation and expert recruitment do not establish unstated data-supply contracts, exclusive authorship, contribution shares or causal supplier performance.

Method: Launch gate appendix. The fixed floors require 39 completed legacy inclusion audits, 60 qualifying suppliers, 150 reviewed paper works, 60 dataset families, 20 benchmark families and disposition of at least 90% of a frozen eligible core. Commercial dispositions are separate; BenchFlow’s pending status does not invalidate its completed inclusion audit. The current bounded frozen core is 416 and its 389/416 deep-disposition rate exceeds the unchanged 90% floor. The source-bound actual finite 200-hypothesis reference audit is accepted against the unchanged 98% precision and 95% recall thresholds with no critical false attribution. The distinct screening audit must recover at least 190 of 200 eligible records and at least 90% in each major stratum. Numerical coverage and an authoring validation do not complete screening, synthesis review, source-cutoff, privacy, archive or technical requirements.

Method: Audit status and limitations. The accepted actual reference panel contains 185 true positives, zero false positives, seven false negatives and eight true negatives. Precision is 185/(185+0) = 1.0 and recall is 185/(185+7) = 0.9635416667. These quantities evaluate the disposition of original source-bound first-peer hypotheses in that finite panel. They do not estimate extraction recall, discovery recall, population recall or the precision of all final corrected outputs. Fifteen inputs were actually withheld, including seven supported opportunities. Original hypotheses, failed earlier validation, truth/prediction records, correction history and genuine review outcomes remain retained. No screening recovery or pass is inferred from the reference result; publication requires the separate accepted, dated exact screening result in the edition audit downloads. Accessible English-language and well-documented material remains overrepresented. The publication state of this exact report/graph pair is recorded separately in the verified edition metadata.

Method: Search and selection bias. This is a selected public-source corpus rather than a probability sample. English-language material, public research artifacts and providers with detailed documentation are easier to discover and inspect. Confidential contracts, unavailable or blocked documents, undocumented suppliers and unobserved language/domain intersections can be missing. Unknown fields are preserved as unknown and do not count as evidence of absence. The report’s comparisons are sample-limited and should not be extrapolated into the entire market.

Method: Source dates and cutoff. published_on records an established source publication date, captured_on its actual capture, reviewed_on its actual review and assertion as_of its declared observation date. This refresh retains those input dates and binds the actual frozen-core captured-source cutoff 2026-11-12T17:32:13.289438Z. Later source-content review can assess a pre-cutoff captured version without relabeling capture time. The first actual edition may follow only after all existing gates pass under the coordinator’s matching one-time launch authorization. Later editions return to Friday 17:00 Pacific cutoffs and Monday 09:00 Pacific publication targets. Delayed work, captures, archive confirmations and public activation are never backdated.

Method: Interpretation and reproducible calculations. Approved counts are computed from qualifying identities and current approved assertions. Sample ratios should identify numerator and denominator before rounding. Project cost and acceptance formulas in chapter 8 are proposed methods with no invented inputs. Paper-reported results in the atlas retain the exact task, version and measurement condition described by the approved statement; they are not independently reproduced results, current model rankings or estimates of supplier performance. No metric is averaged across incompatible benchmark protocols.

Method: Public reuse and downloads. The research hub is https://sciencetostartup.com/research/; the living report route is https://sciencetostartup.com/research/state-of-ai-data-labs/; the evidence and download index is https://sciencetostartup.com/research/data/; source-backed review and correction entry is https://sciencetostartup.com/review/. Download links in the prepared publication must bind its own JSON/CSV/SQLite, schemas, bibliography, checksum manifest, offline evidence appendix, PDF and reproduction package. This source report does not assert that the expanded matching package, DOI, Zenodo archive or Hugging Face mirror already exists.

Method: Publication and corrections. Website preparation, deployment, archive confirmation and mirror synchronization have separate states and receipts. Accounts remain prepare-only under the current policy. A source-backed correction should identify the claim or report block, the source version and locator, the proposed wording and the reason. The research workflow reviews support, privacy and identity before integrating changes. Historical frozen packages retain pinned bytes; a correction belongs in a new version and change log rather than an invisible replacement.

Method: Bibliography reading guide. The entries below provide the reviewed paper-work title, a source URL from its bound evidence and the inspected source version. Claim links carry the exact locator, excerpt hash, qualifications and review identities. A source URL can resolve to a later page; the recorded version and response digest define the reviewed content. Dataset and supplier bibliography records are available through each atlas/dossier’s claim bindings and the evidence ledger; no external affiliation is used as a substitute for a documented contribution.

Method: Formal audit boundaries. The actual reference audit and its independent admission exist separately from this author invocation; the author supplies no new gold labels, evaluated predictions or independent audit approval. Screening is a distinct edition-level audit. No author-only diagnostic, historical screening attempt, family count or reference-panel result establishes its completion. An actual later screening result must bind its own frozen membership, source-bound independent adjudications and evaluated input, with dated exact results supplied in the edition audit downloads. A changed report wording or method addendum requires its own exact synthesis review.

Method: Author/reviewer separation. This refresh is report authorship over existing approved research inputs. It does not create new primary-source reads, source-content review receipts or launch-audit evidence. The later evidence-auditor and adversarial synthesis invocations must assess the exact revised prose against this graph. This author provides no pass on their behalf.

Method: Source-binding correction history. The current correction ledger contains 151 records and preserves earlier history. This refresh uses the selected approved successors for 84 distinct quarantined claim IDs and records a prose/qualification decision for each of the 141 affected old paragraphs. Original candidates and earlier positive/negative receipts are unchanged. Source qualifications can retain original proposal-stage notes; current approved status and actual exact-content receipts determine current admission, while the scientific limitations remain. The unsupported Dataocean general-offer/legal-alias inference stays withheld. Correction history is not a substitute for the pending separate exact synthesis reviews.

Method: Locator reproduction. A locator identifies pinned source bytes and an extraction convention. One-based physical PDF pages differ from printed labels; global marked-text lines count markers and blank lines, while per-page pdftotext-layout coordinates use their declared physical page and tool convention. Preserve source versions, raw-response hashes and supplied derived-text hashes before interpreting a line range. Dataset-card revisions document the inspected card and do not automatically identify every corpus shard. A successful re-extraction is reproducibility evidence, rather than a new capture, scientific finding or independent content review.

Evidence: Reviewed-work bibliography - τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. Source: https://arxiv.org/pdf/2506.07982v1. Inspected version: 2506.07982v1.

Sources: [376]

Evidence: Reviewed-work bibliography - A General Theoretical Paradigm to Understand Learning from Human Preferences. Source: https://arxiv.org/pdf/2310.12036v1. Inspected version: 2310.12036v1.

Sources: [68]

Evidence: Reviewed-work bibliography - AgentBench: Evaluating LLMs as Agents. Source: https://arxiv.org/pdf/2308.03688v1. Inspected version: 2308.03688v1.

Sources: [48]

Evidence: Reviewed-work bibliography - AgentGym: Evolving Large Language Model-based Agents across Diverse Environments. Source: https://arxiv.org/pdf/2406.04151v1. Inspected version: 2406.04151v1.

Sources: [210]

Evidence: Reviewed-work bibliography - AlpaGasus: Training A Better Alpaca with Fewer Data. Source: https://arxiv.org/pdf/2307.08701v1. Inspected version: 2307.08701v1.

Sources: [121]

Evidence: Reviewed-work bibliography - Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. Source: https://arxiv.org/pdf/2402.14740v1. Inspected version: 2402.14740v1.

Sources: [318]

Evidence: Reviewed-work bibliography - BenchGuard: Who Guards the Benchmarks? Automated Auditing of LLM Agent Benchmarks. Source: https://arxiv.org/pdf/2604.24955v1. Inspected version: 2604.24955v1.

Sources: [337]

Evidence: Reviewed-work bibliography - BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models. Source: https://arxiv.org/pdf/2502.07346v1. Inspected version: 2502.07346v1.

Sources: [381]

Evidence: Reviewed-work bibliography - Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. Source: https://arxiv.org/pdf/2206.04615v1. Inspected version: 2206.04615v1.

Sources: [214]

Evidence: Reviewed-work bibliography - BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. Source: https://arxiv.org/pdf/2504.12516v1. Inspected version: 2504.12516v1.

Sources: [338]

Evidence: Reviewed-work bibliography - Consent in Crisis: The Rapid Decline of the AI Data Commons. Source: https://arxiv.org/pdf/2407.14933v1. Inspected version: 2407.14933v1.

Sources: [383]

Evidence: Reviewed-work bibliography - Constitutional AI: Harmlessness from AI Feedback. Source: https://arxiv.org/pdf/2212.08073v1. Inspected version: 2212.08073v1.

Sources: [51]

Evidence: Reviewed-work bibliography - Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. Source: https://arxiv.org/pdf/2501.18837v1. Inspected version: 2501.18837v1.

Sources: [293]

Evidence: Reviewed-work bibliography - CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. Source: https://arxiv.org/pdf/2206.08931v1. Inspected version: 2206.08931v1.

Sources: [401]

Evidence: Reviewed-work bibliography - DAPO: An Open-Source LLM Reinforcement Learning System at Scale. Source: https://arxiv.org/pdf/2503.14476v2. Inspected version: 2503.14476v2.

Sources: [402]

Evidence: Reviewed-work bibliography - Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. Source: https://arxiv.org/pdf/2204.01075v1. Inspected version: 2204.01075v1.

Sources: [403]

Evidence: Reviewed-work bibliography - Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Source: https://aclanthology.org/Q18-1041.pdf. Inspected version: ACL-Q18-1041-published-2018; anthology metadata commit ba3130005282c6474307cceec936373fe55d15ff.

Sources: [386]

Evidence: Reviewed-work bibliography - DataComp-LM: In search of the next generation of training sets for language models. Source: https://arxiv.org/pdf/2406.11794v1. Inspected version: 2406.11794v1.

Sources: [131]

Evidence: Reviewed-work bibliography - DataComp: In search of the next generation of multimodal datasets. Source: https://arxiv.org/pdf/2304.14108v1. Inspected version: 2304.14108v1.

Sources: [130]

Evidence: Reviewed-work bibliography - Datasheets for Datasets. Source: https://arxiv.org/abs/1803.09010v8. Inspected version: 1803.09010v8.

Sources: [721], [730]

Evidence: Reviewed-work bibliography - Deep Reinforcement Learning from Human Preferences. Source: https://arxiv.org/pdf/1706.03741v1. Inspected version: 1706.03741v1.

Sources: [60]

Evidence: Reviewed-work bibliography - DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Source: https://arxiv.org/pdf/2501.12948v1. Inspected version: 2501.12948v1.

Sources: [295]

Evidence: Reviewed-work bibliography - DeepSeek-V3 Technical Report. Source: https://arxiv.org/pdf/2412.19437v1. Inspected version: 2412.19437v1.

Sources: [296]

Evidence: Reviewed-work bibliography - DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. Source: https://arxiv.org/pdf/2402.03300v1. Inspected version: 2402.03300v1.

Sources: [294]

Evidence: Reviewed-work bibliography - Deliberative Alignment: Reasoning Enables Safer Language Models. Source: https://arxiv.org/pdf/2412.16339v1. Inspected version: 2412.16339v1.

Sources: [297]

Evidence: Reviewed-work bibliography - DICES Dataset: Diversity in Conversational AI Evaluation for Safety. Source: https://arxiv.org/pdf/2306.11247v1. Inspected version: 2306.11247v1.

Sources: [132]

Evidence: Reviewed-work bibliography - Direct Preference Optimization: Your Language Model is Secretly a Reward Model. Source: https://arxiv.org/abs/2305.18290v3. Inspected version: 2305.18290v3.

Sources: [725], [732]

Evidence: Reviewed-work bibliography - Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. Source: https://arxiv.org/pdf/2104.08758v2. Inspected version: 2104.08758v2.

Sources: [404]

Evidence: Reviewed-work bibliography - Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. Source: https://arxiv.org/pdf/2402.00159v1. Inspected version: 2402.00159v1.

Sources: [134]

Evidence: Reviewed-work bibliography - Evaluating Large Language Models Trained on Code. Source: https://arxiv.org/pdf/2107.03374v1. Inspected version: 2107.03374v1.

Sources: [224]

Evidence: Reviewed-work bibliography - Extracting Training Data from Large Language Models. Source: https://arxiv.org/pdf/2012.07805v2. Inspected version: 2012.07805v2.

Sources: [387]

Evidence: Reviewed-work bibliography - Fair human-centric image dataset for ethical AI benchmarking. Source: https://www.cs.utexas.edu/~pstone/Papers/bib2html-links/peter_nature_2025.pdf. Inspected version: institutional-copy-sha256-66ac69cbb2d33c170f14ff64072c07c51ab5d3ff66c2357a69dbb59d64fcaf78.

Sources: [388]

Evidence: Reviewed-work bibliography - GAIA: A Benchmark for General AI Assistants. Source: https://arxiv.org/pdf/2311.12983v1. Inspected version: 2311.12983v1.

Sources: [220]

Evidence: Reviewed-work bibliography - GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. Source: https://arxiv.org/pdf/2510.04374v1. Inspected version: 2510.04374v1.

Sources: [350]

Evidence: Reviewed-work bibliography - GPQA: A Graduate-Level Google-Proof Q&A Benchmark. Source: https://arxiv.org/pdf/2311.12022v1. Inspected version: 2311.12022v1.

Sources: [351]

Evidence: Reviewed-work bibliography - GPT-4 Technical Report. Source: https://arxiv.org/pdf/2303.08774v1. Inspected version: 2303.08774v1.

Sources: [304]

Evidence: Reviewed-work bibliography - Group Sequence Policy Optimization. Source: https://arxiv.org/pdf/2507.18071v2. Inspected version: 2507.18071v2.

Sources: [407]

Evidence: Reviewed-work bibliography - HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. Source: https://arxiv.org/pdf/2402.04249v1. Inspected version: 2402.04249v1.

Sources: [223]

Evidence: Reviewed-work bibliography - HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages. Source: https://arxiv.org/abs/2505.11475v2. Inspected version: 2505.11475v2.

Sources: [728]

Evidence: Reviewed-work bibliography - How do data owners say no? A case study of data consent mechanisms in web-scraped vision-language AI training datasets. Source: https://arxiv.org/pdf/2511.08637v1. Inspected version: 2511.08637v1.

Sources: [384]

Evidence: Reviewed-work bibliography - How Reliable is Multilingual LLM-as-a-Judge?. Source: https://aclanthology.org/2025.findings-emnlp.587.pdf. Inspected version: ACL-2025.findings-emnlp.587-published; anthology metadata commit ba3130005282c6474307cceec936373fe55d15ff.

Sources: [396]

Evidence: Reviewed-work bibliography - Humanity's Last Exam. Source: https://arxiv.org/pdf/2501.14249v1. Inspected version: 2501.14249v1.

Sources: [354]

Evidence: Reviewed-work bibliography - HybridFlow: A Flexible and Efficient RLHF Framework. Source: https://arxiv.org/pdf/2409.19256v2. Inspected version: 2409.19256v2.

Sources: [408]

Evidence: Reviewed-work bibliography - If in a Crowdsourced Data Annotation Pipeline, a GPT-4. Source: https://arxiv.org/pdf/2402.16795v2. Inspected version: 2402.16795v2.

Sources: [406]

Evidence: Reviewed-work bibliography - Improving Task Instructions for Data Annotators: How Clear Rules and Higher Pay Increase Performance in Data Annotation in the AI Economy. Source: https://arxiv.org/pdf/2312.14565v2. Inspected version: 2312.14565v2.

Sources: [389]

Evidence: Reviewed-work bibliography - Instruction-Following Evaluation for Large Language Models. Source: https://arxiv.org/pdf/2311.07911v1. Inspected version: 2311.07911v1.

Sources: [355]

Evidence: Reviewed-work bibliography - Is It Thinking or Cheating? Detecting Implicit Reward Hacking by Measuring Reasoning Effort. Source: https://arxiv.org/pdf/2510.01367v1. Inspected version: 2510.01367v1.

Sources: [398]

Evidence: Reviewed-work bibliography - JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. Source: https://arxiv.org/pdf/2404.01318v5. Inspected version: 2404.01318v5.

Sources: [225]

Evidence: Reviewed-work bibliography - KTO: Model Alignment as Prospect Theoretic Optimization. Source: https://arxiv.org/pdf/2402.01306v1. Inspected version: 2402.01306v1.

Sources: [69]

Evidence: Reviewed-work bibliography - Learning from Natural Language Feedback. Source: https://arxiv.org/pdf/2204.14146v1. Inspected version: 2204.14146v1.

Sources: [306]

Evidence: Reviewed-work bibliography - Learning to summarize from human feedback. Source: https://arxiv.org/pdf/2009.01325v3. Inspected version: 2009.01325v3.

Sources: [158]

Evidence: Reviewed-work bibliography - Let's Verify Step by Step. Source: https://arxiv.org/pdf/2305.20050v1. Inspected version: 2305.20050v1.

Sources: [78]

Evidence: Reviewed-work bibliography - LIMA: Less Is More for Alignment. Source: https://arxiv.org/pdf/2305.11206v1. Inspected version: 2305.11206v1.

Sources: [149]

Evidence: Reviewed-work bibliography - Llama 2: Open Foundation and Fine-Tuned Chat Models. Source: https://arxiv.org/pdf/2307.09288v1. Inspected version: 2307.09288v1.

Sources: [309]

Evidence: Reviewed-work bibliography - M-IFEval: Multilingual Instruction-Following Evaluation. Source: https://aclanthology.org/2025.findings-naacl.344.pdf. Inspected version: ACL-2025.findings-naacl.344-published; anthology metadata commit ba3130005282c6474307cceec936373fe55d15ff.

Sources: [391]

Evidence: Reviewed-work bibliography - M-RewardBench: Evaluating Reward Models in Multilingual Settings. Source: https://aclanthology.org/2025.acl-long.3.pdf. Inspected version: ACL-2025.acl-long.3-published; anthology metadata commit ba3130005282c6474307cceec936373fe55d15ff.

Sources: [394]

Evidence: Reviewed-work bibliography - MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. Source: https://arxiv.org/pdf/2310.02255v1. Inspected version: 2310.02255v1.

Sources: [226]

Evidence: Reviewed-work bibliography - Measuring Massive Multitask Language Understanding. Source: https://arxiv.org/pdf/2009.03300v3. Inspected version: 2009.03300v3.

Sources: [714]

Evidence: Reviewed-work bibliography - MENLO: From Preferences to Proficiency -- Evaluating and Modeling Native-like Quality Across 47 Languages. Source: https://arxiv.org/pdf/2509.26601v1. Inspected version: 2509.26601v1.

Sources: [390]

Evidence: Reviewed-work bibliography - Minority Reports: Balancing Cost and Quality in Ground Truth Data Annotation. Source: https://arxiv.org/pdf/2504.09341v1. Inspected version: 2504.09341v1.

Sources: [410]

Evidence: Reviewed-work bibliography - MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. Source: https://arxiv.org/pdf/2410.07095v1. Inspected version: 2410.07095v1.

Sources: [363]

Evidence: Reviewed-work bibliography - MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. Source: https://arxiv.org/pdf/2409.02813v1. Inspected version: 2409.02813v1.

Sources: [364]

Evidence: Reviewed-work bibliography - MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. Source: https://arxiv.org/pdf/2311.16502v1. Inspected version: 2311.16502v1.

Sources: [232]

Evidence: Reviewed-work bibliography - Model Cards for Model Reporting. Source: https://arxiv.org/pdf/1810.03993v2. Inspected version: 1810.03993v2.

Sources: [392]

Evidence: Reviewed-work bibliography - Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. Source: https://arxiv.org/pdf/2503.11926v1. Inspected version: 2503.11926v1.

Sources: [393]

Evidence: Reviewed-work bibliography - MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs. Source: https://arxiv.org/pdf/2501.17399v1. Inspected version: 2501.17399v1.

Sources: [395]

Evidence: Reviewed-work bibliography - On the Reliability of Majority-Vote Pseudo-Labels in Test-Time Reinforcement Learning. Source: https://arxiv.org/pdf/2611.09872v1. Inspected version: 2611.09872v1.

Sources: [416]

Evidence: Reviewed-work bibliography - OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework. Source: https://arxiv.org/pdf/2405.11143v6. Inspected version: 2405.11143v6.

Sources: [412]

Evidence: Reviewed-work bibliography - OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. Source: https://arxiv.org/pdf/2404.07972v1. Inspected version: 2404.07972v1.

Sources: [233]

Evidence: Reviewed-work bibliography - PaperBench: Evaluating AI's Ability to Replicate AI Research. Source: https://arxiv.org/pdf/2504.01848v1. Inspected version: 2504.01848v1.

Sources: [365]

Evidence: Reviewed-work bibliography - Process Reinforcement through Implicit Rewards. Source: https://arxiv.org/pdf/2502.01456v2. Inspected version: 2502.01456v2.

Sources: [413]

Evidence: Reviewed-work bibliography - Program Synthesis with Large Language Models. Source: https://arxiv.org/pdf/2108.07732v1. Inspected version: 2108.07732v1.

Sources: [227]

Evidence: Reviewed-work bibliography - Proximal Policy Optimization Algorithms. Source: https://arxiv.org/pdf/1707.06347v1. Inspected version: 1707.06347v1.

Sources: [77]

Evidence: Reviewed-work bibliography - RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. Source: https://arxiv.org/pdf/2304.06767v1. Inspected version: 2304.06767v1.

Sources: [315]

Evidence: Reviewed-work bibliography - RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. Source: https://arxiv.org/pdf/2411.15114v1. Inspected version: 2411.15114v1.

Sources: [366]

Evidence: Reviewed-work bibliography - REBEL: Reinforcement Learning via Regressing Relative Rewards. Source: https://arxiv.org/pdf/2404.16767v1. Inspected version: 2404.16767v1.

Sources: [316]

Evidence: Reviewed-work bibliography - Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. Source: https://arxiv.org/pdf/2209.07858v1. Inspected version: 2209.07858v1.

Sources: [154]

Evidence: Reviewed-work bibliography - Reference-free Monolithic Preference Optimization with Odds Ratio. Source: https://arxiv.org/pdf/2403.07691v1. Inspected version: 2403.07691v1.

Sources: [76]

Evidence: Reviewed-work bibliography - REINFORCE++: Stabilizing Critic-Free Policy Optimization with Global Advantage Normalization. Source: https://arxiv.org/pdf/2501.03262v9. Inspected version: 2501.03262v9.

Sources: [415]

Evidence: Reviewed-work bibliography - Reinforcement Learning on Web Interfaces Using Workflow-Guided Exploration. Source: https://arxiv.org/pdf/1802.08802v1. Inspected version: 1802.08802v1.

Sources: [231]

Evidence: Reviewed-work bibliography - ReMax: A Simple, Effective, and Efficient Method for Aligning Large Language Models. Source: https://arxiv.org/pdf/2310.10505v1. Inspected version: 2310.10505v1.

Sources: [317]

Evidence: Reviewed-work bibliography - RewardBench 2: Advancing Reward Model Evaluation. Source: https://arxiv.org/pdf/2506.01937v1. Inspected version: 2506.01937v1.

Sources: [367]

Evidence: Reviewed-work bibliography - RLAIF-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness. Source: https://arxiv.org/pdf/2405.17220v1. Inspected version: 2405.17220v1.

Sources: [155]

Evidence: Reviewed-work bibliography - RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. Source: https://arxiv.org/pdf/2309.00267v1. Inspected version: 2309.00267v1.

Sources: [79]

Evidence: Reviewed-work bibliography - RRHF: Rank Responses to Align Language Models with Human Feedback without tears. Source: https://arxiv.org/pdf/2304.05302v1. Inspected version: 2304.05302v1.

Sources: [319]

Evidence: Reviewed-work bibliography - Scaling Instruction-Finetuned Language Models. Source: https://arxiv.org/pdf/2210.11416v5. Inspected version: 2210.11416v5.

Sources: [137]

Evidence: Reviewed-work bibliography - ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. Source: https://arxiv.org/pdf/2410.05080v1. Inspected version: 2410.05080v1.

Sources: [368]

Evidence: Reviewed-work bibliography - ScienceWorld: Is your Agent Smarter than a 5th Grader?. Source: https://arxiv.org/pdf/2203.07540v1. Inspected version: 2203.07540v1.

Sources: [253]

Evidence: Reviewed-work bibliography - SelectTTRL: Representation-Guided Data Selection for Efficient Test-Time Reinforcement Learning. Source: https://arxiv.org/pdf/2610.16104v1. Inspected version: 2610.16104v1.

Sources: [417]

Evidence: Reviewed-work bibliography - SELF-INSTRUCT: Aligning Language Model with Self Generated Instructions. Source: https://arxiv.org/pdf/2212.10560v1. Inspected version: 2212.10560v1.

Sources: [83]

Evidence: Reviewed-work bibliography - Self-Play Preference Optimization for Language Model Alignment. Source: https://arxiv.org/pdf/2405.00675v1. Inspected version: 2405.00675v1.

Sources: [327]

Evidence: Reviewed-work bibliography - Self-Refine: Iterative Refinement with Self-Feedback. Source: https://arxiv.org/pdf/2303.17651v2. Inspected version: 2303.17651v2.

Sources: [156]

Evidence: Reviewed-work bibliography - ShieldAgent: Shielding Agents via Verifiable Safety Policy Reasoning. Source: https://arxiv.org/pdf/2503.22738v1. Inspected version: 2503.22738v1.

Sources: [397]

Evidence: Reviewed-work bibliography - SimPO: Simple Preference Optimization with a Reference-Free Reward. Source: https://arxiv.org/pdf/2405.14734v1. Inspected version: 2405.14734v1.

Sources: [325]

Evidence: Reviewed-work bibliography - SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks. Source: https://arxiv.org/pdf/2602.12670v1. Inspected version: 2602.12670v1.

Sources: [369]

Evidence: Reviewed-work bibliography - SLiC-HF: Sequence Likelihood Calibration with Human Feedback. Source: https://arxiv.org/pdf/2305.10425v1. Inspected version: 2305.10425v1.

Sources: [326]

Evidence: Reviewed-work bibliography - Sources of Disagreement in Data for LLM Instruction Tuning. Source: https://aclanthology.org/2025.comedi-1.3.pdf. Inspected version: acl.2025.comedi-1.3.

Sources: [133]

Evidence: Reviewed-work bibliography - STAMP Your Content: Proving Dataset Membership via Watermarked Rephrasings. Source: https://arxiv.org/pdf/2504.13416v2. Inspected version: 2504.13416v2.

Sources: [418]

Evidence: Reviewed-work bibliography - Statistical Rejection Sampling Improves Preference Optimization. Source: https://arxiv.org/pdf/2309.06657v1. Inspected version: 2309.06657v1.

Sources: [320]

Evidence: Reviewed-work bibliography - Step-DPO: Step-wise Preference Optimization for Long-chain Reasoning of LLMs. Source: https://arxiv.org/pdf/2406.18629v1. Inspected version: 2406.18629v1.

Sources: [328]

Evidence: Reviewed-work bibliography - SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains?. Source: https://arxiv.org/pdf/2410.03859v1. Inspected version: 2410.03859v1.

Sources: [374]

Evidence: Reviewed-work bibliography - SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks?. Source: https://arxiv.org/pdf/2509.16941v1. Inspected version: 2509.16941v1.

Sources: [375]

Evidence: Reviewed-work bibliography - SWE-bench: Can Language Models Resolve Real-World GitHub Issues?. Source: https://arxiv.org/pdf/2310.06770v3. Inspected version: 2310.06770v3.

Sources: [733]

Evidence: Reviewed-work bibliography - SWE-Lancer: Can Frontier LLMs Earn $1 Million from Real-World Freelance Software Engineering?. Source: https://arxiv.org/pdf/2502.12115v1. Inspected version: 2502.12115v1.

Sources: [373]

Evidence: Reviewed-work bibliography - Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. Source: https://arxiv.org/pdf/2601.11868v1. Inspected version: 2601.11868v1.

Sources: [377]

Evidence: Reviewed-work bibliography - The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. Source: https://arxiv.org/pdf/2506.05209v1. Inspected version: 2506.05209v1.

Sources: [382]

Evidence: Reviewed-work bibliography - The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Source: https://arxiv.org/pdf/2310.16787v1. Inspected version: 2310.16787v1.

Sources: [153]

Evidence: Reviewed-work bibliography - The Data-Production Dispositif. Source: https://arxiv.org/pdf/2205.11963v1. Inspected version: 2205.11963v1.

Sources: [385]

Evidence: Reviewed-work bibliography - The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. Source: https://arxiv.org/pdf/1805.03677v1. Inspected version: 1805.03677v1.

Sources: [411]

Evidence: Reviewed-work bibliography - The digital labour of artificial intelligence in Latin America: a comparison of Argentina, Brazil, and Venezuela. Source: https://arxiv.org/pdf/2502.06317v1. Inspected version: 2502.06317v1.

Sources: [409]

Evidence: Reviewed-work bibliography - The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. Source: https://arxiv.org/pdf/2406.17557v2. Inspected version: 2406.17557v2.

Sources: [734]

Evidence: Reviewed-work bibliography - The Llama 3 Herd of Models. Source: https://arxiv.org/pdf/2407.21783v1. Inspected version: 2407.21783v1.

Sources: [310]

Evidence: Reviewed-work bibliography - The Problems with Proxies: Making Data Work Visible through Requester Practices. Source: https://arxiv.org/pdf/2408.11667v1. Inspected version: 2408.11667v1.

Sources: [414]

Evidence: Reviewed-work bibliography - TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. Source: https://arxiv.org/pdf/2412.14161v1. Inspected version: 2412.14161v1.

Sources: [334]

Evidence: Reviewed-work bibliography - ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. Source: https://arxiv.org/pdf/2307.16789v1. Inspected version: 2307.16789v1.

Sources: [255]

Evidence: Reviewed-work bibliography - Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. Source: https://arxiv.org/pdf/2204.05862v1. Inspected version: 2204.05862v1.

Sources: [138]

Evidence: Reviewed-work bibliography - Training language models to follow instructions with human feedback. Source: https://arxiv.org/abs/2203.02155v1. Inspected version: 2203.02155v1.

Sources: [724], [731]

Evidence: Reviewed-work bibliography - Training Verifiers to Solve Math Word Problems. Source: https://arxiv.org/pdf/2110.14168v1. Inspected version: 2110.14168v1.

Sources: [90]

Evidence: Reviewed-work bibliography - TruthfulQA: Measuring How Models Mimic Human Falsehoods. Source: https://arxiv.org/pdf/2109.07958v2. Inspected version: 2109.07958v2.

Sources: [675]

Evidence: Reviewed-work bibliography - TTRL: Test-Time Reinforcement Learning. Source: https://arxiv.org/pdf/2504.16084v3. Inspected version: 2504.16084v3.

Sources: [419]

Evidence: Reviewed-work bibliography - Understanding R1-Zero-Like Training: A Critical Perspective. Source: https://arxiv.org/pdf/2503.20783v2. Inspected version: 2503.20783v2.

Sources: [405]

Evidence: Reviewed-work bibliography - VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks. Source: https://arxiv.org/pdf/2504.05118v3. Inspected version: 2504.05118v3.

Sources: [420]

Evidence: Reviewed-work bibliography - VerifiAgent: a Unified Verification Agent in Language Model Reasoning. Source: https://arxiv.org/pdf/2504.00406v1. Inspected version: 2504.00406v1.

Sources: [399]

Evidence: Reviewed-work bibliography - Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. Source: https://arxiv.org/pdf/2405.21075v1. Inspected version: 2405.21075v1.

Sources: [380]

Evidence: Reviewed-work bibliography - VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. Source: https://arxiv.org/pdf/2401.13649v1. Inspected version: 2401.13649v1.

Sources: [256]

Evidence: Reviewed-work bibliography - WebArena: A Realistic Web Environment for Building Autonomous Agents. Source: https://arxiv.org/pdf/2307.13854v1. Inspected version: 2307.13854v1.

Sources: [91]

Evidence: Reviewed-work bibliography - WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. Source: https://arxiv.org/pdf/2207.01206v1. Inspected version: 2207.01206v1.

Sources: [257]

Evidence: Reviewed-work bibliography - WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. Source: https://arxiv.org/pdf/2304.12244v3. Inspected version: 2304.12244v3.

Sources: [159]

Evidence: Reviewed-work bibliography - World of Bits: An Open-Domain Platform for Web-Based Agents. Source: https://proceedings.mlr.press/v70/shi17a/shi17a.pdf. Inspected version: PMLR-v70-shi17a-2017.

Sources: [400]

Evidence: Reviewed-work bibliography - WPO: Enhancing RLHF with Weighted Preference Optimization. Source: https://arxiv.org/pdf/2406.11827v1. Inspected version: 2406.11827v1.

Sources: [333]

Evidence: Reviewed-work bibliography - τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. Source: https://arxiv.org/pdf/2406.12045v1. Inspected version: 2406.12045v1.

Sources: [254]

Evidence: Reviewed-work bibliography - “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. Source: https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/0d556e45afc54afeb2eb6b51a9bc1827b9961ff4.pdf. Inspected version: doi.10.1145.3411764.3445518.

Sources: [127]

Evidence: Reviewed-work bibliography - AutoLibra: Agent Metric Induction from Open-Ended Human Feedback. Source: https://arxiv.org/pdf/2505.02820v3. Inspected version: 2505.02820v3.

Sources: [421]

Evidence: Reviewed-work bibliography - Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. Source: https://arxiv.org/pdf/2402.06619v1. Inspected version: 2402.06619v1.

Sources: [422]

Evidence: Reviewed-work bibliography - CrowdAgent: Multi-Agent Managed Multi-Source Annotation System. Source: https://aclanthology.org/2025.emnlp-demos.72.pdf. Inspected version: ACL-Anthology-2025.emnlp-demos.72-captured-copy.

Sources: [423]

Evidence: Reviewed-work bibliography - CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. Source: https://arxiv.org/pdf/1808.06080v1. Inspected version: 1808.06080v1.

Sources: [424]

Evidence: Reviewed-work bibliography - A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. Source: https://arxiv.org/pdf/1712.05796v2. Inspected version: 1712.05796v2.

Sources: [425]

Evidence: Reviewed-work bibliography - Experts Rise Where LLMs Disagree: Using Cross-Model Disagreement to Target Expert Effort in LLM Codebook Revision for Large-Scale Annotation. Source: https://arxiv.org/pdf/2609.26926v1. Inspected version: 2609.26926v1.

Sources: [426]

Evidence: Reviewed-work bibliography - Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?. Source: https://arxiv.org/pdf/1912.08320v1. Inspected version: 1912.08320v1.

Sources: [427]

Evidence: Reviewed-work bibliography - ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. Source: https://arxiv.org/pdf/2303.15056v2. Inspected version: 2303.15056v2.

Sources: [428]

Evidence: Reviewed-work bibliography - HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. Source: https://arxiv.org/pdf/2311.09528v1. Inspected version: 2311.09528v1.

Sources: [429]

Evidence: Reviewed-work bibliography - HelpSteer2: Open-source dataset for training top-performing reward models. Source: https://arxiv.org/pdf/2406.08673v1. Inspected version: 2406.08673v1.

Sources: [430]

Evidence: Reviewed-work bibliography - Inherent Disagreements in Human Textual Inferences. Source: https://aclanthology.org/Q19-1043.pdf. Inspected version: ACL-Anthology-Q19-1043-captured-copy.

Sources: [431]

Evidence: Reviewed-work bibliography - LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems. Source: https://arxiv.org/pdf/2403.09810v1. Inspected version: 2403.09810v1.

Sources: [432]

Evidence: Reviewed-work bibliography - Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. Source: https://arxiv.org/pdf/2103.14749v4. Inspected version: 2103.14749v4.

Sources: [433]

Evidence: Reviewed-work bibliography - Efficient Data Labeling by Hierarchical Crowdsourcing with Large Language Models. Source: https://aclanthology.org/2025.coling-main.748.pdf. Inspected version: ACL-Anthology-2025.coling-main.748-captured-copy.

Sources: [434]

Evidence: Reviewed-work bibliography - Quality Assured: Rethinking Annotation Strategies in Imaging AI. Source: https://arxiv.org/pdf/2407.17596v2. Inspected version: 2407.17596v2.

Sources: [435]

Evidence: Reviewed-work bibliography - RECODE-H: A Benchmark for Research Code Development with Interactive Human Feedback. Source: https://arxiv.org/pdf/2510.06186v2. Inspected version: 2510.06186v2.

Sources: [436]

Evidence: Reviewed-work bibliography - Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. Source: https://aclanthology.org/D08-1027.pdf. Inspected version: ACL-Anthology-D08-1027-captured-copy.

Sources: [437]

Evidence: Reviewed-work bibliography - UltraFeedback: Boosting Language Models with Scaled AI Feedback. Source: https://arxiv.org/pdf/2310.01377v2. Inspected version: 2310.01377v2.

Sources: [438]

Evidence: Reviewed-work bibliography - Who Annotates in NLP? A Large-scale Assessment of Human Annotation Reporting between 2018 and 2025. Source: https://arxiv.org/pdf/2606.02255v3. Inspected version: 2606.02255v3.

Sources: [439]

Evidence: Reviewed-work bibliography - WildChat: 1M ChatGPT Interaction Logs in the Wild. Source: https://arxiv.org/pdf/2405.01470v1. Inspected version: 2405.01470v1.

Sources: [440]

Evidence: Reviewed-work bibliography - Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions. Source: https://ojs.aaai.org/index.php/AAAI/article/download/26886/26658. Inspected version: AAAI-37-13-26886-26658-captured-2026-11-10. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [457]

Evidence: Reviewed-work bibliography - Quantifying and Avoiding Unfair Qualification Labour in Crowdsourcing. Source: https://aclanthology.org/2021.acl-short.44.pdf. Inspected version: ACL-Anthology-2021.acl-short.44-captured-2026-11-10. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [458]

Evidence: Reviewed-work bibliography - Red Teaming Language Models with Language Models. Source: https://aclanthology.org/2022.emnlp-main.225.pdf. Inspected version: ACL-Anthology-2022.emnlp-main.225-captured-2026-11-10. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [459]

Evidence: Reviewed-work bibliography - Findings of the WMT’22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages. Source: https://aclanthology.org/2022.wmt-1.72.pdf. Inspected version: ACL-Anthology-2022.wmt-1.72-captured-2026-11-10. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [460]

Evidence: Reviewed-work bibliography - LLMs Simulate Big5 Personality Traits: Further Evidence. Source: https://aclanthology.org/2024.personalize-1.7.pdf. Inspected version: ACL-Anthology-2024.personalize-1.7-captured-2026-11-10. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [461]

Evidence: Reviewed-work bibliography - Learning Whom to Trust with MACE. Source: https://aclanthology.org/N13-1132.pdf. Inspected version: ACL-Anthology-N13-1132-captured-2026-11-10. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [462]

Evidence: Reviewed-work bibliography - Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. Source: https://arxiv.org/pdf/2106.10328v2. Inspected version: 2106.10328v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [463]

Evidence: Reviewed-work bibliography - CrowdSpeech and VoxDIY: Benchmark Datasets for Crowdsourced Audio Transcription. Source: https://arxiv.org/pdf/2107.01091v2. Inspected version: 2107.01091v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [464]

Evidence: Reviewed-work bibliography - Deduplicating Training Data Makes Language Models Better. Source: https://arxiv.org/pdf/2107.06499v2. Inspected version: 2107.06499v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [465]

Evidence: Reviewed-work bibliography - Learning from Crowds with Crowd-Kit. Source: https://arxiv.org/pdf/2109.08584v4. Inspected version: 2109.08584v4. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [466]

Evidence: Reviewed-work bibliography - IMDB-WIKI-SbS: An Evaluation Dataset for Crowdsourced Pairwise Comparisons. Source: https://arxiv.org/pdf/2110.14990v2. Inspected version: 2110.14990v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [467]

Evidence: Reviewed-work bibliography - WebGPT: Browser-assisted question-answering with human feedback. Source: https://arxiv.org/pdf/2112.09332v3. Inspected version: 2112.09332v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [468]

Evidence: Reviewed-work bibliography - Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision. Source: https://arxiv.org/pdf/2203.13270v2. Inspected version: 2203.13270v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [469]

Evidence: Reviewed-work bibliography - Clustering Without Knowing How To: Application and Evaluation. Source: https://arxiv.org/pdf/2209.10267v3. Inspected version: 2209.10267v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [470]

Evidence: Reviewed-work bibliography - Best Prompts for Text-to-Image Models and How to Find Them. Source: https://arxiv.org/pdf/2209.11711v3. Inspected version: 2209.11711v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [471]

Evidence: Reviewed-work bibliography - Learning to Reason With Relational Abstractions. Source: https://arxiv.org/pdf/2210.02615v2. Inspected version: 2210.02615v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [472]

Evidence: Reviewed-work bibliography - Measuring Progress on Scalable Oversight for Large Language Models. Source: https://arxiv.org/pdf/2211.03540v2. Inspected version: 2211.03540v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [473]

Evidence: Reviewed-work bibliography - Discovering Language Model Behaviors with Model-Written Evaluations. Source: https://arxiv.org/pdf/2212.09251v1. Inspected version: 2212.09251v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [474]

Evidence: Reviewed-work bibliography - DOC: Improving Long Story Coherence With Detailed Outline Control. Source: https://arxiv.org/pdf/2212.10077v3. Inspected version: 2212.10077v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [475]

Evidence: Reviewed-work bibliography - StarCoder: may the source be with you!. Source: https://arxiv.org/pdf/2305.06161v2. Inspected version: 2305.06161v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [476]

Evidence: Reviewed-work bibliography - Enabling Large Language Models to Generate Text with Citations. Source: https://arxiv.org/pdf/2305.14627v2. Inspected version: 2305.14627v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [549]

Evidence: Reviewed-work bibliography - Inverse Scaling: When Bigger Isn't Better. Source: https://arxiv.org/pdf/2306.09479v2. Inspected version: 2306.09479v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [550]

Evidence: Reviewed-work bibliography - Universal and Transferable Adversarial Attacks on Aligned Language Models. Source: https://arxiv.org/pdf/2307.15043v2. Inspected version: 2307.15043v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [551]

Evidence: Reviewed-work bibliography - Beemo: Benchmark of Expert-edited Machine-generated Outputs. Source: https://arxiv.org/pdf/2411.04032v3. Inspected version: 2411.04032v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [693]

Evidence: Reviewed-work bibliography - U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs. Source: https://arxiv.org/pdf/2412.03205v4. Inspected version: 2412.03205v4. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [552]

Evidence: Reviewed-work bibliography - Weak-to-Strong Generalization Through the Data-Centric Lens. Source: https://arxiv.org/pdf/2412.03881v2. Inspected version: 2412.03881v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [553]

Evidence: Reviewed-work bibliography - JEEM: Vision-Language Understanding in Four Arabic Dialects. Source: https://arxiv.org/pdf/2503.21910v1. Inspected version: 2503.21910v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [554]

Evidence: Reviewed-work bibliography - Voices of Freelance Professional Writers on AI: Limitations, Expectations, and Fears. Source: https://arxiv.org/pdf/2504.05008v3. Inspected version: 2504.05008v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [555]

Evidence: Reviewed-work bibliography - Shrinking the Generation-Verification Gap with Weak Verifiers. Source: https://arxiv.org/pdf/2506.18203v3. Inspected version: 2506.18203v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [556]

Evidence: Reviewed-work bibliography - AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following. Source: https://arxiv.org/pdf/2511.10507v2. Inspected version: 2511.10507v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [557]

Evidence: Reviewed-work bibliography - Multimodal RewardBench 2: Evaluating Omni Reward Models for Interleaved Text and Image. Source: https://arxiv.org/pdf/2512.16899v3. Inspected version: 2512.16899v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [558]

Evidence: Reviewed-work bibliography - APEX-SWE. Source: https://arxiv.org/pdf/2601.08806v3. Inspected version: 2601.08806v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [559]

Evidence: Reviewed-work bibliography - The Hierarchy of Agentic Capabilities: Evaluating Frontier Models on Realistic RL Environments. Source: https://arxiv.org/pdf/2601.09032v1. Inspected version: 2601.09032v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [560]

Evidence: Reviewed-work bibliography - APEX-Agents. Source: https://arxiv.org/pdf/2601.14242v3. Inspected version: 2601.14242v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [561]

Evidence: Reviewed-work bibliography - Benchmarking Agents in Insurance Underwriting Environments. Source: https://arxiv.org/pdf/2602.00456v1. Inspected version: 2602.00456v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [562]

Evidence: Reviewed-work bibliography - EnterpriseBench Corecraft: Training Generalizable Agents on High-Fidelity RL Environments. Source: https://arxiv.org/pdf/2602.16179v5. Inspected version: 2602.16179v5. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [6]

Evidence: Reviewed-work bibliography - Intent Laundering: AI Safety Datasets Are Not What They Seem. Source: https://arxiv.org/pdf/2602.16729v3. Inspected version: 2602.16729v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [703]

Evidence: Reviewed-work bibliography - τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge. Source: https://arxiv.org/pdf/2603.04370v1. Inspected version: 2603.04370v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [563]

Evidence: Reviewed-work bibliography - PivotRL: High Accuracy Agentic Post-Training at Low Compute Cost. Source: https://arxiv.org/pdf/2603.21383v1. Inspected version: 2603.21383v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [564]

Evidence: Reviewed-work bibliography - The Hitchhikers Guide to Rubric Quality Understanding and Enrichment. Source: https://arxiv.org/pdf/2604.01375v4. Inspected version: 2604.01375v4. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [565]

Evidence: Reviewed-work bibliography - ClawsBench: Evaluating Capability and Safety of LLM Productivity Agents in Simulated Workspaces. Source: https://arxiv.org/pdf/2604.05172v2. Inspected version: 2604.05172v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [701]

Evidence: Reviewed-work bibliography - Riemann-Bench: A Benchmark for Moonshot Mathematics. Source: https://arxiv.org/pdf/2604.06802v3. Inspected version: 2604.06802v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [21]

Evidence: Reviewed-work bibliography - Learning from Less: Measuring the Effectiveness of RLVR in Low Data and Compute Regimes. Source: https://arxiv.org/pdf/2604.18381v1. Inspected version: 2604.18381v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [696]

Evidence: Reviewed-work bibliography - StereoTales: A Multilingual Framework for Open-Ended Stereotype Discovery in LLMs. Source: https://arxiv.org/pdf/2605.10442v2. Inspected version: 2605.10442v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [704]

Evidence: Reviewed-work bibliography - AttuneBench: A Conversation-Based Benchmark for LLM Emotional Intelligence. Source: https://arxiv.org/pdf/2605.21739v2. Inspected version: 2605.21739v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [699]

Evidence: Reviewed-work bibliography - Agents’ Last Exam. Source: https://arxiv.org/pdf/2606.05405v4. Inspected version: 2606.05405v4. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [683]

Evidence: Reviewed-work bibliography - ComplexConstraints and Beyond: Expert Rubrics for RLVR. Source: https://arxiv.org/pdf/2606.09118v3. Inspected version: 2606.09118v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [5]

Evidence: Reviewed-work bibliography - GDP.pdf: Benchmarking Grounded Multimodal Reasoning over Professional PDF Documents. Source: https://arxiv.org/pdf/2607.11192v3. Inspected version: 2607.11192v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [8]

Evidence: Reviewed-work bibliography - HANDBOOK.md: A Benchmark for Long-Context Agentic Instruction Following. Source: https://arxiv.org/pdf/2607.25398v3. Inspected version: 2607.25398v3. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [10]

Evidence: Reviewed-work bibliography - APEX-Accounting. Source: https://arxiv.org/pdf/2607.27189v2. Inspected version: 2607.27189v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [697]

Evidence: Reviewed-work bibliography - Cross-Benchmark Generalization in Long-Horizon Agents. Source: https://arxiv.org/pdf/2608.00181v1. Inspected version: 2608.00181v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [702]

Evidence: Reviewed-work bibliography - Post-Training on Office Work Improves Software Engineering: A Behavioral Account of Cross-Domain Transfer. Source: https://arxiv.org/pdf/2608.01604v1. Inspected version: 2608.01604v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [685], [702]

Evidence: Reviewed-work bibliography - Chartography: A Benchmark for Professional Chart Understanding. Source: https://arxiv.org/pdf/2608.10677v1. Inspected version: 2608.10677v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [3]

Evidence: Reviewed-work bibliography - Cross-Benchmark Transfer from RL on Agentic Coding Tasks. Source: https://arxiv.org/pdf/2610.00890v1. Inspected version: 2610.00890v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [698]

Evidence: Reviewed-work bibliography - DAYJOB: A Benchmark for Long-Horizon Professional Work. Source: https://arxiv.org/pdf/2610.01306v1. Inspected version: 2610.01306v1. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [695]

Evidence: Reviewed-work bibliography - PLE: Physical-Engineering’s Last Exam. Source: https://arxiv.org/pdf/2610.30537v2. Inspected version: 2610.30537v2. The approved observations cover the bounded methods/results or worked-example/limitations scope recorded by their receipts; no experiment reproduction or complete corpus audit is implied.

Sources: [700]

Method: Author synthesis scope for this refresh. The author inspected the complete report section/paragraph inventory, every one of the 141 routed affected paragraphs, the nine executive findings and original ten-pilot structure, the 84 successor records and their qualifications, all added work/family/supplier observations and current identity dispositions, source-version metadata and the local report/graph interfaces. Unchanged retained prose inherits the earlier authoring pass; this refresh does not claim a new cover-to-cover reading of every retained source or every old narrative paragraph. New comparisons and every changed block were checked as prose against their approved inputs. No new primary research, experiment, operational audit, rights clearance, human review or independent synthesis pass is created. Separate reviewers must inspect the complete exact output.

Evidence: Reviewed-work bibliography - CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. Source: https://arxiv.org/pdf/2406.18521v1. Inspected version: arxiv:2406.18521v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [634], [639]

Evidence: Reviewed-work bibliography - Comprehensive Verilog Design Problems: A Next-Generation Benchmark Dataset for Evaluating Large Language Models and Agents on RTL Design and Verification. Source: https://arxiv.org/pdf/2506.14074v1. Inspected version: arxiv:2506.14074v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [643], [644]

Evidence: Reviewed-work bibliography - DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report. Source: https://arxiv.org/pdf/2601.08536v1. Inspected version: arxiv:2601.08536v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [610], [611], [612]

Evidence: Reviewed-work bibliography - DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. Source: https://arxiv.org/pdf/2607.07946v1. Inspected version: arxiv:2607.07946v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [607], [608], [609]

Evidence: Reviewed-work bibliography - GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine. Source: https://www.biorxiv.org/content/10.64898/2026.06.29.735386v2.full.pdf. Inspected version: doi:10.64898/2026.06.29.735386v2. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [673]

Evidence: Reviewed-work bibliography - HealthBench: Evaluating Large Language Models Towards Improved Human Health. Source: https://arxiv.org/pdf/2505.08775v1. Inspected version: arxiv:2505.08775v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [646], [647]

Evidence: Reviewed-work bibliography - Measuring AI Ability to Complete Long Software Tasks. Source: https://arxiv.org/pdf/2503.14499v4. Inspected version: arxiv:2503.14499v4. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [617]

Evidence: Reviewed-work bibliography - MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. Source: https://arxiv.org/pdf/2501.18362v1. Inspected version: arxiv:2501.18362v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [649], [650], [651]

Evidence: Reviewed-work bibliography - MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. Source: https://arxiv.org/pdf/2406.01574v1. Inspected version: arxiv:2406.01574v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [653]

Evidence: Reviewed-work bibliography - ProgramBench: Can Language Models Rebuild Programs From Scratch?. Source: https://arxiv.org/pdf/2605.03546v1. Inspected version: arxiv:2605.03546v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [655]

Evidence: Reviewed-work bibliography - SciCode: A Research Coding Benchmark Curated by Scientists. Source: https://arxiv.org/pdf/2407.13168v1. Inspected version: arxiv:2407.13168v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [637], [638], [659]

Evidence: Reviewed-work bibliography - SUPERChem: A Multimodal Reasoning Benchmark in Chemistry. Source: https://arxiv.org/pdf/2512.01274v1. Inspected version: arxiv:2512.01274v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [662]

Evidence: Reviewed-work bibliography - SWE Atlas: Benchmarking Coding Agents Beyond Issue Resolution. Source: https://arxiv.org/pdf/2605.08366v1. Inspected version: arxiv:2605.08366v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [670]

Evidence: Reviewed-work bibliography - The AI Productivity Index (APEX). Source: https://arxiv.org/pdf/2509.25721v1. Inspected version: arxiv:2509.25721v1. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [605]

Evidence: Reviewed-work bibliography - Towards a Science of Corpus Annotation. Source: https://www.cs.cmu.edu/~hovy/papers/10KNS-annotation-Hovy-Lavid.pdf. Inspected version: captured document response sha256:0ce0c6a23e1ce3a167ecbed07c9d64fb96464dfa06ea56a4d04790f1fdbe531d. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [613]

Evidence: Reviewed-work bibliography - VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions. Source: https://ceur-ws.org/Vol-2932/invited1.pdf. Inspected version: captured document response sha256:fb49f2021942a49d40c699446313ba08397ba84dc721c311b498c032d905bdbc. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [619]

Evidence: Reviewed-work bibliography - WSDM Cup 2023 Challenge on Visual Question Answering. Source: https://ceur-ws.org/Vol-3357/invited1.pdf. Inspected version: captured document response sha256:f72ce0cdc18b596cf8ec5f06f554ca0342e4a066004f6c73123a996afc69d555. The bound claims retain all other inspected paper/card/repository version contexts and their actual capture/review dates; detailed provenance, methods, admitted results or result gaps, limitations and original assessment appear in the work atlas.

Sources: [620]

Evidence: Family-source bibliography - AA-LCR [benchmark.aa-lcr]. https://huggingface.co/datasets/ArtificialAnalysis/AA-LCR/resolve/9a77ef56b717057ade24ceab4d273712a0b4f19e/README.md (inspected version: 9a77ef56b717057ade24ceab4d273712a0b4f19e); https://www.mercor.com/apex/methodology/ (inspected version: body-sha256:fddaa02bee9bb213c886bc27d6e0077b48baeba2e292818fdcc10a1b3dfb9f0c). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [762], [775]

Evidence: Family-source bibliography - AdvancedIF [benchmark.advancedif]. https://arxiv.org/pdf/2511.10507v2 (inspected version: 2511.10507v2); https://www.mercor.com/apex/methodology/ (inspected version: body-sha256:fddaa02bee9bb213c886bc27d6e0077b48baeba2e292818fdcc10a1b3dfb9f0c). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [557], [775]

Evidence: Family-source bibliography - AIME [benchmark.aime]. https://www.mercor.com/apex/ (inspected version: body-sha256:cfd7c2878ffede2a5638fb607319c97e530e207945542566959ab1086f9561a5); https://www.mercor.com/apex/methodology/ (inspected version: body-sha256:fddaa02bee9bb213c886bc27d6e0077b48baeba2e292818fdcc10a1b3dfb9f0c). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [774], [775]

Evidence: Family-source bibliography - AM-DeepSeek-R1-Distilled [dataset.am-deepseek-r1-distilled]. https://huggingface.co/datasets/a-m-team/AM-DeepSeek-R1-Distilled-1.4M/resolve/53531c06634904118a2dcd83961918c4d69d1cdf/README.md (inspected version: hf:53531c06634904118a2dcd83961918c4d69d1cdf). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [441]

Evidence: Family-source bibliography - Antidote [benchmark.surge.antidote]. https://surgehq.ai/blog/introducing-antidote (inspected version: body-sha256:561c2781c4deeb6c32064c90be2bd12c5c57b7cc37a3253c2a980dbac58c9f3f). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [22]

Evidence: Family-source bibliography - APEX-SWE [benchmark.apex-swe]. https://arxiv.org/pdf/2601.08806v3 (inspected version: 2601.08806v3); https://github.com/Mercor-Intelligence/apex-swe (inspected version: body-sha256:50602ac696b779bd4648f2344e53c4a43a7b47c1027c23c1227f453b38b18c50); https://github.com/Mercor-Intelligence/apex-swe/blob/main/observability/README.md (inspected version: git-blob:d7350fa3dc67851d0c45a6b11c197c29e4586c81). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [559], [763], [764]

Evidence: Family-source bibliography - BabaIsBench / Baba Is Harbor [benchmark.quesma.babaisbench]. https://api.github.com/repos/stared/baba-is-harbor/readme (inspected version: GitHub-blob:dee1f7dbee8af3b74106577c49d0cffd91a4c381); https://quesma.com/benchmarks/babaisbench/ (inspected version: body-sha256:3b66bae56f5f7428600a77d3537d0e1f1df57dae643c81bef9553d08d82d7688). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [590], [591]

Evidence: Family-source bibliography - Bespoke-Stratos [dataset.bespoke-stratos]. https://huggingface.co/datasets/bespokelabs/Bespoke-Stratos-17k/resolve/9e9adba943911a9fc44dffcb30aaa18dc96ae6df/README.md (inspected version: hf:9e9adba943911a9fc44dffcb30aaa18dc96ae6df). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [442]

Evidence: Family-source bibliography - BinaryAudit [benchmark.quesma.binaryaudit]. https://api.github.com/repos/QuesmaOrg/BinaryAudit/readme (inspected version: GitHub-blob:1e3a134981efd907a190bab3d7198a7a33a06fc2); https://quesma.com/benchmarks/binaryaudit/ (inspected version: body-sha256:153bda4d7431d0f2eb2bb2d500afaee6903dbeb27f0e479521e4fd7858cdfab6). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [592], [593]

Evidence: Family-source bibliography - BrickVision [benchmark.elorian.brickvision]. https://elorian.ai/blog/why-grounding (inspected version: body-sha256:ada0c46a48d29aba52edde0dd6bcdea57e61b16446b1151bef6784962220f779); https://huggingface.co/datasets/elorian-ai/BrickVision/resolve/ccedb16baadb1c7f2a85c0af04c168d953fa8b75/README.md (inspected version: ccedb16baadb1c7f2a85c0af04c168d953fa8b75). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [585], [586]

Evidence: Family-source bibliography - BrowseComp [benchmark.browsecomp]. https://arxiv.org/pdf/2504.12516v1 (inspected version: 2504.12516v1). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [338]

Evidence: Family-source bibliography - Chartography [benchmark.surge.chartography]. https://arxiv.org/pdf/2608.10677v1 (inspected version: 2608.10677v1); https://github.com/surge-ai/chartography/blob/main/README.md (inspected version: git-blob:dc534fbfff5fa7f94cd62a567dc73e9747666725). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [2], [3]

Evidence: Family-source bibliography - CompileBench [benchmark.quesma.compilebench]. https://api.github.com/repos/QuesmaOrg/CompileBench/contents/LICENSE (inspected version: GitHub-blob:2fb4359106d3bcb47479820d21da2e2bcb5f00b1); https://api.github.com/repos/QuesmaOrg/CompileBench/readme (inspected version: GitHub-blob:902a931b7c33926e03648e39b3c078f0d3b43ebd). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [581], [594]

Evidence: Family-source bibliography - ComplexConstraints [benchmark.surge.complexconstraints]. https://arxiv.org/pdf/2606.09118v3 (inspected version: 2606.09118v3); https://github.com/surge-ai/complex-constraints/blob/main/README.md (inspected version: git-blob:c0eb92b3db35cc3c650a569fb89da72bbe5e5dfc); https://surgehq.ai/blog/complexconstraints-a-benchmark-for-entangled-instruction-following (inspected version: body-sha256:20f4b5b6606d9ec14b60871f380e14cf76460f188e343136d63651eae63c1a23). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [4], [5], [23]

Evidence: Family-source bibliography - DecoyCount [benchmark.elorian.decoycount]. https://api.github.com/repos/elorian-ai/decoycount/contents/DATASET.md (inspected version: GitHub-blob:cfccc6e32582204fd6991cba75495bf9213314d4); https://api.github.com/repos/elorian-ai/decoycount/contents/DATA_LICENSE.md (inspected version: GitHub-blob:071e80fba98a592a53114eb3a287bc2fc0b8b564); https://api.github.com/repos/elorian-ai/decoycount/contents/RELEASE.md (inspected version: GitHub-blob:b4d17df87da04b6bf46e859174262b45298e76e3); https://api.github.com/repos/elorian-ai/decoycount/readme (inspected version: GitHub-blob:fb42ade06749442dcc382778d75d5bce33e215c0); https://elorian.ai/blog/decoycount (inspected version: body-sha256:036234eea02e98a60455ece65f8b9c3ae5ef54c0008fd32a9ddc2e1b7ff8ebcd). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [582], [583], [584], [587], [588]

Evidence: Family-source bibliography - DeepResearch Bench II [benchmark.deepresearch-bench]. https://arxiv.org/pdf/2601.08536v1 (inspected version: arxiv:2601.08536v1); https://github.com/imlrz/DeepResearch-Bench-II/blob/main/README.md (inspected version: git-blob:f3a55c29bae84b5983695105ae3957912c6b1c28); https://www.mercor.com/apex/oss-benchmarks/oss-deepresearch-bench-ii-leaderboard/ (inspected version: body-sha256:da8c828606613b4d2ca8d4a2d6e79010ef66b2ce20fd208c6b5e352b4cc81e80). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [611], [765], [766]

Evidence: Family-source bibliography - EMRBench [benchmark.refresh-emrbench]. https://refresh.dev/blog/emr-bench (inspected version: body-sha256:073c58623e6a84d1301a5e3b18d386e28462ab5419fb18bc1a040ca9c586d56b); https://refresh.dev/one-pager (inspected version: body-sha256:8c21d7374c7c733649582386112ea4750a5185f39c5e3fb9cd70f020141f04c8). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [597], [598]

Evidence: Family-source bibliography - EnterpriseBench / CoreCraft [benchmark.surge.enterprisebench-corecraft]. https://arxiv.org/pdf/2602.16179v5 (inspected version: 2602.16179v5); https://surgehq.ai/blog/enterprisebench-corecraft (inspected version: body-sha256:06e3008dcf989c0d54aa90dda31ca01fc15d476f86fbb0415edb8d1a924c140e). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [6], [24]

Evidence: Family-source bibliography - FrontierCode [benchmark.frontiercode]. https://cognition.com/blog/frontier-code (inspected version: body-sha256:9a52a8b788d74df0e9fea13992e5547f8f4bd88e0b953546024e7ca11972a6bd); https://www.mercor.com/apex/partner-benchmarks/frontiercode-leaderboard/ (inspected version: body-sha256:98368088aa5230bf236c9cfc05553282deb1a2223bb1f6fd18f24346177389f0). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [767], [768]

Evidence: Family-source bibliography - Gauntlet [dataset.refresh.gauntlet]. https://refresh.dev/blog/gauntlet-4k-rlvr (inspected version: body-sha256:f82fb6f79540e02c2980ea20edb9d72ad5b73211af0051681a65ac14d2b868ab); https://refresh.dev/blog/gauntlet-80k (inspected version: body-sha256:47cbdd1710cb41f52aebabea877862dbbcd3fbed089c09681488c4e4d5180da6). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [621], [622]

Evidence: Family-source bibliography - GDP.pdf [benchmark.surge.gdp-pdf]. https://arxiv.org/pdf/2607.11192v3 (inspected version: 2607.11192v3); https://github.com/surge-ai/gdp-pdf/blob/main/README.md (inspected version: git-blob:567b4708e61dfbdca9a740f4ecc5dec55d4ec528); https://surgehq.ai/blog/gdp-pdf-can-100b-ai-models-master-the-documents-that-run-the-world (inspected version: body-sha256:2c518677115b1fe3e4c422353e1c36c46ff52683a38a1590ebab80d55c6ba141). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [7], [8], [25]

Evidence: Family-source bibliography - GDP.xlsx [benchmark.gdp-xlsx]. https://huggingface.co/datasets/surgeai/GDP.xlsx/resolve/7cdb5406bb7bd12cf73cbaaabb0edcf449a473e9/README.md (inspected version: 7cdb5406bb7bd12cf73cbaaabb0edcf449a473e9). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [623]

Evidence: Family-source bibliography - GDPval [benchmark.gdpval]. https://arxiv.org/pdf/2510.04374v1 (inspected version: 2510.04374v1). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [350]

Evidence: Family-source bibliography - HackerRank community graded submissions [dataset.hackerrank-community-graded-submissions]. https://www.hackerrank.com/bench (inspected version: body-sha256:c50227b4688a9e11850631e587d959923580fd9bf11f5e56af3027faf0848752); https://www.hackerrank.com/bench/report (inspected version: body-sha256:c550f513c8286fae1b2cb26bc125fcd29c341387d2d939b826035c4d9f9fbd46). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [769], [770]

Evidence: Family-source bibliography - HackerRank Orchestrate agent traces [dataset.hackerrank-orchestrate-agent-traces]. https://www.hackerrank.com/bench (inspected version: body-sha256:c50227b4688a9e11850631e587d959923580fd9bf11f5e56af3027faf0848752); https://www.hackerrank.com/bench/report (inspected version: body-sha256:c550f513c8286fae1b2cb26bc125fcd29c341387d2d939b826035c4d9f9fbd46). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [769], [770]

Evidence: Family-source bibliography - HANDBOOK.md [benchmark.surge.handbook]. https://arxiv.org/pdf/2607.25398v3 (inspected version: 2607.25398v3); https://github.com/surge-ai/handbook/blob/main/README.md (inspected version: git-blob:34ddc11c8c64f4db7cd1c4ad62c33427993d5940). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [9], [10]

Evidence: Family-source bibliography - Harvey LAB [benchmark.harvey-lab]. https://github.com/harveyai/harvey-labs (inspected version: body-sha256:2577822e195fcc8f4bbbafe5052922ba4f74101333b0f8212565f56b91578cbf); https://github.com/harveyai/harvey-labs/blob/main/docs/architecture.md (inspected version: git-blob:4550b48a7dfeb04567c3c00603c907ad44501e36); https://github.com/harveyai/harvey-labs/blob/main/docs/eval-strategies.md (inspected version: git-blob:190dc2ad7f8503aca6cca316478105723f9ae955); https://www.mercor.com/apex/ (inspected version: body-sha256:cfd7c2878ffede2a5638fb607319c97e530e207945542566959ab1086f9561a5). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [771], [772], [773], [774]

Evidence: Family-source bibliography - Hemingway-bench [benchmark.surge.hemingway]. https://surgehq.ai/blog/hemingway-bench-ai-writing-leaderboard (inspected version: body-sha256:a8c6db13ea98f5cbfa538d6d3f0c8654397dcc2259981308108568ea14b24fee). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [26]

Evidence: Family-source bibliography - HSS [benchmark.elorian.hss]. https://huggingface.co/datasets/ScaleAI/HSS/resolve/9add42c59783cdd0d5b871473374e59f05f2b3cb/README.md (inspected version: 9add42c59783cdd0d5b871473374e59f05f2b3cb). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [600]

Evidence: Family-source bibliography - Light-R1 training data family [dataset.light-r1]. https://huggingface.co/datasets/qihoo360/Light-R1-SFTData/resolve/f939e79f82f0c7c14cbb871ee0b75f8f7226ad73/README.md (inspected version: hf:f939e79f82f0c7c14cbb871ee0b75f8f7226ad73). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [443]

Evidence: Family-source bibliography - MedMosaic [dataset.medmosaic]. https://huggingface.co/datasets/icml-anon-submission/medmosaic-dataset/resolve/a6ea67bd4a65b87248c6651e559656b2c31fa669/README.md (inspected version: a6ea67bd4a65b87248c6651e559656b2c31fa669). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [624]

Evidence: Family-source bibliography - Multi-SWE-bench [benchmark.multi-swe-bench]. https://arxiv.org/pdf/2504.02605v1 (inspected version: 2504.02605v1); https://huggingface.co/datasets/ByteDance-Seed/Multi-SWE-bench/resolve/56ff018c04a38e27ada1e9d0a6d5839a51f88f0d/README.md (inspected version: 56ff018c04a38e27ada1e9d0a6d5839a51f88f0d); https://raw.githubusercontent.com/multi-swe-bench/multi-swe-bench/24f493f8a103e72312ded4f6b9c89f081d69cb09/README.md (inspected version: 24f493f8a103e72312ded4f6b9c89f081d69cb09). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [627], [628], [629]

Evidence: Family-source bibliography - Multi-SWE-RL [dataset.multi-swe-rl]. https://arxiv.org/pdf/2504.02605v1 (inspected version: 2504.02605v1); https://huggingface.co/datasets/ByteDance-Seed/Multi-SWE-RL/resolve/9777648932daa214ba18c70c81e85821b5836f32/README.md (inspected version: 9777648932daa214ba18c70c81e85821b5836f32). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [629], [630]

Evidence: Family-source bibliography - Nexdata product 1463 [dataset.nexdata.product1463]. https://www.nexdata.ai/ (inspected version: body-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65); https://www.nexdata.ai/datasets/embodied-ai/1463 (inspected version: body-sha256:e3d60a197240029b0055ddbb6630342bd8502c69df019db6722e4af519c7a1c4). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [11], [14]

Evidence: Family-source bibliography - Nexdata product 1515 [dataset.nexdata.product1515]. https://www.nexdata.ai/ (inspected version: body-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65); https://www.nexdata.ai/datasets/llm/1515 (inspected version: body-sha256:6e8b506f440911fea8fbab1cc493ebbceafb9d851182391a9f0765b70ffa940c). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [11], [15]

Evidence: Family-source bibliography - Nexdata product 1814 [dataset.nexdata.product1814]. https://www.nexdata.ai/ (inspected version: body-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65); https://www.nexdata.ai/datasets/llm/1814 (inspected version: body-sha256:9ba68417e7e6dcfab54d1d7a2ac88704768d7c1ca9c58c70e36facce9ee5cab1). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [11], [16]

Evidence: Family-source bibliography - Nexdata product 2121 [dataset.nexdata.product2121]. https://www.nexdata.ai/ (inspected version: body-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65); https://www.nexdata.ai/datasets/llm/2121 (inspected version: body-sha256:6635303d177b5e96d948d3b4d1ed6b89ac8cd48b54e3373b47641c8ef449a21d). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [11], [17]

Evidence: Family-source bibliography - Nexdata product 2145 [dataset.nexdata.product2145]. https://www.nexdata.ai/ (inspected version: body-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65); https://www.nexdata.ai/datasets/embodied-ai/2145 (inspected version: body-sha256:9946cc44dcccf95642be79d6bac8e0f25a0dacad80df305df26764738ef82674). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [11], [18]

Evidence: Family-source bibliography - Nexdata product 2185 [dataset.nexdata.product2185]. https://www.nexdata.ai/ (inspected version: body-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65); https://www.nexdata.ai/datasets/llm/2185 (inspected version: body-sha256:83b4b7a89d08228482851d704278f6c40fa0e94a00e119afab8b2a8a799e6e87). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [11], [19]

Evidence: Family-source bibliography - Nexdata product 2236 [dataset.nexdata.product2236]. https://www.nexdata.ai/ (inspected version: body-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65); https://www.nexdata.ai/datasets/embodied-ai/2236 (inspected version: body-sha256:b752361cdb55f340d23ea851fd24c07534d6cc6e3449fa3797d483fe87c00d51). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [11], [20]

Evidence: Family-source bibliography - OTelBench [benchmark.quesma.otelbench]. https://api.github.com/repos/QuesmaOrg/otel-bench/readme (inspected version: GitHub-blob:7b4d9b63a1dda0f4fc91b8fbb68636c849f10e38); https://quesma.com/blog/introducing-otel-bench/ (inspected version: body-sha256:11d5f342de0c6efc3b9adf9e8a788fcfed54879744e027a788e06646a80b1348). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [595], [596]

Evidence: Family-source bibliography - PDFBench [benchmark.refresh-pdfbench]. https://refresh.dev/blog/pdf-bench (inspected version: body-sha256:8e7739c0fccadef5cebecc86394aaeeb0d41284c574dd5354b5119611059ac21); https://refresh.dev/one-pager (inspected version: body-sha256:8c21d7374c7c733649582386112ea4750a5185f39c5e3fb9cd70f020141f04c8). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [598], [599]

Evidence: Family-source bibliography - Physical Engineering's Last Exam (PLE) [benchmark.ple]. https://arxiv.org/pdf/2610.30537v2 (inspected version: 2610.30537v2); https://github.com/google-research/ple (inspected version: body-sha256:ddec3c207071e3059a3f8fd611e8eb9a418c800f6e016b10bebbac6edf9cf4c6). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [700], [776]

Evidence: Family-source bibliography - PostTrain Arena [benchmark.posttrain-arena]. https://github.com/benchflow-ai/posttrainarena/blob/main/README.md (inspected version: git-blob:ecd07be0ee6bd8358865d53309979495b4b63a73); https://posttrain.com (inspected version: body-sha256:ad59b854827516ba198cd0870e6c32e722177e153ff4478692773142a8288e06); https://posttrain.com/docs/spec (inspected version: body-sha256:1a22b683ee06ffc0b1f65c81ccc5f2d7095ad33960f1c728552dc8bc3a39e9fc). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [1], [12], [13]

Evidence: Family-source bibliography - RedlineBench [benchmark.redlinebench]. https://github.com/crosbylegal/redline-bench/blob/main/README.md (inspected version: git-blob:426e74f8777ba93d2e4a4538d23247b97085d94b); https://huggingface.co/datasets/crosbylegal/RedlineBench/resolve/eee1b6790982ed1279e86bec7616b662a61993e6/README.md (inspected version: eee1b6790982ed1279e86bec7616b662a61993e6); https://intelligence.crosby.ai/benchmark/ (inspected version: body-sha256:ac54c61bff51239722b8579ec29b3242be62531e543bbd5ed652cbdbdd4468f1). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [777], [778], [779]

Evidence: Family-source bibliography - Riemann-Bench [benchmark.surge.riemann]. https://arxiv.org/pdf/2604.06802v3 (inspected version: 2604.06802v3); https://surgehq.ai/blog/riemann-bench-a-benchmark-for-moonshot-mathematics (inspected version: body-sha256:4698f9d2b48e2ee48d3d58605a94affbaf0c8a95b0339c18ae8fe52cacd18cf2). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [21], [27]

Evidence: Family-source bibliography - s1K [dataset.s1k]. https://huggingface.co/datasets/simplescaling/s1K-1.1/resolve/96c411f1fe4c49d20f0e2a1565f61e1a28b0b84d/README.md (inspected version: hf:96c411f1fe4c49d20f0e2a1565f61e1a28b0b84d). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [444]

Evidence: Family-source bibliography - ScienceVision [benchmark.elorian.sciencevision]. https://elorian.ai/blog/sciencevision (inspected version: body-sha256:99fd76e1bdbad37eec877d963b99cc111d9c628887897f250200a18e82088c7e). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [589]

Evidence: Family-source bibliography - Search-PI [benchmark.meta.search-pi]. https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report/ (inspected version: body-sha256:1b760ea7770828e223daaa966a547529b4b1398a7ac44b97992488fab740133a). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [625]

Evidence: Family-source bibliography - SmallThoughts [dataset.smallthoughts]. https://huggingface.co/datasets/SmallDoge/SmallThoughts/resolve/8fdc9afef80403b7c50ebb4356dd7878d3218d37/README.md (inspected version: hf:8fdc9afef80403b7c50ebb4356dd7878d3218d37). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [445]

Evidence: Family-source bibliography - Snorkel Finance Reasoning [benchmark.snorkel.finance-reasoning]. https://huggingface.co/datasets/snorkelai/agent-finance-reasoning/resolve/4c89e0aa8358509b6645b87ede329300336e4d3c/README.md (inspected version: 4c89e0aa8358509b6645b87ede329300336e4d3c); https://snorkel.ai/leaderboard/category/financereasoning/ (inspected version: body-sha256:58e9a00deedf172e09f8177d6179f4f08453e24c18041dfaa07fc5eba34ef2d3). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [631], [632]

Evidence: Family-source bibliography - sudo L7 [benchmark.surge.sudo-l7]. https://surgehq.ai/blog/sudo-l7 (inspected version: body-sha256:da59b2f17377c3ea549ea207653e0b368784a1ec82053be7467053663c1dfab9). These are the source versions bound by this family’s approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [28]

Evidence: Family-source bibliography - SWE-Marathon [benchmark.swe-marathon]. https://api.github.com/repos/refreshdotdev/long-horizon/readme (inspected version: GitHub-blob:040356da67f4e9ac7b4f39dc8765116ece5ba1bf; API-body-sha256:d05036cb2b799a06335d70065f851c56502e1058f645c073279e0bc7a1d412d9); https://api.github.com/repositories/1153976336/readme (inspected version: GitHub-blob:9c45212807174aff78b014c9ea1e5ee207756c02; API-body-sha256:9c8b0743ae9875b876c3368947c505aa2762395cd4c77007822112f48e450b22); https://arxiv.org/pdf/2606.07682v1 (inspected version: body-sha256:f0276a983f7f32c5f5bd8aeacb82141282da2bde3cc8da9cc879b9dbe2e20f5f). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [601], [602], [603]

Evidence: Family-source bibliography - Terminal-Bench [benchmark.terminal-bench]. https://arxiv.org/pdf/2601.11868v1 (inspected version: 2601.11868v1). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [377]

Evidence: Family-source bibliography - Trajectory 72-scenario indirect prompt-injection evaluation [benchmark.trajectory-indirect-prompt-injection-72]. https://claude.com/resources/articles/auto-mode-default-in-claude-code (inspected version: body-sha256:01de78d6668c4637906840660344872b288978fafbcffc5ff384322b901d9d68). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [604]

Evidence: Family-source bibliography - U-MATH [benchmark.u-math]. https://arxiv.org/pdf/2412.03205v4 (inspected version: 2412.03205v4); https://huggingface.co/datasets/toloka/u-math/raw/7210f97b3f21122c3a90126935e966ca7d8951b6/README.md (inspected version: 7210f97b3f21122c3a90126935e966ca7d8951b6). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [552], [633]

Evidence: Family-source bibliography - μ-MATH [benchmark.mu-math]. https://arxiv.org/pdf/2412.03205v4 (inspected version: 2412.03205v4); https://huggingface.co/datasets/toloka/mu-math/resolve/c3d791bc7be33bf73a1ae5d8c3084a820cb580c4/README.md (inspected version: c3d791bc7be33bf73a1ae5d8c3084a820cb580c4). These are the source versions bound by this family's approved claims. Their claim records retain exact locators, response hashes, capture dates, bounded reading spans, qualifications and genuine AI review receipts; a later live URL is not a substitute for the pinned version.

Sources: [552], [626]

Method: Bounded core admission and discovery scope. The actual complete-inventory independent scope certificate has SHA-256 49a6561fa9dc40829623cb368d8bd093587126e29e2ee5641f1658d039682783; the accepted stable-identity join has SHA-256 b04e8172f6dc429f82116c91dbbb9bdf9c5e6cdbe9f758d62273ae8a0e33771a. The coordinator subsequently coapproved and actually applied the freeze. Their 532 retained named targets resolve to 416 source-qualified logical units: 84 mandatory organization-audit units, 160 paper works, 85 dataset families and 87 benchmark families. The remaining target states are 102 Unknowns, thirteen nonunit aliases or manifestations and one source-results container. Eligibility is determined independently of scientific completion or achieved coverage. A fully reviewed scientific paper can remain outside this source-qualified denominator; this is why 225 scientifically reviewed works and 160 inventory paper identities are different measures. No prior source proof, failed route, negative review or unresolved target is erased by the 416-unit freeze.

Method: Admission routes and counting rule. Route A requires a captured official intended-supplier direct artifact link and concrete data-workflow relevance. Route B requires creator-primary evidence of a distinct named family or evaluation protocol. Route C requires primary reusable-method evidence and an independently authored field-level origin, reference-method or reusable-baseline designation. Route D is configured intended-supplier/legacy audit membership, with commercial qualification decided separately. A bibliography occurrence, title, scientific completion or corporate byline is insufficient by itself. Shared authors are disclosed rather than subjected to a universal all-author-disjoint rule. Reused examples do not by themselves defeat a separately defined family; releases, aliases, mirrors, formats and subsets do not automatically create one. These are the accepted inventory-policy rules, not new scientific approvals.

Method: Retained discovery frontiers. The complete scope record preserves 145 discovery partitions and all 97,307 original discovery rows, 125 creator-root records and 27 foundation-source functions. The retained foundation accounting includes 3,167 citation occurrences, 134 explicit designation tasks, 87 group designation tasks and 53 identity-precision holds. Each partition retains its query, cursor/capture state, errors and pending action rather than being relabeled complete when a bounded named membership is accepted. The 85-record supplier roster and its known, pending and restricted function URLs are also retained. The 416 resolved identities are a bounded source-qualified subtotal; unresolved pagination, linked child artifacts, private catalogs, missing functions and unexamined queries remain open. No finite capacity envelope is an upper bound on the literature or on inaccessible commercial stock.

Method: Function scope remains bounded. The later captured metadata decisions for micro1, AfterQuery, BenchFlow, Mercor and HackerRank approve only their exact observed listing, controls, UI states or named membership routes. They do not close whole research, creator, supplier or foundation functions. Earlier failed captures and unread panels remain historical unknowns at their original scope. Newly observed UI states form a separate later layer. A contact-form Next control does not prove research pagination, and an empty-state string coexisting with populated cards is not an exclusion. Nested identity, scientific reading, access and rights tasks remain separate from listing enumeration.

Method: Unknown-target register reading rule. The following paragraphs enumerate all 102 retained named Unknown targets: 90 paper works, eleven native/family-identity leads and one organization-scope lead. An Unknown does not mean irrelevant, excluded, scientifically false or zero remaining capacity. An already approved scientific observation remains approved at its actual scope. The open task is the stated source route or logical identity. The stable IDs support reproduction without exposing private reading-ledger paths, raw source bodies, private payloads or contributor records.

Method: Unknown paper-route register. Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions [paper.aaai.26886]; Efficient Data Labeling by Hierarchical Crowdsourcing with Large Language Models [paper.acl.2025.coling-main.748]; Sources of Disagreement in Data for LLM Instruction Tuning [paper.acl.2025.comedi-1.3]; CrowdAgent: Multi-Agent Managed Multi-Source Annotation System [paper.acl.2025.emnlp-demos.72]; How Reliable is Multilingual LLM-as-a-Judge? [paper.acl.2025.findings-emnlp.587]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. M-IFEval: Multilingual Instruction-Following Evaluation [paper.acl.2025.findings-naacl.344]; Reinforcement Learning on Web Interfaces Using Workflow-Guided Exploration [paper.arxiv.1802.08802]; CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement [paper.arxiv.1808.06080]; Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus [paper.arxiv.2104.08758]; ScienceWorld: Is your Agent Smarter than a 5th Grader? [paper.arxiv.2203.07540]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. The Data-Production Dispositif [paper.arxiv.2205.11963]; CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation [paper.arxiv.2206.08931]; WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions [paper.arxiv.2304.12244]; Let's Verify Step by Step [paper.arxiv.2305.20050]; UltraFeedback: Boosting Language Models with Scaled AI Feedback [paper.arxiv.2310.01377]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts [paper.arxiv.2310.02255]; ReMax: A Simple, Effective, and Efficient Method for Aligning Large Language Models [paper.arxiv.2310.10505]; The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI [paper.arxiv.2310.16787]; MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI [paper.arxiv.2311.16502]; Improving Task Instructions for Data Annotators: How Clear Rules and Higher Pay Increase Performance in Data Annotation in the AI Economy [paper.arxiv.2312.14565]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning [paper.arxiv.2402.06619]; If in a Crowdsourced Data Annotation Pipeline, a GPT-4 [paper.arxiv.2402.16795]; LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems [paper.arxiv.2403.09810]; JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models [paper.arxiv.2404.01318]; Self-Play Preference Optimization for Language Model Alignment [paper.arxiv.2405.00675]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. WildChat: 1M ChatGPT Interaction Logs in the Wild [paper.arxiv.2405.01470]; RLAIF-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness [paper.arxiv.2405.17220]; Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis [paper.arxiv.2405.21075]; AgentGym: Evolving Large Language Model-based Agents across Diverse Environments [paper.arxiv.2406.04151]; DataComp-LM: In search of the next generation of training sets for language models [paper.arxiv.2406.11794]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. WPO: Enhancing RLHF with Weighted Preference Optimization [paper.arxiv.2406.11827]; Step-DPO: Step-wise Preference Optimization for Long-chain Reasoning of LLMs [paper.arxiv.2406.18629]; Consent in Crisis: The Rapid Decline of the AI Data Commons [paper.arxiv.2407.14933]; Quality Assured: Rethinking Annotation Strategies in Imaging AI [paper.arxiv.2407.17596]; The Problems with Proxies: Making Data Work Visible through Requester Practices [paper.arxiv.2408.11667]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains? [paper.arxiv.2410.03859]; ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery [paper.arxiv.2410.05080]; MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering [paper.arxiv.2410.07095]; RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts [paper.arxiv.2411.15114]; TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks [paper.arxiv.2412.14161]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. REINFORCE++: Stabilizing Critic-Free Policy Optimization with Global Advantage Normalization [paper.arxiv.2501.03262]; MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs [paper.arxiv.2501.17399]; Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming [paper.arxiv.2501.18837]; Process Reinforcement through Implicit Rewards [paper.arxiv.2502.01456]; The digital labour of artificial intelligence in Latin America: a comparison of Argentina, Brazil, and Venezuela [paper.arxiv.2502.06317]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models [paper.arxiv.2502.07346]; SWE-Lancer: Can Frontier LLMs Earn $1 Million from Real-World Freelance Software Engineering? [paper.arxiv.2502.12115]; ShieldAgent: Shielding Agents via Verifiable Safety Policy Reasoning [paper.arxiv.2503.22738]; VerifiAgent: a Unified Verification Agent in Language Model Reasoning [paper.arxiv.2504.00406]; PaperBench: Evaluating AI's Ability to Replicate AI Research [paper.arxiv.2504.01848]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. Minority Reports: Balancing Cost and Quality in Ground Truth Data Annotation [paper.arxiv.2504.09341]; STAMP Your Content: Proving Dataset Membership via Watermarked Rephrasings [paper.arxiv.2504.13416]; TTRL: Test-Time Reinforcement Learning [paper.arxiv.2504.16084]; AutoLibra: Agent Metric Induction from Open-Ended Human Feedback [paper.arxiv.2505.02820]; The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text [paper.arxiv.2506.05209]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks? [paper.arxiv.2509.16941]; MENLO: From Preferences to Proficiency -- Evaluating and Modeling Native-like Quality Across 47 Languages [paper.arxiv.2509.26601]; Is It Thinking or Cheating? Detecting Implicit Reward Hacking by Measuring Reasoning Effort [paper.arxiv.2510.01367]; RECODE-H: A Benchmark for Research Code Development with Interactive Human Feedback [paper.arxiv.2510.06186]; How do data owners say no? A case study of data consent mechanisms in web-scraped vision-language AI training datasets [paper.arxiv.2511.08637]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. BenchGuard: Who Guards the Benchmarks? Automated Auditing of LLM Agent Benchmarks [paper.arxiv.2604.24955]; Who Annotates in NLP? A Large-scale Assessment of Human Annotation Reporting between 2018 and 2025 [paper.arxiv.2606.02255]; Experts Rise Where LLMs Disagree: Using Cross-Model Disagreement to Target Expert Effort in LLM Codebook Revision for Large-Scale Annotation [paper.arxiv.2609.26926]; SelectTTRL: Representation-Guided Data Selection for Efficient Test-Time Reinforcement Learning [paper.arxiv.2610.16104]; On the Reliability of Majority-Vote Pseudo-Labels in Test-Time Reinforcement Learning [paper.arxiv.2611.09872]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown paper-route register. Fair human-centric image dataset for ethical AI benchmarking [paper.doi.10.1038.s41586-025-09716-2]; Text Recognition Using Anonymous CAPTCHA Answers [paper.doi.10.1145.3336191.3371795]; Prediction of Hourly Earnings and Completion Time on a Crowdsourcing Platform [paper.doi.10.1145.3394486.3403369]; World of Bits: An Open-Domain Platform for Web-Based Agents [paper.pmlr.v70-shi17a]. For each listed work, the retained official-link and field-bibliography crosswalk does not yet authenticate a qualifying direct supplier workflow link or an explicit independent field-level designation. Scientific completion, title, abstract and legacy flags do not supply that route. The next task is to authenticate a contextual official A link or capture the exact independent C designation and reconcile the native work/version while reusing existing primary-method proofs at their actual reading scope.

Method: Unknown-target disposition - artifact.pending.centific.icl-coraal [artifact.pending.centific.icl-coraal]. Two official surfaces describe a CORAAL-based evaluation protocol, but whether this is a separate family, a release or an evaluation subset remains unresolved. Retain the positive protocol lead and capture failures; establish exact creator identity without inventing new annotation credit or merging it with MedMosaic because of a misdirected link. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.hf.bethgelab.curatedthoughts [dataset.hf.bethgelab.curatedthoughts]. The captured source describes filtered OpenR1/OpenThoughts configurations and regex/MCQ filters. Whether the named multi-parent curation is a separate maintained family or filtered manifestations of parents remains unresolved; preserve possible additional identity without inventing mixture weights or independent-example counts. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.hf.secemp9.instruction_solution_thought [dataset.hf.secemp9.instruction_solution_thought]. The short creator card describes a merge of Bespoke-Stratos, OpenThoughts-114k and an OpenThoughts math view, with future expansion. It does not settle distinct curation, protocol or family identity; retain a possible derived family and the need for fuller creator documentation. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.hf.surgeai.chartography [dataset.hf.surgeai.chartography]. Exact primary content and the native-card-to-public-family identity join remain pending. Similar naming alone does not merge this lead with the separately admitted public Chartography benchmark or add a second family. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.hf.surgeai.complexconstraints [dataset.hf.surgeai.complexconstraints]. Exact primary content and the native-card-to-public-family identity join remain pending. Similar naming alone does not merge this lead with the separately admitted public ComplexConstraints benchmark or add a second family. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.hf.surgeai.gdp.pdf [dataset.hf.surgeai.gdp.pdf]. Exact primary content and the native-card-to-public-family identity join remain pending. Similar naming alone does not merge this lead with the separately admitted public GDP.pdf benchmark or add a second family. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.nexdata.product1703 [dataset.nexdata.product1703]. The Japanese 205-hour full-duplex multi-channel speech product shares the source-defined protocol/series with Korean and English products, differing principally by language, region and hours. Exact parent/configuration versus separately maintained family identity is unresolved; retain the positive source-native collection lead. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.nexdata.product1704 [dataset.nexdata.product1704]. The Korean 200-hour full-duplex multi-channel speech product belongs to the same described protocol/series as Japanese and English products, with per-language collection details. Source text does not resolve language configuration versus separately maintained family identity; retain the actual positive lead. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.nexdata.product1770 [dataset.nexdata.product1770]. The English 172-hour full-duplex multi-channel speech product shares the protocol/series with Japanese and Korean products. Language and row/hour changes alone do not settle logical-family cardinality. Preserve positive source content and the exact identity task. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.nexdata.product1785 [dataset.nexdata.product1785]. The 50K editing product and 1.51M editing product share editing targets/procedure and potentially overlapping component scopes. The exact cards do not settle separate-family versus expansion/subset identity. No merge, exclusion or zero is inferred. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - dataset.nexdata.product2039 [dataset.nexdata.product2039]. The 1.51M instruction-editing product and 50K editing product share editing targets/procedure and potentially overlapping component scopes. The source cards do not settle separate-family versus maintained expansion/subset identity. Keep both exact products and seek the two own product definitions/lineage; do not assume all SKUs are families or require example-disjointness. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Upwork [org.upwork]. Exact graph organization and its genuine recruitment-channel evidence retained. No retained intended supplier/legacy D origin is established. A paper recruitment relationship is not transitive commercial supplier qualification or declared supplier scope. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - M-RewardBench: Evaluating Reward Models in Multilingual Settings [paper.acl.2025.acl-long.3]. The supplied specialized reward-benchmark book list does not establish the required independent field-level designation. The book author is also a primary coauthor; that overlap is disclosed without inventing a universal author-disjoint rule or excluding the work. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Inherent Disagreements in Human Textual Inferences [paper.acl.Q19-1043]. The inspected worker-practice and Data Discontents passages cite empirical variability and context effects in human inference judgments. Neither explicitly designates this exact work as an origin, reference method or reusable baseline. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk [paper.arxiv.1712.05796]. The inspected worker-practice source cites empirical wage and worker-rights findings. These genuine findings remain retained, but they do not provide an explicit foundation-method designation. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Model Cards for Model Reporting [paper.arxiv.1810.03993]. The Data Discontents passage supports disaggregated reporting without explicitly designating this exact work as the reusable model-card protocol. The separate primary protocol proof remains genuine; an exact independent designation is still needed. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? [paper.arxiv.1912.08320]. The inspected Data Discontents passage supplies empirical evidence of underspecified Twitter dataset reporting. Its following documentation-framework discussion refers to other works; a citation does not transfer their designation to this target. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Extracting Training Data from Large Language Models [paper.arxiv.2012.07805]. The inspected Data Discontents occurrence is an empirical security/privacy extraction example, rather than an explicit origin, reference-method or baseline designation for this exact work. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks [paper.arxiv.2103.14749]. The inspected annotation-quality and Data Discontents passages describe label-error findings and a short validation procedure. They do not explicitly designate this exact work as a reference method or baseline. Genuine primary confident-learning and validation proofs remain retained. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Learning from Natural Language Feedback [paper.arxiv.2204.14146]. The book cites an A. Chen and colleagues 2024 TMLR publication, while the pinned native version lists Jérémy Scheurer and colleagues in 2022 with Angelica Chen as a coauthor. Exact publication/revision identity remains unresolved; title overlap and author reordering do not establish the join. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models [paper.arxiv.2206.04615]. The inspected LLM-survey narrative cites selected BIG-bench task components without explicitly designating the entire exact work as an origin, reference method or reusable baseline. BIG-bench and later BIG-bench-hard remain distinct. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - GPT-4 Technical Report [paper.arxiv.2303.08774]. The inspected Encord paragraph links GPT-4 as a comparative multimodal-model example without the specific data-workflow, annotation, curation or protocol rationale required for this route. Genuine primary disclosure and withholding facts remain retained; another contextual official route or independent designation is needed. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks [paper.arxiv.2303.15056]. The inspected LLM-survey and worker-practice passages cite comparative annotation-capability findings. These are empirical result/context citations without an explicit method or baseline designation. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - DICES Dataset: Diversity in Conversational AI Evaluation for Safety [paper.arxiv.2306.11247]. The inspected safety survey catalogs DICES with ambiguity and diversity descriptors, while another occurrence is bibliographic. A survey-table entry does not automatically supply foundation membership; an exact reference-method or baseline designation remains unresolved. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - AlpaGasus: Training A Better Alpaca with Fewer Data [paper.arxiv.2307.08701]. The inspected LLM survey gives grouped support for small high-quality instruction-data findings without a located named narrative designation of AlpaGasus. Grouped empirical support does not designate this exact method. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM [paper.arxiv.2311.09528]. The inspected book citation supplies dataset provenance and context for HelpSteer rather than an explicit reference-method or baseline designation. Primary annotation-method proofs remain retained. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks [paper.arxiv.2401.13649]. The inspected occurrence is a Digital Divide Data bibliography entry. Genuine primary human/task-evaluation facts remain retained, but a reference-only occurrence does not authenticate the required contextual direct artifact route. Another relevant official link or exact independent designation is needed. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research [paper.arxiv.2402.00159]. The inspected LLM-survey dataset overview describes Dolma mixture composition without a located explicit origin, reference-method or baseline designation. Relevance and genuine primary curation proof do not alone close route C. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale [paper.arxiv.2406.17557]. The captured LLM survey has a FineWeb bibliography entry without an inspected narrative title/name designation. Bibliography presence does not establish foundation membership; primary curation proofs and the remaining retrieval task stay retained. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - The Llama 3 Herd of Models [paper.arxiv.2407.21783]. The supplied book passages concern implementation, model history and context, rather than a located explicit reference-method or baseline designation for this work. Corporate byline and appendix ownership remain disclosed without becoming an automatic exclusion. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - DeepSeek-V3 Technical Report [paper.arxiv.2412.19437]. The inspected book passages concern compute accounting and GRPO use without explicitly designating the DeepSeek-V3 paper as an origin, reference method or reusable baseline. Genuine primary methods remain retained. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages [paper.arxiv.2505.11475]. The inspected book passages identify HelpSteer3 release size, provenance and preference context without a located explicit origin, reference-method or baseline designation. No new complete primary byline reading is claimed from HTML metadata selection. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Unknown-target disposition - RewardBench 2: Advancing Reward Model Evaluation [paper.arxiv.2506.01937]. The inspected book benchmark list and on-policy result citation do not establish the required independently authored origin, reference-method or baseline designation for RewardBench 2. The shared-author overlap is disclosed without imposing an all-author-disjoint rule or excluding the work. This is a retained inventory-method judgment, with no new scientific approval or deep-review credit.

Method: Nonunit alias/manifestation register - dataset.hf.a-m-team.am-deepseek-r1-distilled-1.4m. The accepted inventory join maps this target to AM-DeepSeek-R1-Distilled; response-source components and the small sample remain within that family. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.benchflow.skillsbench-leaderboard. The accepted inventory join maps this target to SkillsBench leaderboard/submission evidence; trial, trajectory, result and audit exports are benchmark manifestations, with prior version conflicts retained. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.bespokelabs.bespoke-stratos-17k. The accepted inventory join maps this target to Bespoke-Stratos; the named release and model sizes do not add families. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.google.physical-engineerings-last-exam. The accepted inventory join maps this target to the existing PLE benchmark through the exact primary paper and creator project, with gated payload and declared terms remaining separate. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.mercor.apex-accounting. The accepted inventory join maps this target to the existing APEX-Accounting family; a retained unavailable-card response does not erase the separately established public identity. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.open-r1.openr1-math-raw. The accepted inventory join maps this target to the raw manifestation of OpenR1-Math, with correctness/reparsing columns and multiple traces retained within the family. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.open-r1.openthoughts-114k-math. The accepted inventory join maps this target to a filtered and metadata-enriched OpenThoughts-114k view with original-question mapping, rather than a new family. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.open-thoughts.openthoughts-114k. The accepted inventory join maps this target to the original OpenThoughts family’s creator collection and metadata view. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.open-thoughts.openthoughts2-1m. The accepted inventory join maps this target to the second maintained OpenThoughts iteration; its upstream and new-generation lineage remains explicit without an additional family. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.open-thoughts.openthoughts3-1.2m. The accepted inventory join maps this target to the third maintained OpenThoughts iteration; generation and annotation multiplicity do not create another family. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.qihoo360.light-r1-sftdata. The accepted inventory join maps this target to Light-R1; SFT curriculum stages and DPO variants remain manifestations of its coherent maintained collection. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.simplescaling.s1k-1.1. The accepted inventory join maps this target to s1K; the same question set with new teacher traces remains a version of that family. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Nonunit alias/manifestation register - dataset.hf.smalldoge.smallthoughts. The accepted inventory join maps this target to SmallThoughts; split and serialization forms remain one named creator collection, distinct from OpenThoughts without an example-disjointness assumption. It adds no separate eligible unit or deep credit. Underlying source, payload, contributor and rights gaps remain at their actual scope.

Method: Source-results container register - dataset.hf.benchflow.posttrain-arena-results. The retained PostTrain Arena results repository exposes public metadata for organizer-reviewed experiments, including environment, model and evaluation pins, rather than a separately defined training-example corpus. Private logs and collaborator-only execution artifacts remain inaccessible/unknown. The container adds no separate logical unit, deep credit or claim that a complete held-out run succeeded.

Method: Retained identity-capacity component - G2i portfolio and SkillsBench. The named G2i twenty-task portfolio may be a manifestation of SkillsBench or a separate collection. Exact release absence is insufficient to settle the join; retain zero to one additional identity.

Method: Retained identity-capacity component - HackerRank ASTRA and Bench. Both source-defined protocols remain positive. The ASTRA project benchmark and current long-horizon Bench benchmark have unresolved roll-up identity; a business-page redirect does not settle it. Retain one to two possible additional units pending a source-backed join.

Method: Retained identity-capacity component - Apollo private attack collections. The broad-failure, data-exfiltration and safety-related named attack collections retain one to three possible family identities. Shared and held-out variants must roll up appropriately. No private payload, credentials, outreach or automatic exclusion is required.

Method: Retained identity-capacity component - MMMU-Pro and MMMU. The creator’s robust-version description, filtering, answer-option changes and visual renderings do not yet settle maintained-version versus separate-family identity. Retain zero to one additional unit.

Method: Retained identity-capacity component - SWE-bench Multilingual lineage. The supplier-linked multilingual SWE-bench expansion is not merged with ByteDance Multi-SWE-bench from a similar name, nor automatically assigned a separate family. Retain zero to one additional identity pending the exact creator lineage.

Method: Retained identity-capacity component - HealthBench Professional lineage. The named clinician-specific Professional collection and the separately documented HealthBench release retain an unresolved alias/subset or separate-family join. Retain zero to one additional identity; no scientific completion is transferred between them.

Method: Retained identity-capacity component - Mercor ten named Extended collections. BrowseComp Extended, CharXiv Extended, GDPval Extended, HLE Extended, Long Context Reasoning (AA-LCR) Extended, MedXpertQA MM Extended, MMMU-Pro Extended, SWE Atlas Codebase QnA Extended, SWE-bench Verified Extended and Terminal-Bench 2.1 Extended are retained named collection leads. Derived-family versus extension/release identity remains unresolved: zero to ten possible additional identities, with no automatic family or deep-review credit.

Method: Retained identity-capacity component - Mercor nine named off-the-shelf collections. APEX-Agents, APEX-Accounting, GDPval, BrowseComp, HLE, MMMUExpert, SWEBenchExtension, TerminalBenchPro and APEX-SWE remain nine named commercial collection identities whose training-family versus benchmark/extension relationships are unresolved. Retain zero to nine additional identities. They may alias Extended collections or resolved families; the combined Mercor uncertainty is not a count of asserted independent units and does not bound unspecified private stock.

Method: Retained identity-capacity component - CuratedThoughts filtered configurations. The named Unknown card remains a possible derived family or a set of filtered parent manifestations: zero to one possible additional unit, already represented by its Unknown target rather than reserved twice.

Method: Retained identity-capacity component - Instruction-solution-thought merge. The named three-source merge remains zero to one possible derived family, already represented by its Unknown target rather than reserved twice.

Method: Retained identity-capacity component - Centific ICL-CORAAL protocol. The creator protocol remains zero to one possible additional family, already represented by its Unknown target. A misdirected OpenDataset link does not establish identity with MedMosaic or a new CORAAL collection.

Method: Capacity arithmetic and uncardinalized scope. The eleven retained components record two to thirty-one possible additional units beyond resolved identities. Five further Nexdata product leads retain editing-pair and full-duplex language-series identity uncertainty; three Surge native-card joins remain unresolved against separately positive public benchmark families. The scope certificate’s finite capacity envelope of 418 to 546 is a bookkeeping envelope for named unresolved opportunities, not an eligible denominator, a universal upper bound or a deep-review percentage. These capacities overlap existing leads where specified and must not be added independently without identity reconciliation. Inaccessible unspecified commercial stock, unclosed functions and ongoing discovery remain uncardinalized. An actual newly source-qualified stable unit requires additive core expansion with an explicit versioned receipt.

Method: Frozen-core deep-review gap register. AIMMO [org.aimmo] (disposition pending, review depth metadata, status candidate); BenchFlow [org.benchflow] (disposition pending, review depth full_source, status candidate); Daivergent [org.daivergent] (disposition pending, review depth metadata, status candidate); Summa Linguae Technologies [org.summa_linguae_technologies] (disposition pending, review depth metadata, status candidate). These source-qualified identities remain in the 416-unit denominator. The current merged graph does not establish their completed deep disposition under method 1.3.0. Source eligibility, prior partial reading and adjacent artifact completion are not substitutes for their own bounded deep evidence and actual required reviews.

Method: Frozen-core deep-review gap register. paper.arxiv.1701.06049 [paper.arxiv.1701.06049] (disposition pending, review depth metadata, status candidate); paper.arxiv.1709.10163 [paper.arxiv.1709.10163] (disposition pending, review depth metadata, status candidate); paper.arxiv.1811.06521 [paper.arxiv.1811.06521] (disposition pending, review depth metadata, status candidate); paper.arxiv.1811.07871 [paper.arxiv.1811.07871] (disposition pending, review depth metadata, status candidate). These source-qualified identities remain in the 416-unit denominator. The current merged graph does not establish their completed deep disposition under method 1.3.0. Source eligibility, prior partial reading and adjacent artifact completion are not substitutes for their own bounded deep evidence and actual required reviews.

Method: Frozen-core deep-review gap register. paper.arxiv.1909.08593 [paper.arxiv.1909.08593] (disposition pending, review depth metadata, status candidate); Human Feedback is not Gold Standard [paper.arxiv.2309.16349] (disposition included, review depth metadata, status approved); paper.arxiv.2512.02556 [paper.arxiv.2512.02556] (disposition pending, review depth metadata, status candidate); paper.arxiv.2604.13609 [paper.arxiv.2604.13609] (disposition pending, review depth metadata, status candidate). These source-qualified identities remain in the 416-unit denominator. The current merged graph does not establish their completed deep disposition under method 1.3.0. Source eligibility, prior partial reading and adjacent artifact completion are not substitutes for their own bounded deep evidence and actual required reviews.

Method: Frozen-core deep-review gap register. paper.arxiv.2606.07682 [paper.arxiv.2606.07682] (disposition pending, review depth metadata, status candidate); paper.arxiv.2606.12683 [paper.arxiv.2606.12683] (disposition pending, review depth metadata, status candidate); paper.arxiv.2609.34458 [paper.arxiv.2609.34458] (disposition pending, review depth metadata, status candidate); paper.knox-stone.2008.tamer [paper.knox-stone.2008.tamer] (disposition pending, review depth metadata, status candidate). These source-qualified identities remain in the 416-unit denominator. The current merged graph does not establish their completed deep disposition under method 1.3.0. Source eligibility, prior partial reading and adjacent artifact completion are not substitutes for their own bounded deep evidence and actual required reviews.

Method: Frozen-core deep-review gap register. System Card: Claude Fable5.1andClaude Mythos5.1 [paper.report.anthropic.claude-fable-5-1-mythos-5-1-system-card] (disposition pending, review depth metadata, status candidate); System Card: Claude Fable5.2andClaude Mythos5.2 [paper.report.anthropic.claude-fable-5-2-mythos-5-2-system-card] (disposition pending, review depth metadata, status candidate); System Card: Claude Fable5andClaude Mythos5 [paper.report.anthropic.claude-fable-5-mythos-5-system-card] (disposition pending, review depth metadata, status candidate); System Card: Claude Opus5.5 [paper.report.anthropic.claude-opus-5-5-system-card] (disposition pending, review depth metadata, status candidate). These source-qualified identities remain in the 416-unit denominator. The current merged graph does not establish their completed deep disposition under method 1.3.0. Source eligibility, prior partial reading and adjacent artifact completion are not substitutes for their own bounded deep evidence and actual required reviews.

Method: Frozen-core deep-review gap register. System Card: Claude Opus5 [paper.report.anthropic.claude-opus-5-system-card] (disposition pending, review depth metadata, status candidate); Crosby RedlineBench [paper.report.crosby.redlinebench] (disposition pending, review depth metadata, status candidate); Gemini4Argon SafetyandSecurity Report [paper.report.google-deepmind.gemini-4-argon-safety-security] (disposition pending, review depth metadata, status candidate); Muse Spark SafetyandPreparedness Report [paper.report.meta.muse-spark-safety-preparedness] (disposition pending, review depth metadata, status candidate). These source-qualified identities remain in the 416-unit denominator. The current merged graph does not establish their completed deep disposition under method 1.3.0. Source eligibility, prior partial reading and adjacent artifact completion are not substitutes for their own bounded deep evidence and actual required reviews.

Method: Frozen-core deep-review gap register. Compare harnesses, not models: Blitzy vs. GPT-5.4 on SWE-Bench Pro [paper.report.quesma.blitzy-swebench-pro-audit] (disposition pending, review depth metadata, status candidate); report.hackerrank-astra-technical-report [report.hackerrank-astra-technical-report] (disposition pending, review depth metadata, status candidate); report.hackerrank-bench-technical-report [report.hackerrank-bench-technical-report] (disposition pending, review depth metadata, status candidate). These source-qualified identities remain in the 416-unit denominator. The current merged graph does not establish their completed deep disposition under method 1.3.0. Source eligibility, prior partial reading and adjacent artifact completion are not substitutes for their own bounded deep evidence and actual required reviews.

Method: Screening acceptance record - The independently accepted screening run recovered 197 of the frozen 200 eligible reference records: benchmarks 25/25, papers 65/65, supplier leads 50/50 and datasets 57/60. Its selected and full-frame critical-attribution and unresolved-conclusive counters are zero. This is a known-reference post-remediation repeat, not an unseen holdout or an estimate of population precision, discovery recall, final supplier qualification or market-wide performance. The original 200-record sensitivity contract remains false; the original thirteen uncertain cases are not all declared resolved, and prior held runs remain preserved. Separate AI admission review compared 24 source-first cases and nine additional postscore-targeted source-first cases, preserving all frozen truths, rows, membership and thresholds. Two operational deviations remain explicit: one producer authenticated execute-only capsule bytes outside the prescribed helper, and the coordinator hashed sealed files before its opening record. The latter was followed by a separately reviewed, prospective semantic-assembly continuation. Strict access-mode compliance, global blindness, complete historical transcript evidence and independent engine attestation are not claimed. Original-proof materiality review accepted bounded scientific compatibility while retaining these limits. The public first-edition screening review and the edition’s machine-readable audit and acceptance records document the exact scope.

Bibliography and source versions

References group identical URL, version and response digest. Multiple evidence bindings to the same pinned document share a reference number; all source IDs, precise locators and review receipts remain separate in evidence.html and the dataset.

[1] benchflow-posttrain. BenchFlow / PostTrain Arena organizers;
body-sha256:ad59b854827516ba198cd0870e6c32e722177e153ff4478692773142a8288e06
https://posttrain.com
Response SHA-256: ad59b854827516ba198cd0870e6c32e722177e153ff4478692773142a8288e06
Scoped evidence bindings: 3

[2] chartography-creator-readme. Surge AI; git-blob:dc534fbfff5fa7f94cd62a567dc73e9747666725
https://github.com/surge-ai/chartography/blob/main/README.md
Response SHA-256: b588ba1f8ab0ce4926577c8c05312c46d705a85b755d8e8d4ec79a376982852f
Scoped evidence bindings: 2

[3] chartography-paper. Surge AI; 2608.10677v1
https://arxiv.org/pdf/2608.10677v1
Response SHA-256: 77ef3accf5fe10db02baaa5ae6a7e8f9783cea449a1b8c2f9b8ea3bc803990fb
Scoped evidence bindings: 16

[4] complexconstraints-creator-readme. Surge AI; git-blob:c0eb92b3db35cc3c650a569fb89da72bbe5e5dfc
https://github.com/surge-ai/complex-constraints/blob/main/README.md
Response SHA-256: d9e75df9554411068c3e7c6e824e39a9189e6e77738f452d3dec7cbec7c7a9f2
Scoped evidence bindings: 3

[5] complexconstraints-paper. Surge AI; 2606.09118v3
https://arxiv.org/pdf/2606.09118v3
Response SHA-256: cc10bf23f29b145de1b510c2209f2e793d6ec5dfa78953f700556b10c82636bf
Scoped evidence bindings: 13

[6] corecraft-paper. Surge AI; 2602.16179v5
https://arxiv.org/pdf/2602.16179v5
Response SHA-256: 17c8dea1085ca21673f859bc8da9d5a9417302d1338a38b542ec0714fa1f75f8
Scoped evidence bindings: 11

[7] gdp-pdf-creator-readme. Surge AI; git-blob:567b4708e61dfbdca9a740f4ecc5dec55d4ec528
https://github.com/surge-ai/gdp-pdf/blob/main/README.md
Response SHA-256: fda47ebf36bf4299b5d69288978de80e06fa6a001b8bc75550eb1f38d8e9e148
Scoped evidence bindings: 2

[8] gdp-pdf-paper. Surge AI; 2607.11192v3
https://arxiv.org/pdf/2607.11192v3
Response SHA-256: e139828de98e24caee6ba7f749dd39e9c541ff16c7a6c91ddb18dcd5a8defda4
Scoped evidence bindings: 15

[9] handbook-creator-readme. Surge AI; git-blob:34ddc11c8c64f4db7cd1c4ad62c33427993d5940
https://github.com/surge-ai/handbook/blob/main/README.md
Response SHA-256: dc622b23cd777063020645781f58eca79f9e411663a05aa5a97571cc477155e5
Scoped evidence bindings: 2

[10] handbook-paper. Surge AI; 2607.25398v3
https://arxiv.org/pdf/2607.25398v3
Response SHA-256: a3fa318ebbfa6de9a7f3e8ec0c25883e52426dac7d8de6c9735b9dd42214fc12
Scoped evidence bindings: 17

[11] nexdata-official-catalog. Nexdata (catalog publisher; upstream creator not established);
body-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65
https://www.nexdata.ai/
Response SHA-256: 4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65
Scoped evidence bindings: 1

[12] posttrain-arena-creator-readme. BenchFlow / PostTrain Arena organizers;
git-blob:ecd07be0ee6bd8358865d53309979495b4b63a73
https://github.com/benchflow-ai/posttrainarena/blob/main/README.md
Response SHA-256: 6705d0a04541b92026089459bafae0ddf88c045327c489508d23ce6c45e55980
Scoped evidence bindings: 2

[13] posttrain-spec. BenchFlow / PostTrain Arena organizers;
body-sha256:1a22b683ee06ffc0b1f65c81ccc5f2d7095ad33960f1c728552dc8bc3a39e9fc
https://posttrain.com/docs/spec
Response SHA-256: 1a22b683ee06ffc0b1f65c81ccc5f2d7095ad33960f1c728552dc8bc3a39e9fc
Scoped evidence bindings: 10

[14] product1463-official-product. Nexdata (catalog publisher; upstream creator not established);
body-sha256:e3d60a197240029b0055ddbb6630342bd8502c69df019db6722e4af519c7a1c4
https://www.nexdata.ai/datasets/embodied-ai/1463
Response SHA-256: e3d60a197240029b0055ddbb6630342bd8502c69df019db6722e4af519c7a1c4
Scoped evidence bindings: 3

[15] product1515-official-product. Nexdata (catalog publisher; upstream creator not established);
body-sha256:6e8b506f440911fea8fbab1cc493ebbceafb9d851182391a9f0765b70ffa940c
https://www.nexdata.ai/datasets/llm/1515
Response SHA-256: 6e8b506f440911fea8fbab1cc493ebbceafb9d851182391a9f0765b70ffa940c
Scoped evidence bindings: 3

[16] product1814-official-product. Nexdata (catalog publisher; upstream creator not established); body-sha256:9ba68417e7e6dcfab54d1d7a2ac88704768d7c1ca9c58c70e36facce9ee5cab1
https://www.nexdata.ai/datasets/llm/1814
Response SHA-256: 9ba68417e7e6dcfab54d1d7a2ac88704768d7c1ca9c58c70e36facce9ee5cab1
Scoped evidence bindings: 3

[17] product2121-official-product. Nexdata (catalog publisher; upstream creator not established); body-sha256:6635303d177b5e96d948d3b4d1ed6b89ac8cd48b54e3373b47641c8ef449a21d
https://www.nexdata.ai/datasets/llm/2121
Response SHA-256: 6635303d177b5e96d948d3b4d1ed6b89ac8cd48b54e3373b47641c8ef449a21d
Scoped evidence bindings: 3

[18] product2145-official-product. Nexdata (catalog publisher; upstream creator not established); body-sha256:9946cc44dcccf95642be79d6bac8e0f25a0dacad80df305df26764738ef82674
https://www.nexdata.ai/datasets/embodied-ai/2145
Response SHA-256: 9946cc44dcccf95642be79d6bac8e0f25a0dacad80df305df26764738ef82674
Scoped evidence bindings: 4

[19] product2185-official-product. Nexdata (catalog publisher; upstream creator not established); body-sha256:83b4b7a89d08228482851d704278f6c40fa0e94a00e119afab8b2a8a799e6e87
https://www.nexdata.ai/datasets/llm/2185
Response SHA-256: 83b4b7a89d08228482851d704278f6c40fa0e94a00e119afab8b2a8a799e6e87
Scoped evidence bindings: 3

[20] product2236-official-product. Nexdata (catalog publisher; upstream creator not established); body-sha256:b752361cdb55f340d23ea851fd24c07534d6cc6e3449fa3797d483fe87c00d51
https://www.nexdata.ai/datasets/embodied-ai/2236
Response SHA-256: b752361cdb55f340d23ea851fd24c07534d6cc6e3449fa3797d483fe87c00d51
Scoped evidence bindings: 3

[21] riemann-paper. Surge AI; 2604.06802v3
https://arxiv.org/pdf/2604.06802v3
Response SHA-256: 621d8428e724fcbbf292ff334b71e04148820d834b5f483b7a980cd029cbe93b
Scoped evidence bindings: 12

[22] surge-antidote. Surge AI; body-sha256:561c2781c4deeb6c32064c90be2bd12c5c57b7cc37a3253c2a980dbac58c9f3f
https://surgehq.ai/blog/introducing-antidote
Response SHA-256: 561c2781c4deeb6c32064c90be2bd12c5c57b7cc37a3253c2a980dbac58c9f3f
Scoped evidence bindings: 5

[23] surge-complexconstraints. Surge AI; body-sha256:20f4b5b6606d9ec14b60871f380e14cf76460f188e343136d63651eae63c1a23
https://surgehq.ai/blog/complexconstraints-a-benchmark-for-entangled-instruction-following
Response SHA-256: 20f4b5b6606d9ec14b60871f380e14cf76460f188e343136d63651eae63c1a23
Scoped evidence bindings: 1

[24] surge-enterprisebench-corecraft. Surge AI; body-sha256:06e3008dcf989c0d54aa90dda31ca01fc15d476f86fbb0415edb8d1a924c140e
https://surgehq.ai/blog/enterprisebench-corecraft
Response SHA-256: 06e3008dcf989c0d54aa90dda31ca01fc15d476f86fbb0415edb8d1a924c140e
Scoped evidence bindings: 2

[25] surge-gdp-pdf. Surge AI; body-sha256:2c518677115b1fe3e4c422353e1c36c46ff52683a38a1590ebab80d55c6ba141
https://surgehq.ai/blog/gdp-pdf-can-100b-ai-models-master-the-documents-that-run-the-world
Response SHA-256: 2c518677115b1fe3e4c422353e1c36c46ff52683a38a1590ebab80d55c6ba141
Scoped evidence bindings: 1

[26] surge-hemingway. Surge AI;
body-sha256:a8c6db13ea98f5cbfa538d6d3f0c8654397dcc2259981308108568ea14b24fee
https://surgehq.ai/blog/hemingway-bench-ai-writing-leaderboard
Response SHA-256: a8c6db13ea98f5cbfa538d6d3f0c8654397dcc2259981308108568ea14b24fee
Scoped evidence bindings: 3

[27] surge-riemann. Surge AI; body-sha256:4698f9d2b48e2ee48d3d58605a94affbaf0c8a95b0339c18ae8fe52cacd18cf2
https://surgehq.ai/blog/riemann-bench-a-benchmark-for-moonshot-mathematics
Response SHA-256: 4698f9d2b48e2ee48d3d58605a94affbaf0c8a95b0339c18ae8fe52cacd18cf2
Scoped evidence bindings: 2

[28] surge-sudo-l7. Surge AI; body-sha256:da59b2f17377c3ea549ea207653e0b368784a1ec82053be7467053663c1dfab9
https://surgehq.ai/blog/sudo-l7
Response SHA-256: da59b2f17377c3ea549ea207653e0b368784a1ec82053be7467053663c1dfab9
Scoped evidence bindings: 5

[29] AlpacaEval README.md. tatsu-lab/alpaca_eval; cd543a149df89434d8a54582c0151c0b945c3d20
https://raw.githubusercontent.com/tatsu-lab/alpaca_eval/cd543a149df89434d8a54582c0151c0b945c3d20/README.md
Response SHA-256: 27576663fb6f13adf5b3adea8d7840f47ffe366fca74177fe6aac92e8b6ff11c
Scoped evidence bindings: 6

[30] BeaverTails README.md. PKU-Alignment/BeaverTails; 8401fe609d288129cc684a9b3be6a93e41cfe678
https://huggingface.co/datasets/PKU-Alignment/BeaverTails/raw/8401fe609d288129cc684a9b3be6a93e41cfe678/README.md
Response SHA-256: 92bce76909c3abb6e7ca2bbb1240f6b03da0a531cde8c03b3bbf7cb20c9a7bef
Scoped evidence bindings: 6

[31] GPQA LICENSE. Idavidrein/gpqa; 56686c06f5e19865c153de0fdb11be3890014df7
https://raw.githubusercontent.com/Idavidrein/gpqa/56686c06f5e19865c153de0fdb11be3890014df7/LICENSE
Response SHA-256: 09c64e279464e54aa88696bcb73dc77722f905a028313c48a17e16572db00d8a
Scoped evidence bindings: 1

[32] GPQA arXiv HTML. GPQA paper authors / arXiv; 2311.12022v1
https://arxiv.org/html/2311.12022v1
Response SHA-256: 6c4a6426771956afd4b61176bfaecce0d773f4c1fa8dc76e87de33b2081905a9
Scoped evidence bindings: 3

[33] GPQA README.md. Idavidrein/gpqa; 56686c06f5e19865c153de0fdb11be3890014df7
https://raw.githubusercontent.com/Idavidrein/gpqa/56686c06f5e19865c153de0fdb11be3890014df7/README.md
Response SHA-256: 0be2cf410fecf82fb811f02e2314ecc5a19decc1515014498e59bae4680d4d81
Scoped evidence bindings: 2

[34] GSM8K LICENSE. openai/grade-school-math; 3101c7d5072418e28b9008a6636bde82a006892c
https://raw.githubusercontent.com/openai/grade-school-math/3101c7d5072418e28b9008a6636bde82a006892c/LICENSE
Response SHA-256: 86bbb73e855821d7c401912fd4bf82e34313e6e3b6fd6f909f2b6cc9e209a53b
Scoped evidence bindings: 1

[35] GSM8K README.md. openai/grade-school-math; 3101c7d5072418e28b9008a6636bde82a006892c
https://raw.githubusercontent.com/openai/grade-school-math/3101c7d5072418e28b9008a6636bde82a006892c/README.md
Response SHA-256: 27e1676d572f2289210170ff6dbb8c2ef7d9f6950c0f08c978ea8ee6b39dca29
Scoped evidence bindings: 4

[36] Humanity’s Last Exam hle_eval/run_judge_results.py. centerforaisafety/hle; 22ed3074b1e7b134bcbc09028d0ba320839b0655
https://raw.githubusercontent.com/centerforaisafety/hle/22ed3074b1e7b134bcbc09028d0ba320839b0655/hle_eval/run_judge_results.py
Response SHA-256: a796e74d6e58beab3bd88f46182f599443a102308b2d1c7e0b4b22f578468446
Scoped evidence bindings: 2

[37] Humanity’s Last Exam LICENSE. centerforaisafety/hle; 22ed3074b1e7b134bcbc09028d0ba320839b0655
https://raw.githubusercontent.com/centerforaisafety/hle/22ed3074b1e7b134bcbc09028d0ba320839b0655/LICENSE
Response SHA-256: f402878026b8008bd53f2f038b5c9e3e6b7c1a690e05e2ca6ddbc9019aa3c08e
Scoped evidence bindings: 1

[38] Humanity’s Last Exam README.md. centerforaisafety/hle; 22ed3074b1e7b134bcbc09028d0ba320839b0655
https://raw.githubusercontent.com/centerforaisafety/hle/22ed3074b1e7b134bcbc09028d0ba320839b0655/README.md
Response SHA-256: fd3bd0049e1a2b39244d8240a63d0ccfe037b0b6e5979d5826e1d1cf4203ee88
Scoped evidence bindings: 4

[39] MT-Bench LICENSE. lm-sys/FastChat; 587d5cfa1609a43d192cedb8441cac3c17db105d
https://raw.githubusercontent.com/lm-sys/FastChat/587d5cfa1609a43d192cedb8441cac3c17db105d/LICENSE
Response SHA-256: c71d239df91726fc519c6eb72d318ec65820627232b2f796219e87dcf35d0ab4
Scoped evidence bindings: 1

[40] MT-Bench fastchat/llm_judge/README.md. lm-sys/FastChat; 587d5cfa1609a43d192cedb8441cac3c17db105d
https://raw.githubusercontent.com/lm-sys/FastChat/587d5cfa1609a43d192cedb8441cac3c17db105d/fastchat/llm_judge/README.md
Response SHA-256: 4104f9e877fe04a4a87a5b6618114adc8604f09ceca09be198f96189c4c2238e
Scoped evidence bindings: 5

[41] Nectar README.md. berkeley-nest/Nectar; 3c6b4c47fa1cc38869f9f32dce1699f7abad8b06
https://huggingface.co/datasets/berkeley-nest/Nectar/raw/3c6b4c47fa1cc38869f9f32dce1699f7abad8b06/README.md
Response SHA-256: dfd3ff7efc6cf425eb9846a5f49ad4d3997655b45ee0f0b147a7afa59c3a1d9a
Scoped evidence bindings: 8

[42] OpenHermes README.md. teknium/OpenHermes-2.5; b82037821055c377bed0d495e72e46de3bc72e84
https://huggingface.co/datasets/teknium/OpenHermes-2.5/raw/b82037821055c377bed0d495e72e46de3bc72e84/README.md
Response SHA-256: 92fd711b9f0f6f8c7312d18d93ad412af415cd25a207584554c8e72c0cfe9f64
Scoped evidence bindings: 5

[43] PRM800K LICENSE. openai/prm800k; 7ecc794703b2877f63226f2477a49b34f9b25163
https://raw.githubusercontent.com/openai/prm800k/7ecc794703b2877f63226f2477a49b34f9b25163/LICENSE
Response SHA-256: 22d61a2226e035da51c56149c3358c36d1f61716435866bc6e6fe0e54c7270db
Scoped evidence bindings: 1

[44] PRM800K README.md. openai/prm800k; 7ecc794703b2877f63226f2477a49b34f9b25163
https://raw.githubusercontent.com/openai/prm800k/7ecc794703b2877f63226f2477a49b34f9b25163/README.md
Response SHA-256: a721d8d35cd77c128352acb3378dd32c223a41b5d0f6ca2e8b85fb4bfe88b5fe
Scoped evidence bindings: 5

[45] PKU-SafeRLHF README.md. PKU-Alignment/PKU-SafeRLHF; 9421ffafec3fa40a1f1a7d567b4d525079477ecb
https://huggingface.co/datasets/PKU-Alignment/PKU-SafeRLHF/raw/9421ffafec3fa40a1f1a7d567b4d525079477ecb/README.md
Response SHA-256: 77ce394a7406b94e7ebe04b63e25850733286369db2d038936a6114502923579
Scoped evidence bindings: 7

[46] Tülu 3 SFT Mixture README.md. allenai/tulu-3-sft-mixture; b14afda60f1bbebe55d5d2fa1e4df5042f97f8be
https://huggingface.co/datasets/allenai/tulu-3-sft-mixture/raw/b14afda60f1bbebe55d5d2fa1e4df5042f97f8be/README.md
Response SHA-256: 1b6fff37d3f206b70086d9d65015a75dfb98b603a99af7330b0d44cd198464d6
Scoped evidence bindings: 9

[47] UltraInteract README.md. openbmb/UltraInteract_pair; 3692843c72c3a297c7914cdf76c0ffce5e38a234
https://huggingface.co/datasets/openbmb/UltraInteract_pair/raw/3692843c72c3a297c7914cdf76c0ffce5e38a234/README.md
Response SHA-256: 016a683bb4e42f14a8a96963ed541d6be23710a6e9c7c5c0f0bd4bcf7038ee91
Scoped evidence bindings: 7

[48] AgentBench: Evaluating LLMs as Agents. arXiv (author-deposited preprint); 2308.03688v1
https://arxiv.org/pdf/2308.03688v1
Response SHA-256: 5cd8a5ddbbb95a8d156bc167ee640e4338ac56ec2de574c08276cb8bad27eb4e
Scoped evidence bindings: 7

[49] A Global AgTech Company + CloudFactory Partnership. CloudFactory; Exact response sha256:16b1a273e45d983cde6502f101f1af08575d909c089773b94a0db9eab7bae524
https://www.cloudfactory.com/client-stories/global-agtech-company
Response SHA-256: 16b1a273e45d983cde6502f101f1af08575d909c089773b94a0db9eab7bae524
Scoped evidence bindings: 3

[50] Data Security | CloudFactory. CloudFactory; Exact response sha256:9f1252fa9e8074fe3a58cd2dae995b32e212f2ed55eb5f3cbabd56aa1ff09a93
https://www.cloudfactory.com/data-security
Response SHA-256: 9f1252fa9e8074fe3a58cd2dae995b32e212f2ed55eb5f3cbabd56aa1ff09a93
Scoped evidence bindings: 2

[51] Constitutional AI: Harmlessness from AI Feedback. arXiv (author-deposited preprint); 2212.08073v1
https://arxiv.org/pdf/2212.08073v1
Response SHA-256: 9a456a07ad346e3372f9867d346f69f5b0f68b4c65f060aca0b8a13fa9d98e83
Scoped evidence bindings: 5

[52] Empowering Safer AI with IBM | DataForce. DataForce by TransPerfect; Exact response sha256:2de55055344325226fc7e18496f00f74cc0bc83e9af18928370f2da0dba7f6c1
https://www.dataforce.ai/success-stories/empowering-safer-ai-ibm
Response SHA-256: 2de55055344325226fc7e18496f00f74cc0bc83e9af18928370f2da0dba7f6c1
Scoped evidence bindings: 2

[53] Generative AI Training | DataForce. DataForce by TransPerfect; Exact response sha256:3dd24bbc3853acc05871aa870eb98bf4600eeea279bffb191b295c2c86484dd9
https://www.dataforce.ai/services/generative-ai-training
Response SHA-256: 3dd24bbc3853acc05871aa870eb98bf4600eeea279bffb191b295c2c86484dd9
Scoped evidence bindings: 1

[54] AI Home | DataForce. DataForce by TransPerfect; Exact response sha256:45bce3c67c97c278e6a3c7a2479905f354b2c4fb062417c37e1707984a8933ae
https://www.dataforce.ai/
Response SHA-256: 45bce3c67c97c278e6a3c7a2479905f354b2c4fb062417c37e1707984a8933ae
Scoped evidence bindings: 3

[55] DataForce - Data Privacy Notice for Vendors | DataForce. DataForce by TransPerfect; Exact response sha256:562ebd06e7967fd3ffb2d120f642d7577d4f9eeb7747eb4d44f14ad69d5f7646
https://www.dataforce.ai/data-privacy-notice-vendors
Response SHA-256: 562ebd06e7967fd3ffb2d120f642d7577d4f9eeb7747eb4d44f14ad69d5f7646
Scoped evidence bindings: 1

[56] Data Collection Services for AI Training | Defined.ai. Defined.ai; Exact response
sha256:54e3d276bb3b02dbc277951ebe12a659ad6363501aa3e502ddc96682e94ca82d
https://defined.ai/solutions/data-collection
Response SHA-256: 54e3d276bb3b02dbc277951ebe12a659ad6363501aa3e502ddc96682e94ca82d
Scoped evidence bindings: 1

[57] Sensor Data Collection for Scalable Robotics | Defined.ai. Defined.ai; Exact response
sha256:f545e30e2f798bcdd8ae1f2f4d6d938f22bfcd422a5ee314d5a3b49be79f7e52
https://defined.ai/case-study/sensor-data-collection-robotics
Response SHA-256: f545e30e2f798bcdd8ae1f2f4d6d938f22bfcd422a5ee314d5a3b49be79f7e52
Scoped evidence bindings: 2

[58] Ethical AI: Consent, Provenance & Fair Data | Defined.ai. Defined.ai; Exact response
sha256:80aefeebc5a8249da3650cc76788eaa59d02ce96b2c9b657264cff213c5e994f
https://defined.ai/about-us/ethical-ai
Response SHA-256: 80aefeebc5a8249da3650cc76788eaa59d02ce96b2c9b657264cff213c5e994f
Scoped evidence bindings: 2

[59] Supplier Program - General Terms & Conditions | Defined.ai. Defined.ai; Exact response
sha256:ed1d419012a62a80ec3acedf945cdc49e5f8cce77db18623349dd37148418ccc
https://defined.ai/supplier-program
Response SHA-256: ed1d419012a62a80ec3acedf945cdc49e5f8cce77db18623349dd37148418ccc
Scoped evidence bindings: 1

[60] Deep Reinforcement Learning from Human Preferences. arXiv (author-deposited preprint); 1706.03741v1
https://arxiv.org/pdf/1706.03741v1
Response SHA-256: 10d47f151dcc7cb891899e14b8c4bf9e076cb120fbfc95a178cd631b25462f31
Scoped evidence bindings: 5

[61] iMerit: Improving Model Output with RLHF - Case Study. iMerit; Exact response
sha256:47995b478f4083f6615ca09fc079f72ba00ca724a1c96fd964830d672b31f010
https://7583693.fs1.hubspotusercontent-na2.net/hubfs/7583693/Case-Studies/iMerit_Improving%20Model%20Output%20with%20RLHF_Case%20Study.pdf
Response SHA-256: 47995b478f4083f6615ca09fc079f72ba00ca724a1c96fd964830d672b31f010
Scoped evidence bindings: 2

[62] RLHF Services for Expert AI Model Alignment | iMerit. iMerit; Exact response
sha256:5320c7d847e36828809a1eec081c5b3d7c9bac60d59c8f1478ab0990a15c7e20
https://imerit.ai/solutions/generative-ai-data-solutions/rlhf-services/
Response SHA-256: 5320c7d847e36828809a1eec081c5b3d7c9bac60d59c8f1478ab0990a15c7e20
Scoped evidence bindings: 2

[63] How iMerit Ensures Data Security | iMerit. iMerit; Exact response
sha256:a93411cbf2a750204aa771e88ae778cef4c54cc40d5b57ccc4f6498a12a9c692
https://imerit.ai/insights/how-imerit-ensures-data-security/
Response SHA-256: a93411cbf2a750204aa771e88ae778cef4c54cc40d5b57ccc4f6498a12a9c692
Scoped evidence bindings: 1

[64] Model-Ready AI Training Datasets | Invisible Technologies. Invisible Technologies Inc.; Exact response
sha256:e8ea411895933ecb5d491ff8048f35c08d1fb22a60ed1b70a2c1cabffca6b903
https://invisibletech.ai/data-lab/data
Response SHA-256: e8ea411895933ecb5d491ff8048f35c08d1fb22a60ed1b70a2c1cabffca6b903
Scoped evidence bindings: 1

[65] Launching a Leading LLM 6 Months Faster with Invisible AI. Invisible Technologies Inc.; Exact response sha256:c5eb8ca481d14643c9d75c9d8f927b42a86dd2a3bc7ba13d11cf779c9ddce477
https://www.invisible.co/case-studies/launching-a-leading-llm-6-months-faster
Response SHA-256: c5eb8ca481d14643c9d75c9d8f927b42a86dd2a3bc7ba13d11cf779c9ddce477
Scoped evidence bindings: 2

[66] Human Experts for AI Training | Invisible Technologies. Invisible Technologies Inc.; Exact response sha256:2fd7d13864ae3f26327bb534ca19b62468fd3dffea9227bded3db3bcc25646b8
https://invisibletech.ai/data-lab/experts
Response SHA-256: 2fd7d13864ae3f26327bb534ca19b62468fd3dffea9227bded3db3bcc25646b8
Scoped evidence bindings: 1

[67] Privacy policy | Invisible Technologies. Invisible Technologies Inc.; Exact response sha256:ab2a19999e1d1e9aa87963299ee34b67ebc82bb872883ae4e42e4b4211c08378
https://invisibletech.ai/privacy-policy
Response SHA-256: ab2a19999e1d1e9aa87963299ee34b67ebc82bb872883ae4e42e4b4211c08378
Scoped evidence bindings: 1

[68] A General Theoretical Paradigm to Understand Learning from Human Preferences. arXiv (author-deposited preprint); 2310.12036v1
https://arxiv.org/pdf/2310.12036v1
Response SHA-256: 01461de3b0bcb07b069bb7f233785ad719cacb71cfad842867cc79099e02f81a
Scoped evidence bindings: 5

[69] KTO: Model Alignment as Prospect Theoretic Optimization. arXiv (author-deposited preprint); 2402.01306v1
https://arxiv.org/pdf/2402.01306v1
Response SHA-256: 64214b766a9f59bec5fb11cd65412648114a12d32f942e416c694791d9bc2293
Scoped evidence bindings: 5

[70] Alignerr Expert Network - Labelbox. Labelbox, Inc.; Exact response sha256:c64a07a0f7e2ab31d2d21f06868043b0f7d77b1f2c6ff6b0146ea8742dc00145
https://labelbox.com/products/alignerr/
Response SHA-256: c64a07a0f7e2ab31d2d21f06868043b0f7d77b1f2c6ff6b0146ea8742dc00145
Scoped evidence bindings: 3

[71] How a leading text-to-speech AI lab harnesses human-generated data to deliver generative AI faster. Labelbox, Inc.; Exact response sha256:c4c5831c89b4eda8fab3a9d6908a96bbfe9c8058a5508c3e59f00949bb403a24
https://labelbox.com/customers/text-to-speech-app-customer-story/
Response SHA-256: c4c5831c89b4eda8fab3a9d6908a96bbfe9c8058a5508c3e59f00949bb403a24
Scoped evidence bindings: 1

[72] Labelbox raises $100M+ Series E. Labelbox, Inc.; Exact response sha256:fe5502ed3564c6dd9eed2e57ef0113c006442a3a92e9438d0e4206b8d5df143a; 2026-10-14
https://labelbox.com/blog/labelbox-series-e/
Response SHA-256: fe5502ed3564c6dd9eed2e57ef0113c006442a3a92e9438d0e4206b8d5df143a
Scoped evidence bindings: 1

[73] Enhancing foundational Large Language Models for a global technology leader | High-Quality AI Data to Power Innovation | LXT. LXT; Exact response sha256:e500e7e30ef8b189268c456e82d50a6f54d90f32f7d6faf60d463ea3a6464149
https://www.lxt.ai/case-studies/enhancing-foundational-large-language-models-for-a-global-technology-leader/
Response SHA-256: e500e7e30ef8b189268c456e82d50a6f54d90f32f7d6faf60d463ea3a6464149
Scoped evidence bindings: 2

[74] Privacy Policy: High-Quality AI Data | LXT. LXT; Exact response
sha256:d13d553b535e4d30ee2b0f57387aa7572951b6997a8bcb9941de66385d84843d
https://www.lxt.ai/privacy-policy
Response SHA-256: d13d553b535e4d30ee2b0f57387aa7572951b6997a8bcb9941de66385d84843d
Scoped evidence bindings: 1

[75] RLHF Services for AI: Scalable Human Feedback for Alignment. LXT; Exact response
sha256:0e7273806c5265e4042f162d4e15c2ad2d532ea29f351a3da19b2752f1ec0fff
https://www.lxt.ai/services/rlhf-services/
Response SHA-256: 0e7273806c5265e4042f162d4e15c2ad2d532ea29f351a3da19b2752f1ec0fff
Scoped evidence bindings: 3

[76] Reference-free Monolithic Preference Optimization with Odds Ratio. arXiv (author-deposited preprint); 2403.07691v1
https://arxiv.org/pdf/2403.07691v1
Response SHA-256: 595b1bbc563473ba171744d3b7c242e9ad65f5930d654aa49d3d3b31228b52d4
Scoped evidence bindings: 5

[77] Proximal Policy Optimization Algorithms. arXiv (author-deposited preprint); 1707.06347v1
https://arxiv.org/pdf/1707.06347v1
Response SHA-256: c4f1b8b8b2e2a976c2cae67c05a347dac44ae6e63eebed9a66bfb0241bd45c93
Scoped evidence bindings: 5

[78] Let's Verify Step by Step. arXiv (author-deposited preprint); 2305.20050v1
https://arxiv.org/pdf/2305.20050v1
Response SHA-256: fbd170e2042c32950c3fe97d3a558d89e8a8dffaadc942e774ccb4b751abc123
Scoped evidence bindings: 6

[79] RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. arXiv (author-deposited preprint); 2309.00267v1
https://arxiv.org/pdf/2309.00267v1
Response SHA-256: afac8f42dd040e213e0af9a6efd6a5608477c9b9270dbdfb884122aaf9b88207
Scoped evidence bindings: 5

[80] How a Fortune 500 Tech Leader Accelerated Gen AI Model Performance in the U.S. with Sama | Sama. Samasource Impact Sourcing, Inc.; Exact response
sha256:90ba08fd22c2368bb2815001a1d533064e6dc1ed27fc1d8f3eb8558659de6365
https://www.sama.com/case-studies/how-a-fortune-500-tech-leader-accelerated-gen-ai-model-performance-with-sama
Response SHA-256: 90ba08fd22c2368bb2815001a1d533064e6dc1ed27fc1d8f3eb8558659de6365
Scoped evidence bindings: 2

[81] Data Annotation & Labeling Company | Sama. Samasource Impact Sourcing, Inc.; Exact response
sha256:f93112f4673178fea7d3710123a40e409bdc41e056757759705fabfb2b7580a5
https://www.sama.com/
Response SHA-256: f93112f4673178fea7d3710123a40e409bdc41e056757759705fabfb2b7580a5
Scoped evidence bindings: 2

[82] Sama Delivers High-quality & Secure Annotated Data | Sama. Samasource Impact Sourcing, Inc.; Exact response
sha256:73c6dfb759025c07dfcbfcee8f39244695027e5acf8d659a229db6cb1085367c
https://www.sama.com/data-security-and-trust
Response SHA-256: 73c6dfb759025c07dfcbfcee8f39244695027e5acf8d659a229db6cb1085367c
Scoped evidence bindings: 2

[83] SELF-INSTRUCT: Aligning Language Model with Self Generated Instructions. arXiv (author-deposited preprint); 2212.10560v1
https://arxiv.org/pdf/2212.10560v1
Response SHA-256: 08dba30c0ff241ac17b2f7ac31d4202faca0f847d77bc215f9e03dc905da8553
Scoped evidence bindings: 6

[84] Databricks optimizes model evaluation with SuperAnnotate's expertise. SuperAnnotate; Exact response sha256:928302955a172e1d905bf62d2e375ae2824cc413c70b3c03f2796ef201c5fd5a
https://www.superannotate.com/blog/databricks-case-study
Response SHA-256: 928302955a172e1d905bf62d2e375ae2824cc413c70b3c03f2796ef201c5fd5a
Scoped evidence bindings: 4

[85] Trust Center | SuperAnnotate. SuperAnnotate; Exact response sha256:447b8c199555677d168ff1e84c534d8c45bb2c862c054c1cc5c7c1b7aa376c8b
https://www.superannotate.com/trust-center
Response SHA-256: 447b8c199555677d168ff1e84c534d8c45bb2c862c054c1cc5c7c1b7aa376c8b
Scoped evidence bindings: 2

[86] Global AI Policy Guidelines for TaskUs Employees. TaskUs, Inc.; Exact response sha256:63b2e4cb771bea3088f13e6690997afbc173c68cd51cacfc3571fa7a30655108
https://www.taskus.com/global-ai-policy/
Response SHA-256: 63b2e4cb771bea3088f13e6690997afbc173c68cd51cacfc3571fa7a30655108
Scoped evidence bindings: 2

[87] Smarter Data Review for Safer LLMs. TaskUs, Inc.; Exact response sha256:8c2211ab586fccb3a07ac60deea0fc1303cd71f007d23bb4e0961f754bce75ab
https://www.taskus.com/insights/smarter-data-review-for-safer-llms/
Response SHA-256: 8c2211ab586fccb3a07ac60deea0fc1303cd71f007d23bb4e0961f754bce75ab
Scoped evidence bindings: 1

[88] Data Feedback and Evaluations | TaskUs. TaskUs, Inc.; Exact response sha256:b45593e9e4cacaef67c1d49f3d6e072676e31f181537298eda16bac1ef634ac8
https://www.taskus.com/services/ai-data-services/data-feedback-evaluations/
Response SHA-256: b45593e9e4cacaef67c1d49f3d6e072676e31f181537298eda16bac1ef634ac8
Scoped evidence bindings: 2

[89] TELUS Digital AI Data Solutions: High-quality data and human intelligence to test, train and improve AI models. TELUS Digital; Exact response sha256:8e3cde27bf4a0b8b8870af7439d576929aaaedd0dc076e1f7374305aec284243
https://assets.ctfassets.net/3viuren4us1n/7Is3KTyGYuK6wG4BYlfa9C/08840fef56e4eb58b47562b1106dbb62/TELUS_Digital_AI_Data_Solutions_Brochure_2025.pdf
Response SHA-256: 8e3cde27bf4a0b8b8870af7439d576929aaaedd0dc076e1f7374305aec284243
Scoped evidence bindings: 6

[90] Training Verifiers to Solve Math Word Problems. arXiv (author-deposited preprint); 2110.14168v1
https://arxiv.org/pdf/2110.14168v1
Response SHA-256: 9d829ca79818d2d8ddc9c9809348718856789a0f52c143165c8fa5a065948774
Scoped evidence bindings: 6

[91] WebArena: A Realistic Web Environment for Building Autonomous Agents. arXiv (author-deposited preprint); 2307.13854v1
https://arxiv.org/pdf/2307.13854v1
Response SHA-256: 10161cf3b90868aa5564b123103d0fed2fe8babf0a8f3a576d112877a0b7d141
Scoped evidence bindings: 6

[92] CASE STUDY: Accelerating LLM Development & Fine-Tuning - Welocalize. Welocalize / Welo Global; Exact response sha256:88aab62743399d78b8798281e5874427ad31326ec8f2cf1d9252fa7011839928; 2023-12-11
https://www.welocalize.com/insights/case-study-accelerating-llm-development-fine-tuning/
Response SHA-256: 88aab62743399d78b8798281e5874427ad31326ec8f2cf1d9252fa7011839928
Scoped evidence bindings: 3

[93] Information Security Policy | Welo Global. Welocalize / Welo Global; Exact response sha256:1a24478fee68339224094e31d22c6291e942f49a3643ff301eb5e7c242a4df61
https://www.weloglobal.com/information-security-policy/
Response SHA-256: 1a24478fee68339224094e31d22c6291e942f49a3643ff301eb5e7c242a4df61
Scoped evidence bindings: 1

[94] Enterprise AI Training Data & Human-in-the-Loop Evaluation | Welo Data. Welo Data; Exact response sha256:a83856615cb35badea714cfabffd4efbc42185f5a69764e806891ffff09eab80
https://welodata.ai/
Response SHA-256: a83856615cb35badea714cfabffd4efbc42185f5a69764e806891ffff09eab80
Scoped evidence bindings: 3

[95] Align-Anything creator README.md. PKU-Alignment/align-anything; eb9f03cbc9e15d2c6129034d30f77540be386acb
https://huggingface.co/datasets/PKU-Alignment/align-anything/resolve/eb9f03cbc9e15d2c6129034d30f77540be386acb/README.md
Response SHA-256: 35648aab416222f034d4cec17482c9762549256b68e30273edda168e228f93b9
Scoped evidence bindings: 6

[96] Capybara creator README.md. LDJnr/Capybara; c2bc39ac72f24748f60f5fb55b77e08fb0660ba6
https://huggingface.co/datasets/LDJnr/Capybara/resolve/c2bc39ac72f24748f60f5fb55b77e08fb0660ba6/README.md
Response SHA-256: 51459380b8e4f5074fe07d545cd278261443de669aa5c5daffcfabd107152370
Scoped evidence bindings: 6

[97] Chatbot Arena Conversations creator README.md. lmsys/chatbot_arena_conversations; 1b6335d42a1d2c7e34870c905d03ab964f7f2bd8
https://huggingface.co/datasets/lmsys/chatbot_arena_conversations/resolve/1b6335d42a1d2c7e34870c905d03ab964f7f2bd8/README.md
Response SHA-256: 53ef9f134991f1c75a81cd24c7b77a602e78d2bc01c6d9666cac53d4e101c50d
Scoped evidence bindings: 9

[98] Daring-Anteater creator README.md. nvidia/Daring-Anteater; ae79f8ac44cf185fbd3250dbe46057a7f7c4ec40
https://huggingface.co/datasets/nvidia/Daring-Anteater/resolve/ae79f8ac44cf185fbd3250dbe46057a7f7c4ec40/README.md
Response SHA-256: 2fd55bcb2d32a8fa19c3d6ab5f5ea5992c8cbb1cd5f9d511be39a07c2ec69fd3
Scoped evidence bindings: 5

[99] LMSYS-Chat-1M creator README.md. lmsys/lmsys-chat-1m; 200748d9d3cddcc9d782887541057aca0b18c5da
https://huggingface.co/datasets/lmsys/lmsys-chat-1m/resolve/200748d9d3cddcc9d782887541057aca0b18c5da/README.md
Response SHA-256: 02e1793a3f45ac262f56c5b092b3d16e0239ba55d92a76e0e979519bff1c45de
Scoped evidence bindings: 5

[100] Magpie creator README.md. magpie-align/magpie; b734a36818e4ba7ef7f0f582fc55f7e860407d26
https://raw.githubusercontent.com/magpie-align/magpie/b734a36818e4ba7ef7f0f582fc55f7e860407d26/README.md
Response SHA-256: d4f626c9fca26a262cb240b5f201ac7ba8be4d66b86ecea4d90a3a69e9a17903
Scoped evidence bindings: 2

[101] Magpie creator README.md. Magpie-Align/Magpie-Pro-300K-Filtered; 523df96eb7474e97bca6f378b3baa372a4735fcc
https://huggingface.co/datasets/Magpie-Align/Magpie-Pro-300K-Filtered/resolve/523df96eb7474e97bca6f378b3baa372a4735fcc/README.md
Response SHA-256: ee617bdb7e1c45f5980e98aadb6258bb295ceef54aa1aea7384152cbd6bbf6de
Scoped evidence bindings: 5

[102] NuminaMath creator README.md. AI-MO/NuminaMath-CoT; 9d8d210c9f6a36c8f3cd84045668c9b7800ef517
https://huggingface.co/datasets/AI-MO/NuminaMath-CoT/resolve/9d8d210c9f6a36c8f3cd84045668c9b7800ef517/README.md
Response SHA-256: 2e6fead04b0c3a7eb3d219267ca84d81f91d9b90830cfbb1fde73ddd7118a149
Scoped evidence bindings: 2

[103] NuminaMath creator report/numina_dataset.pdf. project-numina/aimo-progress-prize; 712e1abbcbcf870dccf4c6a3e7837e2967bd7d7f
https://raw.githubusercontent.com/project-numina/aimo-progress-prize/712e1abbcbcf870dccf4c6a3e7837e2967bd7d7f/report/numina_dataset.pdf
Response SHA-256: 613b93a38d0fdc24bc2bf4866a9b78a1eaab3855956d273253b461982054c4f6
Scoped evidence bindings: 4

[104] OpenMathInstruct creator README.md. nvidia/OpenMathInstruct-1; 4627efae6bd2ddcebb8acac00d513ffd8e00775c
https://huggingface.co/datasets/nvidia/OpenMathInstruct-1/resolve/4627efae6bd2ddcebb8acac00d513ffd8e00775c/README.md
Response SHA-256: 5839c59deabebeca12291f74706f2cfd67d762948b979177a88700dac5978e75
Scoped evidence bindings: 4

[105] OpenMathInstruct creator LICENSE. nvidia/OpenMathInstruct-1; 4627efae6bd2ddcebb8acac00d513ffd8e00775c
https://huggingface.co/datasets/nvidia/OpenMathInstruct-1/resolve/4627efae6bd2ddcebb8acac00d513ffd8e00775c/LICENSE
Response SHA-256: 117fbe0cc466811cac9488b009a7e5712bb5673c15cbc4f7ab343afcaaccde75
Scoped evidence bindings: 1

[106] OpenOrca creator README.md. Open-Orca/OpenOrca; e9c87b4abb2609913751f9b26553fdb9c061796c
https://huggingface.co/datasets/Open-Orca/OpenOrca/resolve/e9c87b4abb2609913751f9b26553fdb9c061796c/README.md
Response SHA-256: 119592215b487c63f92cdc46c85cdbf5eb1ecf4ce463396c54bec7bc403c96f9
Scoped evidence bindings: 5

[107] Pick-a-Pic creator paper.pdf. NeurIPS 2023 proceedings; NeurIPS2023-73aacd8b3b05b4b503d58310b523553c-Paper-Conference
https://papers.neurips.cc/paper_files/paper/2023/file/73aacd8b3b05b4b503d58310b523553c-Paper-Conference.pdf
Response SHA-256: b83f0f8b9bc43413a8a9d2a3e2d511a7ef13bd847e7d735f961f8caa79562b9a
Scoped evidence bindings: 4

[108] Pick-a-Pic creator README.md. yuvalkirstain/PickScore; f79607ff68609c1e6dfa78cba38afeea4c248df1
https://raw.githubusercontent.com/yuvalkirstain/PickScore/f79607ff68609c1e6dfa78cba38afeea4c248df1/README.md
Response SHA-256: d519113238e7c440ce623ef8732010ecc377e6e39ca7ee2cce519353d4a9effa
Scoped evidence bindings: 1

[109] Pick-a-Pic creator LICENSE. yuvalkirstain/PickScore; f79607ff68609c1e6dfa78cba38afeea4c248df1
https://raw.githubusercontent.com/yuvalkirstain/PickScore/f79607ff68609c1e6dfa78cba38afeea4c248df1/LICENSE
Response SHA-256: 26f64c803b07dfcea82ea31790b87fcf5bfc764d63513cfec6af817b62d6bebc
Scoped evidence bindings: 1

[110] RLAIF-V creator README.md. openbmb/RLAIF-V-Dataset; cdfc8c13778434e38afd538b0641ea942df4af78
https://huggingface.co/datasets/openbmb/RLAIF-V-Dataset/resolve/cdfc8c13778434e38afd538b0641ea942df4af78/README.md
Response SHA-256: 1c19c25807b2b68729210ef2ac6178ffb8ab11413afa5c0edb58100d8bc1c1cc
Scoped evidence bindings: 5

[111] RLHF-V creator README.md. openbmb/RLHF-V-Dataset; 1d8e9804b59e9da64ad7b1e17d505869ab9b2ad3
https://huggingface.co/datasets/openbmb/RLHF-V-Dataset/resolve/1d8e9804b59e9da64ad7b1e17d505869ab9b2ad3/README.md
Response SHA-256: 133d98d77de1b5f57691d10121f0f2fdfc81e5902b88f2d6e5c08614e4a095c6
Scoped evidence bindings: 5

[112] UltraChat creator README.md. stingning/ultrachat; f220fe796ce3ed62fbe1681b45ce6cbc9c6cabe0
https://huggingface.co/datasets/stingning/ultrachat/resolve/f220fe796ce3ed62fbe1681b45ce6cbc9c6cabe0/README.md
Response SHA-256: 1a8af8055f71acc69bac07fb06d0d2de2a1695e8cb590223f5d993c5c20db8e5
Scoped evidence bindings: 5

[113] VideoFeedback creator README.md. TIGER-Lab/VideoFeedback; 4557cb71595098e68790bb35253225c2fe797dcf
https://huggingface.co/datasets/TIGER-Lab/VideoFeedback/resolve/4557cb71595098e68790bb35253225c2fe797dcf/README.md
Response SHA-256: ae697b66418407ba534066f80252f7a9c9553e977755edfa5dfc80be4a9f96ed
Scoped evidence bindings: 4

[114] VideoFeedback creator README.md. TIGER-AI-Lab/VideoScore; f87faf4647e637066bcb74721670cd579e0f4349
https://raw.githubusercontent.com/TIGER-AI-Lab/VideoScore/f87faf4647e637066bcb74721670cd579e0f4349/README.md
Response SHA-256: 0e5765f6e0da01cc47bf8674218e07206da99ed40d39d644510e16ef067e3643
Scoped evidence bindings: 2

[115] VLFeedback creator README.md. MMInstruction/VLFeedback; 137dcea93e7667decc117c54b60f7401c6903310
https://huggingface.co/datasets/MMInstruction/VLFeedback/resolve/137dcea93e7667decc117c54b60f7401c6903310/README.md
Response SHA-256: 457def2f3e4ce7c209c5bfc02d6f78a92aaed7f72d3e5ca07538647b9d9382d6
Scoped evidence bindings: 4

[116] VLFeedback creator README.md. vlf-silkie/VLFeedback; de0bff35dbc6432ccfc214ab6bda61f42d79613f
https://raw.githubusercontent.com/vlf-silkie/VLFeedback/de0bff35dbc6432ccfc214ab6bda61f42d79613f/README.md
Response SHA-256: 1d75774e34d28ebd7ab94a86dcc357a27da565131a6a52cab5f86bdc1b4385ee
Scoped evidence bindings: 1

[117] WildChat creator README.md. allenai/WildChat-1M; 7d6490e462285cf85d91eabea0f9a954fbddcd1f
https://huggingface.co/datasets/allenai/WildChat-1M/resolve/7d6490e462285cf85d91eabea0f9a954fbddcd1f/README.md
Response SHA-256: 0bbf0494bdf60a3d8f644c155cb3c5f8fb21cb247045008f99bfc261e7e15162
Scoped evidence bindings: 4

[118] WildChat creator LICENSE.md. allenai/WildChat-1M; 7d6490e462285cf85d91eabea0f9a954fbddcd1f
https://huggingface.co/datasets/allenai/WildChat-1M/resolve/7d6490e462285cf85d91eabea0f9a954fbddcd1f/LICENSE.md
Response SHA-256: a7c7f6bdb20d261b8726c7594afb5832b36926e0cabfa116c41f58625bbc776c
Scoped evidence bindings: 1

[119] Data Annotation & Collection Services for your AI/ML projects - Alegion. Alegion; Exact response sha256:e932b217d29b716fa0ab1b4b5e81d03769868d9650dbd89b8d2a3ec5a6a29c41
https://www.alegion.com/
Response SHA-256: e932b217d29b716fa0ab1b4b5e81d03769868d9650dbd89b8d2a3ec5a6a29c41
Scoped evidence bindings: 3

[120] Better Annotation Services via our Approach, Process & Workforce - Alegion. Alegion; Exact response
sha256:76009ea75c4f25ccd6f8a842936b15653a34bb3c5b2c1e02b003530cc74066ae
https://alegion.com/annotation-services-approach-process/
Response SHA-256: 76009ea75c4f25ccd6f8a842936b15653a34bb3c5b2c1e02b003530cc74066ae
Scoped evidence bindings: 3

[121] AlpaGasus: Training A Better Alpaca with Fewer Data. arXiv (author-deposited preprint); 2307.08701v1
https://arxiv.org/pdf/2307.08701v1
Response SHA-256: 8ce75ae2f13ee4d18d3789f99627137d7c4f2f458c52989a44c079c5dc126157
Scoped evidence bindings: 5

[122] Aya Data: Your Trusted Global AI Data & Annotation Company. Aya Data; Exact response
sha256:57be9bb32191746c12a72c84bd377f6a62d3d23546721160b20dc661bbbe62e7
https://www.ayadata.ai/
Response SHA-256: 57be9bb32191746c12a72c84bd377f6a62d3d23546721160b20dc661bbbe62e7
Scoped evidence bindings: 1

[123] Trusted RLHF Services and Solution Providers - Aya Data. Aya Data; Exact response
sha256:8c7540874f8d4d2af6a3b65cd0d9f5012e1012465b965e122e71e127483a9f69
https://www.ayadata.ai/service/rlhf-services/
Response SHA-256: 8c7540874f8d4d2af6a3b65cd0d9f5012e1012465b965e122e71e127483a9f69
Scoped evidence bindings: 3

[124] BasicAI | AI Data Annotation Services & Platform. BasicAI; Exact response
sha256:0c73bde5150b43e4b51a3a70ab8b04768a3c4b00b0e2cb44b328d81f6d055609
https://www.basic.ai/
Response SHA-256: 0c73bde5150b43e4b51a3a70ab8b04768a3c4b00b0e2cb44b328d81f6d055609
Scoped evidence bindings: 2

[125] Large Language Models Data Solution | BasicAI. BasicAI; Exact response
sha256:9e7ddee701889f6faf99783c139db01683d1740766231a4088a6ef174ada1843
https://www.basic.ai/data-annotation-service/large-language-models-data-solution
Response SHA-256: 9e7ddee701889f6faf99783c139db01683d1740766231a4088a6ef174ada1843
Scoped evidence bindings: 2

[126] Data Processing and Security | BasicAI. BasicAI; Exact response
sha256:8656dd153cb409b101a68f344b93f3f15c6eb60aa474ea78da9f109ab120a97e
https://www.basic.ai/data-processing-and-security
Response SHA-256: 8656dd153cb409b101a68f344b93f3f15c6eb60aa474ea78da9f109ab120a97e
Scoped evidence bindings: 1

[127] “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. ACM CHI2021 paper, creator Google Research public PDF; doi.10.1145.3411764.3445518
https://storage.googleapis.com/gweb-research2023-media/pubtools/pdf/0d556e45afc54afeb2eb6b51a9bc1827b9961ff4.pdf
Response SHA-256: a449fda9f0673f071fe5e4de5770a29935ed39807d34514d431123cbff36ed3e
Scoped evidence bindings: 5

[128] OneForma | Centific. Centific; Exact response
sha256:2e5a9114d043f8b021103346d06cbc3b9c79ceeb9e0483a35b71d0b818edb027
https://www.centific.com/products/oneforma
Response SHA-256: 2e5a9114d043f8b021103346d06cbc3b9c79ceeb9e0483a35b71d0b818edb027
Scoped evidence bindings: 4

[129] RLHF & Preference Optimization | Centific. Centific; Exact response
sha256:1fce95030f645be3e5173a1fdf791275acc9f9faa1aae7408e696bba1edc36a3
https://www.centific.com/solution/rlhf-preference-optimization
Response SHA-256: 1fce95030f645be3e5173a1fdf791275acc9f9faa1aae7408e696bba1edc36a3
Scoped evidence bindings: 2

[130] DataComp: In search of the next generation of multimodal datasets. arXiv (author-deposited preprint); 2304.14108v1
https://arxiv.org/pdf/2304.14108v1
Response SHA-256: b001791ad499b7aed924de73226bfdcb366bf1c407ff1aa912f8cf78f08363ee
Scoped evidence bindings: 5

[131] DataComp-LM: In search of the next generation of training sets for language models. arXiv (author-deposited preprint); 2406.11794v1
https://arxiv.org/pdf/2406.11794v1
Response SHA-256: f19ea7fda1bf30fd9fe877ef6996c2dbc7f84538d813903b66a7bb7ff8f38bc0
Scoped evidence bindings: 8

[132] DICES Dataset: Diversity in Conversational AI Evaluation for Safety. arXiv (author-deposited preprint); 2306.11247v1
https://arxiv.org/pdf/2306.11247v1
Response SHA-256: 13bacc2af1b33bc79d2b5c0e888e17523885a4f9bfea38bb6ce94d57afb0be77
Scoped evidence bindings: 5

[133] Sources of Disagreement in Data for LLM Instruction Tuning. International Committee on Computational Linguistics / ACL Anthology; acl.2025.comedi-1.3
https://aclanthology.org/2025.comedi-1.3.pdf
Response SHA-256: 3dac1d018c58110a74f757800d4631f213e0871a361f1748c77373db297befe4
Scoped evidence bindings: 5

[134] Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. arXiv (author-deposited preprint); 2402.00159v1
https://arxiv.org/pdf/2402.00159v1
Response SHA-256: d4aa2060184ca64aa2ad81596b94c11ac555ff2420f2fc9f743d0e0ada4f3dbb
Scoped evidence bindings: 6

[135] Data Annotation | Encord. Encord; Exact response
sha256:4cd535bd3e778f6a421c5c9afa3f62bf3eda7fc70cbbe8e6747edc8227e0fd90
https://encord.com/data-annotation/
Response SHA-256: 4cd535bd3e778f6a421c5c9afa3f62bf3eda7fc70cbbe8e6747edc8227e0fd90
Scoped evidence bindings: 5

[136] Security at Encord | Encord. Encord; Exact response
sha256:26504e2a514903ce694a321e2afe71869794bbeb10137100ccf029598abf596a
https://encord.com/security/
Response SHA-256: 26504e2a514903ce694a321e2afe71869794bbeb10137100ccf029598abf596a
Scoped evidence bindings: 1

[137] Scaling Instruction-Finetuned Language Models. arXiv (author-deposited preprint); 2210.11416v5
https://arxiv.org/pdf/2210.11416v5
Response SHA-256: 771f758c1b711c2a63ca2439e80ab90751351d721632897a058c0205ba9e2a22
Scoped evidence bindings: 5

[138] Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. arXiv (author-deposited preprint); 2204.05862v1
https://arxiv.org/pdf/2204.05862v1
Response SHA-256: 26e390d4f76938d2e2591225603637c643cc108e94afb756c65f5d2e0fc9037b
Scoped evidence bindings: 6

[139] Fair work policy | Humans in the Loop. Humans in the Loop; Exact response sha256:8e79584334a9dbefc2f6354124bb950cb29eddd1d857b8312b9fdc2c3be077a4
https://humansintheloop.org/impact/fair-work-policy/
Response SHA-256: 8e79584334a9dbefc2f6354124bb950cb29eddd1d857b8312b9fdc2c3be077a4
Scoped evidence bindings: 3

[140] Humans in the Loop | Ethical AI with 95% Data Accuracy. Humans in the Loop; Exact response sha256:1cfadf2999fda4ea003c6f4d363781f0d866eb911eff4f81cc273f14acb72921
https://humansintheloop.org/
Response SHA-256: 1cfadf2999fda4ea003c6f4d363781f0d866eb911eff4f81cc273f14acb72921
Scoped evidence bindings: 1

[141] How it works | Humans in the Loop | HITL. Humans in the Loop; Exact response sha256:df843974948316b9dd24a55e0c0ca148547ab0839d88a7cc4a5127a3a9899e24
https://humansintheloop.org/how-it-works/
Response SHA-256: df843974948316b9dd24a55e0c0ca148547ab0839d88a7cc4a5127a3a9899e24
Scoped evidence bindings: 2

[142] Human Preference Optimization: RLHF & DPO Services | Innodata. Innodata; Exact response sha256:11d3c6dedc0a8b0854bdf5b91c9ed9fcc35898336fb0aa93cefa9cadcf523613
https://innodata.com/generative-ai/human-preference-optimization/
Response SHA-256: 11d3c6dedc0a8b0854bdf5b91c9ed9fcc35898336fb0aa93cefa9cadcf523613
Scoped evidence bindings: 5

[143] Work for isahit, an impact sourcing platform. isahit; Exact response sha256:61eae5f760696f9f468a3267e9b1e0cf3a6ddca4b80c7f0fd5ad34d54650d17a
https://www.isahit.com/work-for-us
Response SHA-256: 61eae5f760696f9f468a3267e9b1e0cf3a6ddca4b80c7f0fd5ad34d54650d17a
Scoped evidence bindings: 2

[144] Isahit - Diverse and Qualified Workforce to Scale Your Projects. isahit; Exact response sha256:f4a67f380a9014f5d1c0698145fe964d8a96e629d33d20473827506cb6a62077
https://www.isahit.com/workforce
Response SHA-256: f4a67f380a9014f5d1c0698145fe964d8a96e629d33d20473827506cb6a62077
Scoped evidence bindings: 3

[145] SeeChange Retail Theft Detection AI Training Data | Keymakr. Keymakr; Exact response sha256:9e3a75a85e4ec3af0ca468770e796c1a421bd88e4a6db25c5de4521aee07bdc6
https://keymakr.com/seechange-case.html
Response SHA-256: 9e3a75a85e4ec3af0ca468770e796c1a421bd88e4a6db25c5de4521aee07bdc6
Scoped evidence bindings: 4

[146] Image & Video Annotation, Data Collection | Keymakr. Keymakr; Exact response sha256:30a3fbd7cf38a0a058899258de08caa12a54cbbca69d67a64a940ce48f13b545
https://keymakr.com/
Response SHA-256: 30a3fbd7cf38a0a058899258de08caa12a54cbbca69d67a64a940ce48f13b545
Scoped evidence bindings: 1

[147] Aviation Polygon Annotation Case Study: Searidge | Label Your Data. Label Your Data; Exact response
sha256:5d3bff6e47e88ae53c0f55e53735dc172ba1965ae7a5f835fcc5dd1c4a489b9c
https://labelyourdata.com/resources/case-study/searidge-polygon-annotation
Response SHA-256: 5d3bff6e47e88ae53c0f55e53735dc172ba1965ae7a5f835fcc5dd1c4a489b9c
Scoped evidence bindings: 2

[148] Data Annotation Company for Complex AI | Label Your Data. Label Your Data; Exact response
sha256:8ab461a563055575956f4bcff9a84f560eeff674aa2d7617660b2a787e067bb1
https://labelyourdata.com/
Response SHA-256: 8ab461a563055575956f4bcff9a84f560eeff674aa2d7617660b2a787e067bb1
Scoped evidence bindings: 2

[149] LIMA: Less Is More for Alignment. arXiv (author-deposited preprint); 2305.11206v1
https://arxiv.org/pdf/2305.11206v1
Response SHA-256: 759ddebaa24a03ee772a70d3a0ecc546f53310f209de0f8f6052a3dddb289144
Scoped evidence bindings: 5

[150] Data Annotation & Labeling Services for AI and ML Projects. Objectways; Exact response
sha256:5fc4ddac5b14ded16d540056b69383a372b23830740648c2f5b197402bc320aa
https://objectways.com/services/data-annotation/
Response SHA-256: 5fc4ddac5b14ded16d540056b69383a372b23830740648c2f5b197402bc320aa
Scoped evidence bindings: 3

[151] AI Data Solutions | AI Data Services | Objectways. Objectways; Exact response
sha256:2379a296f99b5f81626c6530ce151d0d550b834598634c38146773f127809d5b
https://objectways.com/
Response SHA-256: 2379a296f99b5f81626c6530ce151d0d550b834598634c38146773f127809d5b
Scoped evidence bindings: 1

[152] Data Processing and Security | Objectways. Objectways; Exact response
sha256:80c3d3e59ddac97adb57ef27fe6ccb311ec2f516f38258c4d868c48fa1dea6d2
https://objectways.com/data-security/
Response SHA-256: 80c3d3e59ddac97adb57ef27fe6ccb311ec2f516f38258c4d868c48fa1dea6d2
Scoped evidence bindings: 1

[153] The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. arXiv (author-deposited preprint); 2310.16787v1
https://arxiv.org/pdf/2310.16787v1
Response SHA-256: 8a68980a962a3638cb8f52a47f8b05c95820ad9555655ffce6c610bd5bd3e4c4
Scoped evidence bindings: 5

[154] Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv (author-deposited preprint); 2209.07858v1
https://arxiv.org/pdf/2209.07858v1
Response SHA-256: a0b0eb43b178435dcc1eb0dd8faf5aff8d637c4842664eeb8907db3abd3b12f1
Scoped evidence bindings: 6

[155] RLAIF-V: Aligning MLLMs through Open-Source AI Feedback for Super GPT-4V Trustworthiness. arXiv (author-deposited preprint); 2405.17220v1
https://arxiv.org/pdf/2405.17220v1
Response SHA-256: c9caa5fae980118176a7b3f06a55751110da6de5ed79bcb89784930e759f8ba3
Scoped evidence bindings: 6

[156] Self-Refine: Iterative Refinement with Self-Feedback. arXiv (author-deposited preprint); 2303.17651v2
https://arxiv.org/pdf/2303.17651v2
Response SHA-256: 81e44592314d80218ad108d3490cd0b84ba5962fefd5c84819987afd77a57087
Scoped evidence bindings: 5

[157] RLHF Services: Human Feedback Data from Domain Experts | Shaip. Shaip; Exact response sha256:b1f8d9dbfd04dfbb107b0d71430fe171442b6c4909e16fb31798ee2cfdad04db
https://www.shaip.com/generative-ai/rlhf-solutions/
Response SHA-256: b1f8d9dbfd04dfbb107b0d71430fe171442b6c4909e16fb31798ee2cfdad04db
Scoped evidence bindings: 6

[158] Learning to summarize from human feedback. arXiv (author-deposited preprint); 2009.01325v3
https://arxiv.org/pdf/2009.01325v3
Response SHA-256: 6bb15be37338385cae1d910bcb46bb2cc99e599918a0625c9c1d0a57f90e5224
Scoped evidence bindings: 6

[159] WizardLM: Empowering Large Pre-Trained Language Models to Follow Complex Instructions. arXiv (author-deposited preprint); 2304.12244v3
https://arxiv.org/pdf/2304.12244v3
Response SHA-256: e5fc7a244d34703bd77c4cc73b2bf1daa8d84160a90af37cacabfa8573bdd538
Scoped evidence bindings: 5

[160] Aya Dataset - README.md. CohereForAI/aya_dataset; f9ea04583f02a8f86404ff6c58bf75fe637df8a2
https://huggingface.co/datasets/CohereForAI/aya_dataset/raw/f9ea04583f02a8f86404ff6c58bf75fe637df8a2/README.md
Response SHA-256: 3ac54f63eec84100c155571945da01c7c18cd8ec97ea94639e2b94a6860e8573
Scoped evidence bindings: 6

[161] Code-Feedback - README.md. m-a-p/Code-Feedback; f411b16a97c910ac9acf8b0d0948e340aa77cc34
https://huggingface.co/datasets/m-a-p/Code-Feedback/raw/f411b16a97c910ac9acf8b0d0948e340aa77cc34/README.md
Response SHA-256: 0c77b55d450a864e9e2c5f69ce266e8e2061f52b6c6f8c2c7710dff034f2b70f
Scoped evidence bindings: 2

[162] Code-Feedback - README.md. OpenCodeInterpreter/OpenCodeInterpreter; 9674d8a521df7f92becd8a7a8e60a8792e71869a
https://raw.githubusercontent.com/OpenCodeInterpreter/OpenCodeInterpreter/9674d8a521df7f92becd8a7a8e60a8792e71869a/README.md
Response SHA-256: 43ec99edb44440d79ca2229d64c3b1a165f039eada68647b6f34f9757da9eba1
Scoped evidence bindings: 2

[163] Code-Feedback - README.md. m-a-p/CodeFeedback-Filtered-Instruction; a08c213a9748c66c15d0225814be80a2e77adf4a
https://huggingface.co/datasets/m-a-p/CodeFeedback-Filtered-Instruction/raw/a08c213a9748c66c15d0225814be80a2e77adf4a/README.md
Response SHA-256: 1842787cdd4c324bacbc7e44239af3561d70e32a82b4ff27c2506ebaf1392de8
Scoped evidence bindings: 2

[164] Databricks Dolly 15k - README.md. databricks/databricks-dolly-15k; bdd27f4d94b9c1f951818a7da7fd7aeea5dbff1a
https://huggingface.co/datasets/databricks/databricks-dolly-15k/raw/bdd27f4d94b9c1f951818a7da7fd7aeea5dbff1a/README.md
Response SHA-256: db06c60090a11b20e79d1933b2f316f781605f07761ff2ae2795d6523e3e960b
Scoped evidence bindings: 6

[165] The Flan Collection - flan/v2/README.md. google-research/FLAN; e9e4ec6e2701182c7a91af176f705310da541277
https://raw.githubusercontent.com/google-research/FLAN/e9e4ec6e2701182c7a91af176f705310da541277/flan/v2/README.md
Response SHA-256: bc1037acf7fea6c913352785fd24080bc223336bf21d6617aaa44dcd95794042
Scoped evidence bindings: 6

[166] The Flan Collection - README.md. google-research/FLAN; e9e4ec6e2701182c7a91af176f705310da541277
https://raw.githubusercontent.com/google-research/FLAN/e9e4ec6e2701182c7a91af176f705310da541277/README.md
Response SHA-256: 38a12f7b33784ebcf385436fa9a592074c3ba8d7eaacacf28e6a682786988273
Scoped evidence bindings: 1

[167] HarmBench - docs/evaluation_pipeline.md. CenterForAISafety/HarmBench; 8e1604d1171fe8a48d8febecd22f600e462bdcdd
https://raw.githubusercontent.com/CenterForAISafety/HarmBench/8e1604d1171fe8a48d8febecd22f600e462bdcdd/docs/evaluation_pipeline.md
Response SHA-256: d900abc3ef8ca94179ba247da3a40fcafe4c26bfb6ddcbc5af797ef7f73d6cec
Scoped evidence bindings: 5

[168] HarmBench - LICENSE. CenterForAISafety/HarmBench; 8e1604d1171fe8a48d8febecd22f600e462bdcdd
https://raw.githubusercontent.com/CenterForAISafety/HarmBench/8e1604d1171fe8a48d8febecd22f600e462bdcdd/LICENSE
Response SHA-256: fe29c68dd1216e81e98a6069165e1c720757b23bbcaa6c0cb8f9a957bdaaca80
Scoped evidence bindings: 1

[169] HumanEval - LICENSE. openai/human-eval; b9453e8f1865df0a0057a1383767caa400acfbcf
https://raw.githubusercontent.com/openai/human-eval/b9453e8f1865df0a0057a1383767caa400acfbcf/LICENSE
Response SHA-256: bcba3de214851cce46ed5af42d6698044616eeace887c3231bc7a20474ab639e
Scoped evidence bindings: 1

[170] HumanEval - README.md. openai/human-eval; b9453e8f1865df0a0057a1383767caa400acfbcf
https://raw.githubusercontent.com/openai/human-eval/b9453e8f1865df0a0057a1383767caa400acfbcf/README.md
Response SHA-256: a6a3741a64527ae20026b7880e37af5de88e8ae4dcfed81eb19620c674b6e272
Scoped evidence bindings: 3

[171] HumanEval - human_eval/evaluation.py. openai/human-eval; b9453e8f1865df0a0057a1383767caa400acfbcf
https://raw.githubusercontent.com/openai/human-eval/b9453e8f1865df0a0057a1383767caa400acfbcf/human_eval/evaluation.py
Response SHA-256: e926d5e9b8f1ec040096cd5952e8cf79f8dc0b7fb80ba6667ce604ffea0397bc
Scoped evidence bindings: 2

[172] Intel Orca DPO Pairs - intel_extension_for_transformers/neural_chat/examples/finetuning/finetune_neuralchat_v3/README.md. intel/intel-extension-for-transformers; 087056c348f3321a78bffde6fc35c81a6aaf19d9
https://raw.githubusercontent.com/intel/intel-extension-for-transformers/087056c348f3321a78bffde6fc35c81a6aaf19d9/intel_extension_for_transformers/neural_chat/examples/finetuning/finetune_neuralchat_v3/README.md
Response SHA-256: 32fa604aba553e57270297b868074d87babb1dbcebd16b9c3f6d7e26415c678e
Scoped evidence bindings: 4

[173] Intel Orca DPO Pairs - README.md. Intel/orca_dpo_pairs; 624952e3f420ae18d88b31977ee2ea436c833abb
https://huggingface.co/datasets/Intel/orca_dpo_pairs/raw/624952e3f420ae18d88b31977ee2ea436c833abb/README.md
Response SHA-256: 3059107fa3f36cb558142b2b80e6dfdbaa5d8d6d3a26564487420fe402eceb52
Scoped evidence bindings: 2

[174] JailbreakBench - README.md. JailbreakBench/jailbreakbench; 23dbdf6b19650521604456229bc1d9c4156c85c1
https://raw.githubusercontent.com/JailbreakBench/jailbreakbench/23dbdf6b19650521604456229bc1d9c4156c85c1/README.md
Response SHA-256: aa7d1c4154ac4defadb533a7c686a720a808022f24c73763567e028cd5825d8a
Scoped evidence bindings: 6

[175] JailbreakBench - README.md. JailbreakBench/JBB-Behaviors; 886acc352a31533ffbcf4ef22c744658688086fc
https://huggingface.co/datasets/JailbreakBench/JBB-Behaviors/raw/886acc352a31533ffbcf4ef22c744658688086fc/README.md
Response SHA-256: fc84ea5865e60ba15f3776ff43ace47cbbf3470775f389afbb6bdb4d6333618a
Scoped evidence bindings: 1

[176] Math-Shepherd - README.md. peiyi9979/Math-Shepherd; ae6b0c54fca9fa26096f7de175c747e4b262e01a
https://huggingface.co/datasets/peiyi9979/Math-Shepherd/raw/ae6b0c54fca9fa26096f7de175c747e4b262e01a/README.md
Response SHA-256: 5cd51153b70134d08a71362c29c3d7b8788e14700ff7917bfcc7f172039a2065
Scoped evidence bindings: 6

[177] MathVista - README.md. lupantech/MathVista; 5f6b4deca69ecd1b8a1a4f338cee085e50524e4b
https://raw.githubusercontent.com/lupantech/MathVista/5f6b4deca69ecd1b8a1a4f338cee085e50524e4b/README.md
Response SHA-256: f55ef5c9f27869c73cb180d93cb9ea5681eca056b9312970a8620e27329ddc19
Scoped evidence bindings: 5

[178] MathVista - evaluation/calculate_score.py. lupantech/MathVista; 5f6b4deca69ecd1b8a1a4f338cee085e50524e4b
https://raw.githubusercontent.com/lupantech/MathVista/5f6b4deca69ecd1b8a1a4f338cee085e50524e4b/evaluation/calculate_score.py
Response SHA-256: abcd0ff68ec3ed76c3f343be4d01d0f4b88238b660a64e6b998b5f15e3ef56d1
Scoped evidence bindings: 1

[179] MathVista - README.md. AI4Math/MathVista; 2b6ad69445fbb5695c9b165475e8decdbeb97747
https://huggingface.co/datasets/AI4Math/MathVista/raw/2b6ad69445fbb5695c9b165475e8decdbeb97747/README.md
Response SHA-256: b2c1784e13e6a4bf1935930e20a7472a04b40e2d47b7509508266352d0e121ed
Scoped evidence bindings: 1

[180] MBPP - mbpp/README.md. google-research/google-research; 03ac4dca9776084a3a002d4513a55dc7348f680e
https://raw.githubusercontent.com/google-research/google-research/03ac4dca9776084a3a002d4513a55dc7348f680e/mbpp/README.md
Response SHA-256: 02d63a4ffdad806c3c87b5ea2aa91e946f6eec0c2869643b09831650883f41be
Scoped evidence bindings: 5

[181] MBPP - LICENSE. google-research/google-research; 03ac4dca9776084a3a002d4513a55dc7348f680e
https://raw.githubusercontent.com/google-research/google-research/03ac4dca9776084a3a002d4513a55dc7348f680e/LICENSE
Response SHA-256: cfc7749b96f63bd31c3c42b5c471bf756814053e847c10f3eb003417bc523d30
Scoped evidence bindings: 1

[182] MedInstruct - README.md. xz97/MedInstruct; 2934ad53c8250ba4e4130464a6dae3139d1053de
https://huggingface.co/datasets/xz97/MedInstruct/raw/2934ad53c8250ba4e4130464a6dae3139d1053de/README.md
Response SHA-256: d6a1ca0c60549f9c01b92a316681c527d7f94d94e76c10f6d53e46b2b2445f30
Scoped evidence bindings: 5

[183] MedInstruct - README.md. XZhang97666/AlpaCare; b469c41bcd7ddd57dd0bc072ef99d017c5cb0e4b
https://raw.githubusercontent.com/XZhang97666/AlpaCare/b469c41bcd7ddd57dd0bc072ef99d017c5cb0e4b/README.md
Response SHA-256: 39477f146c3bbae7e8a41601f78ef9b73d14e3c029ee76595dceab533cf63c6b
Scoped evidence bindings: 2

[184] MMMU - README.md. MMMU-Benchmark/MMMU; 268471d0d488258990025331c7528359c324aa25
https://raw.githubusercontent.com/MMMU-Benchmark/MMMU/268471d0d488258990025331c7528359c324aa25/README.md
Response SHA-256: a4e89dd389f9b4c82b92e54b5d50a58e55a53c077a5009f7d4b4e53c09bdcc3a
Scoped evidence bindings: 5

[185] MMMU - README.md. MMMU/MMMU; 876ce5cb130f7f7e290ce4d9984357737d4db5cf
https://huggingface.co/datasets/MMMU/MMMU/raw/876ce5cb130f7f7e290ce4d9984357737d4db5cf/README.md
Response SHA-256: 0c9190b5582c5798f5744d37215858de0df51c74de49fc4e64f925c3efd725a8
Scoped evidence bindings: 1

[186] MMMU - mmmu/README.md. MMMU-Benchmark/MMMU; 268471d0d488258990025331c7528359c324aa25
https://raw.githubusercontent.com/MMMU-Benchmark/MMMU/268471d0d488258990025331c7528359c324aa25/mmmu/README.md
Response SHA-256: 8f0cbea394ed36ce16c5cf92c18247df188c1158590ed56697e506684ed4dfa5
Scoped evidence bindings: 1

[187] No Robots - README.md. HuggingFaceH4/no_robots; e6f9a4ac5c37faeb744ba9ecf0473184d7f8105b
https://huggingface.co/datasets/HuggingFaceH4/no_robots/raw/e6f9a4ac5c37faeb744ba9ecf0473184d7f8105b/README.md
Response SHA-256: 22b642520e082ba1a6c4a93047bd5946aeca099763c6f2d11b59d91e61e79186
Scoped evidence bindings: 6

[188] OpenCodeInstruct - README.md. nvidia/OpenCodeInstruct; 8f3ba5bafe4d6e8db46082cf7ae6741bc370604d
https://huggingface.co/datasets/nvidia/OpenCodeInstruct/raw/8f3ba5bafe4d6e8db46082cf7ae6741bc370604d/README.md
Response SHA-256: 2449b70191f455f0e904ff78fabfd7185f20dbe89165ce70ca39a8193241ebd4
Scoped evidence bindings: 6

[189] OpenThoughts - README.md. open-thoughts/OpenThoughts-114k; bd093c3994fd54d2390985b66988ddf282a55eb6
https://huggingface.co/datasets/open-thoughts/OpenThoughts-114k/raw/bd093c3994fd54d2390985b66988ddf282a55eb6/README.md
Response SHA-256: 87dd0984b1ae912cd50507327d1358e378b3de07740600895734c939d0a55946
Scoped evidence bindings: 6

[190] OSWorld - desktop_env/desktop_env.py. xlang-ai/OSWorld; 8903ebeb0efdf74331bd7dc37ed91ccbbff22317
https://raw.githubusercontent.com/xlang-ai/OSWorld/8903ebeb0efdf74331bd7dc37ed91ccbbff22317/desktop_env/desktop_env.py
Response SHA-256: 169f0f196bfa8083e1e3eba329a64c7a2c00284705afc7d47c381f4b4fd1033c
Scoped evidence bindings: 2

[191] OSWorld - README.md. xlang-ai/OSWorld; 8903ebeb0efdf74331bd7dc37ed91ccbbff22317
https://raw.githubusercontent.com/xlang-ai/OSWorld/8903ebeb0efdf74331bd7dc37ed91ccbbff22317/README.md
Response SHA-256: 9a01ccee12ac531c0a86d54e0d4399cdaae7774eef33c5a0d5a019317471e814
Scoped evidence bindings: 4

[192] OSWorld - LICENSE. xlang-ai/OSWorld; 8903ebeb0efdf74331bd7dc37ed91ccbbff22317
https://raw.githubusercontent.com/xlang-ai/OSWorld/8903ebeb0efdf74331bd7dc37ed91ccbbff22317/LICENSE
Response SHA-256: a8d1d88a0f9cce3cdbff73b03a80dc5e5d69e3ba2443946c268815a550c9e0ec
Scoped evidence bindings: 1

[193] Stanford Alpaca - datasheet.md. tatsu-lab/stanford_alpaca; 761dc5bfbdeeffa89b8bff5d038781a4055f796a
https://raw.githubusercontent.com/tatsu-lab/stanford_alpaca/761dc5bfbdeeffa89b8bff5d038781a4055f796a/datasheet.md
Response SHA-256: 341433dc5888bba8f2492350f3ef9b21afe4aeb5f78e9568bb9c4dc2e0cb2394
Scoped evidence bindings: 4

[194] Stanford Alpaca - README.md. tatsu-lab/stanford_alpaca; 761dc5bfbdeeffa89b8bff5d038781a4055f796a
https://raw.githubusercontent.com/tatsu-lab/stanford_alpaca/761dc5bfbdeeffa89b8bff5d038781a4055f796a/README.md
Response SHA-256: bc335b356f6de0a3294cf65d0a389edc4a796764e31e777b7a8d61a7cb7aa551
Scoped evidence bindings: 3

[195] OpenAI Summarize from Feedback - README.md. openai/summarize-from-feedback; 700967448d10004279f138666442bf1497d0e705
https://raw.githubusercontent.com/openai/summarize-from-feedback/700967448d10004279f138666442bf1497d0e705/README.md
Response SHA-256: 5c5c30159193e8d5d104fa0aeda27cbfdee1ba7053ca43807cc7d59b5af917d0
Scoped evidence bindings: 6

[196] Super-NaturalInstructions - README.md. allenai/natural-instructions; 55a365637381ce7f3748fa2eac7aef1a113bbb82
https://raw.githubusercontent.com/allenai/natural-instructions/55a365637381ce7f3748fa2eac7aef1a113bbb82/README.md
Response SHA-256: bd5c0768d317f0184c00afd3233d166672aa63a6fee52fa368f3aa117111ccd7
Scoped evidence bindings: 6

[197] τ-bench - LICENSE. sierra-research/tau-bench; 59a200c6d575d595120f1cb70fea53cef0632f6b
https://raw.githubusercontent.com/sierra-research/tau-bench/59a200c6d575d595120f1cb70fea53cef0632f6b/LICENSE
Response SHA-256: 243d23d45b80122b5ac575586ccef352fdc9c45e6d7d2605449aca8a17478b42
Scoped evidence bindings: 1

[198] τ-bench - tau_bench/envs/base.py. sierra-research/tau-bench; 59a200c6d575d595120f1cb70fea53cef0632f6b
https://raw.githubusercontent.com/sierra-research/tau-bench/59a200c6d575d595120f1cb70fea53cef0632f6b/tau_bench/envs/base.py
Response SHA-256: 1ad0402073f0ac9ca6b527efa0706e2aa8836dd2e5c1d453ed414ecb768c6d21
Scoped evidence bindings: 2

[199] τ-bench - README.md. sierra-research/tau-bench; 59a200c6d575d595120f1cb70fea53cef0632f6b
https://raw.githubusercontent.com/sierra-research/tau-bench/59a200c6d575d595120f1cb70fea53cef0632f6b/README.md
Response SHA-256: 1a4b1b00634fe46fd677dd11dfa25a9bf357b5d7c3d726ff83025e8f104329f9
Scoped evidence bindings: 4

[200] UltraMedical - README.md. TsinghuaC3I/UltraMedical; e53eab0a5d0bf981c8cb9c55744d61dc0b4d56a9
https://raw.githubusercontent.com/TsinghuaC3I/UltraMedical/e53eab0a5d0bf981c8cb9c55744d61dc0b4d56a9/README.md
Response SHA-256: 479f472c4f06fc4ed9aaeffae3c12b28271b3540806033dea0947860d0071fb6
Scoped evidence bindings: 2

[201] UltraMedical - README.md. TsinghuaC3I/UltraMedical; 8930b349985de8345df14d65eb894cf8fce36d9c
https://huggingface.co/datasets/TsinghuaC3I/UltraMedical/raw/8930b349985de8345df14d65eb894cf8fce36d9c/README.md
Response SHA-256: d1f52a4de8b015cae1386ceee29568855fe229128292d3aaa4d362a0d0026228
Scoped evidence bindings: 4

[202] WebGPT Comparisons - README.md. openai/webgpt_comparisons; f30e1bd2f0a51535b493de36d3686cfcfff47abe
https://huggingface.co/datasets/openai/webgpt_comparisons/raw/f30e1bd2f0a51535b493de36d3686cfcfff47abe/README.md
Response SHA-256: 50ea5bd9737deb4ea6275b7ec9b6080233d4cf10fe1a46fe89c473aa2aca1502
Scoped evidence bindings: 6

[203] WebShop - README.md. princeton-nlp/WebShop; 64fa2a5c15c7daa698b9ac93f5bb5437b634c9bd
https://raw.githubusercontent.com/princeton-nlp/WebShop/64fa2a5c15c7daa698b9ac93f5bb5437b634c9bd/README.md
Response SHA-256: 79b7a90a1413a52deb53142b5bbd81ecf40f61743bd4601c2835e91375898ec0
Scoped evidence bindings: 6

[204] xP3 - README.md. bigscience/xP3; 0ab0411dca6e222e62d210bc681dbcb476d6fe4c
https://huggingface.co/datasets/bigscience/xP3/raw/0ab0411dca6e222e62d210bc681dbcb476d6fe4c/README.md
Response SHA-256: b037cb4eac19ccc852b6570dda2ca0cc5017298407fbcc4aaaebef6382e110ea
Scoped evidence bindings: 9

[205] abundant. Abundant;
response-sha256:6de358eee32fb49a45f1e3ac038016017405529bdd84b54f33dc56f697b177b2
https://www.abundant.ai/
Response SHA-256: 6de358eee32fb49a45f1e3ac038016017405529bdd84b54f33dc56f697b177b2
Scoped evidence bindings: 2

[206] Introducing Foundry - Abundant. Abundant;
response-sha256:7647c497db91cecb3b8df0857d31ec839b77bf178654e4f530250bd215e9f7c2
https://www.abundant.ai/blog/fireworks/
Response SHA-256: 7647c497db91cecb3b8df0857d31ec839b77bf178654e4f530250bd215e9f7c2
Scoped evidence bindings: 3

[207] Expert Data for Frontier AI - AfterQuery. AfterQuery;
response-sha256:f778907ba417eb9d8689afe5bc99caf62e3024e3bfe671cfa5815741d8f53a53
https://www.afterquery.com/
Response SHA-256: f778907ba417eb9d8689afe5bc99caf62e3024e3bfe671cfa5815741d8f53a53
Scoped evidence bindings: 3

[208] AfterQuery. AfterQuery;
response-sha256:c07887aa70f6e967c8e84fd7187568539656ee57b1d8269ad1ce9e6a661bdbaf
https://www.afterquery.com/privacy
Response SHA-256: c07887aa70f6e967c8e84fd7187568539656ee57b1d8269ad1ce9e6a661bdbaf
Scoped evidence bindings: 1

[209] AfterQuery. AfterQuery;
response-sha256:ee69548cbd324b741688409f60bfd38707b9d9552d7e6945eeebe5eabb23063a
https://www.afterquery.com/terms
Response SHA-256: ee69548cbd324b741688409f60bfd38707b9d9552d7e6945eeebe5eabb23063a
Scoped evidence bindings: 1

[210] AgentGym: Evolving Large Language Model-based Agents across Diverse Environments. arXiv (author-deposited preprint); 2406.04151v1
https://arxiv.org/pdf/2406.04151v1
Response SHA-256: 0950510830b43a850bc21763af9a240e83e69ef7efff5258f07d66bc99412b60
Scoped evidence bindings: 5

[211] BenchFlow documentation - BenchFlow. BenchFlow;
response-sha256:824bdc43ad23c0d6f327d7bc75d9392759dcd201f438650a5e6d04de13c863ac
https://www.benchflow.ai/docs/benchflow
Response SHA-256: 824bdc43ad23c0d6f327d7bc75d9392759dcd201f438650a5e6d04de13c863ac
Scoped evidence bindings: 3

[212] BenchFlow - A frontier environment lab for AI agents. BenchFlow;
response-sha256:0b28242ccbca607761c3bbcf5649b82aa240e8a8c3d731085816a6f739a14be0
https://www.benchflow.ai/
Response SHA-256: 0b28242ccbca607761c3bbcf5649b82aa240e8a8c3d731085816a6f739a14be0
Scoped evidence bindings: 2

[213] About - BenchFlow. BenchFlow;
response-sha256:c837863807a9e968ed8c70bf8e9c95a547add56db6073656eafd0190290d1a43
https://www.benchflow.ai/about
Response SHA-256: c837863807a9e968ed8c70bf8e9c95a547add56db6073656eafd0190290d1a43
Scoped evidence bindings: 1

[214] Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. arXiv (author-deposited preprint); 2206.04615v1
https://arxiv.org/pdf/2206.04615v1
Response SHA-256: 5e9388ec6aca9dd0dfa58b78ed056a27ceca1641356dde3e1b79de41af3b0569
Scoped evidence bindings: 5

[215] Research | Datacurve. Datacurve;
response-sha256:fbe545268ac1da5d7665c2edfa8d5d2d9e295bc69318f4dc5709356680b499f5
https://datacurve.ai/research
Response SHA-256: fbe545268ac1da5d7665c2edfa8d5d2d9e295bc69318f4dc5709356680b499f5
Scoped evidence bindings: 1

[216] DeepSWE. Datacurve;
response-sha256:4ad6365bf7625f9fe176a5c6c709855ecaa74d3c829dadfb407546d0aefcfae2
https://deepswe.datacurve.ai/
Response SHA-256: 4ad6365bf7625f9fe176a5c6c709855ecaa74d3c829dadfb407546d0aefcfae2
Scoped evidence bindings: 2

[217] Products | Datacurve. Datacurve;
response-sha256:05381109abef3561e6766112e54720d357560b6d8ce81976da1ddf74c99388fb
https://datacurve.ai/products
Response SHA-256: 05381109abef3561e6766112e54720d357560b6d8ce81976da1ddf74c99388fb
Scoped evidence bindings: 3

[218] G2i.ai. G2i; response-sha256:13e8221994b3d11ed14fa57b8140e4ad7da0c87ecd5e3b6912c8531762463946
https://www.g2i.ai/
Response SHA-256: 13e8221994b3d11ed14fa57b8140e4ad7da0c87ecd5e3b6912c8531762463946
Scoped evidence bindings: 3

[219] Verifying Multi-SWE-bench TypeScript: Quality Analysis of 210 Tasks · G2i.ai. G2i;
response-sha256:d1f7fa03606e6910e4227d4019d29a4fb0ed5df8fc34d9b2e0cc1aa2b8d774f8
https://www.g2i.ai/research/multi-swe-bench-typescript
Response SHA-256: d1f7fa03606e6910e4227d4019d29a4fb0ed5df8fc34d9b2e0cc1aa2b8d774f8
Scoped evidence bindings: 2

[220] GAIA: A Benchmark for General AI Assistants. arXiv (author-deposited preprint); 2311.12983v1
https://arxiv.org/pdf/2311.12983v1
Response SHA-256: 5bf6f968d4f5cd20acdfca6e81e3c9f5277bb66061fad7a02e35442dff812ba0
Scoped evidence bindings: 5

[221] Handshake AI Contractor Agreement | Handshake. Handshake AI;
response-sha256:1f5cbe6b5a3db39b2971605385f43d222d94453c94639415ef58fccbaa39c40b
https://joinhandshake.com/legal/contractor-agreement/
Response SHA-256: 1f5cbe6b5a3db39b2971605385f43d222d94453c94639415ef58fccbaa39c40b
Scoped evidence bindings: 5

[222] Join Handshake | Handshake. Handshake AI;
response-sha256:6ada18100ee3b7bda86680cd48d94f59ee5ebcae59078fb63f7492f350a8496c
https://joinhandshake.com/privacy-policy/
Response SHA-256: 6ada18100ee3b7bda86680cd48d94f59ee5ebcae59078fb63f7492f350a8496c
Scoped evidence bindings: 1

[223] HarmBench: A Standardized Evaluation Framework for Automated Red Teaming and Robust Refusal. arXiv (author-deposited preprint); 2402.04249v1
https://arxiv.org/pdf/2402.04249v1
Response SHA-256: 2f9ac1c3dad14bc3c562827f4029129f6f6aed1e4d2a4c94649c7c3bfa192493
Scoped evidence bindings: 5

[224] Evaluating Large Language Models Trained on Code. arXiv (author-deposited preprint); 2107.03374v1
https://arxiv.org/pdf/2107.03374v1
Response SHA-256: 7085950bcce34fee8ee1821ed1ac0c2672e13708ec866810f997c0bf467d5009
Scoped evidence bindings: 5

[225] JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models. arXiv (author-deposited preprint); 2404.01318v5
https://arxiv.org/pdf/2404.01318v5
Response SHA-256: 78c2f251cc101bce36e3dd3138fc0c89f7b73c599099f79ec55318165fe4a981
Scoped evidence bindings: 5

[226] MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. arXiv (author-deposited preprint); 2310.02255v1
https://arxiv.org/pdf/2310.02255v1
Response SHA-256: e6b25f084c455b9657f5e43b73cdf38b2473bc2db7cf198ab261e78d70ef99b0
Scoped evidence bindings: 5

[227] Program Synthesis with Large Language Models. arXiv (author-deposited preprint); 2108.07732v1
https://arxiv.org/pdf/2108.07732v1
Response SHA-256: 820a8de4a9fe8aef947a40d5da13db0d660baf3658b4a9cebf6d30ac5ba351ab
Scoped evidence bindings: 5

[228] Frontier evaluations and RL environments for AI Labs. micro1;
response-sha256:acc1e3e5ba5a0377d2d95bd919d6ed145a7592aabc3c7abfad9f295723884716
https://www.micro1.ai/realm
Response SHA-256: acc1e3e5ba5a0377d2d95bd919d6ed145a7592aabc3c7abfad9f295723884716
Scoped evidence bindings: 4

[229] Expert support program. micro1;
response-sha256:b9ccbd24127eabb603c142d8b45972e9a0d540292c97748dfd8d80e1d1734321
https://www.micro1.ai/experts/support-program
Response SHA-256: b9ccbd24127eabb603c142d8b45972e9a0d540292c97748dfd8d80e1d1734321
Scoped evidence bindings: 1

[230] MissionReport. micro1;
response-sha256:826072b47c4e813fd646132fadbf3efab01a3d65a003f887add4eb0b4ad9fdf7
https://www.micro1.ai/benchmark/mission-report
Response SHA-256: 826072b47c4e813fd646132fadbf3efab01a3d65a003f887add4eb0b4ad9fdf7
Scoped evidence bindings: 1

[231] Reinforcement Learning on Web Interfaces Using Workflow-Guided Exploration. arXiv (author-deposited preprint); 1802.08802v1
https://arxiv.org/pdf/1802.08802v1
Response SHA-256: 7a7ce9d5f8f31c7f7c01a35271c5bdfdf2697dfe17e72ff090a9465d55f7f60d
Scoped evidence bindings: 6

[232] MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv (author-deposited preprint); 2311.16502v1
https://arxiv.org/pdf/2311.16502v1
Response SHA-256: 16c07f7ef6d8aae648ec80fdc71ce97787178e5b6b183de757247908d0e56f4a
Scoped evidence bindings: 5

[233] OSWorld: Benchmarking Multimodal Agents for Open-Ended Tasks in Real Computer Environments. arXiv (author-deposited preprint); 2404.07972v1
https://arxiv.org/pdf/2404.07972v1
Response SHA-256: 8499a6ec06757448687f9e1a4e10bd11a5a561cbe2e52ecc2643e17b17b524a7
Scoped evidence bindings: 5

[234] Introducing AttuneBench. Pareto AI;
response-sha256:bf4c78302ee38c9f64824173a350344e2e0aef04a5f18c4afe576c7952b7b57a
https://pareto.ai/blog/introducing-attunebench
Response SHA-256: bf4c78302ee38c9f64824173a350344e2e0aef04a5f18c4afe576c7952b7b57a
Scoped evidence bindings: 1

[235] Join as an Expert | Pareto AI. Pareto AI;
response-sha256:e4ef0d8001b6eeef89659c5d5b2798f112328de4ef8e60c136a7d08cea8bb3ce
https://pareto.ai/experts
Response SHA-256: e4ef0d8001b6eeef89659c5d5b2798f112328de4ef8e60c136a7d08cea8bb3ce
Scoped evidence bindings: 3

[236] Terms of Use | Pareto AI. Pareto AI;
response-sha256:7a3f9405ba1ed093c649eda943865848d90ab78fb4980109941fdc58f4605392
https://pareto.ai/terms-of-use
Response SHA-256: 7a3f9405ba1ed093c649eda943865848d90ab78fb4980109941fdc58f4605392
Scoped evidence bindings: 2

[237] Polymath. Polymath;
response-sha256:b7ab1966cac11e01639ff17b5e1a21c08d84b85735b3e91c056b39633befae43
https://polymathlabs.ai/terms
Response SHA-256: b7ab1966cac11e01639ff17b5e1a21c08d84b85735b3e91c056b39633befae43
Scoped evidence bindings: 3

[238] Polymath. Polymath;
response-sha256:aa3a25143abb9c828694be185b89ca05c2962ab5b5c994f1ab42d2af4b00be99
https://polymathlabs.ai/
Response SHA-256: aa3a25143abb9c828694be185b89ca05c2962ab5b5c994f1ab42d2af4b00be99
Scoped evidence bindings: 1

[239] Polymath. Polymath; response-sha256:6bee82d9f5b2ed107faa27dc2c26fec39fe768835a299eff04ac154da47c85c7
https://polymathlabs.ai/blog/horizon-swe
Response SHA-256: 6bee82d9f5b2ed107faa27dc2c26fec39fe768835a299eff04ac154da47c85c7
Scoped evidence bindings: 2

[240] Prime Intellect - The Open Superintelligence Stack. Prime Intellect;
response-sha256:612ae3d4551e969e0fdba93a0667d04f6760cd119001d4e050944772dbeb1b82
https://www.primeintellect.ai/
Response SHA-256: 612ae3d4551e969e0fdba93a0667d04f6760cd119001d4e050944772dbeb1b82
Scoped evidence bindings: 1

[241] Post-Training Nemotron 3.5 Super VL for Long-Horizon Enterprise Work. Prime Intellect;
response-sha256:0d9737a2b383410e8dae1bb26aecb356162be647a391d36a35bac864a1d33d16
https://www.primeintellect.ai/case-study/nvidia
Response SHA-256: 0d9737a2b383410e8dae1bb26aecb356162be647a391d36a35bac864a1d33d16
Scoped evidence bindings: 2

[242] MODEL FACTORY SERVICE TERMS. Prime Intellect;
response-sha256:2ead413b724508b7ae120dad0e4d668725e0a80c549f2d0779682d6357264bbd
https://www.primeintellect.ai/legal/model-factory-terms
Response SHA-256: 2ead413b724508b7ae120dad0e4d668725e0a80c549f2d0779682d6357264bbd
Scoped evidence bindings: 2

[243] Privacy Policy • Proximal. Proximal;
response-sha256:c9fd77546736df0e2f706456ca21576cfa67080e3300ca499c2d1101659c058f
https://www.proximal.ai/privacy/
Response SHA-256: c9fd77546736df0e2f706456ca21576cfa67080e3300ca499c2d1101659c058f
Scoped evidence bindings: 1

[244] Proximal. Proximal; response-sha256:49e57cbfc68fcf0cbbccff8db0ad867ced03ac12d319d7d87276f80aedee3a4b
https://www.proximal.ai/
Response SHA-256: 49e57cbfc68fcf0cbbccff8db0ad867ced03ac12d319d7d87276f80aedee3a4b
Scoped evidence bindings: 1

[245] Master Development Services Agreement • Proximal. Proximal;
response-sha256:02ab2281c5999269875868559ee040f15a2a4d89698d9dc38e84fe5c86fb347e
https://www.proximal.ai/msa/
Response SHA-256: 02ab2281c5999269875868559ee040f15a2a4d89698d9dc38e84fe5c86fb347e
Scoped evidence bindings: 2

[246] New Frontiers. Proximal;
response-sha256:d4ef948c0b00467303ac3401c43278ec141920143c28c94b5430ed84500eaf51
https://www.proximal.ai/blog/new-frontiers/
Response SHA-256: d4ef948c0b00467303ac3401c43278ec141920143c28c94b5430ed84500eaf51
Scoped evidence bindings: 2

[247] AI Coding Agent Analytics & Cost Intelligence | Quesma. Quesma;
response-sha256:223fa6e24accbeefdf985052bae8ee4d8565b745feba301bf629c36c0c776ffc
https://quesma.com/
Response SHA-256: 223fa6e24accbeefdf985052bae8ee4d8565b745feba301bf629c36c0c776ffc
Scoped evidence bindings: 2

[248] Compare harnesses, not models: Blitzy vs. GPT-5.4 on SWE-Bench Pro - Quesma Blog. Quesma;
response-sha256:5c44e6b157d66c556a16597a9031c22c95f36f060d42b93b880496953b825090
https://quesma.com/blog/verifying-blitzy-swe-bench-pro/
Response SHA-256: 5c44e6b157d66c556a16597a9031c22c95f36f060d42b93b880496953b825090
Scoped evidence bindings: 2

[249] Privacy Policy - Quesma. Quesma;
response-sha256:140e90707fe65a6d0aba8a1deac55af6022279854866de61f34388b5e1469d88
https://quesma.com/privacy/
Response SHA-256: 140e90707fe65a6d0aba8a1deac55af6022279854866de61f34388b5e1469d88
Scoped evidence bindings: 1

[250] BlenderBench | Refresh. Refresh;
response-sha256:b2900ee1804bf98405890a9b02b5ae81249e3e34758ff71d06732d026148c8c5
https://refresh.dev/blog/blenderbench
Response SHA-256: b2900ee1804bf98405890a9b02b5ae81249e3e34758ff71d06732d026148c8c5
Scoped evidence bindings: 2

[251] Refresh. Refresh; response-sha256:62c2acc88826f95e5f5f94676c465609df474e0a106ae9be33421f29e94705b3
https://refresh.dev/
Response SHA-256: 62c2acc88826f95e5f5f94676c465609df474e0a106ae9be33421f29e94705b3
Scoped evidence bindings: 3

[252] Terms of Service | Refresh. Refresh;
response-sha256:f3b0b025a49e80523abde67024ef9cfa7180e61b23a9e7aeda698f5d212ecfea
https://refresh.dev/terms
Response SHA-256: f3b0b025a49e80523abde67024ef9cfa7180e61b23a9e7aeda698f5d212ecfea
Scoped evidence bindings: 1

[253] ScienceWorld: Is your Agent Smarter than a 5th Grader?. arXiv (author-deposited preprint); 2203.07540v1
https://arxiv.org/pdf/2203.07540v1
Response SHA-256: c7a60e21ad7bf66a25e04e10d76dc54d26c8dfd2cf776d592554e0b01dc24cde
Scoped evidence bindings: 5

[254] τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. arXiv (author-deposited preprint); 2406.12045v1
https://arxiv.org/pdf/2406.12045v1
Response SHA-256: 0ce66a1763d698c61bb311c3c874bf593d1e9a5bfff11bb35f6f72b981f6da56
Scoped evidence bindings: 5

[255] ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. arXiv (author-deposited preprint); 2307.16789v1
https://arxiv.org/pdf/2307.16789v1
Response SHA-256: 6998d4dc2544f53fc8d1a51e0e76b22e60a31e3fad72bc3991b1d6d26cde6398
Scoped evidence bindings: 5

[256] VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks. arXiv (author-deposited preprint); 2401.13649v1
https://arxiv.org/pdf/2401.13649v1
Response SHA-256: c713cb700d2d6c56e2f6b1a76bb1bbef05e6298de31bd116bfb4a4002ab228b2
Scoped evidence bindings: 5

[257] WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. arXiv (author-deposited preprint); 2207.01206v1
https://arxiv.org/pdf/2207.01206v1
Response SHA-256: 43ac6342de3471d4d7ab94397f583b6d769f52cf85c9daf2ba15e55a4b96afee
Scoped evidence bindings: 5

[258] AgentInstruct (AgentTuning) - README.md. THUDM/AgentTuning; e33a45d7eab2b63cac4d1956da1e6377fca9fcc7
https://raw.githubusercontent.com/THUDM/AgentTuning/e33a45d7eab2b63cac4d1956da1e6377fca9fcc7/README.md
Response SHA-256: ad3e0550e3f0ee4c426f89c60de6ec8d2a99292f37ec33619b3cc11352899ee2
Scoped evidence bindings: 2

[259] AgentInstruct (AgentTuning) - README.md. THUDM/AgentInstruct; e252cf78ced8a0ea5f62cfd591784cdbbddbac8a
https://huggingface.co/datasets/THUDM/AgentInstruct/raw/e252cf78ced8a0ea5f62cfd591784cdbbddbac8a/README.md
Response SHA-256: 3ab41d3ba06250f6c8f703b7b16fce7330b402bed8570a70ce3ccd4a49f11788
Scoped evidence bindings: 4

[260] APPS - README.md. codeparrot/apps; 21e74ddf8de1a21436da12e3e653065c5213e9d1
https://huggingface.co/datasets/codeparrot/apps/raw/21e74ddf8de1a21436da12e3e653065c5213e9d1/README.md
Response SHA-256: bc954bda94e94e9ce92d80ec16d69607444fcdff240cae89df6ca84ff497e846
Scoped evidence bindings: 4

[261] APPS - LICENSE. hendrycks/apps; 362aedc3c71cd7d9bd2fc96a6c80e11dbc38c7a5
https://raw.githubusercontent.com/hendrycks/apps/362aedc3c71cd7d9bd2fc96a6c80e11dbc38c7a5/LICENSE
Response SHA-256: 8f6401fb0b1fbfcac17aac8d61e4e1b0d2de9aa190d8aa655fedc5fd783b2ace
Scoped evidence bindings: 1

[262] Civil Comments / Jigsaw toxicity annotations - tensorflow_datasets/text/civil_comments.py. tensorflow/datasets; 9ada602af8ce7c8122cdd749c80453f637531acb
https://raw.githubusercontent.com/tensorflow/datasets/9ada602af8ce7c8122cdd749c80453f637531acb/tensorflow_datasets/text/civil_comments.py
Response SHA-256: 64c981f437bb045e0a61b0ac2742e710c13dec26274ae796c8fe40d287f0ed59
Scoped evidence bindings: 5

[263] Civil Comments / Jigsaw toxicity annotations - README.md. google/civil_comments; f2970eb3a55777454c94069077cc8d9b5866312d
https://huggingface.co/datasets/google/civil_comments/raw/f2970eb3a55777454c94069077cc8d9b5866312d/README.md
Response SHA-256: 8ce0fee0cce0d54ddd5daaad7ba7cd5ed31d78b8e8f2d5cd2345b0dd015932c6
Scoped evidence bindings: 2

[264] CodeContests - README.md. google-deepmind/code_contests; fa7a4f8139aab08362503f3344778eb86901709a
https://raw.githubusercontent.com/google-deepmind/code_contests/fa7a4f8139aab08362503f3344778eb86901709a/README.md
Response SHA-256: a58e6866fa383903939ae85131ab1d294af2392c6faf9789054dd38fcb23210d
Scoped evidence bindings: 5

[265] CodeContests - contest_problem.proto. google-deepmind/code_contests; fa7a4f8139aab08362503f3344778eb86901709a
https://raw.githubusercontent.com/google-deepmind/code_contests/fa7a4f8139aab08362503f3344778eb86901709a/contest_problem.proto
Response SHA-256: c44121c7a16b7b8f0356a8941cee42bf50fa3eb76d4729006f1b25b9f5ec9903
Scoped evidence bindings: 2

[266] CoT Collection - README.md. kaist-ai/CoT-Collection; c9d352cdc119df4a4f7526d100e4acb4a72a7a5c
https://huggingface.co/datasets/kaist-ai/CoT-Collection/raw/c9d352cdc119df4a4f7526d100e4acb4a72a7a5c/README.md
Response SHA-256: 4f7d853300ca337bcb1b1da64713956ed3c18ccc50e0b9d51aa5c8f87b1e5304
Scoped evidence bindings: 5

[267] CoT Collection - README.md. kaistAI/CoT-Collection; afce9fe8291b25b04b497d9a767f358d053786b1
https://raw.githubusercontent.com/kaistAI/CoT-Collection/afce9fe8291b25b04b497d9a767f358d053786b1/README.md
Response SHA-256: 2198fabac4b189287c65e0f9742d9aa11ad85d342f9cc9a2fc1bc2e1f8426361
Scoped evidence bindings: 2

[268] HotpotQA - public project page. HotpotQA creators;
body-sha256-8786505bfdd092637429ffc0ed7a8e3bf0948b8f7a73c1b665b9a5684d4dc0f9
https://hotpotqa.github.io/blog
Response SHA-256: 8786505bfdd092637429ffc0ed7a8e3bf0948b8f7a73c1b665b9a5684d4dc0f9
Scoped evidence bindings: 2

[269] HotpotQA - README.md. hotpotqa/hotpot; 3635853403a8735609ee997664e1528f4480762a
https://raw.githubusercontent.com/hotpotqa/hotpot/3635853403a8735609ee997664e1528f4480762a/README.md
Response SHA-256: 2a0a1758bb9a2e52b52e6a6528bbf72d46a5a01e0318cdea0473d0660d555998
Scoped evidence bindings: 4

[270] Human Preference Dataset (HPD/HPS) - README.md. tgxs002/HPSv2;
866735ecaae999fa714bd9edfa05aa2672669ee3
https://raw.githubusercontent.com/tgxs002/HPSv2/866735ecaae999fa714bd9edfa05aa2672669ee3/README.md
Response SHA-256: e57b9518ff21030b354b12857d79530808829d5149afaab24ddc8b4af0ec2e19
Scoped evidence bindings: 6

[271] Human Preference Dataset (HPD/HPS) - README.md. ymhao/HPDv2;
bb0f2563e451e11c3f376b2ba93615db00e042e1
https://huggingface.co/datasets/ymhao/HPDv2/raw/bb0f2563e451e11c3f376b2ba93615db00e042e1/README.md
Response SHA-256: 09f4bab376cedcb1fa165a81a31afccb5ee2e816402970adba3c9ffe795b238b
Scoped evidence bindings: 1

[272] ImageRewardDB - README.md. THUDM/ImageRewardDB; c493721e19e296eb615420036f2a2eed08412bb4
https://huggingface.co/datasets/THUDM/ImageRewardDB/raw/c493721e19e296eb615420036f2a2eed08412bb4/README.md
Response SHA-256: 2965f9f966d8aa56ebb070d7b92bf83ea4072a9e34011eddf5e3fcdc690c6b44
Scoped evidence bindings: 6

[273] MuSiQue - README.md. stonybrooknlp/musique; 922ac98f19a201998dbdae6d7f2887a5258dbdeb
https://raw.githubusercontent.com/stonybrooknlp/musique/922ac98f19a201998dbdae6d7f2887a5258dbdeb/README.md
Response SHA-256: 7a3fb6a960f3fa122d2a1a81b73b9cf5a39ac983b11735d158490947960b9359
Scoped evidence bindings: 6

[274] MuSiQue - original dataset construction methods. TACL / MuSiQue authors;
ACL-2022.tacl-1.31-body-sha256-c0d8b89636bcbca4972b55ab68f014eb20f74c4b9a4fcf126ca68d7f2ec1e753
https://aclanthology.org/2022.tacl-1.31.pdf
Response SHA-256: c0d8b89636bcbca4972b55ab68f014eb20f74c4b9a4fcf126ca68d7f2ec1e753
Scoped evidence bindings: 2

[275] OpenR1-Math - README.md. open-r1/OpenR1-Math-220k; e4e141ec9dea9f8326f4d347be56105859b2bd68
https://huggingface.co/datasets/open-r1/OpenR1-Math-220k/raw/e4e141ec9dea9f8326f4d347be56105859b2bd68/README.md
Response SHA-256: 87a0390ba632618254d50c1614feb6a5a7803a28710118e1705bab18bda5e64e
Scoped evidence bindings: 7

[276] PRISM Alignment Dataset - README.md. HannahRoseKirk/prism-alignment; 18ab5cfb37456f4ec8cbc00212ce54cf7b1239f6
https://huggingface.co/datasets/HannahRoseKirk/prism-alignment/raw/18ab5cfb37456f4ec8cbc00212ce54cf7b1239f6/README.md
Response SHA-256: 5cfc798de48c958f16605743101c9aee4fc52a5c45a591eed2bb09b417349199
Scoped evidence bindings: 7

[277] ProsocialDialog - README.md. allenai/prosocial-dialog; d77d7ad3c624c51030f2f32c83e892b3d620b3d4
https://huggingface.co/datasets/allenai/prosocial-dialog/raw/d77d7ad3c624c51030f2f32c83e892b3d620b3d4/README.md
Response SHA-256: d472c5e3315b789e545eddc0d21adb4347658ccfa830402ade71446c5472b050
Scoped evidence bindings: 6

[278] ProsocialDialog - README.md. skywalker023/prosocial-dialog; 77a5e2fd363af976c06c4408fc83cd4bb302944d
https://raw.githubusercontent.com/skywalker023/prosocial-dialog/77a5e2fd363af976c06c4408fc83cd4bb302944d/README.md
Response SHA-256: dd8cf4902ab5b9e21a147c81521c1a656e60318ea126af60c009edd11f671d21
Scoped evidence bindings: 1

[279] RichHF-18K - README.md. google-research-datasets/richhf-18k; 65290d3ce0e8c9cd407e25e54574cad8589353b6
https://raw.githubusercontent.com/google-research-datasets/richhf-18k/65290d3ce0e8c9cd407e25e54574cad8589353b6/README.md
Response SHA-256: 2e407495e779761d24073fa3d0d6d2f7daf6cdd5f608f2d41142ea5e35944635
Scoped evidence bindings: 6

[280] Social Bias Inference Corpus (SBIC) - README.md. allenai/social_bias_frames; a73510df3ec15f488636be8b66292077719aaa55
https://huggingface.co/datasets/allenai/social_bias_frames/raw/a73510df3ec15f488636be8b66292077719aaa55/README.md
Response SHA-256: d75fdaf98e26e6e96f55819d25f0c5741a995942473412156724bf54239986a0
Scoped evidence bindings: 7

[281] TabMWP - README.md. lupantech/PromptPG; 5a1a52214521590b075f545b76a4f5ce666345e3
https://raw.githubusercontent.com/lupantech/PromptPG/5a1a52214521590b075f545b76a4f5ce666345e3/README.md
Response SHA-256: dd61c99cca31e38ac48341f8daabe5b834187f76c37ca44ca5f57ff35f32a340
Scoped evidence bindings: 5

[282] TabMWP - public project page. PromptPG creators; body-sha256-cc663a9a827d07b4052940a74fa59dc6b1f73f39cc13d184e45800e45733f1b3
https://promptpg.github.io/
Response SHA-256: cc663a9a827d07b4052940a74fa59dc6b1f73f39cc13d184e45800e45733f1b3
Scoped evidence bindings: 1

[283] TACO - README.md. BAAI/TACO; d593ed0a2becbbc952230bb89be09189bf1056dc
https://huggingface.co/datasets/BAAI/TACO/raw/d593ed0a2becbbc952230bb89be09189bf1056dc/README.md
Response SHA-256: 4d087f7330c80401684459d8b59626d88cdfc5cd36a34c9b39a5cc7241be9b34
Scoped evidence bindings: 6

[284] ToxicChat - README.md. lmsys/toxic-chat; 29df8e4dba60e1f4af4b4075c0705c5b313548a8
https://huggingface.co/datasets/lmsys/toxic-chat/raw/29df8e4dba60e1f4af4b4075c0705c5b313548a8/README.md
Response SHA-256: 70c63cac8fff2a4934c85b3d3b52d2e169b8af9fe893dd9de1ea5e4abdb51441
Scoped evidence bindings: 7

[285] Hire Senior RL & Environments Engineers | A.Team. A.Team;
response-sha256:7f364b55e5b5caba2342075f21dfd1b93aca81602621fe813af58e705d49e15f
https://www.a.team/hire-rl-engineers
Response SHA-256: 7f364b55e5b5caba2342075f21dfd1b93aca81602621fe813af58e705d49e15f
Scoped evidence bindings: 5

[286] Home | AIXI Labs. AIXI Labs;
response-sha256:7233b124d8fce21906f7520e9f2ed26c5fad8b6cb4d4200dcbacc8377cfa663a
https://www.aixi.uk/
Response SHA-256: 7233b124d8fce21906f7520e9f2ed26c5fad8b6cb4d4200dcbacc8377cfa663a
Scoped evidence bindings: 3

[287] Overview | AIXI Labs. AIXI Labs;
response-sha256:626714452bdba98703697793ce11664baabf1e60043733c344736c07a1244eb2
https://www.aixi.uk/research/
Response SHA-256: 626714452bdba98703697793ce11664baabf1e60043733c344736c07a1244eb2
Scoped evidence bindings: 2

[288] Alinia - Deploy your AI without breaking rules. Alinia AI;
response-sha256:4f1c713b5af4ce21fd68d40b86dec3a21c15cd1e7d6d47d95ec33c866fd8ea0e
https://www.alinia.ai/
Response SHA-256: 4f1c713b5af4ce21fd68d40b86dec3a21c15cd1e7d6d47d95ec33c866fd8ea0e
Scoped evidence bindings: 4

[289] Alinia Privacy Policy - How we protect your data. Alinia AI;
response-sha256:84f15c9004a9b11ba9d051443d1c2b917ae59877b34dd8754ba8991f44e7ffd5
https://www.alinia.ai/privacy-policy
Response SHA-256: 84f15c9004a9b11ba9d051443d1c2b917ae59877b34dd8754ba8991f44e7ffd5
Scoped evidence bindings: 1

[290] APTO | AI Training Data for LLM, Multimodal & Physical AI. APTO;
response-sha256:706156597d85151553cfbd490f4943c42d0de3dfd0296d0d8477f5fbc65c6825
https://apto.co.jp/en/
Response SHA-256: 706156597d85151553cfbd490f4943c42d0de3dfd0296d0d8477f5fbc65c6825
Scoped evidence bindings: 4

[291] Bespoke Labs: Ship Reliable AI Agents with SOTA Data Curation. Bespoke Labs;
response-sha256:3ba83e61316e7ceba1a5356e9a8c5310c21fdb040d3be4f9880fd985b033d869
https://bespokelabs.ai/
Response SHA-256: 3ba83e61316e7ceba1a5356e9a8c5310c21fdb040d3be4f9880fd985b033d869
Scoped evidence bindings: 4

[292] Terms of Use · Bespoke Labs. Bespoke Labs;
response-sha256:13305d1ae7ae8e20ebde4a81ea2eb92f718321c45dcd86d99d05e93a67eeabb9
https://bespokelabs.ai/legal/terms-of-use
Response SHA-256: 13305d1ae7ae8e20ebde4a81ea2eb92f718321c45dcd86d99d05e93a67eeabb9
Scoped evidence bindings: 1

[293] Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming. arXiv (author-deposited versioned preprint); 2501.18837v1
https://arxiv.org/pdf/2501.18837v1
Response SHA-256: 9eeb537ebf4bf47a76284f5c82fe37c6c71f5ca57d93dedc7a9a269beb3c1a7e
Scoped evidence bindings: 7

[294] DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv (author-deposited versioned preprint); 2402.03300v1
https://arxiv.org/pdf/2402.03300v1
Response SHA-256: 500ce2fae51dd4e06d27423c9b8ec6ca2013eb577e74e5e15fd87510018c6e53
Scoped evidence bindings: 7

[295] DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv (author-deposited versioned preprint); 2501.12948v1
https://arxiv.org/pdf/2501.12948v1
Response SHA-256: 52d8ca3ac93e88cef9944e1fd03b0e04aec5954495a8250fb2fadf8fa20a4dad
Scoped evidence bindings: 6

[296] DeepSeek-V3 Technical Report. arXiv (author-deposited versioned preprint); 2412.19437v1
https://arxiv.org/pdf/2412.19437v1
Response SHA-256: a67e18a4ca6e2c81abc030239a3ee53705f656b79efab8ce1d6b5ac25630c2a5
Scoped evidence bindings: 6

[297] Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv (author-deposited versioned preprint); 2412.16339v1
https://arxiv.org/pdf/2412.16339v1
Response SHA-256: ce80073ea89c3d3ec419fb01c692f3a818fa28ae7709da951820497242f606f7
Scoped evidence bindings: 7

[298] Privacy Policy - Elorian. Elorian;
response-sha256:26d478005d324859eabada8d7e7bf4449025cda01940a165497b8513ad650186
https://elorian.ai/privacy
Response SHA-256: 26d478005d324859eabada8d7e7bf4449025cda01940a165497b8513ad650186
Scoped evidence bindings: 1

[299] Introducing Glimpse - Elorian. Elorian;
response-sha256:ae94606ae312346d7845660314ea3b6211272d54c08af5a032b9d87b6637d8a2
https://elorian.ai/blog/introducing-glimpse
Response SHA-256: ae94606ae312346d7845660314ea3b6211272d54c08af5a032b9d87b6637d8a2
Scoped evidence bindings: 3

[300] Elorian - Frontier Intelligence. Elorian;
response-sha256:ad8d6b8092dad35a18a8b229e07d922b2f63773f6db7ea89a21bd926c60315d4
https://elorian.ai/
Response SHA-256: ad8d6b8092dad35a18a8b229e07d922b2f63773f6db7ea89a21bd926c60315d4
Scoped evidence bindings: 1

[301] Continuous Red Teaming v2026. Giskard;
response-sha256:51fa260e12b111987c760305ed161de53481901eb31d9881232f2f3aeb0c57c4
https://www.giskard.ai/products/continuous-red-teaming
Response SHA-256: 51fa260e12b111987c760305ed161de53481901eb31d9881232f2f3aeb0c57c4
Scoped evidence bindings: 1

[302] LLM Evaluation Platform | Giskard AI Agent Testing. Giskard;
response-sha256:bfab0f50fdb6d40fa09a5d4202254064c4a9212551315c6406b7eb260385744f
https://www.giskard.ai/products/llm-evaluation
Response SHA-256: bfab0f50fdb6d40fa09a5d4202254064c4a9212551315c6406b7eb260385744f
Scoped evidence bindings: 3

[303] AI Red Teaming & LLM Security Platform | Giskard. Giskard;
response-sha256:cc248c0b2d7cc09d50d9876fc3b9c6a7bcbad22458722a1c30dcb38d255022d7
https://www.giskard.ai/
Response SHA-256: cc248c0b2d7cc09d50d9876fc3b9c6a7bcbad22458722a1c30dcb38d255022d7
Scoped evidence bindings: 1

[304] GPT-4 Technical Report. arXiv (author-deposited versioned preprint); 2303.08774v1
https://arxiv.org/pdf/2303.08774v1
Response SHA-256: 053056a10114d22e4c47b6b5be25e54c320b5f1beeae7466e8638dac0f5f5f66
Scoped evidence bindings: 6

[305] HackerRank Bench | Long-horizon evaluation for frontier agents. HackerRank Bench;
response-sha256:c50227b4688a9e11850631e587d959923580fd9bf11f5e56af3027faf0848752
https://www.hackerrank.com/bench
Response SHA-256: c50227b4688a9e11850631e587d959923580fd9bf11f5e56af3027faf0848752
Scoped evidence bindings: 5

[306] Learning from Natural Language Feedback. arXiv (author-deposited versioned preprint); 2204.14146v1
https://arxiv.org/pdf/2204.14146v1
Response SHA-256: 7f0316658554df7a538ca89a6e44438c36cf31c9bc76b0246964b42f4824ee0d
Scoped evidence bindings: 8

[307] Laude Institute // Ship Your Research. Laude Institute / Terminal-Bench;
response-sha256:f19fddb1745d05102a744c883d89f2c2f607c9641d23353739358aa432d02763
https://www.laude.org/about
Response SHA-256: f19fddb1745d05102a744c883d89f2c2f607c9641d23353739358aa432d02763
Scoped evidence bindings: 3

[308] Terms of Service // Laude. Laude Institute / Terminal-Bench;
response-sha256:b2071a240fd98c9d1d01cedc8db721e0bf480832f17104af1c3d7ceaf8145cac
https://www.laude.org/terms
Response SHA-256: b2071a240fd98c9d1d01cedc8db721e0bf480832f17104af1c3d7ceaf8145cac
Scoped evidence bindings: 2

[309] Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv (author-deposited versioned preprint); 2307.09288v1
https://arxiv.org/pdf/2307.09288v1
Response SHA-256: 03c022db7d7ff846e0a1ada4c8be936416f7f7f9c680c64d9494d7b30ebfdd80
Scoped evidence bindings: 6

[310] The Llama 3 Herd of Models. arXiv (author-deposited versioned preprint); 2407.21783v1
https://arxiv.org/pdf/2407.21783v1
Response SHA-256: d7df3d6f537f025037f32ebe13198e3ea5500ea3e0967abdcceb80e31262d386
Scoped evidence bindings: 7

[311] Ooak Data - Applied AI Research Lab. Ooak Data;
response-sha256:ee41c3b90be07d56a848a933911654b7ac8e620a8e3740baac09ba5ba613dd00
https://ooakdata.com/
Response SHA-256: ee41c3b90be07d56a848a933911654b7ac8e620a8e3740baac09ba5ba613dd00
Scoped evidence bindings: 2

[312] How we protect a company's history when training AI - Ooak Data Research. Ooak Data;
response-sha256:9929aa558f4fabc93cc658e13cef12e156b47849a6ba4825699935d64099f41d
https://ooakdata.com/research/how-we-protect-company-history-when-training-ai
Response SHA-256: 9929aa558f4fabc93cc658e13cef12e156b47849a6ba4825699935d64099f41d
Scoped evidence bindings: 1

[313] Privacy Policy - Ooak Data. Ooak Data;
response-sha256:e863478b7f92280d22b108d6d0a94f61fc5a1c9e818554b90a4ec1e1a18a83b7
https://ooakdata.com/privacy
Response SHA-256: e863478b7f92280d22b108d6d0a94f61fc5a1c9e818554b90a4ec1e1a18a83b7
Scoped evidence bindings: 1

[314] Terms of Service - Ooak Data. Ooak Data;
response-sha256:3deaf0844fb2b61171cc2be615b7f9412122c78b452ea93c24a33e8e36389938
https://ooakdata.com/terms
Response SHA-256: 3deaf0844fb2b61171cc2be615b7f9412122c78b452ea93c24a33e8e36389938
Scoped evidence bindings: 1

[315] RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment. arXiv (author-deposited versioned preprint); 2304.06767v1
https://arxiv.org/pdf/2304.06767v1
Response SHA-256: 2fdceb79aacdbd845e28af3c336ad3e38e399a2009d0ef5a5cb8a06b5fc8d5a1
Scoped evidence bindings: 6

[316] REBEL: Reinforcement Learning via Regressing Relative Rewards. arXiv (author-deposited versioned preprint); 2404.16767v1
https://arxiv.org/pdf/2404.16767v1
Response SHA-256: cddccef39dde4d25ded91ef5d46e99dd27e313792f92bfa97c583213d3ae0d30
Scoped evidence bindings: 6

[317] ReMax: A Simple, Effective, and Efficient Method for Aligning Large Language Models. arXiv (author-deposited versioned preprint); 2310.10505v1
https://arxiv.org/pdf/2310.10505v1
Response SHA-256: 72f3579fc27190ba9206ce9945bdb396ef6058c0c072b15edbc5fd2482eea234
Scoped evidence bindings: 7

[318] Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. arXiv (author-deposited versioned preprint); 2402.14740v1
https://arxiv.org/pdf/2402.14740v1
Response SHA-256: ce67f21ed19ca1084e41539e6798ce91ac7fef5483dd029ed2bc719f68e91611
Scoped evidence bindings: 6

[319] RRHF: Rank Responses to Align Language Models with Human Feedback without tears. arXiv (author-deposited versioned preprint); 2304.05302v1
https://arxiv.org/pdf/2304.05302v1
Response SHA-256: 9ae8beae125e30250c9ff596fcb5b72780adbdbf30be296644b78723bf9ea716
Scoped evidence bindings: 6

[320] Statistical Rejection Sampling Improves Preference Optimization. arXiv (author-deposited versioned preprint); 2309.06657v1
https://arxiv.org/pdf/2309.06657v1
Response SHA-256: c0c82da5528a7b18643dad2052f7ba2544ec74f0ba76dbafaece284ae06823eb
Scoped evidence bindings: 7

[321] Data Annotation and Data Labelling, TrainAI - RWS. RWS / TrainAI;
response-sha256:60998fb956a60ffe0ea2971d583749814ff0808cb45868b7014fa7aa4a6d24e5
https://www.rws.com/artificial-intelligence/train-ai-data-services/data-annotation-and-labeling/
Response SHA-256: 60998fb956a60ffe0ea2971d583749814ff0808cb45868b7014fa7aa4a6d24e5
Scoped evidence bindings: 1

[322] About Train AI - FAQ - TrainAI - RWS Community. RWS / TrainAI;
response-sha256:1a9d36be13cba265451910e5a6500d7a98da3ed2fbd238d3ab840cf76d4e26d3
https://community.rws.com/community-hub/resources/rws-services-train-ai-others/trainai/w/faq/6806/about-train-ai
Response SHA-256: 1a9d36be13cba265451910e5a6500d7a98da3ed2fbd238d3ab840cf76d4e26d3
Scoped evidence bindings: 2

[323] RWS expands AI training data services with launch of TrainAI. RWS / TrainAI;
response-sha256:ab05244659fea82e3a0e23150391007494e67e003c771b706e4d18abd9f348e8
https://www.rws.com/about/news/2023/launches-train-ai/
Response SHA-256: ab05244659fea82e3a0e23150391007494e67e003c771b706e4d18abd9f348e8
Scoped evidence bindings: 1

[324] Human-in-the-Loop Data Validation, TrainAI - RWS. RWS / TrainAI;
response-sha256:d3f5260fa752b621be29fb46e8650bb811f3f7e9317dae095e4b4d4eecdd88ef
https://www.rws.com/artificial-intelligence/train-ai-data-services/human-in-the-loop-data-validation/
Response SHA-256: d3f5260fa752b621be29fb46e8650bb811f3f7e9317dae095e4b4d4eecdd88ef
Scoped evidence bindings: 1

[325] SimPO: Simple Preference Optimization with a Reference-Free Reward. arXiv (author-deposited versioned preprint); 2405.14734v1
https://arxiv.org/pdf/2405.14734v1
Response SHA-256: e295b0f6068088631540502988 29c782647c38dbc2d559fd63bb8fe5c21da0d8
Scoped evidence bindings: 6

[326] SLiC-HF: Sequence Likelihood Calibration with Human Feedback. arXiv (author-deposited versioned preprint); 2305.10425v1
https://arxiv.org/pdf/2305.10425v1
Response SHA-256: fb093c095c211c3d75edf75e7c355573e8772ea0a67f2b0c6720d849f7b44af9
Scoped evidence bindings: 6

[327] Self-Play Preference Optimization for Language Model Alignment. arXiv (author-deposited versioned preprint); 2405.00675v1
https://arxiv.org/pdf/2405.00675v1
Response SHA-256: afd1452ae311176bab821a7386048af15390f7233b95f7ca27f8739e739655f7
Scoped evidence bindings: 7

[328] Step-DPO: Step-wise Preference Optimization for Long-chain Reasoning of LLMs. arXiv (author-deposited versioned preprint); 2406.18629v1
https://arxiv.org/pdf/2406.18629v1
Response SHA-256: b3950e36f4fadb2e4acd69477a8fd410f5a36b98422d21da4fbeb9f37e2fd4fb
Scoped evidence bindings: 6

[329] For Researchers | Terac. Terac;
response-sha256:d11fef543f4e98e1b06f21685ac00c76dbb3d651fd99b01f30137b6aaeade1da
https://terac.com/researchers
Response SHA-256: d11fef543f4e98e1b06f21685ac00c76dbb3d651fd99b01f30137b6aaeade1da
Scoped evidence bindings: 3

[330] Data Partnerships | Turn Your Business Workflows into Revenue | Terac. Terac;
response-sha256:ce2f1ef66cef88f8ff0770e20cf5df2a1b2ef7b2bb54b90834baf52204364194
https://terac.com/data-partnerships
Response SHA-256: ce2f1ef66cef88f8ff0770e20cf5df2a1b2ef7b2bb54b90834baf52204364194
Scoped evidence bindings: 2

[331] Trajectory Labs, PBC | AI safety evals, red teaming, RL environments. Trajectory Labs, PBC;
response-sha256:173c525018408f07cd5445c7165cea3fa9b7620db0ac17eff4b176f3b06be6aa
https://www.trajectorylabs.com/
Response SHA-256: 173c525018408f07cd5445c7165cea3fa9b7620db0ac17eff4b176f3b06be6aa
Scoped evidence bindings: 4

[332] Privacy policy | Trajectory Labs, PBC. Trajectory Labs, PBC;
response-sha256:729207139834bab601560e2a6bc1af9bd3a821db7c1ead01aa6b90de61bb7464
https://www.trajectorylabs.com/privacy
Response SHA-256: 729207139834bab601560e2a6bc1af9bd3a821db7c1ead01aa6b90de61bb7464
Scoped evidence bindings: 1

[333] WPO: Enhancing RLHF with Weighted Preference Optimization. arXiv (author-deposited versioned preprint);
2406.11827v1
https://arxiv.org/pdf/2406.11827v1
Response SHA-256: 3ab2d0145eaeca4eb1d8311dcff2eb30af724e39ae3261d57e54f1a817e86755
Scoped evidence bindings: 6

[334] TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. arXiv (author-deposited versioned preprint); 2412.14161v1; 2024-12-18
https://arxiv.org/pdf/2412.14161v1
Response SHA-256: 3fd93217ad441ee1dbd3f7195d012093751c6feae50b2547c2891643255b4d15
Scoped evidence bindings: 5

[335] benchflow-07. BenchFlow;
response-sha256:a219db1e4b81ca5ad3b4ebda64993d1f3edd47099a1041806f965c055e17ba1c
https://raw.githubusercontent.com/benchflow-ai/benchflow/a61a26f4db4d566de16928565ea58abf3dbaf064/LICENSE
Response SHA-256: a219db1e4b81ca5ad3b4ebda64993d1f3edd47099a1041806f965c055e17ba1c
Scoped evidence bindings: 1

[336] benchflow-06. BenchFlow;
response-sha256:d033efcc20df1dbc9d6249c877141e61fc5117c788ac06f17ebe385fd9076757
https://raw.githubusercontent.com/benchflow-ai/benchflow/a61a26f4db4d566de16928565ea58abf3dbaf064/README.md
Response SHA-256: d033efcc20df1dbc9d6249c877141e61fc5117c788ac06f17ebe385fd9076757
Scoped evidence bindings: 4

[337] BenchGuard: Who Guards the Benchmarks? Automated Auditing of LLM Agent Benchmarks. arXiv (author-deposited versioned preprint); 2604.24955v1; 2026-04-27
https://arxiv.org/pdf/2604.24955v1
Response SHA-256: 72ea4e0ebe96f5d9f9f24693e1bb16876f559aa2ba2801f8506e4a68de5fd424
Scoped evidence bindings: 8

[338] BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. arXiv (author-deposited versioned preprint); 2504.12516v1; 2025-04-16
https://arxiv.org/pdf/2504.12516v1
Response SHA-256: 8de2e7a024b43ff7fbab29e0bb6dbdc131ad07206e6c186c2122502d69477760
Scoped evidence bindings: 14

[339] Data Annotation and Labeling Services for Enterprise AI Models (https://www.cogitotech.com/data-labeling/). Cogito Tech; web-extraction-sha256:dc4fca4750d8b97f168c345c2b43e11fe6495c8bea06a5fbf9bc2970bc4ed409
https://www.cogitotech.com/data-labeling/
Response SHA-256: dc4fca4750d8b97f168c345c2b43e11fe6495c8bea06a5fbf9bc2970bc4ed409
Scoped evidence bindings: 4

[340] Register Now - Cogitotech (https://www.cogitotech.com/register-now/). Cogito Tech;
web-extraction-sha256:6e073ec6c1db4cbf211e84f2ecb33e86f8e9f3b8c31dea3fe276546711ec5cc8
https://www.cogitotech.com/register-now/
Response SHA-256: 6e073ec6c1db4cbf211e84f2ecb33e86f8e9f3b8c31dea3fe276546711ec5cc8
Scoped evidence bindings: 1

[341] Human Feedback | Dataloop. Dataloop;
response-sha256:dd10226bbb83bd5d6647445d234d536c97bad3345cd5a77026c3a87092bfeb87
https://dataloop.ai/platform/human-feedback/
Response SHA-256: dd10226bbb83bd5d6647445d234d536c97bad3345cd5a77026c3a87092bfeb87
Scoped evidence bindings: 4

[342] Privacy Notice | Dataloop. Dataloop;
response-sha256:d91d5f6d2f80736bef295d4a6fd42ef8d32c8ee9847dd42e6cecf6712b01b6a0
https://dataloop.ai/privacy-policy/
Response SHA-256: d91d5f6d2f80736bef295d4a6fd42ef8d32c8ee9847dd42e6cecf6712b01b6a0
Scoped evidence bindings: 1

[343] Datasaur Terms of Service. Datasaur;
response-sha256:d79109156cde4979808c22445140f44943f43e01f02c044b593a8fec4bbc9dfd
https://datasaur.ai/terms-of-service
Response SHA-256: d79109156cde4979808c22445140f44943f43e01f02c044b593a8fec4bbc9dfd
Scoped evidence bindings: 2

[344] About NLP Labeling | Datasaur. Datasaur;
response-sha256:0e1301e572597808e07dd7d6f7ac0e194c9624910e84b8c29958f2745513b357
https://datasaur.ai/studio
Response SHA-256: 0e1301e572597808e07dd7d6f7ac0e194c9624910e84b8c29958f2745513b357
Scoped evidence bindings: 3

[345] Annotate as a Service | Deepen AI. Deepen AI;
response-sha256:467899df84c51b09240f0648471bb47510fbaad15dc66fe20798148664a6ed73
https://www.deepen.ai/services/annotate/
Response SHA-256: 467899df84c51b09240f0648471bb47510fbaad15dc66fe20798148664a6ed73
Scoped evidence bindings: 4

[346] Data Protection Addendum | Deepen AI. Deepen AI;
response-sha256:6dd06c690724ab4e1ded9c47341d7926d427bcc40f24eba099f4ba9a2cbdde93
https://www.deepen.ai/data-protection/
Response SHA-256: 6dd06c690724ab4e1ded9c47341d7926d427bcc40f24eba099f4ba9a2cbdde93
Scoped evidence bindings: 1

[347] AI Data Solutions & Model Fine Tuning Company | DDD. Digital Divide Data;
response-sha256:17abdc61a4c10e2a87c81e4acf6dc624ec85bc0fa41906a8862de200475a70d0
https://www.digitaldividedata.com/
Response SHA-256: 17abdc61a4c10e2a87c81e4acf6dc624ec85bc0fa41906a8862de200475a70d0
Scoped evidence bindings: 3

[348] Terms Of Use - Digitaldividedata.com. Digital Divide Data;
response-sha256:6e4ba2d23425fe4fb219c201c887c42c694cae694989bbddd9d850893ac5ac89
https://www.digitaldividedata.com/terms-of-use
Response SHA-256: 6e4ba2d23425fe4fb219c201c887c42c694cae694989bbddd9d850893ac5ac89
Scoped evidence bindings: 1

[349] Data Annotation Solutions For AI & ML Models | DDD. Digital Divide Data;
response-sha256:e916c9379943948fce543d65a30b5a0f5a072199554e932e06646d9b6df17efa
https://www.digitaldividedata.com/data-annotation-solutions
Response SHA-256: e916c9379943948fce543d65a30b5a0f5a072199554e932e06646d9b6df17efa
Scoped evidence bindings: 2

[350] GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks. arXiv (author-deposited versioned preprint); 2510.04374v1; 2025-10-05
https://arxiv.org/pdf/2510.04374v1
Response SHA-256: a04cca8451c486d871f8d2c6c57b8bc03b45daf1b5b0269146929bc8bf688dec
Scoped evidence bindings: 18

[351] GPQA: A Graduate-Level Google-Proof Q&A Benchmark. arXiv (author-deposited versioned preprint); 2311.12022v1; 2023-11-20
https://arxiv.org/pdf/2311.12022v1
Response SHA-256: 58a85a21b653a01d5794f84952936f58e6a5218d781588f44ae183bb749bf023
Scoped evidence bindings: 5

[352] Terms of Use | Deploy. Deploy / Hive;
response-sha256:3fbfefc60b83567fb85eeea584e31e71775e34952bb9ca3a2603895d13919847
https://deploy.com/terms-of-use
Response SHA-256: 3fbfefc60b83567fb85eeea584e31e71775e34952bb9ca3a2603895d13919847
Scoped evidence bindings: 4

[353] AI to Understand, Search, and Generate Content | Hive. Deploy / Hive;
response-sha256:293b7cc6b2de23fcc64ea6e12d30d92c0ee453e90c8a78bc19eea591db9d2a17
https://thehive.ai/
Response SHA-256: 293b7cc6b2de23fcc64ea6e12d30d92c0ee453e90c8a78bc19eea591db9d2a17
Scoped evidence bindings: 1

[354] Humanity's Last Exam. arXiv (author-deposited versioned preprint); 2501.14249v1; 2025-01-24
https://arxiv.org/pdf/2501.14249v1
Response SHA-256: 7a91dd3ee597fd64f44053675bc87dfa308b8da538e5e8e48ea115db9db22d8f
Scoped evidence bindings: 5

[355] Instruction-Following Evaluation for Large Language Models. arXiv (author-deposited versioned preprint); 2311.07911v1; 2023-11-14
https://arxiv.org/pdf/2311.07911v1
Response SHA-256: 5d7cedfd762af6d9ad13d8768de64fa65895b8ebc3f5e454454c5ffbd48124f2
Scoped evidence bindings: 5

[356] AI Data Services | iSoftStone. iSoftStone;
response-sha256:edcc4b7486af9db839f150a11e3aa67c62d9aac51515cfbf8ab02eadfa67ae4a
https://www.isoftstoneinc.com/ai-data-services/
Response SHA-256: edcc4b7486af9db839f150a11e3aa67c62d9aac51515cfbf8ab02eadfa67ae4a
Scoped evidence bindings: 4

[357] Privacy Policy | iSoftStone. iSoftStone;
response-sha256:99878b0150530e1a4a31feee69f2fede0b47586d5e7ee99e2f3c9485ebc41dbf
https://www.isoftstoneinc.com/privacy-policy/
Response SHA-256: 99878b0150530e1a4a31feee69f2fede0b47586d5e7ee99e2f3c9485ebc41dbf
Scoped evidence bindings: 1

[358] Quality-First AI Evaluation and Post-Training | Kili. Kili Technology;
response-sha256:e4f0a3493d09bbf71c1ae0b2494d1b0284a4dff7c847e7747b902322db99e313
https://kili-technology.com/product/quality
Response SHA-256: e4f0a3493d09bbf71c1ae0b2494d1b0284a4dff7c847e7747b902322db99e313
Scoped evidence bindings: 1

[359] AI Evaluation, Post-Training and Data Annotation Platform | Kili. Kili Technology;
response-sha256:74abc99e07361cc7ba3ad07be9f3b72c5ff00de7073f9e5d967403aab4a7420d
https://kili-technology.com/
Response SHA-256: 74abc99e07361cc7ba3ad07be9f3b72c5ff00de7073f9e5d967403aab4a7420d
Scoped evidence bindings: 3

[360] Terms and Conditions. Kili Technology;
response-sha256:89e7c218c76aeb3b5c6425173c6ff608c5f37f8df5a5d9c5d7114f124ad562ed
https://kili-technology.com/terms-and-conditions
Response SHA-256: 89e7c218c76aeb3b5c6425173c6ff608c5f37f8df5a5d9c5d7114f124ad562ed
Scoped evidence bindings: 1

[361] Privacy Policy. Kili Technology;
response-sha256:924b3f336da026b5c0d2deb5c2c8a2867700de55ff8f9689533d98f96c9ae804
https://kili-technology.com/legal-notice
Response SHA-256: 924b3f336da026b5c0d2deb5c2c8a2867700de55ff8f9689533d98f96c9ae804
Scoped evidence bindings: 1

[362] Häufig gestellte Fragen zum Outsourcing | Mindy Support (https://mindy-support.com/about-us/faq/). Mindy Support;
web-extraction-sha256:7669e5eba0efaa93d1296034977ed1d65f2ef3f328e23455f53c50447114c583
https://mindy-support.com/about-us/faq/
Response SHA-256: 7669e5eba0efaa93d1296034977ed1d65f2ef3f328e23455f53c50447114c583
Scoped evidence bindings: 5

[363] MLE-bench: Evaluating Machine Learning Agents on Machine Learning Engineering. arXiv (author-deposited versioned preprint); 2410.07095v1; 2024-10-09
https://arxiv.org/pdf/2410.07095v1
Response SHA-256: f5f904e91f28c460397a0d7acbeb40aba19bae561093effd71c7af32cf8b7539
Scoped evidence bindings: 6

[364] MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. arXiv (author-deposited versioned preprint); 2409.02813v1; 2024-09-04
https://arxiv.org/pdf/2409.02813v1
Response SHA-256: a7175191c00a58f7ab5f5c7db047c8e50d7707fb1cf6fe52cb77cbfe9de4acbd
Scoped evidence bindings: 5

[365] PaperBench: Evaluating AI's Ability to Replicate AI Research. arXiv (author-deposited versioned preprint); 2504.01848v1; 2025-04-02
https://arxiv.org/pdf/2504.01848v1
Response SHA-256: 4f2f4a48dd27a1790ec322df759f5e319ece547227fc1b7d307fb36c10d31e79
Scoped evidence bindings: 6

[366] RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts. arXiv (author-deposited versioned preprint); 2411.15114v1; 2024-11-22
https://arxiv.org/pdf/2411.15114v1
Response SHA-256: be1b2037a29da00ec2b6964258e50060a3ff4504703cf4835554cba8f47ed05e
Scoped evidence bindings: 8

[367] RewardBench 2: Advancing Reward Model Evaluation. arXiv (author-deposited versioned preprint); 2506.01937v1; 2025-06-02
https://arxiv.org/pdf/2506.01937v1
Response SHA-256: f6d5781356315a7c33eff8b4810d876961861d7b3494037d31acd36496b43830
Scoped evidence bindings: 7

[368] ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. arXiv (author-deposited versioned preprint); 2410.05080v1; 2024-10-07
https://arxiv.org/pdf/2410.05080v1
Response SHA-256: 5b4b8be139d6f88d53acaf89b18938d0c1749fc20f5d50db6989324772d993bc
Scoped evidence bindings: 8

[369] SkillsBench: Benchmarking How Well Agent Skills Work Across Diverse Tasks. arXiv (author-deposited versioned preprint); 2602.12670v1; 2026-02-13
https://arxiv.org/pdf/2602.12670v1
Response SHA-256: fa853b3a44282a28bd816b46572e8fbab945834f01fb5f055ec8847ba1c31dae
Scoped evidence bindings: 11

[370] AI Leading Data Provider | AI Data Solutions | DataOcean AI. Dataocean AI / Speechocean;
response-sha256:61aca77dd331982f847f39c81620ce28ca47d0d69586be82d774ea908469024f
https://dataoceanai.com/
Response SHA-256: 61aca77dd331982f847f39c81620ce28ca47d0d69586be82d774ea908469024f
Scoped evidence bindings: 2

[371] 登录. Dataocean AI / Speechocean;
response-sha256:23205ecaaa52cc56ef62cd411ff69e25c830e41c6db6cba4ce487c78d1141f79
https://people.speechocean.com/login
Response SHA-256: 23205ecaaa52cc56ef62cd411ff69e25c830e41c6db6cba4ce487c78d1141f79
Scoped evidence bindings: 1

[372] Our Privacy Policy | AI Data Solutions | DataOcean AI. Dataocean AI / Speechocean;
response-sha256:25a0666bf1aa7b71f01536280422f109b5d98a624a50c89c3219b9a1debf7f27
https://dataoceanai.com/privacy-policy/
Response SHA-256: 25a0666bf1aa7b71f01536280422f109b5d98a624a50c89c3219b9a1debf7f27
Scoped evidence bindings: 1

[373] SWE-Lancer: Can Frontier LLMs Earn $1 Million from Real-World Freelance Software Engineering?. arXiv (author-deposited versioned preprint); 2502.12115v1; 2025-02-17
https://arxiv.org/pdf/2502.12115v1
Response SHA-256: 428f0f36520a41aad156800c2d0520074dd4ca1d57ea9a4abd8d90807affbd1b
Scoped evidence bindings: 6

[374] SWE-bench Multimodal: Do AI Systems Generalize to Visual Software Domains?. arXiv (author-deposited versioned preprint); 2410.03859v1; 2024-10-04
https://arxiv.org/pdf/2410.03859v1
Response SHA-256: 4fa09c3734547e8ba4a5288e7e8d9273d1136c70295e4f91ef3592df0ead2a71
Scoped evidence bindings: 6

[375] SWE-Bench Pro: Can AI Agents Solve Long-Horizon Software Engineering Tasks?. arXiv (author-deposited versioned preprint); 2509.16941v1; 2025-09-21
https://arxiv.org/pdf/2509.16941v1
Response SHA-256: fa282d00fd32225fced1328838b65bcefb8947b0f5a54b7601f8f2ea27359ebb
Scoped evidence bindings: 6

[376] τ^2-Bench: Evaluating Conversational Agents in a Dual-Control Environment. arXiv (author-deposited versioned preprint); 2506.07982v1; 2025-06-09
https://arxiv.org/pdf/2506.07982v1
Response SHA-256: 0817e3fd33915326180d548caa900dcc5cba42ded27688105d8ce2f7e73aad84
Scoped evidence bindings: 6

[377] Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in Command Line Interfaces. arXiv (author-deposited versioned preprint); 2601.11868v1; 2026-01-17
https://arxiv.org/pdf/2601.11868v1
Response SHA-256: 4430e04a507b2442bb7b24a88899b750948fa0a6b3bbc33a8d541d6d4debcc1e
Scoped evidence bindings: 19

[378] V7 Darwin | AI Data Labeling & ML Training Data Platform. V7 / Darwin; response-sha256:2f20b45d011729f66df22aa79f735c409ab27ce324b4897a2e671a7af6ab8ac1
https://www.v7darwin.com/
Response SHA-256: 2f20b45d011729f66df22aa79f735c409ab27ce324b4897a2e671a7af6ab8ac1
Scoped evidence bindings: 3

[379] Master Service Agreement (MSA) - Darwin - V7. V7 / Darwin; response-sha256:434dd5f26d86be318c40ded57485b0b1a3f6298e0be9153dfdf82c2d8b07e84c
https://www.v7labs.com/terms/msa-darwin
Response SHA-256: 434dd5f26d86be318c40ded57485b0b1a3f6298e0be9153dfdf82c2d8b07e84c
Scoped evidence bindings: 2

[380] Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. arXiv (author-deposited versioned preprint); 2405.21075v1; 2024-05-31
https://arxiv.org/pdf/2405.21075v1
Response SHA-256: 8f86ad54e3017f81b0738ab0a4fa373f63334f2be4e5f0983d401b824f1821aa
Scoped evidence bindings: 5

[381] BenchMAX: A Comprehensive Multilingual Evaluation Suite for Large Language Models. arXiv (author-deposited versioned preprint); 2502.07346v1; 2025-02-11
https://arxiv.org/pdf/2502.07346v1
Response SHA-256: f8e1b140b3eb7d1bf2bd2db8bef76a52179637c11c8ecc533278d7f398eb6f4c
Scoped evidence bindings: 5

[382] The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. arXiv (author-deposited versioned preprint); 2506.05209v1; 2025-06-05
https://arxiv.org/pdf/2506.05209v1
Response SHA-256: 24b26eb7b9bba731af9746974c452a23f870814d31b7d55cd681b771dfe02fdc
Scoped evidence bindings: 5

[383] Consent in Crisis: The Rapid Decline of the AI Data Commons. arXiv (author-deposited versioned preprint); 2407.14933v1; 2024-07-20
https://arxiv.org/pdf/2407.14933v1
Response SHA-256: 890f0591edfaf9938f9e72046b1d7fe2c476c94a9911712921bde59b5dd9cf38
Scoped evidence bindings: 5

[384] How do data owners say no? A case study of data consent mechanisms in web-scraped vision-language AI training datasets. arXiv (author-deposited versioned preprint); 2511.08637v1; 2025-11-10
https://arxiv.org/pdf/2511.08637v1
Response SHA-256: 4351599edbaa15c5e242a577431760799f6784af2dc983cd5c3f3347b3116ea1
Scoped evidence bindings: 5

[385] The Data-Production Dispositif. arXiv (author-deposited versioned preprint); 2205.11963v1; 2022-05-24
https://arxiv.org/pdf/2205.11963v1
Response SHA-256: 926a27314518a1350d7463d30cc996c6d42f4567b04f2228a0433fca3cfa2746
Scoped evidence bindings: 5

[386] Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. Association for Computational Linguistics / ACL Anthology; ACL-Q18-1041-published-2018; anthology metadata commit ba3130005282c6474307cceec936373fe55d15ff
https://aclanthology.org/Q18-1041.pdf
Response SHA-256: da123cf0cc7dfb32479066ce01e0d107c312cff0c2818f9abfccb101c6005ff2
Scoped evidence bindings: 5

[387] Extracting Training Data from Large Language Models. arXiv (author-deposited versioned preprint); 2012.07805v2; 2021-06-15
https://arxiv.org/pdf/2012.07805v2
Response SHA-256: a3d1496575fd199e5b8535a78e69b528731c9895c1c0a3304848ef005593b473
Scoped evidence bindings: 5

[388] Fair human-centric image dataset for ethical AI benchmarking. Nature article; institutional PDF copy; institutional-copy-sha256-66ac69cbb2d33c170f14ff64072c07c51ab5d3ff66c2357a69dbb59d64fcaf78
https://www.cs.utexas.edu/~pstone/Papers/bib2html-links/peter_nature_2025.pdf
Response SHA-256: 66ac69cbb2d33c170f14ff64072c07c51ab5d3ff66c2357a69dbb59d64fcaf78
Scoped evidence bindings: 5

[389] Improving Task Instructions for Data Annotators: How Clear Rules and Higher Pay Increase Performance in Data Annotation in the AI Economy. arXiv (author-deposited versioned preprint); 2312.14565v2; 2024-08-16
https://arxiv.org/pdf/2312.14565v2
Response SHA-256: 1cfa3f27249f3c0abc80d1f27ef481ee54283545936c9dd074021ec5bd5dd042
Scoped evidence bindings: 5

[390] MENLO: From Preferences to Proficiency -- Evaluating and Modeling Native-like Quality Across 47 Languages. arXiv (author-deposited versioned preprint); 2509.26601v1; 2025-09-30
https://arxiv.org/pdf/2509.26601v1
Response SHA-256: 91cbda09646ae708e519c49891979a23e88cbcaec880c25e0c2380dce8a944ba
Scoped evidence bindings: 5

[391] M-IFEval: Multilingual Instruction-Following Evaluation. Association for Computational Linguistics / ACL Anthology; ACL-2025.findings-naacl.344-published; anthology metadata commit ba3130005282c6474307cceec936373fe55d15ff
https://aclanthology.org/2025.findings-naacl.344.pdf
Response SHA-256: 1b3f332d3d79359eec16db2b55c8cfd6bb92cdcff973f5df2d87fe17e27fde94
Scoped evidence bindings: 5

[392] Model Cards for Model Reporting. arXiv (author-deposited versioned preprint); 1810.03993v2; 2019-01-14
https://arxiv.org/pdf/1810.03993v2
Response SHA-256: a4049cdc78c3d635bdfb5165e495b0672d85690d66ccd6f149079d6e11e64aa1
Scoped evidence bindings: 5

[393] Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation. arXiv (author-deposited versioned preprint); 2503.11926v1; 2025-03-14
https://arxiv.org/pdf/2503.11926v1
Response SHA-256: 2547162c549d70e02beca2687903a8fce75fd950269a3ee479d5b8ebbb53ff38
Scoped evidence bindings: 5

[394] M-RewardBench: Evaluating Reward Models in Multilingual Settings. Association for Computational Linguistics / ACL Anthology; ACL-2025.acl-long.3-published; anthology metadata commit ba3130005282c6474307cceec936373fe55d15ff
https://aclanthology.org/2025.acl-long.3.pdf
Response SHA-256: 2da13cb132c7933424941b495ac9b7a94e9659d04030a79eb78310266be7418e
Scoped evidence bindings: 5

[395] MultiChallenge: A Realistic Multi-Turn Conversation Evaluation Benchmark Challenging to Frontier LLMs. arXiv (author-deposited versioned preprint); 2501.17399v1; 2025-01-29
https://arxiv.org/pdf/2501.17399v1
Response SHA-256: 5046c8bbe1bf77ab61092346d7fdfe05a25875a61b3c6e5a998313909dfd0074
Scoped evidence bindings: 5

[396] How Reliable is Multilingual LLM-as-a-Judge?. Association for Computational Linguistics / ACL Anthology; ACL-2025.findings-emnlp.587-published; anthology metadata commit ba3130005282c6474307cceec936373fe55d15ff
https://aclanthology.org/2025.findings-emnlp.587.pdf
Response SHA-256: 3c7c41868590e6b1d97123127e362036ae031bd315b0734e665202ee43cad137
Scoped evidence bindings: 5

[397] ShieldAgent: Shielding Agents via Verifiable Safety Policy Reasoning. arXiv (author-deposited versioned preprint); 2503.22738v1; 2025-03-26
https://arxiv.org/pdf/2503.22738v1
Response SHA-256: 2bfa433e30445ff10a687413891d5e132498e5e3715834574b023d614c6a6bb8
Scoped evidence bindings: 5

[398] Is It Thinking or Cheating? Detecting Implicit Reward Hacking by Measuring Reasoning Effort. arXiv (author-deposited versioned preprint); 2510.01367v1; 2025-10-01
https://arxiv.org/pdf/2510.01367v1
Response SHA-256: 3d87b2a6c49cce0adefcccdbfa54f8f101d1c87c69c131768086a8971c7eb1a6
Scoped evidence bindings: 5

[399] VerifiAgent: a Unified Verification Agent in Language Model Reasoning. arXiv (author-deposited versioned preprint); 2504.00406v1; 2025-04-01
https://arxiv.org/pdf/2504.00406v1
Response SHA-256: 2ffccf4b668733a49bf375e7836419f3447fda6a8c28d19110b6b024f1aa64f5
Scoped evidence bindings: 5

[400] World of Bits: An Open-Domain Platform for Web-Based Agents. Proceedings of Machine Learning Research; PMLR-v70-shi17a-2017; 2017-07-17
https://proceedings.mlr.press/v70/shi17a/shi17a.pdf
Response SHA-256: 087008efb9d056df5fd9fbc724413d6a3f20cb14c7f1597d00bacac53d8342d6
Scoped evidence bindings: 5

[401] CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2206.08931v1; 2022-06-09
https://arxiv.org/pdf/2206.08931v1
Response SHA-256: 5249ebf13529cc89a0e72c4c19833355d1b4811104a0102c807d2ce2e13bef89
Scoped evidence bindings: 5

[402] DAPO: An Open-Source LLM Reinforcement Learning System at Scale. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2503.14476v2; 2025-05-20
https://arxiv.org/pdf/2503.14476v2
Response SHA-256: f01e4fd347530cadd68e5c36b1998532a6d1adb272c817e73b927453c26e9d79
Scoped evidence bindings: 5

[403] Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2204.01075v1; 2022-04-03
https://arxiv.org/pdf/2204.01075v1
Response SHA-256: ffe747c69935602e9a721480310f0ab8d70a590e4f7fc9ab3edaa2ab10b18118
Scoped evidence bindings: 5

[404] Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2104.08758v2; 2021-09-30
https://arxiv.org/pdf/2104.08758v2
Response SHA-256: 721d817b6e6743ae7829bc4b77ea518cd585b18520ee716d3a11781d750d920c
Scoped evidence bindings: 5

[405] Understanding R1-Zero-Like Training: A Critical Perspective. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2503.20783v2; 2025-10-06
https://arxiv.org/pdf/2503.20783v2
Response SHA-256: 98243d51297f011fb5baad8a70a972d061ffbef618f1d7cfc10deea37c5887d0
Scoped evidence bindings: 5

[406] If in a Crowdsourced Data Annotation Pipeline, a GPT-4. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2402.16795v2; 2024-06-28
https://arxiv.org/pdf/2402.16795v2
Response SHA-256: ba747597683b1f80c2fbbfbb07308baa5bd206f13aa61e8d493e6f74c81b96f8
Scoped evidence bindings: 5

[407] Group Sequence Policy Optimization. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2507.18071v2; 2025-07-28
https://arxiv.org/pdf/2507.18071v2
Response SHA-256: 67221ec313f42d8b3029686b604b2621a311ae0d74e38524092e7c7a94613bd0
Scoped evidence bindings: 5

[408] HybridFlow: A Flexible and Efficient RLHF Framework. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2409.19256v2; 2024-10-02
https://arxiv.org/pdf/2409.19256v2
Response SHA-256: 4807a008bdea448c7a6aecc3b0aa65ffc45f1f4ca83c056bfe562b5e2397ce98
Scoped evidence bindings: 5

[409] The digital labour of artificial intelligence in Latin America: a comparison of Argentina, Brazil, and Venezuela. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2502.06317v1; 2025-02-10
https://arxiv.org/pdf/2502.06317v1
Response SHA-256: 457d52643f47daa6634839c2a3bf1c8e7289709f8453d4741acc14d37b6b996b
Scoped evidence bindings: 5

[410] Minority Reports: Balancing Cost and Quality in Ground Truth Data Annotation. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2504.09341v1; 2025-04-12
https://arxiv.org/pdf/2504.09341v1
Response SHA-256: b27178636b4cc8f56b7a65247e3ba6ee9ec6d9b4905f67d32802ba036721db19
Scoped evidence bindings: 5

[411] The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 1805.03677v1; 2018-05-09
https://arxiv.org/pdf/1805.03677v1
Response SHA-256: 8be208360982f7b916b3e259b08d8cd3c4f93e5e9a8fae2425a1912c5ff4a331
Scoped evidence bindings: 5

[412] OpenRLHF: An Easy-to-use, Scalable and High-performance RLHF Framework. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2405.11143v6; 2025-10-09
https://arxiv.org/pdf/2405.11143v6
Response SHA-256: b9797d8cb971705064d2aec706671cf44457d90b860910b9aae4cf5839e8750c
Scoped evidence bindings: 5

[413] Process Reinforcement through Implicit Rewards. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2502.01456v2; 2025-09-26
https://arxiv.org/pdf/2502.01456v2
Response SHA-256: b9b0d4ee041d5f21dec9489a67582e3775c3900a61bad979814522ab8b3faef6
Scoped evidence bindings: 5

[414] The Problems with Proxies: Making Data Work Visible through Requester Practices. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2408.11667v1; 2024-08-21
https://arxiv.org/pdf/2408.11667v1
Response SHA-256: 41f6aff403ca4cff9622bcec6e366b0530e074a5bf9e69b62d7895ce2f14fefe
Scoped evidence bindings: 5

[415] REINFORCE++: Stabilizing Critic-Free Policy Optimization with Global Advantage Normalization. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2501.03262v9; 2025-11-10
https://arxiv.org/pdf/2501.03262v9
Response SHA-256: b868a07e9d068dd4bc13eca28e7ce4afc380e26380b05c741d5dd2cc6821a59f
Scoped evidence bindings: 5

[416] On the Reliability of Majority-Vote Pseudo-Labels in Test-Time Reinforcement Learning. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2611.09872v1; 2026-11-06
https://arxiv.org/pdf/2611.09872v1
Response SHA-256: 6ae04bce5f0f268407f8d96ed4e16e6ffa39e0c193cb2e2b91c709c96cd7ba77
Scoped evidence bindings: 5

[417] SelectTTRL: Representation-Guided Data Selection for Efficient Test-Time Reinforcement Learning. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2610.16104v1; 2026-10-12
https://arxiv.org/pdf/2610.16104v1
Response SHA-256: 598550ad70ebfc786fd27b18f2383d54d433bee3c7a6d2abb82f72060d2b94f1
Scoped evidence bindings: 5

[418] STAMP Your Content: Proving Dataset Membership via Watermarked Rephrasings. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2504.13416v2; 2025-06-08
https://arxiv.org/pdf/2504.13416v2
Response SHA-256: 0c48ba4d0980e7fbd9b79b63333965ebf2dbb117580d4c0c080ecee4397c24fd
Scoped evidence bindings: 5

[419] TTRL: Test-Time Reinforcement Learning. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2504.16084v3; 2025-06-30
https://arxiv.org/pdf/2504.16084v3
Response SHA-256: 3828cc886a0133749fe857dc12fff190f76c17dd7421dccfca01c96ec1dd5882
Scoped evidence bindings: 6

[420] VAPO: Efficient and Reliable Reinforcement Learning for Advanced Reasoning Tasks. arXiv (author-deposited versioned preprint; not publisher VoR byte-verified); 2504.05118v3; 2025-04-11
https://arxiv.org/pdf/2504.05118v3
Response SHA-256: a10387626b86a8f9a6a80d17e727eb5a5036dc237a81d370c26e60d7c4168457
Scoped evidence bindings: 5

[421] AutoLibra: Agent Metric Induction from Open-Ended Human Feedback. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2505.02820v3; 2025-10-29
https://arxiv.org/pdf/2505.02820v3
Response SHA-256: 584079f8651492bceb2bb86fdf8ae9f3cb36633833eb7b3551761ac099a620cc
Scoped evidence bindings: 4

[422] Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2402.06619v1; 2024-02-09
https://arxiv.org/pdf/2402.06619v1
Response SHA-256: 521d8a1c9cdfc466def9745499de330c6bc4d5b4e7c48609437462b18410b6a6
Scoped evidence bindings: 4

[423] CrowdAgent: Multi-Agent Managed Multi-Source Annotation System. ACL Anthology (exact captured publication copy); ACL-Anthology-2025.emnlp-demos.72-captured-copy
https://aclanthology.org/2025.emnlp-demos.72.pdf
Response SHA-256: db55042820e04c3c7ffd5d0b62aaffde024486f1811f9b0508f08a791218500e
Scoped evidence bindings: 4

[424] CrowdTruth 2.0: Quality Metrics for Crowdsourcing with Disagreement. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 1808.06080v1; 2018-08-18
https://arxiv.org/pdf/1808.06080v1
Response SHA-256: 5b734de88e5c471eb1017dcc3d25e06a901d7564678fd9c535579d75f2905613
Scoped evidence bindings: 4

[425] A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 1712.05796v2; 2017-12-28
https://arxiv.org/pdf/1712.05796v2
Response SHA-256: 25b098a1baad2d4214dd9c4dac4290fac40d1217231a1d90d6627373c125060c
Scoped evidence bindings: 4

[426] Experts Rise Where LLMs Disagree: Using Cross-Model Disagreement to Target Expert Effort in LLM Codebook Revision for Large-Scale Annotation. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2609.26926v1; 2026-09-22
https://arxiv.org/pdf/2609.26926v1
Response SHA-256: b227f50171f022ebae29b77f68753a690130c045b7d04bffc2e6cb999c628da1
Scoped evidence bindings: 4

[427] Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 1912.08320v1; 2019-12-17
https://arxiv.org/pdf/1912.08320v1
Response SHA-256: 68da81c0d496d9b8e1c0dfc43e5ec65f0de83faa38575bbe5f4bfac0cbb61222
Scoped evidence bindings: 4

[428] ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2303.15056v2; 2023-07-19
https://arxiv.org/pdf/2303.15056v2
Response SHA-256: da6865c3e57e8a956f2aeeadef0c814b2ab4a2db4753ef50037a4074b6e3d26f
Scoped evidence bindings: 4

[429] HelpSteer: Multi-attribute Helpfulness Dataset for SteerLM. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2311.09528v1; 2023-11-16
https://arxiv.org/pdf/2311.09528v1
Response SHA-256: af48d77f75761d69e1d083fc5a2168351d81be30b47138bdb9b5178a6bc025f5
Scoped evidence bindings: 4

[430] HelpSteer2: Open-source dataset for training top-performing reward models. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2406.08673v1; 2024-06-12
https://arxiv.org/pdf/2406.08673v1
Response SHA-256: 9c7754b5d099528a25e99c68c240474da68a1f4b287f753fdcc19139b691c943
Scoped evidence bindings: 4

[431] Inherent Disagreements in Human Textual Inferences. ACL Anthology (exact captured publication copy); ACL-Anthology-Q19-1043-captured-copy
https://aclanthology.org/Q19-1043.pdf
Response SHA-256: 6145dd60a513f3426d50718a2301f4189fce10735935ae86d8c01c826f6a6933
Scoped evidence bindings: 4

[432] LabelAId: Just-in-time AI Interventions for Improving Human Labeling Quality and Domain Knowledge in Crowdsourcing Systems. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2403.09810v1; 2024-03-14
https://arxiv.org/pdf/2403.09810v1
Response SHA-256: 52030878aded4636165666cbafc58d62dcfe5f76d29fbe94af57f030b63eed4a
Scoped evidence bindings: 4

[433] Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2103.14749v4; 2021-11-07
https://arxiv.org/pdf/2103.14749v4
Response SHA-256: 605c6a72f3a970dbbdd02622079c32d687862455a1152da9f2075e35fad2bf1d
Scoped evidence bindings: 4

[434] Efficient Data Labeling by Hierarchical Crowdsourcing with Large Language Models. ACL Anthology (exact captured publication copy); ACL-Anthology-2025.coling-main.748-captured-copy
https://aclanthology.org/2025.coling-main.748.pdf
Response SHA-256: bc7719b1a7fa895353df574927de6d585fc654d30c3c181e24bb14c07c3c87ec
Scoped evidence bindings: 4

[435] Quality Assured: Rethinking Annotation Strategies in Imaging AI. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2407.17596v2; 2024-07-26
https://arxiv.org/pdf/2407.17596v2
Response SHA-256: f196d771167aa4e43d887b59b99c35a8a05d4ffd075b75de60891b1c9ad2665d
Scoped evidence bindings: 4

[436] RECODE-H: A Benchmark for Research Code Development with Interactive Human Feedback. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2510.06186v2; 2025-10-24
https://arxiv.org/pdf/2510.06186v2
Response SHA-256: 26b048307df93645e02a2eb852b1cb0ea1687c0f8ccd7ec66935fe649f903efb
Scoped evidence bindings: 4

[437] Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. ACL Anthology (exact captured publication copy); ACL-Anthology-D08-1027-captured-copy
https://aclanthology.org/D08-1027.pdf
Response SHA-256: 19f42e03a5f3a0d4401c75c7fade0040dd1f890c86fd6b8c1c4fec644bf6b7f4
Scoped evidence bindings: 4

[438] UltraFeedback: Boosting Language Models with Scaled AI Feedback. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2310.01377v2; 2024-07-16
https://arxiv.org/pdf/2310.01377v2
Response SHA-256: 63a40ed6d1603fc07c09f113042f75619c1c17c0a0c8f1ea421477c0facbc8ca
Scoped evidence bindings: 4

[439] Who Annotates in NLP? A Large-scale Assessment of Human Annotation Reporting between 2018 and 2025. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2606.02255v3; 2026-09-02
https://arxiv.org/pdf/2606.02255v3
Response SHA-256: 6fb65a7478f3d8a481bd2d97ab980509fc05ce24e3abb466a2752637a09a77b3
Scoped evidence bindings: 4

[440] WildChat: 1M ChatGPT Interaction Logs in the Wild. arXiv (author-deposited versioned preprint; publisher-copy equivalence not asserted); 2405.01470v1; 2024-05-02
https://arxiv.org/pdf/2405.01470v1
Response SHA-256: 932316edcaf09c144358c0c0bf6ad48a82f315c189e944c4afa0f90835bcf85e
Scoped evidence bindings: 4

[441] AM-DeepSeek-R1-Distilled-1.4M - configs. a-m-team; hf:53531c06634904118a2dcd83961918c4d69d1cdf
https://huggingface.co/datasets/a-m-team/AM-DeepSeek-R1-Distilled-1.4M/resolve/53531c06634904118a2dcd83961918c4d69d1cdf/README.md
Response SHA-256: bde91302b06cfce799ef2c153ac84a9275daa3a16a42a084ee12fee1f0d239b7
Scoped evidence bindings: 10

[442] Bespoke-Stratos-17k - license. bespokelabs; hf:9e9adba943911a9fc44dffcb30aaa18dc96ae6df
https://huggingface.co/datasets/bespokelabs/Bespoke-Stratos-17k/resolve/9e9adba943911a9fc44dffcb30aaa18dc96ae6df/README.md
Response SHA-256: a9b9b2b2a59c9aed6d759800dc5dae3d3cfd99bed88a0374afa6d8b290b78382
Scoped evidence bindings: 5

[443] Light-R1-SFTData - base release. qihoo360; hf:f939e79f82f0c7c14cbb871ee0b75f8f7226ad73
https://huggingface.co/datasets/qihoo360/Light-R1-SFTData/resolve/f939e79f82f0c7c14cbb871ee0b75f8f7226ad73/README.md
Response SHA-256: 429cdbc8f91fdc59f029609babba5fbac849966796c9c662538d3b3228f195fa
Scoped evidence bindings: 8

[444] s1K-1.1 - credit. simplescaling; hf:96c411f1fe4c49d20f0e2a1565f61e1a28b0b84d
https://huggingface.co/datasets/simplescaling/s1K-1.1/resolve/96c411f1fe4c49d20f0e2a1565f61e1a28b0b84d/README.md
Response SHA-256: 7b89481543d085362c222e96563a606f640fa6d2ef3471039ea188a22a1bf70d
Scoped evidence bindings: 7

[445] SmallThoughts - full card. SmallDoge; hf:8fdc9afef80403b7c50ebb4356dd7878d3218d37
https://huggingface.co/datasets/SmallDoge/SmallThoughts/resolve/8fdc9afef80403b7c50ebb4356dd7878d3218d37/README.md
Response SHA-256: 418a5bf4693434e84cf991bf30a75e93ccab7c6e7bfb3c88ec6e57e334c1bf0d
Scoped evidence bindings: 6

[446] openai/genebench-pro-public-package creator dataset card. OpenAI / Hugging Face; huggingface-dataset:eb75a3c0996b3cedcc9af685bad02fd166848fa2
https://huggingface.co/datasets/openai/genebench-pro-public-package/resolve/eb75a3c0996b3cedcc9af685bad02fd166848fa2/README.md
Response SHA-256: 2a8b80a47d8e5d567f4fb9110286ad578818e3974b6c13b00a6a51aad8ab3f1e
Scoped evidence bindings: 1

[447] GAIA - README.md. gaia-benchmark/GAIA; 682dd723ee1e1697e00360edccf2366dc8418dd9
https://huggingface.co/datasets/gaia-benchmark/GAIA/blob/682dd723ee1e1697e00360edccf2366dc8418dd9/README.md
Response SHA-256: 8a160e2df36b843c33ec12fdecd731ae8281e0dae553375207707b506a36c11e
Scoped evidence bindings: 2

[448] Helpful, Honest & Harmless pragmatic alignment evaluation README. Creator-maintained repository; 092b196c1f8f14a54bbc62f24759d43bde46dd3b
https://raw.githubusercontent.com/google/BIG-bench/092b196c1f8f14a54bbc62f24759d43bde46dd3b/bigbench/benchmark_tasks/hhh_alignment/README.md
Response SHA-256: a9026f2a9a4c8900ff2c605ba8c7fe5907fc09d36c55d31609ddc214def9c7e2
Scoped evidence bindings: 3

[449] RewardBench 1 evaluation dataset card. Ai2 / RewardBench creators; 168d848cdbbea9764fae4a544dc9ca1e6cca4931
https://huggingface.co/datasets/allenai/reward-bench/resolve/168d848cdbbea9764fae4a544dc9ca1e6cca4931/README.md
Response SHA-256: 819086ae803ffe3d3126eb2c188b8c67e2a0acb91b67096727f59f5aff423256
Scoped evidence bindings: 3

[450] APPS - README.md. hendrycks/apps; 362aedc3c71cd7d9bd2fc96a6c80e11dbc38c7a5
https://raw.githubusercontent.com/hendrycks/apps/362aedc3c71cd7d9bd2fc96a6c80e11dbc38c7a5/README.md
Response SHA-256: 9a301ef59a87611c90441299b114964c126a6112e0074e308d06d60df3981185
Scoped evidence bindings: 1

[451] NVIDIA HelpSteer2 dataset card. NVIDIA; 990b2711a36180dd19d9c94b8627844866f8982a
https://huggingface.co/datasets/nvidia/HelpSteer2/resolve/990b2711a36180dd19d9c94b8627844866f8982a/README.md
Response SHA-256: 835effb9e7d9cd0e8b7b8c1816d1d97a8961a036543108e4a0b6a5e22712ff7b
Scoped evidence bindings: 6

[452] Anthropic HH-RLHF dataset card. Anthropic; 09be8c5bbc57cb3887f3a9732ad6aa7ec602a1fa
https://huggingface.co/datasets/Anthropic/hh-rlhf/resolve/09be8c5bbc57cb3887f3a9732ad6aa7ec602a1fa/README.md
Response SHA-256: f75f40db0268656ba07736ec8e59a9720c1910ce554c85354cec74b1c8bda175
Scoped evidence bindings: 3

[453] Stanford Human Preferences dataset card. Stanford NLP / SHP creators; e94b5f32602712d78ed494fe79105b1959396686
https://huggingface.co/datasets/stanfordnlp/SHP/resolve/e94b5f32602712d78ed494fe79105b1959396686/README.md
Response SHA-256: 533784230dff4330073a69fec9464de80d01f9285f8e88e41f25bb033d407508
Scoped evidence bindings: 4

[454] UltraFeedback dataset card. OpenBMB / UltraFeedback creators; 40b436560ca83a8dba36114c22ab3c66e43f6d5e
https://huggingface.co/datasets/openbmb/UltraFeedback/resolve/40b436560ca83a8dba36114c22ab3c66e43f6d5e/README.md
Response SHA-256: 70a22b5659215ab738b9dfd0dc6bd40d3bb040bf846b4e34955358f30625a4f7
Scoped evidence bindings: 3

[455] WildJailbreak - README.md (public blob visible text). allenai/wildjailbreak; 5ddc12a7894f842b0619b8e1c7ee496b198af009
https://huggingface.co/datasets/allenai/wildjailbreak/blob/5ddc12a7894f842b0619b8e1c7ee496b198af009/README.md
Response SHA-256: f9e6161b38b2acc5bab0307ee7e835715fb51f2eddbeb130039b47201a4bf28b
Scoped evidence bindings: 2

[456] ScienceToStartup canonical dataset 2.5.0 (historical local snapshot). ScienceToStartup; 2.5.0
sha256:948e69229fa24a4a1e6b5eed6d9c11570568a3c05bf9512e442d19611061a061
https://sciencetostartup.com/research.json
Response SHA-256: 948e69229fa24a4a1e6b5eed6d9c11570568a3c05bf9512e442d19611061a061
Scoped evidence bindings: 1

[457] Data Labeling for Machine Learning Engineers: Project-Based Curriculum and Data-Centric Competitions - limitation context. Association for the Advancement of Artificial Intelligence; AAAI-37-13-26886-26658-captured-2026-11-10
https://ojs.aaai.org/index.php/AAAI/article/download/26886/26658
Response SHA-256: 5a9acc7aa55b5e2e7784626b99a1631859e28bd04679cf06939b673dcd0df2ad
Scoped evidence bindings: 4

[458] Quantifying and Avoiding Unfair Qualification Labour in Crowdsourcing - limitation context. Association for Computational Linguistics; ACL-Anthology-2021.acl-short.44-captured-2026-11-10
https://aclanthology.org/2021.acl-short.44.pdf
Response SHA-256: 1d69e7ee438858b5962e24f72c603f9bfc3dacd163b813f013c6cfbfff64dc44
Scoped evidence bindings: 4

[459] Red Teaming Language Models with Language Models - limitation context. Association for Computational Linguistics; ACL-Anthology-2022.emnlp-main.225-captured-2026-11-10
https://aclanthology.org/2022.emnlp-main.225.pdf
Response SHA-256: f1f838d6488b9634a7650ecf096ad6be5fe583f377dd45253c547faa312fc0c0
Scoped evidence bindings: 4

[460] Findings of the WMT’22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages - limitation context. Association for Computational Linguistics; ACL-Anthology-2022.wmt-1.72-captured-2026-11-10
https://aclanthology.org/2022.wmt-1.72.pdf
Response SHA-256: ab329fcc933ee0eb4d687acf665d14d69b94d0da587c3532997caa616d458bf5
Scoped evidence bindings: 4

[461] LLMs Simulate Big5 Personality Traits: Further Evidence - limitation context. Association for Computational Linguistics; ACL-Anthology-2024.personalize-1.7-captured-2026-11-10
https://aclanthology.org/2024.personalize-1.7.pdf
Response SHA-256: 0580ffa488a59bf7a3f95b358984e20717e11219d20c1926f256401f3915cff2
Scoped evidence bindings: 4

[462] Learning Whom to Trust with MACE - limitation context. Association for Computational Linguistics; ACL-Anthology-N13-1132-captured-2026-11-10
https://aclanthology.org/N13-1132.pdf
Response SHA-256: 7e1d2a30a1a584e0a163468fd1f0c2e89d2ed5dc58c8e30f04adf79d69326f3a
Scoped evidence bindings: 4

[463] Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets - limitation context. arXiv; 2106.10328v2
https://arxiv.org/pdf/2106.10328v2
Response SHA-256: 68d04d53073d8fe4cf7408f8e257caddcab0905ef5d5cd509c52ffcf04a87b8c
Scoped evidence bindings: 4

[464] CrowdSpeech and VoxDIY: Benchmark Datasets for Crowdsourced Audio Transcription - limitation context. arXiv; 2107.01091v2
https://arxiv.org/pdf/2107.01091v2
Response SHA-256: 87440d688c0e8ee622b5ec5489cbec5f681a461630a82ebdbca57856e9fac742
Scoped evidence bindings: 4

[465] Deduplicating Training Data Makes Language Models Better - limitation context. arXiv; 2107.06499v2
https://arxiv.org/pdf/2107.06499v2
Response SHA-256: dcfa46d7aced43a97b9cf2ebd705919b65d4dd8a44d0914b1e646fe2d59ea8d3
Scoped evidence bindings: 4

[466] Learning from Crowds with Crowd-Kit - limitation context. arXiv; 2109.08584v4
https://arxiv.org/pdf/2109.08584v4
Response SHA-256: b0582360a956660e51cb04b71e06da72693ade5138be2c97b108578d75f3f0a2
Scoped evidence bindings: 4

[467] IMDB-WIKI-SbS: An Evaluation Dataset for Crowdsourced Pairwise Comparisons - limitation context. arXiv; 2110.14990v2
https://arxiv.org/pdf/2110.14990v2
Response SHA-256: 4ab845a9a86cf82994c94f749fe155e076437354ba8b900f56ce3b52a5c73de8
Scoped evidence bindings: 4

[468] WebGPT: Browser-assisted question-answering with human feedback - limitation context. arXiv; 2112.09332v3
https://arxiv.org/pdf/2112.09332v3
Response SHA-256: acad8d9fd55a30b94de9dfdae413e10b21e07bde54d767b2405b31341739cedb
Scoped evidence bindings: 4

[469] Shoring Up the Foundations: Fusing Model Embeddings and Weak Supervision - limitation context. arXiv; 2203.13270v2
https://arxiv.org/pdf/2203.13270v2
Response SHA-256: 35755d29f1a7d56bb85460f0e35ef842fe4fd9a4e2eda6a704b4ee7eb2e0fbc4
Scoped evidence bindings: 4

[470] Clustering Without Knowing How To: Application and Evaluation - limitation context. arXiv; 2209.10267v3
https://arxiv.org/pdf/2209.10267v3
Response SHA-256: 34c083fe845a4e4c86a9dc6ea6eb29b8366877fca10066697049950c32411d36
Scoped evidence bindings: 4

[471] Best Prompts for Text-to-Image Models and How to Find Them - limitation context. arXiv; 2209.11711v3
https://arxiv.org/pdf/2209.11711v3
Response SHA-256: bd3569de00701caefe6d951420781ce23aa7406de82d227534bb850617e91d2d
Scoped evidence bindings: 4

[472] Learning to Reason With Relational Abstractions - limitation context. arXiv; 2210.02615v2
https://arxiv.org/pdf/2210.02615v2
Response SHA-256: 86dd5e04730654f6677cc19f5b972177fbf1644432a41a91c7553ae4677ca930
Scoped evidence bindings: 4

[473] Measuring Progress on Scalable Oversight for Large Language Models - limitation context. arXiv; 2211.03540v2
https://arxiv.org/pdf/2211.03540v2
Response SHA-256: 534ef4f8a41b2053736fdf7cdbb40de3a582e5baa176fc3ea6629708451546c6
Scoped evidence bindings: 4

[474] Discovering Language Model Behaviors with Model-Written Evaluations - limitation context. arXiv; 2212.09251v1
https://arxiv.org/pdf/2212.09251v1
Response SHA-256: 7fd476342abaa95cbde50e1c5108765f2c53c3c84a7d55c4f2584fb8aaa01b10
Scoped evidence bindings: 4

[475] DOC: Improving Long Story Coherence With Detailed Outline Control - limitation context. arXiv; 2212.10077v3
https://arxiv.org/pdf/2212.10077v3
Response SHA-256: 6cf384b0f2d886e4a463f92062f0ef2e13de2db7edaac605805236df0ba972dc
Scoped evidence bindings: 3

[476] StarCoder: may the source be with you! - limitation context. arXiv; 2305.06161v2
https://arxiv.org/pdf/2305.06161v2
Response SHA-256: 549e079ffd6104417f8443570954b232d2ed58886cf45c60482e631cb36e1351
Scoped evidence bindings: 5

[477] Terms & Privacy Policy at clickworker.com. Clickworker;
response-sha256:c09660e6644708e172db450f88305a9bddf38956e837de70cccc2560b8031fa0
https://www.clickworker.com/terms-privacy-policy/
Response SHA-256: c09660e6644708e172db450f88305a9bddf38956e837de70cccc2560b8031fa0
Scoped evidence bindings: 1

[478] Data Services by clickworker. Clickworker;
response-sha256:24e23fde31dd2854455e7f27968a9bf3fde1d7e9f905d9fb24c28a96397afdf3
https://www.clickworker.com/
Response SHA-256: 24e23fde31dd2854455e7f27968a9bf3fde1d7e9f905d9fb24c28a96397afdf3
Scoped evidence bindings: 1

[479] Our Clickworker community - Information about the crowd. Clickworker;
response-sha256:09e8dd894c8bee1d022ea696be1906ee76c9992e4e04ed4d1a4bee5c006cf7ca
https://www.clickworker.com/clickworker-crowd/
Response SHA-256: 09e8dd894c8bee1d022ea696be1906ee76c9992e4e04ed4d1a4bee5c006cf7ca
Scoped evidence bindings: 1

[480] About clickworker - Your Data and Content Provider. Clickworker;
response-sha256:b2f13f61d2e01188f2897aa119999d9cc2ed73cceaf897ed6a37c3c874d1aacd
https://www.clickworker.com/about-us/
Response SHA-256: b2f13f61d2e01188f2897aa119999d9cc2ed73cceaf897ed6a37c3c874d1aacd
Scoped evidence bindings: 2

[481] Photo Data Sets - Online - Face recognition - Training. Clickworker;
response-sha256:6c3ddd685a711fd7afe312254f502263637b16158f77dc70a9c63052ffe534f6
https://www.clickworker.com/case-studies/photo-data-sets-face-recognition-training/
Response SHA-256: 6c3ddd685a711fd7afe312254f502263637b16158f77dc70a9c63052ffe534f6
Scoped evidence bindings: 1

[482] DataPure. DataPure; response-sha256:301bb99233d1bfc9cda51263faee30edf1e3122a097d032fe036f702843d2d07
https://www.datapure.ai/
Response SHA-256: 301bb99233d1bfc9cda51263faee30edf1e3122a097d032fe036f702843d2d07
Scoped evidence bindings: 3

[483] Blogs and Case Studies. DataPure;
response-sha256:c4ad34b412941f99a35346bb227dd0b0c529a4641dc22ac1c15f418ed285356b
https://www.datapure.ai/blogs-and-case-studies
Response SHA-256: c4ad34b412941f99a35346bb227dd0b0c529a4641dc22ac1c15f418ed285356b
Scoped evidence bindings: 1

[484] Datatang2026H1 issuer report. Datatang;
response-sha256:e32dd90603fcd871b820743402cdd8fc5235e01429961eb1e125d8ea327196b3
https://pdf.dfcfw.com/pdf/H2_AN202608271828548712_1.pdf?1787853970000.pdf=
Response SHA-256: e32dd90603fcd871b820743402cdd8fc5235e01429961eb1e125d8ea327196b3
Scoped evidence bindings: 7

[485] 会社概要 - Nexdata. Nexdata / Datatang株式会社;
response-sha256:cfa17b301f5269a9d13cd29c1132e6016a5ebc9be9754eb4c8fe8360fb669de2
https://nexdata.jp/company/about-us
Response SHA-256: cfa17b301f5269a9d13cd29c1132e6016a5ebc9be9754eb4c8fe8360fb669de2
Scoped evidence bindings: 1

[486] About FutureBeeAI | Your AI Data Partner. FutureBeeAI;
response-sha256:fdac56968ec47b8af899762149e3a5da7472f519399a51ee4c45d806e6c6b6ae
https://www.futurebeeai.com/about-futurebeeai
Response SHA-256: fdac56968ec47b8af899762149e3a5da7472f519399a51ee4c45d806e6c6b6ae
Scoped evidence bindings: 1

[487] AI Data Collection & Creation Services for ML Models. FutureBeeAI;
response-sha256:db904d7d49e65646fe81f98c19326a74a94dad8f766d3e0301eec9672096f2f5
https://www.futurebeeai.com/ai-data-collection-services
Response SHA-256: db904d7d49e65646fe81f98c19326a74a94dad8f766d3e0301eec9672096f2f5
Scoped evidence bindings: 2

[488] FutureBeeAI | Pre-Labeled Datasets & Custom AI Data Solutions for Robust AI Development. FutureBeeAI;
response-sha256:8b996dbff001dc8fa11731143295a2b53e81117fb148987e73c92e438ad39a62
https://www.futurebeeai.com/
Response SHA-256: 8b996dbff001dc8fa11731143295a2b53e81117fb148987e73c92e438ad39a62
Scoped evidence bindings: 1

[489] Monetize Your Dataset | License Your Data with FutureBeeAI. FutureBeeAI;
response-sha256:ddc37c8d5eaef993ce2bb4f2a85669fd2186a4bb50f7b44a4d76ea76a8243abe
https://www.futurebeeai.com/sell-ai-assets
Response SHA-256: ddc37c8d5eaef993ce2bb4f2a85669fd2186a4bb50f7b44a4d76ea76a8243abe
Scoped evidence bindings: 1

[490] About- AI Dataset Collection Services for Machine Learning | GTS. Globose Technology Solutions;
response-sha256:a0f5f4e84731f7838ba1fc14fdc053836e0cf286099dbcdd258cda7088385e41
https://gts.ai/about-us/
Response SHA-256: a0f5f4e84731f7838ba1fc14fdc053836e0cf286099dbcdd258cda7088385e41
Scoped evidence bindings: 3

[491] AI Dataset Collection Company | Dataset for Machine Learning. Globose Technology Solutions;
response-sha256:c466ae7864737cf653ba894ef85590f3fa52621fcab67d736c6e18c4775465a8
https://gts.ai/
Response SHA-256: c466ae7864737cf653ba894ef85590f3fa52621fcab67d736c6e18c4775465a8
Scoped evidence bindings: 2

[492] Text Annotation of News Articles for AI Models. HabileData;
response-sha256:3bb70c337f8a8655f3cd795f6ecbcfcb6726fcdf25b18ea885915ed5b0267d5f
https://www.habiledata.com/case-studies/text-annotation-of-news-articles.php
Response SHA-256: 3bb70c337f8a8655f3cd795f6ecbcfcb6726fcdf25b18ea885915ed5b0267d5f
Scoped evidence bindings: 1

[493] AI Data Preparation Services for AI Model Training. HabileData;
response-sha256:ae684a465c73199e2cd00aea05aad1d4a326817daff9fa20acff7f63a5caccf1
https://www.habiledata.com/ai-data-services.php
Response SHA-256: ae684a465c73199e2cd00aea05aad1d4a326817daff9fa20acff7f63a5caccf1
Scoped evidence bindings: 2

[494] Your Data Security & Confidentiality Matters | HabileData. HabileData;
response-sha256:667b2427cb9883802d287cfaf42c22d3d3f054c0f2cda28bcaeb1aace0f0f5ed
https://www.habiledata.com/data-security-confidentiality.php
Response SHA-256: 667b2427cb9883802d287cfaf42c22d3d3f054c0f2cda28bcaeb1aace0f0f5ed
Scoped evidence bindings: 2

[495] Agent Trajectory Dataset for Tool-Use Training and AI Agent Evaluation. Nexdata;
response-sha256:2d4ddce427733d27101c7d16cd4e2384523e7e18f50db787178e9bdc116bf1c4
https://www.nexdata.ai/datasets/llm/2185
Response SHA-256: 2d4ddce427733d27101c7d16cd4e2384523e7e18f50db787178e9bdc116bf1c4
Scoped evidence bindings: 1

[496] Sharpen Your AI with Better Data - Nexdata. Nexdata;
response-sha256:4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65
https://www.nexdata.ai/
Response SHA-256: 4c00514bd9e94083624a3e5cd0e4dba5b8b0bd7710ba72fa984a348e6e54cf65
Scoped evidence bindings: 1

[497] Data Annotation Platform - Nexdata. Nexdata;
response-sha256:c0d2c417f9f68f0dc0e7a3394dbe5bac58698a452e9b6612c1bc321dd3eb15ad
https://www.nexdata.ai/platform
Response SHA-256: c0d2c417f9f68f0dc0e7a3394dbe5bac58698a452e9b6612c1bc321dd3eb15ad
Scoped evidence bindings: 1

[498] About us - AI Training Data Company Since 2011 | Nexdata. Nexdata;
response-sha256:e97bcee54ad694e41b87cd08f2e78354b587ff37c2c41d087568a0219161035c
https://www.nexdata.ai/company/about-us
Response SHA-256: e97bcee54ad694e41b87cd08f2e78354b587ff37c2c41d087568a0219161035c
Scoped evidence bindings: 1

[499] Quality & Security - Nexdata. Nexdata;
response-sha256:87dd72bd7b4998029fe8f33c9e6b7834b9bfac4d3afca342bab8d5e54c449309
https://www.nexdata.ai/company/data-security
Response SHA-256: 87dd72bd7b4998029fe8f33c9e6b7834b9bfac4d3afca342bab8d5e54c449309
Scoped evidence bindings: 1

[500] Data Annotation Services for AI & ML | SunTec.ai. SunTec.AI;
response-sha256:d3188611ac2af1ce3a2bd8afa4a72b8662e739383955d3bdb32642c557f2c2e2
https://www.suntec.ai/data-annotation-services.html
Response SHA-256: d3188611ac2af1ce3a2bd8afa4a72b8662e739383955d3bdb32642c557f2c2e2
Scoped evidence bindings: 4

[501] SunTec India Expands AI/ML Capabilities through Strategic Evolution of SunTec.ai Division -- SunTec India | PRLog. SunTec India; response-sha256:86aecb2330afc4102b371ca58c2fe578f3c18d0adcdcf136493f6374b7d8c8c5
https://www.prlog.org/13098736-suntec-india-expands-aiml-capabilities-through-strategic-evolution-of-suntecai-division.html
Response SHA-256: 86aecb2330afc4102b371ca58c2fe578f3c18d0adcdcf136493f6374b7d8c8c5
Scoped evidence bindings: 1

[502] Privacy & Data Ethics | No Personal Data, No Login-Walled Data | TagX. TagX;
response-sha256:08f9421d475c24f5f8c5321bdfd9310a4fa4bc9be4bb3b8d57234c17839fa482
https://www.tagxdata.com/privacy
Response SHA-256: 08f9421d475c24f5f8c5321bdfd9310a4fa4bc9be4bb3b8d57234c17839fa482
Scoped evidence bindings: 1

[503] About TagX | Managed Web Data Provider. TagX;
response-sha256:075ce8d4eaf6a1412295a548bd25657955aa023fa3e626b106c3640bd2f3bdd6
https://www.tagxdata.com/about
Response SHA-256: 075ce8d4eaf6a1412295a548bd25657955aa023fa3e626b106c3640bd2f3bdd6
Scoped evidence bindings: 2

[504] Data APIs & Web Scraping Solutions | Finance, Ecommerce, Travel, Search & Social - TagX. TagX;
response-sha256:35f181231cc3527d720d7081abc078db1dc86546cace617bd185d3bd9cc7e6e4
https://www.tagxdata.com/
Response SHA-256: 35f181231cc3527d720d7081abc078db1dc86546cace617bd185d3bd9cc7e6e4
Scoped evidence bindings: 2

[505] Trupp Global Successfully Completes VAPT for Secure Operations. Trupp Global;
response-sha256:efca447576726299079d1f0bef1616f8e389a96de7e6c049f2fd7a1d4c6a4666
https://www.truppglobal.com/press-and-news/trupp-global-successful-vapt-completion/
Response SHA-256: efca447576726299079d1f0bef1616f8e389a96de7e6c049f2fd7a1d4c6a4666
Scoped evidence bindings: 1

[506] Business Process Outsourcing Company India | Trupp Global. Trupp Global;
response-sha256:19223ee4b2462e4d78bca9982be1b9bd30559fd1ccc544fee750ff591164c08b
https://www.truppglobal.com/
Response SHA-256: 19223ee4b2462e4d78bca9982be1b9bd30559fd1ccc544fee750ff591164c08b
Scoped evidence bindings: 1

[507] Data Annotation: Definition, Types, & Benefits | Trupp Global. Trupp Global;
response-sha256:884113813f9dc90f37b9e3e0f3e77c22a61292e52715d0a4ddd46b4c8a0015e3
https://www.truppglobal.com/what-we-do/data-annotation/
Response SHA-256: 884113813f9dc90f37b9e3e0f3e77c22a61292e52715d0a4ddd46b4c8a0015e3
Scoped evidence bindings: 1

[508] Trupp Global Privacy Policy: Safeguarding Your Information. Trupp Global;
response-sha256:bd28f80adfbaaa41a33d51c68b6685661bb3a357aa05e3e19ff9ddf3c484a3af
https://www.truppglobal.com/privacy-policy/
Response SHA-256: bd28f80adfbaaa41a33d51c68b6685661bb3a357aa05e3e19ff9ddf3c484a3af
Scoped evidence bindings: 1

[509] APEX-Accounting primary paper context: expert-construction. APEX-Accounting creator source; 2607.27189v1
https://arxiv.org/pdf/2607.27189v1
Response SHA-256: b76334035e09371892a490a9ff3a0366877224aa3f9b9c6a549a40b43c4a1b4f
Scoped evidence bindings: 5

[510] APEX-Accounting primary provider_documentation context: expert-construction. APEX-Accounting creator source;
body-sha256:b8296849701e00ebb69d95b6792541655da2e899416d17d5e1d8b85fbb36323b
https://www.mercor.com/blog/introducing-the-ai-productivity-index-for-accounting/
Response SHA-256: b8296849701e00ebb69d95b6792541655da2e899416d17d5e1d8b85fbb36323b
Scoped evidence bindings: 2

[511] APEX-Accounting primary dataset_card context: judge-and-reliability. APEX-Accounting creator source;
1662b7e4c3b04b8717f2f723df823e7b6cbe2118
https://huggingface.co/datasets/mercor/apex-accounting-harbor/resolve/1662b7e4c3b04b8717f2f723df823e7b6cbe2118/README.md
Response SHA-256: af6c3db7b7a3fed38d6a8df850992c89045cc65f41953aea65c948753d679777
Scoped evidence bindings: 5

[512] APEX-Accounting primary provider_documentation context: unresolved-version-counts. APEX-Accounting creator source; body-sha256:7538208bff5f7efaa2e91e217719dbe7297c3a87eb8a70f8e249ad50d4649776
https://www.mercor.com/apex/apex-accounting-leaderboard/
Response SHA-256: 7538208bff5f7efaa2e91e217719dbe7297c3a87eb8a70f8e249ad50d4649776
Scoped evidence bindings: 1

[513] APEX-Agents primary paper context: expert-world-construction. APEX-Agents creator source; 2601.14242v1
https://arxiv.org/pdf/2601.14242v1
Response SHA-256: 581d77904b0b6ff5671b6a110d81f3eb4e4ceb2da78c25746169ccc4ddd9a408
Scoped evidence bindings: 4

[514] APEX-Agents primary dataset_card context: expert-world-construction. APEX-Agents creator source; 941dac17c596b2bda9bd06b5c9f517d69c422abf
https://huggingface.co/datasets/mercor/apex-agents-v1.1/resolve/941dac17c596b2bda9bd06b5c9f517d69c422abf/README.md
Response SHA-256: 8fe1d0d1cb3c75ec8948b7155d2080329b2bab5ac809ec63e3fa456bd4d834f7
Scoped evidence bindings: 5

[515] APEX-Agents primary provider_documentation context: metric-and-release-audits. APEX-Agents creator source; body-sha256:43928e13914303e50f832607d4d1b83308beea7b90074f5f9957fdee8a8d9bc0
https://www.mercor.com/blog/introducing-apex-agents-1-1/
Response SHA-256: 43928e13914303e50f832607d4d1b83308beea7b90074f5f9957fdee8a8d9bc0
Scoped evidence bindings: 1

[516] ClawsBench primary paper context: experimental-design. ClawsBench creator source; 2604.05172v1
https://arxiv.org/pdf/2604.05172v1
Response SHA-256: e03bc02209a3eda17a23cf95c882c370cb463af1b6081eae08b3d1c0806c5a76
Scoped evidence bindings: 5

[517] ClawsBench primary repository context: noncommercial-rights. ClawsBench creator source; c9e4dbaa2efb9856e6b997a3c5f42fb93a2c5bfd
https://raw.githubusercontent.com/benchflow-ai/ClawsBench/c9e4dbaa2efb9856e6b997a3c5f42fb93a2c5bfd/README.md
Response SHA-256: 88a03f22b8725629023bc87bfaf9ee9ce39cfe500666ee0d09e3e557406f90c7
Scoped evidence bindings: 1

[518] ClawsBench primary repository context: noncommercial-rights. ClawsBench creator source; c9e4dbaa2efb9856e6b997a3c5f42fb93a2c5bfd
https://raw.githubusercontent.com/benchflow-ai/ClawsBench/c9e4dbaa2efb9856e6b997a3c5f42fb93a2c5bfd/LICENSE
Response SHA-256: 69a7b701ced8e8a6d2a285de79f3b469fd6fd865a7061106e39e3dbc5914746b
Scoped evidence bindings: 1

[519] ClawsBench primary dataset_card context: task.and-grading-units. ClawsBench creator source; e7c45cc9ff486502176267c1294ac5809cf0700a
https://huggingface.co/datasets/benchflow/ClawsBench/resolve/e7c45cc9ff486502176267c1294ac5809cf0700a/README.md
Response SHA-256: 27e4676086a73d25afd092dc14201c2da3d5487a0b3aee9ee47c72dfc1de7526
Scoped evidence bindings: 3

[520] DAYJOB primary repository context: current-dense-score-protocol. DAYJOB creator source; e32374883b9dcf97e6d0ac94c467235acbaa15cd
https://raw.githubusercontent.com/surge-ai/dayjob/e32374883b9dcf97e6d0ac94c467235acbaa15cd/README.md
Response SHA-256: d45da50d06604f3215fb5e3063c0a78b3ceacaaec2f09b88018c686e2333c404
Scoped evidence bindings: 4

[521] DAYJOB primary dataset_card context: domain-pass-thresholds. DAYJOB creator source; 308ec7ffb5c5447067f22b0b0e98d82b995615fa
https://huggingface.co/datasets/surgeai/DAYJOB-finance/resolve/308ec7ffb5c5447067f22b0b0e98d82b995615fa/README.md
Response SHA-256: b493b4ef5486e7f43d1b392003b66bd2e79c22ee2c98bb030cddbd5c74176582
Scoped evidence bindings: 7

[522] DAYJOB primary dataset_card context: domain-pass-thresholds. DAYJOB creator source; 0aaa171a0e628cd86a134b3b4a076a9829944fae
https://huggingface.co/datasets/surgeai/DAYJOB-healthcare/resolve/0aaa171a0e628cd86a134b3b4a076a9829944fae/README.md
Response SHA-256: d808539106147378ecf6a4db549b9f03192b61ed3f3a2406d27ed01af41c8e5f
Scoped evidence bindings: 3

[523] DAYJOB primary dataset_card context: domain-pass-thresholds. DAYJOB creator source; 91237b657aded987efb0180acaffff11c2cb3418
https://huggingface.co/datasets/surgeai/DAYJOB-tax/resolve/91237b657aded987efb0180acaffff11c2cb3418/README.md
Response SHA-256: 4164dc18768687b9ce67e3424ceebc02a5474fdada8ab9ca57c9e2cde55442ac
Scoped evidence bindings: 3

[524] Toloka Visual Question Answering / WSDMCup2023 primary paper context: crowd-construction. Toloka Visual Question Answering / WSDMCup2023 creator source; body-sha256:f72ce0cdc18b596cf8ec5f06f554ca0342e4a066004f6c73123a996afc69d555
https://ceur-ws.org/Vol-3357/invited1.pdf
Response SHA-256: f72ce0cdc18b596cf8ec5f06f554ca0342e4a066004f6c73123a996afc69d555
Scoped evidence bindings: 5

[525] Toloka Visual Question Answering / WSDMCup2023 primary dataset_card context: crowd-construction. Toloka Visual Question Answering / WSDMCup2023 creator source; 8afd3c854c3ac246be112e1a055fb5efaefb0e96
https://huggingface.co/datasets/toloka/WSDMCup2023/resolve/8afd3c854c3ac246be112e1a055fb5efaefb0e96/README.md
Response SHA-256: ea54f2f44899857704079cebfc89234a27d93ae1734dda422efa3f5858703bac
Scoped evidence bindings: 3

[526] Regulated Advice Bench primary dataset_card context: agreement-denominators. Regulated Advice Bench creator source; 8bc98898cb27af227328b93c9a368b47aca52bca
https://huggingface.co/datasets/alinia/regulated-advice-bench/resolve/8bc98898cb27af227328b93c9a368b47aca52bca/README.md
Response SHA-256: 1d327284db0f333296b3352a27db1e7405b1b79a7bd35c8164c731acc2656d6f
Scoped evidence bindings: 7

[527] Pareto AttuneBench primary repository context: code-data-and-use-terms. Pareto AttuneBench creator source; 37c429ab84c1f80e2ba5ac312a4eab30cb0260fc
https://raw.githubusercontent.com/Thoughtful-Lab/attunebench/37c429ab84c1f80e2ba5ac312a4eab30cb0260fc/README.md
Response SHA-256: ec88e4bb93934b568c777a6847d0e4289d0c68af2453f627a1f4ffe3cc1b97fd
Scoped evidence bindings: 2

[528] Pareto AttuneBench primary repository context: code-data-and-use-terms. Pareto AttuneBench creator source; 37c429ab84c1f80e2ba5ac312a4eab30cb0260fc
https://raw.githubusercontent.com/Thoughtful-Lab/attunebench/37c429ab84c1f80e2ba5ac312a4eab30cb0260fc/LICENSE
Response SHA-256: ed6285b47bd6cad9ef3f53ff63d5b2347e5b313062c04542eb8945b791fb1049
Scoped evidence bindings: 1

[529] Pareto AttuneBench primary paper context: composite-and-judge-scope. Pareto AttuneBench creator source; 2605.21739v1
https://arxiv.org/pdf/2605.21739v1
Response SHA-256: 79386149663a73c8869afdb086597bca67dda69fb71a42d504bd7cc575b2957c
Scoped evidence bindings: 7

[530] Pareto AttuneBench primary provider_documentation context: participant-and-task.units. Pareto AttuneBench creator source; body-sha256:bf4c78302ee38c9f64824173a350344e2e0aef04a5f18c4afe576c7952b7b57a
https://pareto.ai/blog/introducing-attunebench
Response SHA-256: bf4c78302ee38c9f64824173a350344e2e0aef04a5f18c4afe576c7952b7b57a
Scoped evidence bindings: 1

[531] Refresh BlenderBench primary provider_documentation context: access-and-read-scope. Refresh BlenderBench creator source; body-sha256:b2900ee1804bf98405890a9b02b5ae81249e3e34758ff71d06732d026148c8c5
https://refresh.dev/blog/blenderbench
Response SHA-256: b2900ee1804bf98405890a9b02b5ae81249e3e34758ff71d06732d026148c8c5
Scoped evidence bindings: 6

[532] Datacurve DeepSWE primary provider_documentation context: ai-analyzer-calibration-boundary. Datacurve DeepSWE creator source; body-sha256:709ccf8744a14c7033f07ec3a4f700e28d62156ae0480d7cb30453a563a5a485
https://deepswe.datacurve.ai/blog/deepswe
Response SHA-256: 709ccf8744a14c7033f07ec3a4f700e28d62156ae0480d7cb30453a563a5a485
Scoped evidence bindings: 7

[533] Datacurve DeepSWE primary repository context: original-task.and-repository-units. Datacurve DeepSWE creator source; 0b9fabbb63b9104d678fe965e1632f2dd9eaa2ea
https://raw.githubusercontent.com/datacurve-ai/deep-swe/0b9fabbb63b9104d678fe965e1632f2dd9eaa2ea/README.md
Response SHA-256: 87ac640023163ef48aa6c59b2e1806d34b68c4e1c3cf7d61597236c8c159f95f
Scoped evidence bindings: 3

[534] Proximal FrontierSWE primary repository context: access-attribution-and-limit. Proximal FrontierSWE creator source; da83f84f8fbcec3cbf0f2b17c98ebc811355c2df
https://raw.githubusercontent.com/Proximal-Labs/frontier-swe-v2/da83f84f8fbcec3cbf0f2b17c98ebc811355c2df/README.md
Response SHA-256: 7f54b8e27fbce695ba2172ef129a6f96212f819817dd1c0313532a422534f329
Scoped evidence bindings: 3

[535] Proximal FrontierSWE primary provider_documentation context: access-attribution-and-limit. Proximal FrontierSWE creator source; body-sha256:b89440dddd1333477eb9b1ee23c4cb2a8bb89255eb7906ca80d737c83bfb1add
https://www.frontierswe.com/blog/v2
Response SHA-256: b89440dddd1333477eb9b1ee23c4cb2a8bb89255eb7906ca80d737c83bfb1add
Scoped evidence bindings: 9

[536] Proximal FrontierSWE primary provider_documentation context: v2-family-identity. Proximal FrontierSWE creator source; body-sha256:95a6ba12041abf59c10f87d7d1d4482ecac94eda371be4d37eabed37f956f8ba
https://www.proximal.ai/blog/frontierswe/
Response SHA-256: 95a6ba12041abf59c10f87d7d1d4482ecac94eda371be4d37eabed37f956f8ba
Scoped evidence bindings: 1

[537] Polymath Horizon-SWE primary provider_documentation context: app-and-task.units. Polymath Horizon-SWE creator source; body-sha256:6bee82d9f5b2ed107faa27dc2c26fec39fe768835a299eff04ac154da47c85c7
https://polymathlabs.ai/blog/horizon-swe
Response SHA-256: 6bee82d9f5b2ed107faa27dc2c26fec39fe768835a299eff04ac154da47c85c7
Scoped evidence bindings: 4

[538] micro1 MissionReport primary dataset_card context: nonoperational-license-scope. micro1 MissionReport creator source; 2f06f512689ac5e99b2ab99895a49087bfb6ce42
https://huggingface.co/datasets/micro1-inc/MissionReport/resolve/2f06f512689ac5e99b2ab99895a49087bfb6ce42/README.md
Response SHA-256: 5c5584f8662c062f096e63104af41a5c7238c71db41453e8e846b4b8e38adcaa
Scoped evidence bindings: 9

[539] BenchFlow SkillsBench primary dataset_card context: paired-skill-protocol. BenchFlow SkillsBench creator source; be2a6ce2cb1f4ff67ce937307cade0c5a0477a13
https://huggingface.co/datasets/benchflow/SkillsBench/resolve/be2a6ce2cb1f4ff67ce937307cade0c5a0477a13/README.md
Response SHA-256: 62bb9c4a57242e7ca204fa2aa14209eee0c2a3fb69c012c208511790f1647837
Scoped evidence bindings: 3

[540] BenchFlow SkillsBench primary repository context: roster-and-release-conflicts. BenchFlow SkillsBench creator source; 27738384b1df694ea2ae466e416f476e94d8fab9
https://raw.githubusercontent.com/benchflow-ai/skillsbench/27738384b1df694ea2ae466e416f476e94d8fab9/CONTRIBUTING.md
Response SHA-256: 2e15f14168e55362a120abb1237da971a0b9ad7b84257c11f15bdd43dc3aeaee
Scoped evidence bindings: 3

[541] BenchFlow SkillsBench primary repository context: runtime-and-license-boundary. BenchFlow SkillsBench creator source; 27738384b1df694ea2ae466e416f476e94d8fab9
https://raw.githubusercontent.com/benchflow-ai/skillsbench/27738384b1df694ea2ae466e416f476e94d8fab9/README.md
Response SHA-256: 2dc46941b71ffddf5b17a78167a62eb1e903772bfeedf0d86b8553cf01c518fe
Scoped evidence bindings: 2

[542] Chatbot Arena Conversations primary paper context: crowd-versus-model-roles. Chatbot Arena Conversations creator source; 2306.05685v1
https://arxiv.org/pdf/2306.05685v1
Response SHA-256: 7f798ed891c080bab68553b3da6dfc4b4e13bae931e2140976b463b32fc6992b
Scoped evidence bindings: 7

[543] ELOQUENT Preference Prediction primary dataset_card context: card-metadata-limits. ELOQUENT Preference Prediction creator source; 8697e18df763d788b11001b06da9f67849fead53
https://huggingface.co/datasets/Eloquent/preference_prediction/resolve/8697e18df763d788b11001b06da9f67849fead53/README.md
Response SHA-256: 1cf182d2ae7600bbb40da423ed6681f7ca9dd2f0797aae44be874ce7f2f07cdf
Scoped evidence bindings: 4

[544] ELOQUENT Preference Prediction primary repository context: classification-versus-explanations. ELOQUENT Preference Prediction creator source; acc022759861adf402fa3c78aa1ef22cd5ca1909
https://raw.githubusercontent.com/eloquent-lab/eloquent-lab.github.io/acc022759861adf402fa3c78aa1ef22cd5ca1909/2025archive/task%2Dpreference-prediction/index.md
Response SHA-256: 86c1db73e62ba7264a3b81419167228a02b4d468f35550994ca0caf8a68166af
Scoped evidence bindings: 5

[545] The Flan Collection primary paper context: composite-family-identity. The Flan Collection creator source; 2301.13688v1
https://arxiv.org/pdf/2301.13688v1
Response SHA-256: 1dedd2a2a9602490d9ffd60ca0212cd0f959608695d9199287bc8ee9cb1a7ca0
Scoped evidence bindings: 5

[546] The Flan Collection primary repository context: composite-family-identity. The Flan Collection creator source; e9e4ec6e2701182c7a91af176f705310da541277
https://raw.githubusercontent.com/google-research/FLAN/e9e4ec6e2701182c7a91af176f705310da541277/flan/v2/run_example.py
Response SHA-256: 3f5530b3c9d208ac89b7405b032368c1bc3530e2fcc8318ff1cfb383e8884ac9
Scoped evidence bindings: 1

[547] The Flan Collection primary repository context: public-recipe-versus-paper. The Flan Collection creator source; e9e4ec6e2701182c7a91af176f705310da541277
https://raw.githubusercontent.com/google-research/FLAN/e9e4ec6e2701182c7a91af176f705310da541277/flan/v2/mixtures.py
Response SHA-256: 3cc2790bdc605d1315008b6ba2ed067c3be618d9b609cd590699f299938c2895
Scoped evidence bindings: 2

[548] xP3 primary paper context: evaluation-and-limits. xP3 creator source; 2211.01786v1
https://arxiv.org/pdf/2211.01786v1
Response SHA-256: f7df4420a37e7ee3da1e7fceb730e6cb6588c0dec8e759af7b988b1d4ded9a12
Scoped evidence bindings: 6

[549] Enabling Large Language Models to Generate Text with Citations - limitation context. arXiv; 2305.14627v2
https://arxiv.org/pdf/2305.14627v2
Response SHA-256: ac8809fe3ca1d52a413f41ef97afd71c43924944d512c90477ec0c28f2ea3a4d
Scoped evidence bindings: 4

[550] Inverse Scaling: When Bigger Isn't Better - limitation context. arXiv; 2306.09479v2
https://arxiv.org/pdf/2306.09479v2
Response SHA-256: 825e7953b35eb5676a40bf4552f169fbccc542eeec0b680860da2ae15a7ad9d7
Scoped evidence bindings: 4

[551] Universal and Transferable Adversarial Attacks on Aligned Language Models - limitation context. arXiv; 2307.15043v2
https://arxiv.org/pdf/2307.15043v2
Response SHA-256: ab8745c2e3c6dfd939d5a590760bbae4682d92db010947cbb306688ad086bfd1
Scoped evidence bindings: 4

[552] U-MATH: A University-Level Benchmark for Evaluating Mathematical Skills in LLMs - limitation context. arXiv; 2412.03205v4
https://arxiv.org/pdf/2412.03205v4
Response SHA-256: 713e0833c7646a528d927289a8fe502fffdd815c889802b724425bb72cec18e8
Scoped evidence bindings: 14

[553] Weak-to-Strong Generalization Through the Data-Centric Lens - limitation context. arXiv; 2412.03881v2
https://arxiv.org/pdf/2412.03881v2
Response SHA-256: 8ce1e4983d85bdbd12bc2dbd53fa81dc533d4ee48fed8f9e9b59e63ac6cab0bc
Scoped evidence bindings: 4

[554] JEEM: Vision-Language Understanding in Four Arabic Dialects - limitation context. arXiv; 2503.21910v1
https://arxiv.org/pdf/2503.21910v1
Response SHA-256: 68cf137b2a9ca67796af3c519268bc62ae0af67f5c0019d37299c46a40dd2146
Scoped evidence bindings: 4

[555] Voices of Freelance Professional Writers on AI: Limitations, Expectations, and Fears - limitation context. arXiv; 2504.05008v3
https://arxiv.org/pdf/2504.05008v3
Response SHA-256: dea8f8e609b8da3248f9dd66080a2b8c6a3753eed39f57ae04ac2c33d224ab49
Scoped evidence bindings: 4

[556] Shrinking the Generation-Verification Gap with Weak Verifiers - limitation context. arXiv; 2506.18203v3
https://arxiv.org/pdf/2506.18203v3
Response SHA-256: 397f44a0afa0d17d3929aae4e56bf0e2972c2c0b0278084f70d9132313969dc9
Scoped evidence bindings: 4

[557] AdvancedIF: Rubric-Based Benchmarking and Reinforcement Learning for Advancing LLM Instruction Following - limitation context. arXiv; 2511.10507v2
https://arxiv.org/pdf/2511.10507v2
Response SHA-256: 7dfb8637b357aa801b42c00067d9cd442cf3702e01c523583c0a0b0e95bc93da
Scoped evidence bindings: 16

[558] Multimodal RewardBench 2: Evaluating Omni Reward Models for Interleaved Text and Image - limitation context. arXiv; 2512.16899v3
https://arxiv.org/pdf/2512.16899v3
Response SHA-256: df5fc1c4706cd6b26f592bfbf60afb110c714218ba095dcc6b82e5c113e733d5
Scoped evidence bindings: 4

[559] APEX-SWE - limitation context. arXiv; 2601.08806v3
https://arxiv.org/pdf/2601.08806v3
Response SHA-256: f0188398c7c89d38d1c4ec6b5d05f585155b919978f782fc95a4e52b9b6733cb
Scoped evidence bindings: 11

[560] The Hierarchy of Agentic Capabilities: Evaluating Frontier Models on Realistic RL Environments - limitation context. arXiv; 2601.09032v1
https://arxiv.org/pdf/2601.09032v1
Response SHA-256: 3f8056e0bb613061f93002845c178f513f54314978d4c289fa9b504e1e7732a5
Scoped evidence bindings: 4

[561] APEX-Agents - limitation context. arXiv; 2601.14242v3
https://arxiv.org/pdf/2601.14242v3
Response SHA-256: 4f253c14f71cb9a1d17c9204e5b0812f42280f867d1f189f817286029e63594a
Scoped evidence bindings: 4

[562] Benchmarking Agents in Insurance Underwriting Environments - limitation context. arXiv; 2602.00456v1
https://arxiv.org/pdf/2602.00456v1
Response SHA-256: d6c158563cfc95f8a3ae15aa85c71b1f3d6f192b1d5e04df4988a0c3e24b7a35
Scoped evidence bindings: 4

[563] τ-Knowledge: Evaluating Conversational Agents over Unstructured Knowledge - limitation context. arXiv; 2603.04370v1
https://arxiv.org/pdf/2603.04370v1
Response SHA-256: f6fbe657daa349b1495bef6fecd7b1a3c845da3bf296d2589eedb45e051613bd
Scoped evidence bindings: 4

[564] PivotRL: High Accuracy Agentic Post-Training at Low Compute Cost - limitation context. arXiv; 2603.21383v1
https://arxiv.org/pdf/2603.21383v1
Response SHA-256: 9707e5c2f854e462f1b1849708eed1f1eb77724277332c9b54851ab53a3eeb09
Scoped evidence bindings: 5

[565] The Hitchhikers Guide to Rubric Quality Understanding and Enrichment - limitation context. arXiv; 2604.01375v4
https://arxiv.org/pdf/2604.01375v4
Response SHA-256: ac90febe84230e92244f48dcf25a640fd16ab96b18289c139f8be329645e0e64
Scoped evidence bindings: 4

[566] MMLU creator repository evaluate.py. Creator-maintained repository; 4450500f923c49f1fb1dd3d99108a0bd9717b660
https://raw.githubusercontent.com/hendrycks/test/4450500f923c49f1fb1dd3d99108a0bd9717b660/evaluate.py
Response SHA-256: 8b7d84e5a2e9290df0d43d1c060aa68ecb638c4666415b3a58d709c18eace4ec
Scoped evidence bindings: 3

[567] SWE-bench collection guide. Creator-maintained repository; 02e7a74ffd0b707aab73d203fe87bdc7c76afc8e
https://raw.githubusercontent.com/SWE-bench/SWE-bench/02e7a74ffd0b707aab73d203fe87bdc7c76afc8e/docs/assets/collection.md
Response SHA-256: b8ae0f6ae22121b75437dd8f643382b719aec84f877740d5b062f37490d25ea7
Scoped evidence bindings: 1

[568] SWE-bench evaluation guide. Creator-maintained repository; 02e7a74ffd0b707aab73d203fe87bdc7c76afc8e
https://raw.githubusercontent.com/SWE-bench/SWE-bench/02e7a74ffd0b707aab73d203fe87bdc7c76afc8e/docs/guides/evaluation.md
Response SHA-256: 770c8d6c45d507cd40a74297c2c7d779a5ab1f4f82da8a1a9dfae4d6bb03ff6e
Scoped evidence bindings: 2

[569] SWE-bench collection procedure README. Creator-maintained repository; 02e7a74ffd0b707aab73d203fe87bdc7c76afc8e
https://raw.githubusercontent.com/SWE-bench/SWE-bench/02e7a74ffd0b707aab73d203fe87bdc7c76afc8e/swebench/collect/README.md
Response SHA-256: e859628a4e5a29a12bdef112ca1bd4cb59d9de663dbf5f0247a311115a5e6ba6
Scoped evidence bindings: 1

[570] NVIDIA HelpSteer3 dataset card. NVIDIA; f6d145777bcbde96137596340fab89793acd1031
https://huggingface.co/datasets/nvidia/HelpSteer3/resolve/f6d145777bcbde96137596340fab89793acd1031/README.md
Response SHA-256: d5d46db58e8374937ea284b33b1583a4a465b232bf74b941745d80e3c35fd8ae
Scoped evidence bindings: 3

[571] CAIS MMLU dataset card. CAIS / MMLU creators; c30699e8356da336a370243923dbaf21066bb9fe
https://huggingface.co/datasets/cais/mmlu/resolve/c30699e8356da336a370243923dbaf21066bb9fe/README.md
Response SHA-256: 665eef35ecd8a89037c1e4f1e1a0d7fd4c154edc04c81cea5634ace2d3e96953
Scoped evidence bindings: 2

[572] OpenAssistant Conversations OASST1 dataset card. OpenAssistant / LAION-AI; fdf72ae0827c1cda404aff25b6603abec9e3399b
https://huggingface.co/datasets/OpenAssistant/oasst1/resolve/fdf72ae0827c1cda404aff25b6603abec9e3399b/README.md
Response SHA-256: 68483ac2fcc2f3f7779f453352363827678070ef35b6d746ccb2ca6958540fff
Scoped evidence bindings: 4

[573] RewardBench 2 evaluation dataset card. Ai2 / RewardBench creators; 7ff08853b0d5686e79b13fda8677024f566a104a
https://huggingface.co/datasets/allenai/reward-bench-2/resolve/7ff08853b0d5686e79b13fda8677024f566a104a/README.md
Response SHA-256: 4f9f92600767160c6b7b1a3c822361dabfea9ac254f8a152a02cb96ab0650911
Scoped evidence bindings: 1

[574] Original SWE-bench dataset card. SWE-bench creators; e48e2bd1e9fecd5bbd641e9414ac59da9f2e69f6
https://huggingface.co/datasets/princeton-nlp/SWE-bench/resolve/e48e2bd1e9fecd5bbd641e9414ac59da9f2e69f6/README.md
Response SHA-256: f0fb86ca5917025357564e0735f484bcfd340cc10445ca4eac4d866e9a164d52
Scoped evidence bindings: 2

[575] TruthfulQA HF distribution card. TruthfulQA repository creators; HF card community-contributed; 741b8276f2d1982aa3d5b832d3ee81ed3b896490
https://huggingface.co/datasets/truthfulqa/truthful_qa/resolve/741b8276f2d1982aa3d5b832d3ee81ed3b896490/README.md
Response SHA-256: 9e06075e7a580802a227e31cc456b37bdf5a174637f22e142b7609d709dd5f95
Scoped evidence bindings: 2

[576] HHH alignment parent task configuration. Creator-maintained repository; 092b196c1f8f14a54bbc62f24759d43bde46dd3b
https://raw.githubusercontent.com/google/BIG-bench/092b196c1f8f14a54bbc62f24759d43bde46dd3b/bigbench/benchmark_tasks/hhh_alignment/task.json
Response SHA-256: a49a91d541cc917e6d70e9ba7d2462b0609532b550279fccf51f2a0f77a61da5
Scoped evidence bindings: 1

[577] Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, v1. Paper authors / arXiv; 2204.05862v1
https://arxiv.org/html/2204.05862v1
Response SHA-256: c9a87e489b7059afc0b31288e8f1a7e6ed246d74dcc9409e49ebc004a2b87593
Scoped evidence bindings: 4

[578] SWE-bench creator repository README. SWE-bench creators; 02e7a74ffd0b707aab73d203fe87bdc7c76afc8e
https://raw.githubusercontent.com/SWE-bench/SWE-bench/02e7a74ffd0b707aab73d203fe87bdc7c76afc8e/README.md
Response SHA-256: 06841624d7b5130dfe3cc6714642e9a8fd2a608c95c66ae0180e409f12b73aaa
Scoped evidence bindings: 1

[579] TruthfulQA creator repository README. TruthfulQA repository creators; HF card community-contributed; d71c110897f5d31c5d7f309e7bc316c152f6f031
https://raw.githubusercontent.com/sylinrl/TruthfulQA/d71c110897f5d31c5d7f309e7bc316c152f6f031/README.md
Response SHA-256: 23834ca77c34e909cbdcceeaea89135c65b713a5d0c77be6bced7a637763ea44
Scoped evidence bindings: 3

[580] UltraFeedback creator repository README. OpenBMB / UltraFeedback creators; bf80fd46a8c6ceecc86e8babb1ae8771f26a3cbb
https://raw.githubusercontent.com/OpenBMB/UltraFeedback/bf80fd46a8c6ceecc86e8babb1ae8771f26a3cbb/README.md
Response SHA-256: ecf2a9052bbdcf4630fc97c400866d733643b9084c64b04c7f6bce73dbcc356a
Scoped evidence bindings: 1

[581] compilebench-code-license. Quesma; GitHub-blob:2fb4359106d3bcb47479820d21da2e2bcb5f00b1
https://api.github.com/repos/QuesmaOrg/CompileBench/contents/LICENSE
Response SHA-256: 1738d98cb93fa7c0c26170e5edbae4ba95c08a2194e641c1662d17c8447070f2
Scoped evidence bindings: 1

[582] decoycount-data-terms. Elorian; GitHub-blob:071e80fba98a592a53114eb3a287bc2fc0b8b564
https://api.github.com/repos/elorian-ai/decoycount/contents/DATA_LICENSE.md
Response SHA-256: 668e507d470deace52001df61da0538e366d8f7395cfc79563d87f6091499047
Scoped evidence bindings: 1

[583] decoycount-format-and-limits. Elorian; GitHub-blob:cfccc6e32582204fd6991cba75495bf9213314d4
https://api.github.com/repos/elorian-ai/decoycount/contents/DATASET.md
Response SHA-256: 5d0168df2fa38873d18ffa52eee82328f5bf84d6a272c80d9bb55c339a755997
Scoped evidence bindings: 2

[584] decoycount-release. Elorian; GitHub-blob:b4d17df87da04b6bf46e859174262b45298e76e3
https://api.github.com/repos/elorian-ai/decoycount/contents/RELEASE.md
Response SHA-256: f34cb5560bd5e754408179153e84f2c12d4ed51c530382087abacffdafb779f3
Scoped evidence bindings: 2

[585] elorian-brickvision-card. Elorian; ccedb16baadb1c7f2a85c0af04c168d953fa8b75
https://huggingface.co/datasets/elorian-ai/BrickVision/resolve/ccedb16baadb1c7f2a85c0af04c168d953fa8b75/README.md
Response SHA-256: 2ef3e30dadc72ea9bd1aa4d8ed620545c1ce9799d3fbc29be1b975625f54bd76
Scoped evidence bindings: 5

[586] elorian-brickvision. Elorian;
body-sha256:ada0c46a48d29aba52edde0dd6bcdea57e61b16446b1151bef6784962220f779
https://elorian.ai/blog/why-grounding
Response SHA-256: ada0c46a48d29aba52edde0dd6bcdea57e61b16446b1151bef6784962220f779
Scoped evidence bindings: 1

[587] elorian-decoycount-repo. Elorian; GitHub-blob:fb42ade06749442dcc382778d75d5bce33e215c0
https://api.github.com/repos/elorian-ai/decoycount/readme
Response SHA-256: 0c85d623b42a38521bc25dec609c48c5b0c7da0efd0369b032656f306eedad56
Scoped evidence bindings: 2

[588] elorian-decoycount. Elorian;
body-sha256:036234eea02e98a60455ece65f8b9c3ae5ef54c0008fd32a9ddc2e1b7ff8ebcd
https://elorian.ai/blog/decoycount
Response SHA-256: 036234eea02e98a60455ece65f8b9c3ae5ef54c0008fd32a9ddc2e1b7ff8ebcd
Scoped evidence bindings: 6

[589] elorian-sciencevision. Elorian;
body-sha256:99fd76e1bdbad37eec877d963b99cc111d9c628887897f250200a18e82088c7e
https://elorian.ai/blog/sciencevision
Response SHA-256: 99fd76e1bdbad37eec877d963b99cc111d9c628887897f250200a18e82088c7e
Scoped evidence bindings: 8

[590] quesma-babaisbench-repo. Quesma; GitHub-blob:dee1f7dbee8af3b74106577c49d0cffd91a4c381
https://api.github.com/repos/stared/baba-is-harbor/readme
Response SHA-256: fef54eb76d1c2a464451333fe89886a43b371b572763eec8bcb788bccd02ec45
Scoped evidence bindings: 6

[591] quesma-babaisbench. Quesma;
body-sha256:3b66bae56f5f7428600a77d3537d0e1f1df57dae643c81bef9553d08d82d7688
https://quesma.com/benchmarks/babaisbench/
Response SHA-256: 3b66bae56f5f7428600a77d3537d0e1f1df57dae643c81bef9553d08d82d7688
Scoped evidence bindings: 2

[592] quesma-binaryaudit-repo. Quesma; GitHub-blob:1e3a134981efd907a190bab3d7198a7a33a06fc2
https://api.github.com/repos/QuesmaOrg/BinaryAudit/readme
Response SHA-256: cf049820a91903abc8eaded402942d0e9e69402edeba62b9a9cc231dcaf77a39
Scoped evidence bindings: 4

[593] quesma-binaryaudit. Quesma;
body-sha256:153bda4d7431d0f2eb2bb2d500afaee6903dbeb27f0e479521e4fd7858cdfab6
https://quesma.com/benchmarks/binaryaudit/
Response SHA-256: 153bda4d7431d0f2eb2bb2d500afaee6903dbeb27f0e479521e4fd7858cdfab6
Scoped evidence bindings: 3

[594] quesma-compilebench-repo. Quesma; GitHub-blob:902a931b7c33926e03648e39b3c078f0d3b43ebd
https://api.github.com/repos/QuesmaOrg/CompileBench/readme
Response SHA-256: e540f3ae25ac637e2c231cf3f72f7594e96ab8adcd8a206f563d95f67ca6a05f
Scoped evidence bindings: 5

[595] quesma-otel-methods-article. Quesma;
body-sha256:11d5f342de0c6efc3b9adf9e8a788fcfed54879744e027a788e06646a80b1348
https://quesma.com/blog/introducing-otel-bench/
Response SHA-256: 11d5f342de0c6efc3b9adf9e8a788fcfed54879744e027a788e06646a80b1348
Scoped evidence bindings: 4

[596] quesma-otelbench-repo. Quesma; GitHub-blob:7b4d9b63a1dda0f4fc91b8fbb68636c849f10e38
https://api.github.com/repos/QuesmaOrg/otel-bench/readme
Response SHA-256: d9741f33a82c4452ea838cd9b44a10ca0afd7f893fb37118942173fd6cfe340a
Scoped evidence bindings: 3

[597] refresh-emrbench. Refresh;
body-sha256:073c58623e6a84d1301a5e3b18d386e28462ab5419fb18bc1a040ca9c586d56b
https://refresh.dev/blog/emr-bench
Response SHA-256: 073c58623e6a84d1301a5e3b18d386e28462ab5419fb18bc1a040ca9c586d56b
Scoped evidence bindings: 8

[598] refresh-one-pager. Refresh;
body-sha256:8c21d7374c7c733649582386112ea4750a5185f39c5e3fb9cd70f020141f04c8
https://refresh.dev/one-pager
Response SHA-256: 8c21d7374c7c733649582386112ea4750a5185f39c5e3fb9cd70f020141f04c8
Scoped evidence bindings: 2

[599] refresh-pdfbench. Refresh;
body-sha256:8e7739c0fccadef5cebecc86394aaeeb0d41284c574dd5354b5119611059ac21
https://refresh.dev/blog/pdf-bench
Response SHA-256: 8e7739c0fccadef5cebecc86394aaeeb0d41284c574dd5354b5119611059ac21
Scoped evidence bindings: 7

[600] scale-hss-card. Scale AI; 9add42c59783cdd0d5b871473374e59f05f2b3cb
https://huggingface.co/datasets/ScaleAI/HSS/resolve/9add42c59783cdd0d5b871473374e59f05f2b3cb/README.md
Response SHA-256: a5e7d9f8b06b79a97b8fa4289c623ba2fc50ae04bfd254491d48cb986b95cd8f
Scoped evidence bindings: 5

[601] swe-marathon-paper-v1. SWE-Marathon creators / Abundant;
body-sha256:f0276a983f7f32c5f5bd8aeacb82141282da2bde3cc8da9cc879b9dbe2e20f5f
https://arxiv.org/pdf/2606.07682v1
Response SHA-256: f0276a983f7f32c5f5bd8aeacb82141282da2bde3cc8da9cc879b9dbe2e20f5f
Scoped evidence bindings: 11

[602] swe-marathon-primary-readme. SWE-Marathon creators / Abundant;
GitHub-blob:9c45212807174aff78b014c9ea1e5ee207756c02;
API-body-sha256:9c8b0743ae9875b876c3368947c505aa2762395cd4c77007822112f48e450b22
https://api.github.com/repositories/1153976336/readme
Response SHA-256: 9c8b0743ae9875b876c3368947c505aa2762395cd4c77007822112f48e450b22
Scoped evidence bindings: 2

[603] swe-marathon-refresh-fork-readme. SWE-Marathon creators / Abundant;
GitHub-blob:040356da67f4e9ac7b4f39dc8765116ece5ba1bf;
API-body-sha256:d05036cb2b799a06335d70065f851c56502e1058f645c073279e0bc7a1d412d9
https://api.github.com/repos/refreshdotdev/long-horizon/readme
Response SHA-256: d05036cb2b799a06335d70065f851c56502e1058f645c073279e0bc7a1d412d9
Scoped evidence bindings: 1

[604] trajectory72-commissioner-methods. Anthropic (commissioning publisher);
body-sha256:01de78d6668c4637906840660344872b288978fafbcffc5ff384322b901d9d68
https://claude.com/resources/articles/auto-mode-default-in-claude-code
Response SHA-256: 01de78d6668c4637906840660344872b288978fafbcffc5ff384322b901d9d68
Scoped evidence bindings: 5

[605] The AI Productivity Index (APEX). Mercor and named primary-document authors; arxiv:2509.25721v1
https://arxiv.org/pdf/2509.25721v1
Response SHA-256: b14916ed871e2cc0d01f9009d83628454c237b3949bd3184cb6ec5e0e80880c1
Scoped evidence bindings: 15

[606] Mercor/APEX-v1 creator dataset card. Mercor and named primary-document authors;
huggingface-dataset:e0db9513115f8d0449591fac3d77d4bdc1a98fef
https://huggingface.co/datasets/Mercor/APEX-v1/resolve/e0db9513115f8d0449591fac3d77d4bdc1a98fef/README.md
Response SHA-256: bf58471b2b23e90e7902f15f6595845ebaaf9531f68a55f74668d980a728e851
Scoped evidence bindings: 3

[607] DeepSWE: Measuring Frontier Coding Agents on Original, Long-Horizon Engineering Tasks. Datacurve and named DeepSWE authors; arxiv:2607.07946v1
https://arxiv.org/pdf/2607.07946v1
Response SHA-256: 1d988c90737299c6d844694ba967ff63f4e2aa13c66be20b97b8fa4bd89754ce
Scoped evidence bindings: 22

[608] datacurve-ai/deep-swe LICENSE. Datacurve and named DeepSWE authors;
git:0b9fabbb63b9104d678fe965e1632f2dd9eaa2ea
https://raw.githubusercontent.com/datacurve-ai/deep-swe/0b9fabbb63b9104d678fe965e1632f2dd9eaa2ea/LICENSE
Response SHA-256: c71d239df91726fc519c6eb72d318ec65820627232b2f796219e87dcf35d0ab4
Scoped evidence bindings: 2

[609] datacurve-ai/deep-swe README.md. Datacurve and named DeepSWE authors;
git:0b9fabbb63b9104d678fe965e1632f2dd9eaa2ea
https://raw.githubusercontent.com/datacurve-ai/deep-swe/0b9fabbb63b9104d678fe965e1632f2dd9eaa2ea/README.md
Response SHA-256: 87ac640023163ef48aa6c59b2e1806d34b68c4e1c3cf7d61597236c8c159f95f
Scoped evidence bindings: 2

[610] imlrz/DeepResearch-Bench-II LICENSE. Ruizhe Li and coauthors; DeepResearch Bench II creators;
git:7c7a9a8cda9dad4e6728f4d6a89bb04541feb222
https://raw.githubusercontent.com/imlrz/DeepResearch-Bench-II/7c7a9a8cda9dad4e6728f4d6a89bb04541feb222/LICENSE
Response SHA-256: 80bc9fe3eb9ed21cfd800df9d36304c05c3d95d4db2053e17989fe5d72cebcdc
Scoped evidence bindings: 2

[611] DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Report. Ruizhe Li and coauthors; DeepResearch Bench II creators; arxiv:2601.08536v1
https://arxiv.org/pdf/2601.08536v1
Response SHA-256: 49f1c582a2ed26260197e25b9aca1661a9bbc52676e62863d5ace6b4f7279023
Scoped evidence bindings: 22

[612] imlrz/DeepResearch-Bench-II README.md. Ruizhe Li and coauthors; DeepResearch Bench II creators;
git:7c7a9a8cda9dad4e6728f4d6a89bb04541feb222
https://raw.githubusercontent.com/imlrz/DeepResearch-Bench-II/7c7a9a8cda9dad4e6728f4d6a89bb04541feb222/README.md
Response SHA-256: 0f696da1e57ea96fa30b3eb4e59d8ad365239d6a1a7ff06245b8a6fdcad4dd20
Scoped evidence bindings: 5

[613] Towards a ‘Science’ of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. Eduard Hovy and Julia Lavid; International Journal of Translation22(1), Jan-Jun2010; captured document response sha256:0ce0c6a23e1ce3a167ecbed07c9d64fb96464dfa06ea56a4d04790f1fdbe531d
https://www.cs.cmu.edu/~hovy/papers/10KNS-annotation-Hovy-Lavid.pdf
Response SHA-256: 0ce0c6a23e1ce3a167ecbed07c9d64fb96464dfa06ea56a4d04790f1fdbe531d
Scoped evidence bindings: 11

[614] HCAST primary documentation. METR and primary-document authors; arxiv:2503.17354v1
https://arxiv.org/pdf/2503.17354v1
Response SHA-256: a0db39c1ee307d08c4349057c7d165ad8be0d637fdf23e2f47a14558aaf824b5
Scoped evidence bindings: 14

[615] METR/hcast-public LICENSE. METR and primary-document authors;
git:03438239d200fb3e89a5ddfc62ab4d23cc1866fb
https://raw.githubusercontent.com/METR/hcast-public/03438239d200fb3e89a5ddfc62ab4d23cc1866fb/LICENSE
Response SHA-256: 166dac7e22cca09ab287fc5b746ff60e4a341b720aca808ad7d3ce1cc43aceff
Scoped evidence bindings: 1

[616] METR/hcast-public README.md. METR and primary-document authors;
git:03438239d200fb3e89a5ddfc62ab4d23cc1866fb
https://raw.githubusercontent.com/METR/hcast-public/03438239d200fb3e89a5ddfc62ab4d23cc1866fb/README.md
Response SHA-256: 20f8414f446b10f2c491315de0236922125bb5f03dac250896b30ae15457b354
Scoped evidence bindings: 1

[617] Measuring AI Ability to Complete Long Software Tasks. METR and primary-document authors; arxiv:2503.14499v4
https://arxiv.org/pdf/2503.14499v4
Response SHA-256: 29216eb3af4daa417666757a462ab7c7b5f8c18c188ee5a56e258b04116211f7
Scoped evidence bindings: 24

[618] METR Time Horizon 1.1 creator release document. METR and primary-document authors; captured response
sha256:02fb9809775dba58897de77d92d7e1cec2f35ecd17907ea028bf5cf65ffb5aa5
https://metr.org/blog/2026-01-29-time-horizon-1-1/
Response SHA-256: 02fb9809775dba58897de77d92d7e1cec2f35ecd17907ea028bf5cf65ffb5aa5
Scoped evidence bindings: 5

[619] VLDB 2021 Crowd Science Challenge on Aggregating Crowdsourced Audio Transcriptions. Ustalov, Pavlichenko, Stelmakh and Kuznetsov; CEUR Workshop Proceedings; captured document response
sha256:fb49f2021942a49d40c699446313ba08397ba84dc721c311b498c032d905bdbc
https://ceur-ws.org/Vol-2932/invited1.pdf
Response SHA-256: fb49f2021942a49d40c699446313ba08397ba84dc721c311b498c032d905bdbc
Scoped evidence bindings: 8

[620] WSDM Cup 2023 Challenge on Visual Question Answering. Ustalov, Pavlichenko, Likhobaba and Smirnova; CEUR Workshop Proceedings; captured document response
sha256:f72ce0cdc18b596cf8ec5f06f554ca0342e4a066004f6c73123a996afc69d555
https://ceur-ws.org/Vol-3357/invited1.pdf
Response SHA-256: f72ce0cdc18b596cf8ec5f06f554ca0342e4a066004f6c73123a996afc69d555
Scoped evidence bindings: 10

[621] Gauntlet 4K-RLVR creator subset article. Refresh;
body-sha256:f82fb6f79540e02c2980ea20edb9d72ad5b73211af0051681a65ac14d2b868ab
https://refresh.dev/blog/gauntlet-4k-rlvr
Response SHA-256: f82fb6f79540e02c2980ea20edb9d72ad5b73211af0051681a65ac14d2b868ab
Scoped evidence bindings: 2

[622] Gauntlet 80K creator construction article. Refresh;
body-sha256:47cbdd1710cb41f52aebabea877862dbbcd3fbed089c09681488c4e4d5180da6
https://refresh.dev/blog/gauntlet-80k
Response SHA-256: 47cbdd1710cb41f52aebabea877862dbbcd3fbed089c09681488c4e4d5180da6
Scoped evidence bindings: 4

[623] GDP.xlsx creator dataset card. Surge AI; 7cdb5406bb7bd12cf73cbaaabb0edcf449a473e9
https://huggingface.co/datasets/surgeai/GDP.xlsx/resolve/7cdb5406bb7bd12cf73cbaaabb0edcf449a473e9/README.md
Response SHA-256: ca1d638949580d6ef8608ca076a7d0f400cf94130e8a4dae346ce80bb862b341
Scoped evidence bindings: 7

[624] MedMosaic anonymous dataset card. icml-anon-submission (anonymous creator account); a6ea67bd4a65b87248c6651e559656b2c31fa669
https://huggingface.co/datasets/icml-anon-submission/medmosaic-dataset/resolve/a6ea67bd4a65b87248c6651e559656b2c31fa669/README.md
Response SHA-256: b5efa8829dd6b4da6867da9de0b77a440278a26da9e96cd86630c7fa064ea0c3
Scoped evidence bindings: 5

[625] Muse Spark safety and preparedness report: Search-PI methodology. Meta; body-sha256:1b760ea7770828e223daaa966a547529b4b1398a7ac44b97992488fab740133a
https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report/
Response SHA-256: 1b760ea7770828e223daaa966a547529b4b1398a7ac44b97992488fab740133a
Scoped evidence bindings: 8

[626] mu-math-supplement-card. Toloka; c3d791bc7be33bf73a1ae5d8c3084a820cb580c4
https://huggingface.co/datasets/toloka/mu-math/resolve/c3d791bc7be33bf73a1ae5d8c3084a820cb580c4/README.md
Response SHA-256: 98ad3e0aa46f4fa2ddffa76dfbf582b5c69c2c25d6ae24ecbac7865ee1e2ca28
Scoped evidence bindings: 4

[627] Multi-SWE-bench. ByteDance Seed; 56ff018c04a38e27ada1e9d0a6d5839a51f88f0d
https://huggingface.co/datasets/ByteDance-Seed/Multi-SWE-bench/resolve/56ff018c04a38e27ada1e9d0a6d5839a51f88f0d/README.md
Response SHA-256: 26638a5dc8d8c10e04de4578c904beecefb584d44e618cfe2c33fd350ca9810d
Scoped evidence bindings: 1

[628] Multi-SWE-bench creator repository. Multi-SWE-bench creators; 24f493f8a103e72312ded4f6b9c89f081d69cb09
https://raw.githubusercontent.com/multi-swe-bench/multi-swe-bench/24f493f8a103e72312ded4f6b9c89f081d69cb09/README.md
Response SHA-256: f1bef6ea0ba9b9062ba9e70ec5aa92d1b8dca795be1eb4adba68a299189a369c
Scoped evidence bindings: 1

[629] multi-swe-paper-v1. ByteDance Seed / Multi-SWE-bench authors; 2504.02605v1
https://arxiv.org/pdf/2504.02605v1
Response SHA-256: 4de921d8d88a016762ad1680ec8f5b24381c8c69442e96a4265a50f1bce31d92
Scoped evidence bindings: 10

[630] Multi-SWE-RL. ByteDance Seed; 9777648932daa214ba18c70c81e85821b5836f32
https://huggingface.co/datasets/ByteDance-Seed/Multi-SWE-RL/resolve/9777648932daa214ba18c70c81e85821b5836f32/README.md
Response SHA-256: 1053af7c55dc154042891670d4e1b97707af59010a898d5ba2a1e7184e66b412
Scoped evidence bindings: 2

[631] Snorkel Finance Reasoning methodology. Snorkel AI; body-sha256:58e9a00deedf172e09f8177d6179f4f08453e24c18041dfaa07fc5eba34ef2d3
https://snorkel.ai/leaderboard/category/financereasoning/
Response SHA-256: 58e9a00deedf172e09f8177d6179f4f08453e24c18041dfaa07fc5eba34ef2d3
Scoped evidence bindings: 2

[632] snorkel-finance-supplement-card. Snorkel AI; 4c89e0aa8358509b6645b87ede329300336e4d3c
https://huggingface.co/datasets/snorkelai/agent-finance-reasoning/resolve/4c89e0aa8358509b6645b87ede329300336e4d3c/README.md
Response SHA-256: a3cf96396ae41d71b7f5d4f5c3c77a209da7bbb7be5e080b009c1438892995d5
Scoped evidence bindings: 6

[633] Toloka U-MATH creator card. Toloka; 7210f97b3f21122c3a90126935e966ca7d8951b6
https://huggingface.co/datasets/toloka/u-math/raw/7210f97b3f21122c3a90126935e966ca7d8951b6/README.md
Response SHA-256: c13bf3579ababfb2510ca76ca0de6185ef2ff97caaf69ebf6a89d761c8c42871
Scoped evidence bindings: 1

[634] CharXiv primary manuscript v1. CharXiv paper authors via arXiv; arxiv:2406.18521v1
https://arxiv.org/pdf/2406.18521v1
Response SHA-256: 696ad9488b3d5f6d842fd1278d5be0fa2263676740d65bf1469bdd618af8a36c
Scoped evidence bindings: 18

[635] Immutable CharXiv creator repository README. CharXiv creators; 7ebe88f78dee387691551f071abcb2b9e1a8025b
https://raw.githubusercontent.com/princeton-nlp/CharXiv/7ebe88f78dee387691551f071abcb2b9e1a8025b/README.md
Response SHA-256: f6aa799fd28e2a90b5d6fe03b51f08583d6c7c42d6edbbfc9cae86975313919a
Scoped evidence bindings: 5

[636] SciCode captured official project overview. SciCode creators;
response-sha256:e769e53ee838c8abf3bb676ff0ca9c620d83135c9204857990a7883e2ac356de
https://scicode-bench.github.io/
Response SHA-256: e769e53ee838c8abf3bb676ff0ca9c620d83135c9204857990a7883e2ac356de
Scoped evidence bindings: 1

[637] SciCode creator readme. SciCode creators; e3158ea011d4235245a547460d3688d7ccbf9900
https://raw.githubusercontent.com/scicode-bench/SciCode/e3158ea011d4235245a547460d3688d7ccbf9900/README.md
Response SHA-256: a06846f318d4d269449d83201169058bc4e1ee64f3375a6104c4477d6ef03799
Scoped evidence bindings: 4

[638] SciCode creator faq. SciCode creators;
response-sha256:cf673feddd183a6beb87387289add8a3e905bfdb24d71510a6b4dc1de6be0cef
https://scicode-bench.github.io/faq/
Response SHA-256: cf673feddd183a6beb87387289add8a3e905bfdb24d71510a6b4dc1de6be0cef
Scoped evidence bindings: 4

[639] Immutable CharXiv creator dataset card. CharXiv creators; f441eb632fc62f6f777830a0f47619e6e86459b0
https://huggingface.co/datasets/princeton-nlp/CharXiv/resolve/f441eb632fc62f6f777830a0f47619e6e86459b0/README.md
Response SHA-256: 1874b3229ec3fa78abd5aa8ef51548b673ad44cefacd79cf632ebd3e299b1f51
Scoped evidence bindings: 3

[640] CVDP immutable creator data_license. CVDP authors / NVIDIA; 5b807d945f6a99aa645f7e43a64a2115e281b4bf
https://huggingface.co/datasets/nvidia/cvdp-benchmark-dataset/resolve/5b807d945f6a99aa645f7e43a64a2115e281b4bf/LICENSE
Response SHA-256: cedcd612607018ad841d87d7f1c877630778a50c71692610fe76acdc22700719
Scoped evidence bindings: 1

[641] CVDP immutable creator notice. CVDP authors / NVIDIA; 5b807d945f6a99aa645f7e43a64a2115e281b4bf
https://huggingface.co/datasets/nvidia/cvdp-benchmark-dataset/resolve/5b807d945f6a99aa645f7e43a64a2115e281b4bf/NOTICE
Response SHA-256: 3d8753e57eab52910ccb61a1ee09113c43e9382ccd98d5860b5621aff8d932dd
Scoped evidence bindings: 2

[642] CVDP immutable creator code_license. CVDP authors / NVIDIA; 8e894cf74414ab1eaea1e2b4e80a02f123df07b6
https://raw.githubusercontent.com/NVlabs/cvdp_benchmark/8e894cf74414ab1eaea1e2b4e80a02f123df07b6/LICENSE
Response SHA-256: 94fc060effda8b8b950639c87ece92aa212ab25024434350b5bd6623f6d26a88
Scoped evidence bindings: 1

[643] CVDP primary manuscript v1. CVDP authors / NVIDIA; arxiv:2506.14074v1
https://arxiv.org/pdf/2506.14074v1
Response SHA-256: d85e66e7fa3ae80e40c810c45877ff3e2dbed515d47f2c0326779e9ee973e58d
Scoped evidence bindings: 13

[644] CVDP immutable creator readme. CVDP authors / NVIDIA; 8e894cf74414ab1eaea1e2b4e80a02f123df07b6
https://raw.githubusercontent.com/NVlabs/cvdp_benchmark/8e894cf74414ab1eaea1e2b4e80a02f123df07b6/README.md
Response SHA-256: 41ad24d1b03880c9260d0ac1b92fbe705e4e71936de099186d49510fc80afe8a
Scoped evidence bindings: 3

[645] CVDP immutable creator changelog. CVDP authors / NVIDIA; 5b807d945f6a99aa645f7e43a64a2115e281b4bf
https://huggingface.co/datasets/nvidia/cvdp-benchmark-dataset/resolve/5b807d945f6a99aa645f7e43a64a2115e281b4bf/CHANGELOG
Response SHA-256: f8fe17ca9a38abcc83f46a1eef5ab38e4fe48d76327573c84be6cfeb1152c039
Scoped evidence bindings: 2

[646] HealthBench primary manuscript v1. HealthBench authors / OpenAI; arxiv:2505.08775v1
https://arxiv.org/pdf/2505.08775v1
Response SHA-256: f61c53f6bde84a3964531efe04a75389f9bba9e70a35b4012c871c9ec0b512f3
Scoped evidence bindings: 20

[647] HealthBench immutable creator readme. HealthBench authors / OpenAI; 652c89d0ca9df547706735883097e9537d40dc47
https://raw.githubusercontent.com/openai/simple-evals/652c89d0ca9df547706735883097e9537d40dc47/README.md
Response SHA-256: d9647c4a0e34cd9b2b528ff6986141734aac6f5f2f18a068eeadb2af0b29786f
Scoped evidence bindings: 3

[648] HealthBench immutable creator license. HealthBench authors / OpenAI; 652c89d0ca9df547706735883097e9537d40dc47
https://raw.githubusercontent.com/openai/simple-evals/652c89d0ca9df547706735883097e9537d40dc47/LICENSE
Response SHA-256: d831db55645e47ca8e491c5a0e37f1ee744d7b10bf5aa8d50146c795ac0176c0
Scoped evidence bindings: 1

[649] MedXpertQA immutable creator readme. MedXpertQA creators; 5f1db162d0d2b03d62da2d4fe086076a36105e18
https://raw.githubusercontent.com/TsinghuaC3I/MedXpertQA/5f1db162d0d2b03d62da2d4fe086076a36105e18/README.md
Response SHA-256: 63b725f43345e9f7c69468d6efd93371d73fbfc72302097d7c8ab6e7882d3745
Scoped evidence bindings: 5

[650] MedXpertQA primary manuscript v1. MedXpertQA creators; arxiv:2501.18362v1
https://arxiv.org/pdf/2501.18362v1
Response SHA-256: 20dc75d6a2a7a4117c7c8882671a794d88c4e43c759f619454fcfaf06fc123a5
Scoped evidence bindings: 23

[651] MedXpertQA immutable creator card. MedXpertQA creators; 7e7c465a68eb2b866926bfa59c8c9d17a8daba65
https://huggingface.co/datasets/TsinghuaC3I/MedXpertQA/resolve/7e7c465a68eb2b866926bfa59c8c9d17a8daba65/README.md
Response SHA-256: 7deb132dc3745ad22c8e0c3726e7f40cd30675ba5b400e876b85b138fa40c25a
Scoped evidence bindings: 4

[652] MedXpertQA immutable creator license. MedXpertQA creators; 5f1db162d0d2b03d62da2d4fe086076a36105e18
https://raw.githubusercontent.com/TsinghuaC3I/MedXpertQA/5f1db162d0d2b03d62da2d4fe086076a36105e18/LICENSE
Response SHA-256: 381f24d4c372f6cd7d1f730ba8d877f9217cf9ac0cb32a97211327c22b2da8aa
Scoped evidence bindings: 1

[653] MMLU-Pro primary manuscript v1. MMLU-Pro paper authors via arXiv; arxiv:2406.01574v1; 2024-06-03
https://arxiv.org/pdf/2406.01574v1
Response SHA-256: 9bf7b08bf1c220cb6c1ed0dcafb7d10848a21178137e13dd5e13981d8ecd15e3
Scoped evidence bindings: 10

[654] MMLU-Pro creator dataset card at immutable b189ec765aa7ed75c8acfea42df31fdae71f97be. TIGER-Lab via Hugging Face; b189ec765aa7ed75c8acfea42df31fdae71f97be
https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro/resolve/b189ec765aa7ed75c8acfea42df31fdae71f97be/README.md
Response SHA-256: 4bd710f67da3fa359a33edce1b4b5816b3de416c823c2624ba5e89c2557d2a47
Scoped evidence bindings: 4

[655] ProgramBench primary manuscript v1. ProgramBench authors / Meta; arxiv:2605.03546v1
https://arxiv.org/pdf/2605.03546v1
Response SHA-256: 87c67873f06f48044ef3a1bcd5bed21be2c90d92090bfb3b73b6ca7b9b13781c
Scoped evidence bindings: 21

[656] ProgramBench immutable usage guide. ProgramBench authors / Meta; 895f65916bc5ba34a915d36da321fde0301125ad
https://raw.githubusercontent.com/facebookresearch/ProgramBench/895f65916bc5ba34a915d36da321fde0301125ad/docs/README.md
Response SHA-256: be6a66eadee0df33deac4bed3b8d104b09e4080870e326bc615c6b8f0a7f39ba
Scoped evidence bindings: 3

[657] ProgramBench immutable Generated Tests dataset card. ProgramBench authors / Meta; 5954a1957b5e0d770ec2c5523506abd212806315
https://huggingface.co/datasets/programbench/ProgramBench-Tests/raw/5954a1957b5e0d770ec2c5523506abd212806315/README.md
Response SHA-256: 4d3647474a0d60f8e4bfd1a4c1edbea63ec5885b678106173b3168dcfeaf48a2
Scoped evidence bindings: 3

[658] ProgramBench immutable software LICENSE. ProgramBench authors / Meta; 895f65916bc5ba34a915d36da321fde0301125ad
https://raw.githubusercontent.com/facebookresearch/ProgramBench/895f65916bc5ba34a915d36da321fde0301125ad/LICENSE
Response SHA-256: e16b8d6ba8de2b8bfe25ed3d641cb3311e1cde53c31a16eebf03dccd162cb057
Scoped evidence bindings: 1

[659] SciCode primary manuscript v1. SciCode creators; arxiv:2407.13168v1
https://arxiv.org/pdf/2407.13168v1
Response SHA-256: 29d33f99c87112b3f27eaebe134733d2afcf67bfd3be81df7fa5c9457e252974
Scoped evidence bindings: 11

[660] SciCode creator card. SciCode creators; 4510f6a6aa27c43fad7b43da2c59602a86e88480
https://huggingface.co/datasets/SciCode1/SciCode/resolve/4510f6a6aa27c43fad7b43da2c59602a86e88480/README.md
Response SHA-256: f9349885a434d018008ddc95d079e00782c70fa7823d80214cb3ee3bd3e7827b
Scoped evidence bindings: 1

[661] SciCode creator license. SciCode creators; e3158ea011d4235245a547460d3688d7ccbf9900
https://raw.githubusercontent.com/scicode-bench/SciCode/e3158ea011d4235245a547460d3688d7ccbf9900/LICENSE
Response SHA-256: c71d239df91726fc519c6eb72d318ec65820627232b2f796219e87dcf35d0ab4
Scoped evidence bindings: 1

[662] SUPERChem primary manuscript v1. SUPERChem authors; arxiv:2512.01274v1; 2025-12-01
https://arxiv.org/pdf/2512.01274v1
Response SHA-256: 676b26f54f82cf1a66b0385e51b8a7de1d8608b33e735697749fe88babbdda2e
Scoped evidence bindings: 19

[663] SUPERChem immutable dataset card. SUPERChem authors; a455fb0517ec6f444666df77157a4a044660130b
https://huggingface.co/datasets/ZehuaZhao/SUPERChem/raw/a455fb0517ec6f444666df77157a4a044660130b/README.md
Response SHA-256: add0599bd83beb52ecb93b1b81fb2e7b26e378b1ee46f69bc5daeec9935e2bbd
Scoped evidence bindings: 3

[664] SUPERChem immutable evaluation repository README. SUPERChem authors; 2a8c8491ae5a2bb86f40114b6c17a40962770226
https://raw.githubusercontent.com/catalystforyou/SUPERChem_eval/2a8c8491ae5a2bb86f40114b6c17a40962770226/README.md
Response SHA-256: 540d5c92cf00bae6b82b74a94b07498f61f8831ac9d2a89e0aa1faae58e7591e
Scoped evidence bindings: 2

[665] SUPERChem immutable software LICENSE. SUPERChem authors; 2a8c8491ae5a2bb86f40114b6c17a40962770226
https://raw.githubusercontent.com/catalystforyou/SUPERChem_eval/2a8c8491ae5a2bb86f40114b6c17a40962770226/LICENSE
Response SHA-256: 581098d5eff3eed26a28226aa99cacad8b8f79d81251413e90ffd6463b899a47
Scoped evidence bindings: 1

[666] SUPERChem immutable current release protocol. SUPERChem authors; 2a8c8491ae5a2bb86f40114b6c17a40962770226
https://raw.githubusercontent.com/catalystforyou/SUPERChem_eval/2a8c8491ae5a2bb86f40114b6c17a40962770226/docs/release_protocol.md
Response SHA-256: a4656254c102e22a82a14f3d6c2bd51d8e4a2220e65a5fc5de425ac17b8faeaa
Scoped evidence bindings: 1

[667] SWE-Atlas-QnA-AA immutable creator card. Artificial Analysis; 1f28d51b6e191e9f2bfdc22b09d92eecdab6654e
https://huggingface.co/datasets/ArtificialAnalysis/SWE-Atlas-QnA-AA/raw/1f28d51b6e191e9f2bfdc22b09d92eecdab6654e/README.md
Response SHA-256: 222220083f51659e3fce9be85063a65efa3663e8c6ba69112a710259dada2b52
Scoped evidence bindings: 2

[668] SWE Atlas immutable v1.1 repository README. SWE Atlas authors / Scale AI; b3cdddf44414c0ccca50c6e89fe978bb4d844d26
https://raw.githubusercontent.com/scaleapi/SWE-Atlas/b3cdddf44414c0ccca50c6e89fe978bb4d844d26/README.md
Response SHA-256: cfc9f580b0c7cd38ba28c364a4aa9f190188ed227312c9f8227537efbc4085f0
Scoped evidence bindings: 6

[669] SWE Atlas immutable reused-criteria NOTICE. SWE Atlas authors / Scale AI; b3cdddf44414c0ccca50c6e89fe978bb4d844d26
https://raw.githubusercontent.com/scaleapi/SWE-Atlas/b3cdddf44414c0ccca50c6e89fe978bb4d844d26/NOTICE
Response SHA-256: dac7af473ac75034271edf8756bd8ed45f7fdc6b9c06c36003091f5cbeeeca7d
Scoped evidence bindings: 2

[670] SWE Atlas primary manuscript v1. SWE Atlas authors / Scale AI; arxiv:2605.08366v1; 2026-05-08
https://arxiv.org/pdf/2605.08366v1
Response SHA-256: af39ccd39558020ce2229788542294d61342d791ef5e01547b78f160bf4dd5a7
Scoped evidence bindings: 18

[671] SWE-Atlas-QnA immutable creator card. SWE Atlas authors / Scale AI;
fcdc7f8a9814e4c678c9a66d938d57afd67bc335
https://huggingface.co/datasets/ScaleAI/SWE-Atlas-QnA/raw/fcdc7f8a9814e4c678c9a66d938d57afd67bc335/README.md
Response SHA-256: 45f5a32ae5bfc1d5d50391850d9833ba27b7e3287e4188dad06870b9cc9fda4f
Scoped evidence bindings: 1

[672] SWE Atlas immutable software LICENSE. SWE Atlas authors / Scale AI;
b3cdddf44414c0ccca50c6e89fe978bb4d844d26
https://raw.githubusercontent.com/scaleapi/SWE-Atlas/b3cdddf44414c0ccca50c6e89fe978bb4d844d26/LICENSE
Response SHA-256: c71d239df91726fc519c6eb72d318ec65820627232b2f796219e87dcf35d0ab4
Scoped evidence bindings: 1

[673] GeneBench-Pro: Evaluating Multistage Statistical Reasoning in Genomics, Quantitative Biology, and Translational Biomedicine. bioRxiv / author-funder; doi:10.64898/2026.06.29.735386v2; 2026-06-30
https://www.biorxiv.org/content/10.64898/2026.06.29.735386v2.full.pdf
Response SHA-256: d4a17d6385a065abc9d6d91b94111f3621d5028305950dff4187ad10d659e858
Scoped evidence bindings: 1

[674] FrontierSWE. Proximal;
response-sha256:95a6ba12041abf59c10f87d7d1d4482ecac94eda371be4d37eabed37f956f8ba
https://www.proximal.ai/blog/frontierswe/
Response SHA-256: 95a6ba12041abf59c10f87d7d1d4482ecac94eda371be4d37eabed37f956f8ba
Scoped evidence bindings: 2

[675] TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv / paper authors; 2109.07958v2
https://arxiv.org/pdf/2109.07958v2
Response SHA-256: 6e86175861d24d075bbd8fec987cbdabd0d6332700b001a1ccf335fa72f541b2
Scoped evidence bindings: 6

[676] VideoScore repository LICENSE. TIGER-AI-Lab/VideoScore; f87faf4647e637066bcb74721670cd579e0f4349
https://raw.githubusercontent.com/TIGER-AI-Lab/VideoScore/f87faf4647e637066bcb74721670cd579e0f4349/LICENSE
Response SHA-256: 821d038afa755049f68fff5a0bcb4db416f16426b39447c095b80670be795fc8
Scoped evidence bindings: 1

[677] About Us | APTO. APTO;
response-sha256:e9e373c333e7a87335b2a2944706aea5b4653f0df94d254804ed0db63e492185
https://apto.co.jp/en/company/
Response SHA-256: e9e373c333e7a87335b2a2944706aea5b4653f0df94d254804ed0db63e492185
Scoped evidence bindings: 1

[678] Privacy Policy | APTO. APTO;
response-sha256:769a197c93259fa75670798ac4c3f713cc3a26e833bb93eb3099fcc715984b9c
https://apto.co.jp/en/privacy-policy/
Response SHA-256: 769a197c93259fa75670798ac4c3f713cc3a26e833bb93eb3099fcc715984b9c
Scoped evidence bindings: 1

[679] WebShop - LICENSE.md. princeton-nlp/WebShop; 64fa2a5c15c7daa698b9ac93f5bb5437b634c9bd
https://raw.githubusercontent.com/princeton-nlp/WebShop/64fa2a5c15c7daa698b9ac93f5bb5437b634c9bd/LICENSE.md
Response SHA-256: 8872dbf8660b00890b5e07ce1ea1f7a44fa7ae4e1857da56caba172b51dab3cb
Scoped evidence bindings: 1

[680] Privacy Notice - Labelbox. Labelbox, Inc.; Exact response
sha256:66ec60b1d22a6eb140d6d54608222e1b4dd1600b7bf98e4652b88eb510d845c0
https://docs.labelbox.com/page/privacy-notice
Response SHA-256: 66ec60b1d22a6eb140d6d54608222e1b4dd1600b7bf98e4652b88eb510d845c0
Scoped evidence bindings: 1

[681] AI Coding Agent Benchmarks & Evaluations | Quesma. Quesma;
response-sha256:b8e2233698ec81e4515a9385bca254456c6be5b90f3430541283f2b06817c362
https://quesma.com/benchmarks/
Response SHA-256: b8e2233698ec81e4515a9385bca254456c6be5b90f3430541283f2b06817c362
Scoped evidence bindings: 1

[682] Join our Expert Community | Snorkel AI - eligibility. Snorkel AI, Inc.;
response-sha256:0cccfbc0dd20cbfe01b86ff2b1dd6a3bc8922027e43278a35694ea776278da0c;
text-sha256:0eb547f1fdd67837692b5f2179d5a8864be307b1aa8d9026801278484c0dc860
https://snorkel.ai/expert-community/
Response SHA-256: 0cccfbc0dd20cbfe01b86ff2b1dd6a3bc8922027e43278a35694ea776278da0c
Scoped evidence bindings: 4

[683] Agents' Last Exam - physical page 4. arXiv; 2606.05405v4
https://arxiv.org/pdf/2606.05405v4
Response SHA-256: 300b7cf737cd52288e9bf34c8400adfa8a2331a99f05878681c7117ba67e68b3
Scoped evidence bindings: 11

[684] Frequently Asked Questions | Snorkel AI - quality-experts. Snorkel AI, Inc.;
response-sha256:3050d01c8521d07cd613c366efdef416508ce04bccbb91dd1be6829e6060de20;
text-sha256:946b548a95cc33c27395eab05cc873c86b61e844423498704df91b7a64d94bdd; 2026-06-17
https://snorkel.ai/frequently-asked-questions/
Response SHA-256: 3050d01c8521d07cd613c366efdef416508ce04bccbb91dd1be6829e6060de20
Scoped evidence bindings: 3

[685] Post-Training on Office Work Improves Software Engineering: A Behavioral Account of Cross-Domain Transfer - physical page 8. arXiv; 2608.01604v1
https://arxiv.org/pdf/2608.01604v1
Response SHA-256: 442cf8f4cc2d6d30dadfabccf5be0d25544afc6e9d6b87faa55e2b6ba9cdbdf4
Scoped evidence bindings: 6

[686] Human expertise for AI training. - humanus-heading. TrustScale;
response-sha256:4d4794da57f6c656a2b1a6c58156763e7d4a3d26c6135e04690d8e24246c3a19;
text-sha256:9a66efef99febe00fd58812e08992df5520226f8973fbf68c635d764c7e37130
https://www.trustscale.ai/assets/index-CUMNnpew.js
Response SHA-256: 4d4794da57f6c656a2b1a6c58156763e7d4a3d26c6135e04690d8e24246c3a19
Scoped evidence bindings: 3

[687] Why We Became TrustScale - official app-bundle article. TrustScale;
response-sha256:4d4794da57f6c656a2b1a6c58156763e7d4a3d26c6135e04690d8e24246c3a19;
text-sha256:3cacb9151ca3986550d693a697bb67efbfba713b7874dda2dc594fa9350d4ca2
https://www.trustscale.ai/assets/index-CUMNnpew.js
Response SHA-256: 4d4794da57f6c656a2b1a6c58156763e7d4a3d26c6135e04690d8e24246c3a19
Scoped evidence bindings: 1

[688] Investors - An independent company. TrustScale;
response-sha256:4d4794da57f6c656a2b1a6c58156763e7d4a3d26c6135e04690d8e24246c3a19;
text-sha256:de0f59f872a1449b7f3a0872df8e741ff9af114f9db726a64a505f6c53b23025
https://www.trustscale.ai/assets/index-CUMNnpew.js
Response SHA-256: 4d4794da57f6c656a2b1a6c58156763e7d4a3d26c6135e04690d8e24246c3a19
Scoped evidence bindings: 1

[689] Trust Center Privacy - Only what we need. TrustScale;
response-sha256:4d4794da57f6c656a2b1a6c58156763e7d4a3d26c6135e04690d8e24246c3a19;
text-sha256:94f88641db11841bc080ad34f244bd51e090a595b8f2bc72ded9c07ae2225662
https://www.trustscale.ai/assets/index-CUMNnpew.js
Response SHA-256: 4d4794da57f6c656a2b1a6c58156763e7d4a3d26c6135e04690d8e24246c3a19
Scoped evidence bindings: 1

[690] Off-The-Shelf AI Training Datasets (Licence Ready) | Mercor - catalog. Mercor.io Corporation;
response-sha256:4e4f3f637d3d9528053ccb9def276e1a17445d50cb51a0f179fb8121c02251bf;
text-sha256:5be1ab32971c9b806bb1e13c9394f8695d0babf34a13743295ed7baf1148ed3d
https://www.mercor.com/apex/off-the-shelf-data/
Response SHA-256: 4e4f3f637d3d9528053ccb9def276e1a17445d50cb51a0f179fb8121c02251bf
Scoped evidence bindings: 6

[691] Prime Intellect - The Open Superintelligence Stack - compute. Prime Intellect;
response-sha256:612ae3d4551e969e0fdba93a0667d04f6760cd119001d4e050944772dbeb1b82;
text-sha256:ee6427e56a80b70b6bc9ea2dadc891ed49b1a052cc56f4ae17ef956d756e9e74
https://www.primeintellect.ai/
Response SHA-256: 612ae3d4551e969e0fdba93a0667d04f6760cd119001d4e050944772dbeb1b82
Scoped evidence bindings: 4

[692] RL environment creation is becoming continuous QA - oddish-introduction. Abundant;
response-sha256:6d837ab47a1cc210ffca01dca18fcfba7aa84cd48b498ab963408045878492cd;
text-sha256:976eef85612c6bac0aa44727a4da1735eec7809867f287c03811410c5fda61a6
https://www.abundant.ai/blog/env-creation-loop/
Response SHA-256: 6d837ab47a1cc210ffca01dca18fcfba7aa84cd48b498ab963408045878492cd
Scoped evidence bindings: 2

[693] Beemo: Benchmark of Expert-edited Machine-generated Outputs - physical page 4. arXiv; 2411.04032v3
https://arxiv.org/pdf/2411.04032v3
Response SHA-256: 74964e5b6c1b7312c56ffe1448c5edb9b849ee9ffd843082344c58349b8a4f25
Scoped evidence bindings: 5

[694] GAIA - benchmark-and-redistribution. gaia-benchmark/GAIA;
response-sha256:8a160e2df36b843c33ec12fdecd731ae8281e0dae553375207707b506a36c11e;
text-sha256:9e375470a47d8d8d1e97bfb780d6615cc892e2b4a2bcbf61b9fe1ffe488247ba
https://huggingface.co/datasets/gaia-benchmark/GAIA/blob/682dd723ee1e1697e00360edccf2366dc8418dd9/README.md
Response SHA-256: 8a160e2df36b843c33ec12fdecd731ae8281e0dae553375207707b506a36c11e
Scoped evidence bindings: 3

[695] DAYJOB: A Benchmark for Long-Horizon Professional Work - physical page 1. arXiv; 2610.01306v1
https://arxiv.org/pdf/2610.01306v1
Response SHA-256: 9bbda84cf47ce3f33a93aeaf528c6a0c1de317d7e9354fde3c9ec394ccfff721
Scoped evidence bindings: 6

[696] Learning from Less: Measuring the Effectiveness of RLVR in Low Data and Compute Regimes - physical page 3. arXiv; 2604.18381v1
https://arxiv.org/pdf/2604.18381v1
Response SHA-256: a25428f3a66f56f3cd5d8c31a4c54c809521019cc23009aa369d1c6384e2ad76
Scoped evidence bindings: 7

[697] APEX-Accounting - physical page 1. arXiv; 2607.27189v2
https://arxiv.org/pdf/2607.27189v2
Response SHA-256: ada539d8e53a9436007d5ae13d14cf782cb1f93ef0b95603c45594fcb9569d5d
Scoped evidence bindings: 13

[698] Cross-Benchmark Transfer from RL on Agentic Coding Tasks - physical page 1. arXiv; 2610.00890v1
https://arxiv.org/pdf/2610.00890v1
Response SHA-256: afb236bcad430528cb991e0f7fb1a96ec92cf7471fe24dd8534a015d6d76930a
Scoped evidence bindings: 5

[699] AttuneBench: A Conversation-Based Benchmark for LLM Emotional Intelligence - physical page 3. arXiv; 2605.21739v2
https://arxiv.org/pdf/2605.21739v2
Response SHA-256: b3dddb3eb9932ffd593053854359eb51d13d8e93f663b63224f656e037f876bb
Scoped evidence bindings: 8

[700] PLE: Physical-Engineering’s Last Exam - physical page 1. arXiv; 2610.30537v2
https://arxiv.org/pdf/2610.30537v2
Response SHA-256: c3a57cb4e99e00b32da841dda22faee4d7481de89be4ba12704c1cc8a8e3c478
Scoped evidence bindings: 27

[701] ClawsBench: Evaluating Capability and Safety of LLM Productivity Agents in Simulated Workspaces - physical page 3. arXiv; 2604.05172v2
https://arxiv.org/pdf/2604.05172v2
Response SHA-256: cf055518aeb4542331708e70fd0ec3b3e4fadf977456f119239d88e1fe7098b8
Scoped evidence bindings: 8

[702] Cross-Benchmark Generalization in Long-Horizon Agents - physical page 3. arXiv; 2608.00181v1
https://arxiv.org/pdf/2608.00181v1
Response SHA-256: d6781d4c7f50cc94f6e9b9067743d62a7eca32d53b5565e290ed054cdd09af06
Scoped evidence bindings: 6

[703] Intent Laundering: AI Safety Datasets Are Not What They Seem - physical page 6. arXiv; 2602.16729v3
https://arxiv.org/pdf/2602.16729v3
Response SHA-256: f017d06307d0aa6a4ee74a2f09c50b6976ffe499413bc480a17e84a8f7ff1db2
Scoped evidence bindings: 6

[704] StereoTales: A Multilingual Framework for Open-Ended Stereotype Discovery in LLMs - physical page 2. arXiv; 2605.10442v2
https://arxiv.org/pdf/2605.10442v2
Response SHA-256: f8d0d15832df51670c9b52562a234ffa56b61e427a7dff0a537aab369ac28379
Scoped evidence bindings: 11

[705] WildJailbreak - composition. allenai/wildjailbreak;
response-sha256:f9e6161b38b2acc5bab0307ee7e835715fb51f2eddbeb130039b47201a4bf28b;
text-sha256:14608dc8570f1441aafeffbbdaa0eaa1dfff8595623f493596dac7721db6fe33
https://huggingface.co/datasets/allenai/wildjailbreak/blob/5ddc12a7894f842b0619b8e1c7ee496b198af009/README.md
Response SHA-256: f9e6161b38b2acc5bab0307ee7e835715fb51f2eddbeb130039b47201a4bf28b
Scoped evidence bindings: 5

[706] Magpie repository software LICENSE. magpie-align/magpie; b734a36818e4ba7ef7f0f582fc55f7e860407d26
https://raw.githubusercontent.com/magpie-align/magpie/b734a36818e4ba7ef7f0f582fc55f7e860407d26/LICENSE
Response SHA-256: cf6d60c842b1a19cb96774baa0f1badac4440a03eeecf026e47b59beb904400a
Scoped evidence bindings: 1

[707] OpenAssistant data collection task guidelines. Creator-maintained repository;
f1e6ed9526f5817531f3ab85441a40b3671ddccb
https://raw.githubusercontent.com/LAION-AI/Open-Assistant/f1e6ed9526f5817531f3ab85441a40b3671ddccb/docs/docs/guides/guidelines.md
Response SHA-256: 0ee4c2c2bb3fc94c0221dd2f3b1f1dfd3618c56f4f03afb4a9238157408bd329
Scoped evidence bindings: 4

[708] Turing announces Turing Frontier, connecting AI labs with elite domain experts. Turing Enterprises, Inc.; Captured response sha256:544ca3d9d1905acc800bd4f5df1992f0eac1591bd044ce96bd0c10498512f085
https://www.turing.com/press/turing-announces-turing-frontier
Response SHA-256: 544ca3d9d1905acc800bd4f5df1992f0eac1591bd044ce96bd0c10498512f085
Scoped evidence bindings: 3

[709] Label Your Data’s Notice of Certification Under the Data Privacy Framework. Label Your Data; Exact response sha256:e8351f6a590dc22ad1b03d911426c00befbeb1ae5ee99fb1da3e71cb838474de
https://labelyourdata.com/data-privacy-framework-notice
Response SHA-256: e8351f6a590dc22ad1b03d911426c00befbeb1ae5ee99fb1da3e71cb838474de
Scoped evidence bindings: 2

[710] AI Data Security & Compliance | Appen. Appen Limited; Captured response sha256:ef83e1b7fbc0dbf662a79e43363cdbde916e75c44523f8ed879dc91c3973d777
https://www.appen.com/data-security
Response SHA-256: ef83e1b7fbc0dbf662a79e43363cdbde916e75c44523f8ed879dc91c3973d777
Scoped evidence bindings: 1

[711] Frequently Asked Questions | Snorkel AI. Snorkel AI, Inc.; Captured response sha256:3050d01c8521d07cd613c366efdef416508ce04bccbb91dd1be6829e6060de20; 2026-06-17
https://snorkel.ai/frequently-asked-questions/
Response SHA-256: 3050d01c8521d07cd613c366efdef416508ce04bccbb91dd1be6829e6060de20
Scoped evidence bindings: 1

[712] Participants skilled at AI tasks | Prolific Research. Prolific; Captured response sha256:118c59b6b8c84e52b76decfd3f3efded727244a7e3c4bb7b13a533e5499a9d17; 2026-03-18
https://researcher-help.prolific.com/en/articles/445229-participants-skilled-at-ai-tasks
Response SHA-256: 118c59b6b8c84e52b76decfd3f3efded727244a7e3c4bb7b13a533e5499a9d17
Scoped evidence bindings: 2

[713] Modern Slavery Statement | Scale AI. Scale AI; Captured response sha256:70bc87f043c892daa4b7fc963f3139d81325f4ce64f48428ebd3720bfc9a9386
https://scale.com/legal/modern-slavery-statement
Response SHA-256: 70bc87f043c892daa4b7fc963f3139d81325f4ce64f48428ebd3720bfc9a9386
Scoped evidence bindings: 5

[714] Measuring Massive Multitask Language Understanding. arXiv / paper authors; 2009.03300v3
https://arxiv.org/pdf/2009.03300v3
Response SHA-256: d3a7b0f8589f5365ebfd34a3876a9a8b3fde83e0adb9ed19ccf8f0034bb48fc1
Scoped evidence bindings: 5

[715] HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages. arXiv / paper authors; 2505.11475v2
https://arxiv.org/pdf/2505.11475v2
Response SHA-256: ced74aea64c773933b2a0c60e4bb76edeff6313944c255f2dba373cc5d343563
Scoped evidence bindings: 5

[716] Expert Medical Image Data Labeling - Aya Data. Aya Data; Exact response sha256:f14489b15b890569a7f17740b1ffdff5865a5494b848763ec791fd4f7ba119ef
https://www.ayadata.ai/case-study/precision-medical-labeling-a-cydar-success-story/
Response SHA-256: f14489b15b890569a7f17740b1ffdff5865a5494b848763ec791fd4f7ba119ef
Scoped evidence bindings: 1

[717] Human Data Terms of Service · G2i.ai. G2i;
response-sha256:10739415ead464d17cbf43336d82b181a624558ff8a1caaa68fa6a8702d28b29
https://www.g2i.ai/human-data-terms-of-service
Response SHA-256: 10739415ead464d17cbf43336d82b181a624558ff8a1caaa68fa6a8702d28b29
Scoped evidence bindings: 1

[718] Expert human data is now available to all, introducing the new Toloka Platform. Toloka AI B.V.; Captured response sha256:9bae9f5ebef0dfa4ce2c5b6454ebe870b62954f02bd0602dc5d157a274242384; 2026-01-19
https://toloka.ai/blog/introducing-the-new-toloka-platform/
Response SHA-256: 9bae9f5ebef0dfa4ce2c5b6454ebe870b62954f02bd0602dc5d157a274242384
Scoped evidence bindings: 2

[719] How AfterQuery Helped NVIDIA Hill-Climb GDPval - AfterQuery. AfterQuery;
response-sha256:4e2a2dbec0e136088751239cb3246c1569b828a3d840ddd051de28042d93b979
https://www.afterquery.com/blog/how-afterquery-helped-nvidia-hill-climb-gdpval
Response SHA-256: 4e2a2dbec0e136088751239cb3246c1569b828a3d840ddd051de28042d93b979
Scoped evidence bindings: 1

[720] Prolific Pricing | Platform & managed services | Prolific. Prolific; Captured response sha256:0c3c2338fa2ae161efa6c95e62f2830ceefe370da90ad95f7fab7a9759c9acfb
https://www.prolific.com/pricing
Response SHA-256: 0c3c2338fa2ae161efa6c95e62f2830ceefe370da90ad95f7fab7a9759c9acfb
Scoped evidence bindings: 4

[721] Datasheets for Datasets. arXiv / paper authors; 1803.09010v8
https://arxiv.org/abs/1803.09010v8
Response SHA-256: 6ff49b1a436386614b2172e351b60d45e01d5a48c678f8731c2b6c0cdd3947a1
Scoped evidence bindings: 1

[722] Measuring Massive Multitask Language Understanding. arXiv / paper authors; 2009.03300v3
https://arxiv.org/abs/2009.03300v3
Response SHA-256: c2f09edccaea2fa2d18edf4c88ebaaa2531d6b6fb4509ef5deb9bf0b55071f93
Scoped evidence bindings: 1

[723] TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv / paper authors; 2109.07958v2
https://arxiv.org/abs/2109.07958v2
Response SHA-256: f6076f5fb265c258f41325aa571f25d65ed8c73744c654f60bd9095228141c32
Scoped evidence bindings: 1

[724] Training language models to follow instructions with human feedback. arXiv / paper authors; 2203.02155v1
https://arxiv.org/abs/2203.02155v1
Response SHA-256: 350f6cfe6cd4047f95c8c598839df74f9a13d5441bd75115fd87bd23d9a8174a
Scoped evidence bindings: 1

[725] Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv / paper authors; 2305.18290v3
https://arxiv.org/abs/2305.18290v3
Response SHA-256: 08edfdec10277f58ef5b670abe5d4c12772011b485501756e09527c895948d2b
Scoped evidence bindings: 1

[726] SWE-bench: Can Language Models Resolve Real-World GitHub Issues?. arXiv / paper authors; 2310.06770v3
https://arxiv.org/abs/2310.06770v3
Response SHA-256: 55b7b98eefc6bcc924cd4c4978ba276469e321f521400c345e46abd37f19f5a9
Scoped evidence bindings: 1

[727] The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. arXiv / paper authors; 2406.17557v2
https://arxiv.org/abs/2406.17557v2
Response SHA-256: 8b45f0e0560aeba3d03908a052037b79dadc400dd94d9561f64b226826a3909e
Scoped evidence bindings: 1

[728] HelpSteer3-Preference: Open Human-Annotated Preference Data across Diverse Tasks and Languages. arXiv / paper authors; 2505.11475v2
https://arxiv.org/abs/2505.11475v2
Response SHA-256: 68b5071fdac5cd44109bfcbc9dc03d22c15881b22e336ab74c70d5844311a957
Scoped evidence bindings: 1

[729] NVIDIA HelpSteer3 dataset card. NVIDIA; HF dataset card response
sha256:827e51e12a2537403f516e7e5299455c9de36994b621deab2b9a7524e01b29b7
https://huggingface.co/datasets/nvidia/HelpSteer3
Response SHA-256: 827e51e12a2537403f516e7e5299455c9de36994b621deab2b9a7524e01b29b7
Scoped evidence bindings: 1

[730] Datasheets for Datasets. arXiv / paper authors; 1803.09010v8
https://arxiv.org/pdf/1803.09010v8
Response SHA-256: c7f595988ee2c109631d5173fdae37ff1774c68ec17add51bd012f2a6e8792d0
Scoped evidence bindings: 5

[731] Training language models to follow instructions with human feedback. arXiv / paper authors; 2203.02155v1
https://arxiv.org/pdf/2203.02155v1
Response SHA-256: c1984bb50a5b90fddb895fdc3a0f72e5bc977148c9f63ef6040cbe7a3e1f0d98
Scoped evidence bindings: 6

[732] Direct Preference Optimization: Your Language Model is Secretly a Reward Model. arXiv / paper authors; 2305.18290v3
https://arxiv.org/pdf/2305.18290v3
Response SHA-256: 92cb3a2b71362acda98a789b03d88688fd33cf5fcf13f81d2b1de30ee7d3b67a
Scoped evidence bindings: 5

[733] SWE-bench: Can Language Models Resolve Real-World GitHub Issues?. arXiv / paper authors; 2310.06770v3
https://arxiv.org/pdf/2310.06770v3
Response SHA-256: f7e8e1df64129742b8199a21a042734519a823a1dafd6f48f8f3ddcfb48ee296
Scoped evidence bindings: 6

[734] The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. arXiv / paper authors; 2406.17557v2
https://arxiv.org/pdf/2406.17557v2
Response SHA-256: ae632bd6618d2566e708962670f27ab1fbc5ba9bc8de3acaceee8fbda85fbdd4
Scoped evidence bindings: 6

[735] AI training data services | Appen. Appen Limited; Captured response
sha256:2be61f7100fdb4e6cd4d9687fddc37b14b65625cf9f0cf029c099191655d3af5
https://www.appen.com/ai-data
Response SHA-256: 2be61f7100fdb4e6cd4d9687fddc37b14b65625cf9f0cf029c099191655d3af5
Scoped evidence bindings: 5

[736] Appen Annual Report 2025. Appen Limited; Captured response
sha256:93d65cd1fff04294c7847f20d6577b7073c5964bb78df7101e5ff44554657b38
https://cdn.prod.website-files.com/656a605609413dfd3feb9d34/699fbc072974a573b2b7a48c_Appen%20Annual%20Report%202025.pdf
Response SHA-256: 93d65cd1fff04294c7847f20d6577b7073c5964bb78df7101e5ff44554657b38
Scoped evidence bindings: 5

[737] Privacy Statement | Appen. Appen Limited; Captured response
sha256:7e7624f5f9e9fe19a478ba600c60bbc81ab303e81bdcdca5d1e993c0a0b63a50
https://www.appen.com/legal-policies/privacy-statement
Response SHA-256: 7e7624f5f9e9fe19a478ba600c60bbc81ab303e81bdcdca5d1e993c0a0b63a50
Scoped evidence bindings: 2

[738] Introducing APEX-Agents 1.1. Mercor.io Corporation; Captured response
sha256:43928e13914303e50f832607d4d1b83308beea7b90074f5f9957fdee8a8d9bc0; 2026-09-08
https://www.mercor.com/blog/introducing-apex-agents-1-1/
Response SHA-256: 43928e13914303e50f832607d4d1b83308beea7b90074f5f9957fdee8a8d9bc0
Scoped evidence bindings: 3

[739] Mercor | Website Privacy Notice. Mercor.io Corporation; Captured response
sha256:b22118719b4cabf509c51eafe24a3267b68040e4ebb6cd1fb924907f687dd023; 2026-08-26
https://www.mercor.com/privacy-notice/
Response SHA-256: b22118719b4cabf509c51eafe24a3267b68040e4ebb6cd1fb924907f687dd023
Scoped evidence bindings: 2

[740] Mercor Data Breach: Investigation Findings and Updates. Mercor.io Corporation; Captured response
sha256:2648cabcf7aca1dd54bf4d2bb1891d05abda335fc32099dd1a8fe423e43659fd; 2026-06-25
https://www.mercor.com/blog/update-on-mercor-security-incident/
Response SHA-256: 2648cabcf7aca1dd54bf4d2bb1891d05abda335fc32099dd1a8fe423e43659fd
Scoped evidence bindings: 1

[741] Prolific Case Study | Cohere and University of Edinburgh | Prolific. Prolific; Captured response
sha256:5546b2dfc28f559910240a410a4d020d8cdc362f7396cdafc5969d5307ae187b; 2024-06-07
https://www.prolific.com/resources/tom-hosking-navigating-biases-in-human-feedback-for-ai-training
Response SHA-256: 5546b2dfc28f559910240a410a4d020d8cdc362f7396cdafc5969d5307ae187b
Scoped evidence bindings: 2

[742] Prolific sets standards for authentic human data collection | Prolific. Prolific; Captured response
sha256:f7eca38c949ba6dfe1a251ce42579ab97872fd8032a8c8a04c35358e22c4be97; 2026-10-20
https://www.prolific.com/resources/prolific-setting-standards-for-authentic-human-data-collection
Response SHA-256: f7eca38c949ba6dfe1a251ce42579ab97872fd8032a8c8a04c35358e22c4be97
Scoped evidence bindings: 1

[743] Data Security and Privacy at Prolific | Prolific | Prolific. Prolific; Captured response
sha256:677597f1729cdd67ea45f172647fa0a8f546bd370712eaf82a01942b0d043061
https://www.prolific.com/prolific-security-overview
Response SHA-256: 677597f1729cdd67ea45f172647fa0a8f546bd370712eaf82a01942b0d043061
Scoped evidence bindings: 1

[744] Are there types of studies that aren't suitable for Prolific? | Prolific Research. Prolific; Captured response
sha256:bcb70625ccff5f1c0a530a8534fcf80008310bb91a09bde12d6e1bb9fb5526f6; 2026-08-05
https://researcher-help.prolific.com/en/articles/445105-are-there-types-of-studies-that-aren-t-suitable-for-prolific
Response SHA-256: bcb70625ccff5f1c0a530a8534fcf80008310bb91a09bde12d6e1bb9fb5526f6
Scoped evidence bindings: 1

[745] scale.helpsteer3-card pinned README.md. NVIDIA; Captured response
sha256:d5d46db58e8374937ea284b33b1583a4a465b232bf74b941745d80e3c35fd8ae
https://huggingface.co/datasets/nvidia/HelpSteer3/raw/f6d145777bcbde96137596340fab89793acd1031/README.md
Response SHA-256: d5d46db58e8374937ea284b33b1583a4a465b232bf74b941745d80e3c35fd8ae
Scoped evidence bindings: 1

[746] Generative AI Data Engine | Scale AI. Scale AI; Captured response
sha256:ce5e1dabfafc8a607d1ef7ece805150a6ee0c7b71878ff6b6ed3c8eccdfac2c4
https://scale.com/generative-ai-data-engine
Response SHA-256: ce5e1dabfafc8a607d1ef7ece805150a6ee0c7b71878ff6b6ed3c8eccdfac2c4
Scoped evidence bindings: 2

[747] How We Engineer World-Class Data at Scale | Scale AI. Scale AI; Captured response
sha256:86363dc9d41257758e86093047af5ce76d86b41876149801a21c01a2d198bb36; 2026-06-29
https://scale.com/blog/data-quality
Response SHA-256: 86363dc9d41257758e86093047af5ce76d86b41876149801a21c01a2d198bb36
Scoped evidence bindings: 2

[748] Security at Scale.com | Scale AI. Scale AI; Captured response
sha256:25ddfbad6b4a78b7d787f526ca9fea44811500fd737168a3265926c0afaa37b8
https://scale.com/security
Response SHA-256: 25ddfbad6b4a78b7d787f526ca9fea44811500fd737168a3265926c0afaa37b8
Scoped evidence bindings: 1

[749] Snorkel AI Leaderboards | Snorkel AI. Snorkel AI, Inc.; Captured response
sha256:58e9a00deedf172e09f8177d6179f4f08453e24c18041dfaa07fc5eba34ef2d3
https://snorkel.ai/leaderboard/category/financereasoning/
Response SHA-256: 58e9a00deedf172e09f8177d6179f4f08453e24c18041dfaa07fc5eba34ef2d3
Scoped evidence bindings: 1

[750] Privacy Policy | Snorkel AI. Snorkel AI, Inc.; Captured response
sha256:5a1e9296ec15d80d6d7148730b3cc05bd3e7b6abdb08d739fdf8cdeb5a50b0a6; 2026-10-19
https://snorkel.ai/privacy/
Response SHA-256: 5a1e9296ec15d80d6d7148730b3cc05bd3e7b6abdb08d739fdf8cdeb5a50b0a6
Scoped evidence bindings: 3

[751] surge.gsm8k-card pinned README.md. OpenAI; Captured response
sha256:a17e882503578c9e324560630e31017617714ee025c87cf4fea6fd916895f3c1
https://huggingface.co/datasets/openai/gsm8k/raw/740312add88f781978c0658806c59bc2815b9866/README.md
Response SHA-256: a17e882503578c9e324560630e31017617714ee025c87cf4fea6fd916895f3c1
Scoped evidence bindings: 2

[752] Privacy Policy | Surge AI. Surge Labs, Inc.; Captured response
sha256:adc676a5886b92cc0d0375ca66d9136e0b058f26f6963fc89e504aaa6a4cbdf6; 2026-07-01
https://surgehq.ai/privacy-policy
Response SHA-256: adc676a5886b92cc0d0375ca66d9136e0b058f26f6963fc89e504aaa6a4cbdf6
Scoped evidence bindings: 2

[753] Products | Surge AI. Surge Labs, Inc.; Captured response
sha256:26650a8c884ede1498a46b449955dec2fd9166ed844407fa7396319102b32e7d
https://surgehq.ai/products
Response SHA-256: 26650a8c884ede1498a46b449955dec2fd9166ed844407fa7396319102b32e7d
Scoped evidence bindings: 4

[754] Terms of Service | Surge AI. Surge Labs, Inc.; Captured response
sha256:affca92e60b1c466c5c6ca96973171a2a4b7a13503f522566ef4513ccab7421d; 2021-12-07
https://surgehq.ai/terms-of-service
Response SHA-256: affca92e60b1c466c5c6ca96973171a2a4b7a13503f522566ef4513ccab7421d
Scoped evidence bindings: 1

[755] Guidelines for posting compliant tasks | Toloka documentation. Toloka AI B.V.; Captured response
sha256:4ee70477eccc2d50a3a0e14f0d39b3545e363f70ae3d84c02527645e86d94c20
https://toloka.ai/docs/guide/unwanted/
Response SHA-256: 4ee70477eccc2d50a3a0e14f0d39b3545e363f70ae3d84c02527645e86d94c20
Scoped evidence bindings: 4

[756] AI + Humans: Data to Power LLMs & VLMs. Toloka AI B.V.; Captured response
sha256:aa0ae077444a739a14d84d0c9d5abc5f778eb8949e6953af96e5415b8918471f
https://toloka.ai/hybrid-data
Response SHA-256: aa0ae077444a739a14d84d0c9d5abc5f778eb8949e6953af96e5415b8918471f
Scoped evidence bindings: 4

[757] Privacy Notice. Toloka AI B.V.; Captured response
sha256:1d6d372770d126648c7c347fd4a7b8dfd565046fbc9f1206dd0ea084bf84f212; 2026-10-12
https://toloka.ai/legal/privacy-notice
Response SHA-256: 1d6d372770d126648c7c347fd4a7b8dfd565046fbc9f1206dd0ea084bf84f212
Scoped evidence bindings: 4

[758] toloka.umath-card pinned README.md. Toloka AI B.V.; Captured response
sha256:c13bf3579ababfb2510ca76ca0de6185ef2ff97caaf69ebf6a89d761c8c42871
https://huggingface.co/datasets/toloka/u-math/raw/7210f97b3f21122c3a90126935e966ca7d8951b6/README.md
Response SHA-256: c13bf3579ababfb2510ca76ca0de6185ef2ff97caaf69ebf6a89d761c8c42871
Scoped evidence bindings: 1

[759] 35K+ Agent Trajectories and 120K+ RLHF Pairs Across 150+ Real Tools. Turing Enterprises, Inc.; Captured response
sha256:5b0083910fdb37276d7d65eafb7f6f8b1f93889e0a06bb0f220b7dd8ec9b5891
https://www.turing.com/blog/frontier-ai-newsletter-79
Response SHA-256: 5b0083910fdb37276d7d65eafb7f6f8b1f93889e0a06bb0f220b7dd8ec9b5891
Scoped evidence bindings: 2

[760] Experts: Vetted Domain Talent for Frontier AI Labs | Turing. Turing Enterprises, Inc.; Captured response
sha256:58e8b558b1d70033ab022c8b99aab7c477b88508881ed7bb75b882e7e8e21acb
https://www.turing.com/frontier-ai/experts
Response SHA-256: 58e8b558b1d70033ab022c8b99aab7c477b88508881ed7bb75b882e7e8e21acb
Scoped evidence bindings: 1

[761] Turing Privacy Policy | For Visitors, Users & Others. Turing Enterprises, Inc.; Captured response
sha256:45306dbc6a1f708028e93ca241e5033d8a6b37cbac888df8da001c403fa8e0dd; 2025-06-27
https://www.turing.com/policy
Response SHA-256: 45306dbc6a1f708028e93ca241e5033d8a6b37cbac888df8da001c403fa8e0dd
Scoped evidence bindings: 3

[762] aa-lcr-creator-card. Artificial Analysis; 9a77ef56b717057ade24ceab4d273712a0b4f19e
https://huggingface.co/datasets/ArtificialAnalysis/AA-LCR/resolve/9a77ef56b717057ade24ceab4d273712a0b4f19e/README.md
Response SHA-256: a07bb0e510eebe33be1290770efe29981374544212f897e68b753a62ca6aeb16
Scoped evidence bindings: 7

[763] apex-observability-doc. Mercor / Cognition; git-blob:d7350fa3dc67851d0c45a6b11c197c29e4586c81
https://github.com/Mercor-Intelligence/apex-swe/blob/main/observability/README.md
Response SHA-256: 109d917acf82ee385620ffaebcd1876d5b15b2120658efeb3b5c8c3fe86adcab
Scoped evidence bindings: 1

[764] apex-swe-creator-repo. Mercor / Cognition;
body-sha256:50602ac696b779bd4648f2344e53c4a43a7b47c1027c23c1227f453b38b18c50
https://github.com/Mercor-Intelligence/apex-swe
Response SHA-256: 50602ac696b779bd4648f2344e53c4a43a7b47c1027c23c1227f453b38b18c50
Scoped evidence bindings: 1

[765] deepresearch-creator-readme. DeepResearch Bench II authors;
git-blob:f3a55c29bae84b5983695105ae3957912c6b1c28
https://github.com/imlrz/DeepResearch-Bench-II/blob/main/README.md
Response SHA-256: 0f696da1e57ea96fa30b3eb4e59d8ad365239d6a1a7ff06245b8a6fdcad4dd20
Scoped evidence bindings: 5

[766] deepresearch-supplier-methods. Mercor;
body-sha256:da8c828606613b4d2ca8d4a2d6e79010ef66b2ce20fd208c6b5e352b4cc81e80
https://www.mercor.com/apex/oss-benchmarks/oss-deepresearch-bench-ii-leaderboard/
Response SHA-256: da8c828606613b4d2ca8d4a2d6e79010ef66b2ce20fd208c6b5e352b4cc81e80
Scoped evidence bindings: 1

[767] frontiercode-creator-blog. Cognition;
body-sha256:9a52a8b788d74df0e9fea13992e5547f8f4bd88e0b953546024e7ca11972a6bd
https://cognition.com/blog/frontier-code
Response SHA-256: 9a52a8b788d74df0e9fea13992e5547f8f4bd88e0b953546024e7ca11972a6bd
Scoped evidence bindings: 8

[768] frontiercode-supplier-methods. Mercor;
body-sha256:98368088aa5230bf236c9cfc05553282deb1a2223bb1f6fd18f24346177389f0
https://www.mercor.com/apex/partner-benchmarks/frontiercode-leaderboard/
Response SHA-256: 98368088aa5230bf236c9cfc05553282deb1a2223bb1f6fd18f24346177389f0
Scoped evidence bindings: 2

[769] hackerrank-bench-canonical. HackerRank;
body-sha256:c50227b4688a9e11850631e587d959923580fd9bf11f5e56af3027faf0848752
https://www.hackerrank.com/bench
Response SHA-256: c50227b4688a9e11850631e587d959923580fd9bf11f5e56af3027faf0848752
Scoped evidence bindings: 3

[770] hackerrank-bench-report. HackerRank;
body-sha256:c550f513c8286fae1b2cb26bc125fcd29c341387d2d939b826035c4d9f9fbd46
https://www.hackerrank.com/bench/report
Response SHA-256: c550f513c8286fae1b2cb26bc125fcd29c341387d2d939b826035c4d9f9fbd46
Scoped evidence bindings: 2

[771] harvey-architecture-doc. Harvey; git-blob:4550b48a7dfeb04567c3c00603c907ad44501e36
https://github.com/harveyai/harvey-labs/blob/main/docs/architecture.md
Response SHA-256: ba3febacbd2b69e51ad5b12d01f50aec97d6e6dcdbb34a54e2f7666918689d6c
Scoped evidence bindings: 4

[772] harvey-evaluation-doc. Harvey; git-blob:190dc2ad7f8503aca6cca316478105723f9ae955
https://github.com/harveyai/harvey-labs/blob/main/docs/eval-strategies.md
Response SHA-256: 4d54d1eb0e2a72964d7beb1bfab18224342e49fd62286c2500b4bfc633c322e6
Scoped evidence bindings: 5

[773] harvey-lab-creator-repo. Harvey;
body-sha256:2577822e195fcc8f4bbbafe5052922ba4f74101333b0f8212565f56b91578cbf
https://github.com/harveyai/harvey-labs
Response SHA-256: 2577822e195fcc8f4bbbafe5052922ba4f74101333b0f8212565f56b91578cbf
Scoped evidence bindings: 2

[774] mercor-catalog. Mercor; body-sha256:cfd7c2878ffede2a5638fb607319c97e530e207945542566959ab1086f9561a5
https://www.mercor.com/apex/
Response SHA-256: cfd7c2878ffede2a5638fb607319c97e530e207945542566959ab1086f9561a5
Scoped evidence bindings: 2

[775] mercor-methodology. Mercor;
body-sha256:fddaa02bee9bb213c886bc27d6e0077b48baeba2e292818fdcc10a1b3dfb9f0c
https://www.mercor.com/apex/methodology/
Response SHA-256: fddaa02bee9bb213c886bc27d6e0077b48baeba2e292818fdcc10a1b3dfb9f0c
Scoped evidence bindings: 6

[776] ple-creator-repo. Google PLE authors;
body-sha256:ddec3c207071e3059a3f8fd611e8eb9a418c800f6e016b10bebbac6edf9cf4c6
https://github.com/google-research/ple
Response SHA-256: ddec3c207071e3059a3f8fd611e8eb9a418c800f6e016b10bebbac6edf9cf4c6
Scoped evidence bindings: 2

[777] redline-creator-card. Crosby Legal / micro1; eee1b6790982ed1279e86bec7616b662a61993e6
https://huggingface.co/datasets/crosbylegal/RedlineBench/resolve/eee1b6790982ed1279e86bec7616b662a61993e6/README.md
Response SHA-256: 7499baec91f2be8eba5d461178904fad77a17e7f208c69f47fc4ea29628ce97e
Scoped evidence bindings: 6

[778] redline-creator-readme. Crosby Legal / micro1; git-blob:426e74f8777ba93d2e4a4538d23247b97085d94b
https://github.com/crosbylegal/redline-bench/blob/main/README.md
Response SHA-256: a9054e53faa6535f39ad32236cf2c4848532c91eb07107abc800e1d54aa936f2
Scoped evidence bindings: 4

[779] redline-creator-report. Crosby Legal / micro1;
body-sha256:ac54c61bff51239722b8579ec29b3242be62531e543bbd5ed652cbdbdd4468f1
https://intelligence.crosby.ai/benchmark/
Response SHA-256: ac54c61bff51239722b8579ec29b3242be62531e543bbd5ed652cbdbdd4468f1
Scoped evidence bindings: 2

Evidence and reproduction

The complete public evidence graph accompanies this report in the offline archive. Paragraph evidence IDs refer to that graph. The JSON, CSV and SQLite contain equivalent privacy-filtered projection 1.2.0. Canonical dataset 2.5.0 remains unchanged outside this archive. Raw source text and personal contact details are excluded.

Verify MANIFEST.json, then run python3 reproduce.py --write-tables reproduced in the extracted archive. LOCATORS.md documents evidence coordinates. Owned research and text use CC BY 4.0; third-party material retains its rights. Embedded DejaVu fonts and the Noto Sans SC derived CJK subset retain their upstream terms in DejaVu-LICENSE.txt and CJK-OFL.txt and are excluded from the project's CC BY 4.0 and MIT licenses. Review scope and research gaps remain explicit. This PDF's status does not establish website or external mirror publication.