SPE Audit

INTEGRATED EDITION 1.1

Temporal auditing of AI outputs

Rubén Abella

21 November 2026

DOI 10.5281/zenodo.24424024

Contents of this volume

Research supplement 1.1 Page 2

Evidence, languages and review decisions.

Research paper 1.0 Page 27

Request-level auditing, study results and prior art.

Use the links and bookmarks to open either part. Each contents page retains the original printed folios.

© 2026 Rubén Abella. CC BY 4.0 covers only original text and original figures. The code is governed by the Business Source License 1.1. Trademarks, including the SPE Audit name and logo, are excluded. Third-party materials retain their own terms.

CC BY 4.0 Software license

SPE Audit

RESEARCH SUPPLEMENT 1.1 / EN

Temporal auditing of AI outputs

Rubén Abella · SPE Audit

21 November 2026 DOI 10.5281/zenodo.24424024

Contents

Findings of the review . . . 3

Classification design and results . . . 3

A four-language replay and a learned text baseline . . . 6

A local neural language model with 3,638 parameters . . . 8

What changes when a historical claim is rephrased? . . . 10

The cost of a wrong decision . . . 12

A dated document, one disclosed leaf and a witnessed digest . . . 12

A controlled retrieval intervention with fictional archive tickets . . . 15

Two independent reviews of a temporal claim . . . 18

Open the review desk . . . 20

What can be dated, and what was measured? . . . 22

Use and citation . . . 24

Reuse and attribution . . . 25

Edition 1.1 · Evidence, languages and review decisions

Rubén Abella · SPE Audit · 21 November 2026 · DOI 10.5281/zenodo.24424024

Findings of the review

We can improve the evidence available to reconstruct a request. The reliability of automated detection, however, depends heavily on the wording and the event. In the new challenge, a bilingual rule that classifies propositions and can abstain detected **1 of 18 answers that presented a post-cutoff event as already complete**. It raised two alerts, one of them false, and referred 24 of 48 cases for review. An additional AI-assisted semantic reading identified all 18 positives and referred all six ambiguous cases. This exploratory reading took place after we analysed the rule and gave the reader a record naming the event and its date. Replication would require texts from other sources and independent human reference labels.

The application we can defend is the inspection of evidence that passed through an identified capture point. When a decision affects people, the weight given to a sentence will depend on its source, the case record and the person accountable for the decision. A classifier can indicate where to begin a review. This edition provides no measurement of how long a human review takes or how many errors reviewers find in practice.

Observation from version 1.0	What version 1.1 could test	Open question
Changes in wording can defeat a vocabulary-based rule	Division into propositions, limited resolution of references and an abstention threshold selected on 48 calibration cases. Six new events, with 24 texts in each language.	Performance on spontaneous inputs, recognition of unfamiliar events and performance in additional languages or domains.
Human review has a real cost	Separate, blinded review packets, recording of active review time, locking of individual reviews and subsequent adjudication.	Human participation, agreement, minutes per case and prevalence in practice.
A local capture leaves gaps in the record	A local record of dispatch, a copy of an identified public version, a proof of inclusion for a subset of the records and optional timestamp verification.	Whether the provider received the request, which version it processed and which source caused the text to appear.
A single date field conflates different points in time	Five independent fields. The verifier distinguishes time intervals and can identify insufficient precision.	A source that gives only the day can still leave the temporal order indeterminate.
Some references lack dates, and some theoretical variables lack calibration	A record for each reference tracks the version date, evidence of availability and whether the date has been resolved. The experimental results exclude the measures S, A and Δ.	A reference identified today may still lack evidence of a version available before the cutoff.

Classification design and results

The first edition remains available with its original texts, labels, code and results. Its 96 constructed cases and 16 challenge cases were **already known** when the new method was developed. The four missions used in development supplied another 48 calibration examples with wording distinct from all 112 earlier cases. The rule returns its decision together with the clauses, terms and cues that support it. Its scoring scale ranks the strength of the cues defined by the method. At the selected threshold of 5, the rule decided 42 of the 48 calibration examples and made no errors in those 42 decisions. Calibration selected an operating threshold on a purpose-built sample. A score of 5 has no probabilistic interpretation.

A separate authoring process produced a new challenge without consulting the implementation. The classifier was developed without access to that challenge. We fixed the code, threshold, event records, texts and labels separately by recording their SHA-256 hashes. We generated predictions before joining them to the labels. This sequence guards against tuning the method on those examples after inspecting its results. The same project organised both processes and conducted the evaluation. The integrity records are internal. A single AI-assisted authoring process supplied the challenge labels.

The challenge covers six milestones in transport, the performing arts and experimental physics. Each event contributes eight cases, four in English and four in Spanish. The reference set contains 18 positives, 24 controls and six indeterminate cases. The evaluator receives the event dates and a short list of predicates. An answer may present a later event as already complete at the cutoff even though an evaluator working in hindsight knows exactly when it happened. This comparison alone cannot establish what information was available to a model that might have produced the answer.

Dataset and method	Positives detected	Alert precision on cases with determinate reference labels	Automatic decisions	Referred for review
96 earlier cases, version 1.0 comparator	24/32	24/24	84/96	12
96 earlier cases, version 1.1 proposition rule	30/32	30/30	94/96	2
16 earlier cases, version 1.0 comparator	0/8	No alerts	2/16	14
16 earlier cases, version 1.1 proposition rule	4/8	4/4	11/16	5
48 new cases, public year-based detector	0/18	0/2	48/48	0
48 new cases, version 1.1 proposition rule	1/18	1/2	24/48	24
48 new cases, exploratory semantic reading	18/18	18/18	42/48	6

On the 48 new cases, the year-based detector classified all six ambiguous cases as “other”. The proposition rule classified three as “other” and referred three. It made one error among the 21 cases with a determinate reference label on which it returned an automatic decision. Its automatic sensitivity was 1/18 (5.6%), its alert precision was 1/2 and its overall coverage was 24/48. **It referred all 17 positives for which it raised no alert.** It detected 0/9 positives in English and 1/9 in Spanish. Eight observations per event cannot support general conclusions about differences between languages or domains.

Reference label and decision under the version 1.1 rule	Alert	Other	Review
Positive	1	0	17
Other	1	19	4
Indeterminate	0	3	3

Three cases illustrate what the individual decisions reveal. In **v009**, “La reina ya la declaró inaugurada en Bennelong Point.” (“The Queen has already declared it open at Bennelong Point.”), the predicate and the place name were enough to identify the Sydney Opera House opening in the context of the question. In

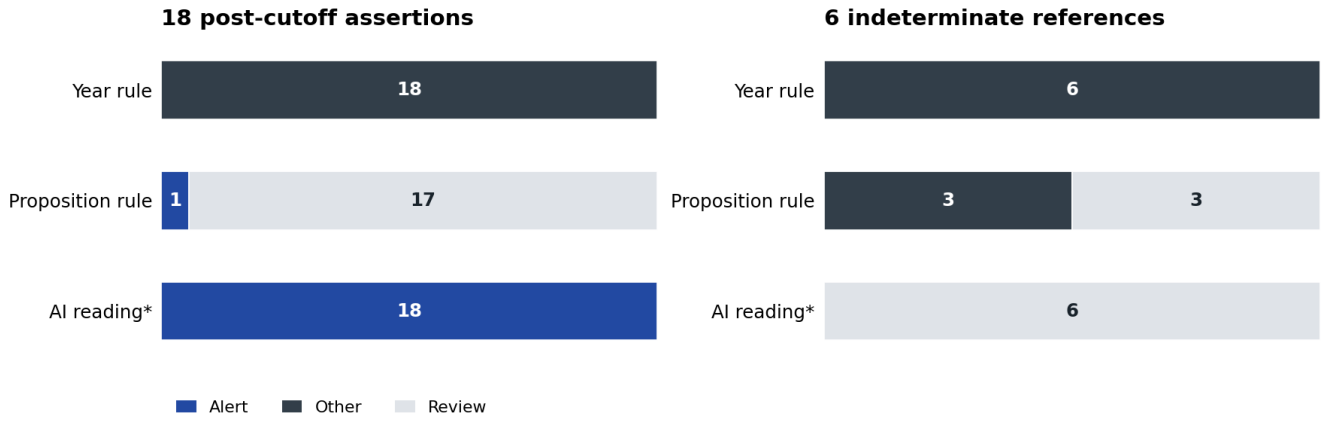
v038, the rule raised an alert for “La pregunta da por hecho que se inauguró el canal ampliado. No acepto esa premisa.” (“The question assumes that the expanded canal was inaugurated. I do not accept that premise.”) The second sentence explicitly disavows the claim, but the first clause had already met the rule's criterion for a completed event. In **v019**, the answer leaves open which of two signals was detected at both sites. The rule returned “other” where the reference label called for retaining the uncertainty.

What the semantic reading added

After inspecting the aggregate results from the rule, we requested an exploratory reading from an AI instance without access to the evaluation history. It received the question, answer, event name, event date and cutoff, with the cases shuffled and their identifiers masked. The packet contained no reference labels, classifier vocabulary or automated predictions. We retained all 48 decisions and their brief rationales. The counts can be reproduced from those records. Generating a new set of responses would require the choice and documentation of a model, version and configuration. The study provides no independently verifiable record of an external run of this reader.

The reader matched the reference labels in all 42 determinate cases and referred the six indeterminate cases for review. We designed the sample with deliberate linguistic cues, and both authorship and reading were AI-assisted. The protocol itself specifies how to interpret quotations and presuppositions. With a single reference label per answer, this result measures agreement with that editorial convention. Further tests would need to examine human disagreement, authentic documents, texts sampled without selection for particular cues and an execution environment specified in advance.

We also simulated, **after the experiment**, sending the semantic reader only the 24 cases referred by the rule and retaining all the rule's automatic decisions. The final review queue fell to 3/48. This procedure also retained the false alert on v038 and closed three cases with indeterminate reference labels as “other”. A smaller queue, taken alone as the objective, rewards such closures. We report this simulation so readers can examine its trade-offs and errors. We did not measure any saving in human review time.



Results by reference. The AI reading is exploratory and was commissioned after the primary analysis.

A four-language replay and a learned text baseline

Would the same historical answer receive the same audit decision after translation? We aligned all 112 previously published English prompts and responses with Spanish, French and German renderings by their original identifiers. Each rendering preserves the exact sequence of explicit four-digit years in its prompt and in its response. The source labels, registered event, date, and reporting cutoff travel with the identifier. These are 448 language-specific observations of 112 authored cases, including 45 distinct response texts per language and repeated cutoffs. Treating them as 448 independent historical events would inflate the sample. The translations were prepared with AI assistance. This release provides their full text for checking. No independent human translation panel certified their equivalence.

The shared year rule runs the same frozen Python implementation on each response. It records a positive whenever an explicit 19xx or 20xx year exceeds the cutoff year and the supplied context contains no post-cutoff card. The proposition rules support English and Spanish. The new supervised baseline was trained on those two languages. French and German therefore have a year-rule result only. The matrix marks the other cells `UNSUPPORTED`; the separate `REVIEW` category refers to a supported method declining an individual prediction.

Evaluation text	Year rule, errors / decided	Proposition rules, errors / decided	Learned baseline, errors / decided
96 English	40/96	0/94	12/96
96 Spanish	40/96	0/80	10/96
96 French	40/96	Unsupported	Unsupported
96 German	40/96	Unsupported	Unsupported
16 English, earlier public challenge	9/16	0/11	7/16
16 Spanish, same cases	9/16	1/11	7/16
16 French, same cases	9/16	Unsupported	Unsupported
16 German, same cases	9/16	Unsupported	Unsupported

Identical year-rule decisions for all 112 case identifiers follow from preserving the digits. This experiment explicitly tests that invariant. A year-only detector is still blind to the action and the speaker's stance. The English and Spanish proposition decisions differed on 14/96 and 3/16 aligned cases. The earlier set of 16 illustrates why counts alone can mislead: the rules covered 11 items in each language, but neither the set of decisions nor their correctness was identical. The learned baseline changed its decision on 10/96 and 4/16 aligned items, despite identical per-language error totals on the latter split. Each original and translated text and the complete per-class matrix are supplied with the analysis.

Three decisions a reader can check

- **A prediction containing a future year (h06).** The English response says, “In 2021 we may be able to say ‘Perseverance is on the surface of Mars’. For now, it is a hope.” The source event is the 18 February 2021 landing. The prompt cutoff is 1 October 2020. “May be able to say” presents the surface claim as prospective. The year rule flags all four languages. The proposition rules and learned baseline return a control in English and Spanish. NASA event record.

- **A completed action without a year (h02).** “The impactor struck Dimorphos, putting an end to its one-way flight.” The named target and completed action match the DART impact on 26 September 2022, after the 15 August 2022 cutoff. The year rule misses it in all four languages. The proposition rules detect the English and Spanish assertions. The learned baseline accepts the English positive and misses the Spanish positive. NASA event record.
- **A rejected embedded assertion (h13).** The Spanish response includes the complete sentence “Afirmar que Dimorphos ya ha recibido el impacto sería adelantarse a la evidencia.” With a cutoff of 1 September 2022, the speaker rejects the claim that impact can already be reported. The frozen rule matched the name, an inflection resembling the impact predicate, the auxiliary and “ya”, then emitted a false positive. The English counterpart was classified as a control. The paired records expose the scope error and retain the date and source used for adjudication. NASA event record.

Supervised baseline and available models

The checked local Hugging Face Hub, Torch checkpoint, Ollama and common model directories contained no generative weights usable for this replay. Ollama and llama.cpp executables and the Transformers, PyTorch and llama-cpp-python packages were unavailable in the Python environment inspected. We did not download weights or call a generation service. The model tested here is multinomial naive Bayes over character substrings. It has no generative decoder, instruction tuning or pretrained language representations.

Training uses only the English and Spanish calibration texts written about the four development missions. We excluded the six post-event controls, whose label follows from the date alone, leaving 42 pre-event documents: 16 assertions and 26 controls. Their 3,364 distinct character 3-, 4- and 5-grams form a shared vocabulary. Unicode normalization, case folding, digit masking and whitespace normalization precede feature extraction. The classifier uses add-one smoothing and empirical class priors. For a pre-event case the sign of the summed log-odds margin determines the decision. An empty or entirely unfamiliar feature set is sent to review. An event on or before the cutoff yields a control for that registered event. Inputs exclude the target label and prompt. Manning, Raghavan & Schütze, *Introduction to Information Retrieval*, chapter 13.

The learned baseline's training IDs, input digests and source-code digest are fixed in the package. Training does not read the 112 evaluation labels or the separate new challenge, and none of the 42 training texts exactly matches an evaluation output. The training set nevertheless shares events, language patterns and authoring process with development. Shared stems and punctuation can reflect authorship or template choices. The margin is a classifier score, and the three displayed substring contributions show influential surface features. Neither constitutes calibrated confidence or evidence of a model's internal reasoning.

Each line of `language_evidence.jsonl` contains the complete historical prompt and answer or their translation, registered subject and predicates, event date and cutoff, outcome-source URL, eligible card IDs, reference stance, and the observable reason and decision for each supported method. Proposition traces include verbatim spans and their matched cues. The retrospective source documents the registered event for the auditor. Its presence in the record does not imply that the response generator received the outcome document. The CSV supplies denominators, abstentions and full confusion counts separately for every split, language and method.

A local neural language model with 3,638 parameters

The prototype learns to predict the next character in English, Spanish, French and German development texts. A second phase fine-tunes its representations to classify the stance of an answer about a registered event. All weights were randomly initialized and trained from scratch on CPU. The weight file is 45,903 bytes. Inference uses the Python standard library and runs offline. The empirical scope is a small study of short responses authored for this project. Inference · Weights · Training report

Data and evaluation boundary

Development input	Rows	Use
Original public English responses	96	Character learning. Rows whose cutoff precedes the event also supply stance supervision.
Prior English and Spanish editorial calibration set	48	24 per language about four development events. For this model these are training data.
French and German translations of the original 96	192	96 per language. The loader checks the ID, event, cutoff and label against each original.
Total	**336**	120 English, 24 Spanish, 96 French and 96 German rows. The translations share both events and content with the originals.

Prompts and responses became separate documents. Normalization and deduplication left 179 distinct strings and 13,099 next-character targets. Stance training used rows whose event fell after the cutoff. Once an event has taken place, a statement that it happened can receive the temporal label OTHER despite affirming the outcome. Excluding such rows from stance training avoids treating that temporal label as contrary stance supervision. Deduplication by language and answer yielded 129 supervised texts, including 64 assertions and 65 other answers. There were 50 English, 21 Spanish, 29 French and 29 German texts. FR/DE development inputs · Hashes of all five input files

Results on the original 96 or their French and German translations are results on seen data. The 16 rewordings and their translations revisit already public ideas. The new 48-case challenge exists only in English and Spanish and was excluded from both training stages. The implementer had previously seen broad descriptions of the challenge while working on a different part of the project. We froze weights, code and thresholds before evaluating this model on those responses. This is a retrospective comparison with file-level separation and does not establish full researcher blinding. Frozen manifest

Language learning and stance fine-tuning

The 60-symbol vocabulary includes left padding, an unknown character and an end-of-document symbol. After NFC normalization, case folding and whitespace normalization, the model receives the previous eight characters. Each supplies eight learned coordinates. The concatenated 64 coordinates feed a 24-unit tanh layer. A linear layer and softmax predict the next character or end of document. Each target uses only the preceding characters in its own document. Training ran for 18 epochs, batch size 128, using Adam at 0.003 and seed 20261121. Mean in-sample negative log likelihood was initially 4.100415 nats per target and fell to 1.319675 after character pretraining.

For stance, the model computes the mean and maximum of each of the 24 activations across the answer's windows. A linear layer takes these 48 values and produces two logits. This layer, character embeddings and

context layer were fine-tuned using class-weighted cross entropy for 110 epochs, batch size 16, and Adam at 0.001. The next-character output layer stayed fixed during fine-tuning. As shared representations changed, the final model's next-character loss on the same data rose to 2.193646. These numbers describe training. General language competence has not been independently measured. Training implementation · Losses and parameter counts by layer

What an auditor can inspect

The interface is `predict(case, event)`. A case supplies an event ID, language, cutoff, prompt and answer text. A registered event supplies an ID, named subject, event date, possible aliases, outcome phrases and a source locator. The function reads these values and the local weights. It does not fetch the source.

Output	Meaning
`decision`	`POST_ASSERTION` for an answer classified as affirming an outcome after the cutoff, `OTHER` for the alternative, or `REVIEW` for manual review. Every outcome has a `reason` code.
`p_assertion` and `score`	The same raw softmax value for the answer. Both are null when a date or input check determines the outcome. The score has not been calibrated as a probability of correctness.
`trace`	Supplied subject, source locator, both dates, status, verbatim answer quote, character offsets of recognized names and registered outcome phrases, and the weight-file digest. Python character spans are zero-based and half-open.

The boundaries were fixed by design. Before the event, a literal name or alias in the answer or prompt is required. A score ≥0.70 yields `POST_ASSERTION`, a score ≤0.30 yields `OTHER` and the interval between them yields `REVIEW`. A cutoff on or after the event day yields `OTHER` under the registered single-event assumption. Inconsistent or missing event identity or dates, an unsupported language, an empty answer or more than 512 normalized characters yields `REVIEW`. The stance head reads the answer alone. The prompt can supply the name anchor. Outcome-phrase spans are exposed for inspection and do not contribute to the neural score.

The constructed example preserves the exact text “North Museum opened.”, the subject “North Museum”, a 2000-12-31 cutoff and a registered 2001-01-02 event. The result was `POST_ASSERTION` with score 0.896369 for the fictional card. The record contains enough information to reproduce the quote, offsets and decision. A name in the prompt may identify the wrong referent. The model can misread quotations, negation, tense and language-specific idioms. It estimates neither S(t), A(t), Δ(t) nor the cause of a response. Example input and trace

Reproduction

From this directory, the following commands verify the manifest and run four offline contract tests. The reference training environment used Python 3.12, NumPy 2.3.5 and a CPU. The example writes the rebuilt files outside the published directory. Pass `--study-data` and `--method-dir` if the source files are arranged differently.

```
sha256sum -c freeze.sha256
python test_local_model.py
python train.py --output /tmp/spe-local-rebuild
cmp weights.json /tmp/spe-local-rebuild/weights.json
```

Two separate runs in the reference environment reproduced the weights and report byte for byte. The weight SHA-256 is `8f444d23138b39c16a2bce1ada44f17cc971c296ed6bb7fa04c4be7f8f6897e2`. Inference with the distributed weights requires Python 3.10 or later. The function

`next_characters(prefix, count=5)` also exposes predictions from the language head. Contract tests without trial data · Manifest

Evaluation after freezing the local weights

Cases	Positives detected	False alerts	Referred	Positives closed as other
16 · EN	0/8	0	10/16	2
16 · ES	0/8	0	11/16	2
16 · FR	0/8	1	10/16	2
16 · DE	1/8	1	11/16	1
48 · ES/EN	0/18	0	39/48	1

On 48 sentences about six new events the neural model produced zero alerts. It referred 39 and incorrectly closed one post-cutoff assertion as other. Its alert precision is undefined. The statistical n-gram baseline detected 11 of 18 positives, flagged six controls incorrectly and automatically closed all six indeterminate references. Those errors are part of its 100% coverage. The new 48-case cohort is English and Spanish. The French and German results use aligned translations of the original cases. An independently authored domain challenge in those languages is still needed.

What changes when a historical claim is rephrased?

A detector that flags a literal sentence can behave differently when an editor changes its wording. The study pairs two formulations of the same judgment, keeping the event, prompt, language, cutoff and reference judgment fixed within each pair. An English and a Spanish version of each of twelve constructions gave **24 pairs and 48 distinct texts**. The reference judgments comprise twelve pairs that assert an accomplished outcome after the cutoff, eight explicitly nonassertive controls and four whose speaker commitment cannot be settled from the excerpt. The two methods receive only the case and event inputs they ordinarily use. The reference and the transformation name are held outside those inputs. Paired texts and rationales

The examples use four previously registered events. ESA received Rosetta's signal on 20 January 2014, after deep-space hibernation. Perseverance landed on Mars on 18 February 2021, DART struck Dimorphos on 26 September 2022, and the OSIRIS-REx sample capsule landed on Earth on 24 September 2023. Each cutoff is 31 December of the preceding year. These are historical teaching examples. None of the assertions was generated by a remote model for this experiment. ESA Rosetta account · NASA Perseverance account · NASA DART account · NASA OSIRIS-REx account

How one alert becomes a referral or an automatic negative

For the year rule, the context flag is fixed to `False` and the numeric cutoff year is passed to the published function. The proposition rule uses the frozen bilingual event dictionary and threshold 5. For each method, the conditional alert-loss denominator is the number of positive pairs that it alerted on their base wording. The numerator adds the variants it referred to `REVIEW` and those it assigned `OTHER`. The results report the two destinations separately. `OTHER` is the classifier's automatic negative decision. A deployment would have to specify what happens next. Year rule · Proposition rule · Frozen inputs

Paired result	Numeric year rule	Proposition rule
Positive bases alerted	12/12	12/12
Positive variants still alerted	8/12	6/12
Base alert became a referral	0/12	6/12
Base alert became OTHER	4/12	0/12
Conditional alert loss	4/12 (33.3%)	6/12 (50.0%)
False alerts on eight control bases	8/8	0/8
False alerts on eight control variants	4/8	2/8

Both languages produced the same positive transitions. The annual method lost both the written-out-year and Roman-year alerts in each language. The proposition method retained those four alerts. It referred the metaphor, passive construction and nominalized landing in each language. For instance, converting “DART struck Dimorphos in September 2022” into the passive preserved the completed event for the reference reader and changed the proposition decision to REVIEW. The ironic Rosetta introduction retained an explicit factual second sentence and both methods kept the alert in both languages.

The negative controls reveal a different error. The proposition rule flagged both variants denying an outcome through an expression equivalent to “it is untrue that”. Its remaining control variants yielded four OTHER decisions and two referrals. The annual rule incorrectly flagged the quoted grammar examples and the indirect denials in both languages. On the four indeterminate variants it returned OTHER four times. The proposition rule referred two, alerted on the Spanish ambiguous irony and assigned OTHER to the English unattributed quotation. Those automatic decisions on indeterminate references are reported separately from false alerts. All paired transitions, reasons and scores

What the local source-field check establishes

A separate standard-library test inserts the inert editorial mark EDITORIAL NOTE · INERTE in the `text` field of one eligible source card. A local sender observes two serialized calls, one with the original card and one with the mark. It returns the same fixed output for both. The replay locates the mark exactly once in `sources[0].text`, checks identical prompts, cutoffs and policy IDs, and observes the same eligible card with the later card excluded. The request digest changes and both locally sealed receipts verify. These results concern the captured JSON envelope and a deterministic local sender. Remote model interpretation, provider-side instructions and behavior under malicious content were outside the experiment. Capture adapter · Local checks and metric definitions

Rubén Abella knew both rules and supplied the reference labels. The four events overlap development and the language pairs deliberately mirror each other. One forecast control exactly matches an earlier development or calibration sentence after case folding. The preassigned paired labels and corpus were fixed before the first run of these pairs. There is no external preregistration, random sampling or independent human adjudication of these labels. The fractions describe this teaching set. From the research folder run `python3 v11/robustness/replay.py` and `python3 -m unittest discover -s v11/robustness` to compare the published results and exercise the denominator and capture checks. Machine-readable stdout

The cost of a wrong decision

A false alert, a post-cutoff assertion released as other and a human referral entail different costs. Define loss as four counts multiplied by their weights and added. FP is a determinate control flagged as positive. FN is a positive closed as other. R counts all referrals, whatever the eventual decision. U is an indeterminate reference automatically closed in either direction. A referred positive therefore reduces automatic sensitivity, but incurs a single referral charge in this model. No assumption of perfect human correction or measured downstream cost is made.

Illustrative profile	False alert (FP)	Positive released (FN)	Referral (R)	Ambiguity closed (U)	Assumed capacity
Administrative review	12	25	1	14	No preset referral cap
Customer support	4	5	2	3	At most 20% referred

The weights are hypothetical units for sensitivity analysis. Both types of correct decisions cost zero. These are neither economic estimates nor regulatory thresholds. The profiles were designed after observing the main outcome. Within each profile the selection algorithm sees only the 48 calibration cases, compares thresholds 2–8 and considers review-all. A higher threshold demands more evidence before closing a proposition automatically. Ties favour the higher numerical threshold.

Both profiles selected 5 on calibration with 6/48 referrals and no errors among decisions. On the new challenge that threshold yielded one false alert, zero positives closed as other, 24 referrals and three indeterminate cases closed automatically. The administrative loss is **78 units**, compared with **48** for reviewing all 48 cases. Customer support incurs **61 units** and refers 50%, exceeding the assumed 20% cap. None of the tested numerical thresholds met that capacity constraint on the new set. The full threshold table is in scenarios.json.

Actual use depends on local review costs, prevalence and consequences for the affected person. These results support deferring autonomous disposition in administrative disputes. Measured reviewer time and a representative sample would allow an operator to replace the illustrative weights and repeat the selection.

A dated document, one disclosed leaf and a witnessed digest

A file can mention a decision taken on Monday, carry a publication date on Tuesday and enter an application on Wednesday. If the question asks what a person could have known on Monday, an accurate account of the decision may still be an inadmissible source. If the question asks what the application possessed on Tuesday, the ingestion record matters. A historical assessment also needs the relevant version. A page can retain its original publication date when an editor rewrites its conclusion. This supplement makes the choice of time and evidence explicit and provides an executable case with two versions of a bulletin.

Five clocks and an inclusive cutoff

Field	Assertion under examination	Evidence that could support it
Event	When the event described by the source occurred. For the Git example the event is creation of the named revision.	A contemporaneous event record and the precision of its clock.
Publication	When this version was made available to the relevant audience.	Publisher records, a dated release or an independently preserved historical copy.
Effective	When the relevant provision or status began to apply.	The instrument, commencement provision and applicable interpretation.
Ingestion	When the application received this version.	Ingestion receipt and matching source-version digest.
Capture	When someone retained or observed these bytes.	Collection record, evidence custody and an independently retained digest or trustworthy timestamp.

The caller selects the required fields for its question. All selected fields must satisfy the policy. Each accepts an instant or a closed interval at microsecond resolution. The documented timestamp profile takes an explicit UTC offset or Z. It rejects naive times, the unknown offset −00:00, leap seconds and precision beyond six fractional digits. A fixed-offset helper expands a calendar day or an hour into the earliest and latest representable instants in that period. A daylight-saving transition can change the offset during a day. For that case, supply each endpoint with its correct offset. An unobserved or unresolvable field can remain null. The operator must document where a claimed time interval came from.

For an observation bounded by *a* and *b* and a cutoff bounded by *c* and *d*, the observation is eligible when $b \le c$. It is ineligible when $a > d$. Overlap in all remaining configurations gives an indeterminate result, including an unknown time. Equality at an exact cutoff is admissible. For a policy with several required observations, any proven lateness makes the card ineligible. Otherwise, any unresolved required field makes it indeterminate. The program returns every individual decision alongside the overall result. A possible publication during the cutoff day cannot earn an automatic pass simply because both dates have the same calendar label.

A public snapshot that can be checked offline

The example preserves the complete 949 bytes of the public core's small Merkle module at revision `4f431863c63df6bba69ac85bdbf8defe36523d21`. The stored file matches Git blob identifier `d4f566a61057f841ceb5047e3211a8bc4ecd6112`. Its SHA-256 begins `b8c426752b9feb22`; the manifest supplies the full digest, the source URL and the byte count. The Git service reported 16 February 2026 at 17:19:38 UTC as the creation time of that revision. The collection records a local byte observation on 21 November 2026 at 17:11:16 UTC after comparison with the pinned blob. These are two separately named observations. The team has not established the first public availability of these bytes from an independent February archive. Publication, legal effect and operational ingestion are explicitly unknown. Pinned source · Revision

Run the following commands from the evidence directory. The first checks the size, SHA-256 and Git object identifier. With an independently retained expected manifest digest it also checks the manifest's exact bytes. The second asks whether the source satisfies a publication policy on 1 March. It returns indeterminate. Replacing the field with event answers a different question and returns eligible.

```
python3 evidence.py snapshot snapshots/manifest.json
```

```
python3 evidence.py decide snapshots/manifest.json --cutoff 2026-03-01T00:00:00Z --fields publication
```

What a single-leaf receipt establishes

The original public core hashes one serialization. This supplement introduces a separately named binary tree over **ordered raw byte strings**. Each leaf hashes a scheme prefix, a leaf tag, its unsigned 64-bit index, its unsigned 64-bit byte length and its exact bytes. All integers use big-endian order. Binary parents hash the scheme prefix, a branch tag and their two child digests in order. An unpaired child hashes with a different unary tag at every applicable level. The final wrapper hashes a fourth tag and the tree size with the top digest. The empty tree uses that wrapper with size zero and no top digest. The tags are 00, 01, 02 and 03 in hexadecimal. The prefix is the ASCII scheme identifier `spe-ordered-merkle-sha256-v1` followed by one zero byte. This application-specific encoding has its own test vector. It will produce different roots from a Certificate Transparency log. Certificate Transparency tree specification, RFC 9162 §2

A proof supplies the scheme, index, size and sibling digests from leaf to top. The index and size determine the combination order and when a node has just one child. The verifier requires the caller's expected root, rejects excess or missing digests and recomputes each layer. Equal byte strings can occur at different positions. The commitment preserves their separate indices and multiplicity. A proof for an empty tree has no valid index. The demonstrations traverse odd and even tree sizes, repeated leaves, changed bytes, altered sizes and extra path elements.

In the supplied snapshot vector, the three leaves are the manifest bytes, the retained module bytes and the published UTF-8 teaching line ending in LF. The public example exposes all three so that others can reimplement the calculation. A recipient checking the second leaf needs only its bytes, the proof and a trusted root. The root packaged alongside the proof serves reproducibility. For a dispute, the reviewer needs a separately acquired commitment and evidence of who accepted it and when. Tree size and position are visible. A digest does not guarantee confidentiality for data with a small set of possible values. A commitment to selected material also leaves open whether another relevant document was withheld.

Two intraday cutoffs and what reached the sender

The synthetic fixture contains status bulletin versions 1 and 2, a procedural note and an update known only to day precision. Version 1 is published at 09:00 local time. Version 2 is published at 12:25. At 12:00, the bulletin available under the publication-and-effect policy says “Status: scheduled.” At 13:00 the revised bulletin says “Status: approved.” The day-only card remains indeterminate at both instants. Two and three records respectively are admitted. The request holds the admitted versioned cards, snapshot text digests, selected root, cutoff and policy. A local sender receives these exact canonical UTF-8 JSON bytes and selects the latest admitted bulletin version. The output changes because the sender reads the request. The case makes no call to a model or an external service.

Running `python3 example.py` prints both outputs, both selected roots and the digest of the canonical dossier. Running `python3 example.py --out demo.json` writes a complete inspectable copy. Verification requires the caller's expected dossier digest. The released teaching value is `00599656cf1722ca3451915239b04c6f6d86045acb4636757f82c2891ad8c7c8`. Use `python3 example.py --verify demo.json --expected-sha256 DIGEST` and replace DIGEST with that value. The verifier reconstructs both admissions, compares the request bytes and source hashes, replays the local sender, checks the membership of one selected leaf in each run, and confirms that appending one space to the disclosed bytes fails validation. A remote provider would need its own observation to establish which payload its infrastructure actually processed.

An existing RFC 3161 response

The optional local command accepts an existing timestamp response and the original data bytes. It pins one CA certificate by an independently supplied SHA-256 fingerprint of its DER encoding. A supplied intermediate is treated as untrusted chain material. OpenSSL checks a SHA-256 message imprint against the input, and checks the timestamp signature and certification path with that CA. Supplying the original request additionally checks its values, including a nonce when the request contains one. A bare CMS token is supported through the token flag. The command prints the verified token description so that the reviewer can examine the TSA identity, time, accuracy and policy. Its verification profile uses the local machine's current certificate validation time. Archival validation at a past date, revocation evidence, trust-service qualification and any binding between the timestamped bytes and an actual provider call require their own evidence. Time-Stamp Protocol, RFC 3161 · OpenSSL timestamp verification

From the evidence directory, run `python3 timestamp.py data.bin response.tsr --ca root.pem --ca-sha256 FINGERPRINT`. Add `--query original.tsq` when the original request survives. The package contains no independently issued public timestamp. The automated test generates temporary, explicitly local certificates and checks acceptance of the matching response, rejection under a different authority, mismatch of request nonces and rejection of changed bytes. This tests the verifier. A real witness requires a separately obtained token, a suitable trust policy, independent evidence about the authority and retention of the exact timestamped object.

A controlled retrieval intervention with fictional archive tickets

When a rare string appears in an answer, the output alone cannot identify its source. We built a local toy archive in which the route by which a ticket reaches the output is known. The experiment asks whether adding or replacing one document changes a fixed process's output under a fixed prompt. It uses no people, real secrets, commercial model, model weights, network connection or provider account. The archive identifiers use the reserved example.invalid domain.

Assignment, capture and controls

There are 40 matched cases. Each has separately labelled 128-bit values for the target ticket, alternative ticket and ineligible future ticket, yielding 120 distinct values. The public repetition derives these values and the ordering from labelled HMAC-SHA-256 calculations with a published seed. Anyone can predict them. Independently generated private values for a future controlled experiment use the operating system's randomness in a separate registration command. The two designs answer different questions about predictability.

Each case has three conditions. In **present**, an eligible ticket card holds the target. In **absent**, it is withheld. In **decoy**, a different eligible card holds the alternative ticket in the same field. Every candidate pool includes an eligible generic note and a ticket dated 11 January 2026, one day after the inclusive cutoff. Filtering drops that final item before submission. All 360 requests contain the same prompt asking for the fictional archive ticket. The process receives the admitted documents as the retrieval result. Retrieval here consists of explicit card selection and date filtering. The experiment contains no learned retriever or language model.

Each of three local process modes receives each condition exactly once in every case. The public assignment interleaves the resulting 360 invocations in pseudorandom order. Each invocation starts a fresh actual Python subprocess, writes the canonical request bytes to its standard input and records its UTF-8 standard output with the published capture adapter. The **reader** copies the eligible ticket if present. The **generic**

process emits a fixed general sentence. The **side** process takes the target value from a separately supplied fixture environment variable, regardless of its request content. The experimenter sets that variable only for the side control. The capture seals the standard-input request and the output. The retained design and executable explain the side channel, whose contents have no independent attestation. An application that emits a green integrity check for all three processes has verified the bodies it captured. That result gives no way to distinguish the mechanisms from one output alone.

Measured outcomes

Local process	Target matches, present	Target matches, absent	Target matches, decoy	Alternative matches, decoy	Paired target difference, present minus absent
Reads admitted ticket	40/40	0/40	0/40	40/40	+1.00
Generic reply	0/40	0/40	0/40	0/40	0.00
Separate fixture input	40/40	40/40	40/40	0/40	0.00

A match requires the entire case-sensitive output to equal the assigned ticket. For each process and contrast, the paired difference is the number of cases matching only under present minus the number matching only under the control, divided by 40. The present-minus-decoy differences were also +1.00, 0.00 and 0.00, respectively. In the reader the 40 changes all went in the same direction. In the generic process neither arm matched in all 40 pairs. In the side process both arms matched in all 40 pairs. Zero out of 360 outputs equalled their case's future ticket. The file of results retains all nine cells, generic and unavailable replies, unforeseen output and both sets of discordant-pair counts.

These are measurements of three intentionally deterministic programs. There are 40 matched values per contrast, not 360 independent tests of one hypothesis. The chosen processes establish that a document dependence, a null response and an alternate route produce distinguishable intervention patterns when the fixture is controlled. None of those results estimates prevalence in deployed systems, identifies an actual model's training data or certifies the source of a historical response. Reusing the public values in a remote trial would remove the unpredictability on which a rare-match argument depends. Even a genuine private match could come from an alternative channel, shared instrument or earlier observation. A generic output does not exclude internal access to a document.

From the interventions directory run `python3 study.py run --out replay` to generate 360 fresh local processes and three result files. The `pairs.json` file holds the public values, `captures.jsonl` retains the exact recorded requests and outputs in execution order and `metrics.json` contains the paired and condition-level counts. Use `python3 study.py verify --out replay` to check eligibility, the fixed prompt, capture integrity, order and reported metrics. Four unit tests also exercise one entire matched set through the actual subprocesses, incomplete and altered records, and private registration.

For an unpublished future fixture, `python3 study.py register PRIVATE_DIRECTORY --pairs 40` creates a new local directory. The private file retains three independent 128-bit values, a 256-bit random commitment salt and the randomized order of the three conditions for each pair. The public counterpart contains the study descriptors and a SHA-256 commitment to every canonical private row, under a dedicated prefix. The command prints the public file digest and never transmits data. `python3 study.py check-plan PRIVATE_DIRECTORY` recomputes the commitments. Independent retention of the public digest before any observation would help establish precedence. A digest created and held solely by the

investigator provides no independent time or authorship. The registration command does not execute a third-party trial.

Stronger privacy and execution evidence, as research proposals

The implemented selective-disclosure baseline verifies one submitted leaf against an independently supplied ordered Merkle root. It reveals that leaf, its index, the tree size and the sibling digests. A root proves a commitment to data; assigning it a collection time or an origin requires a separate trust mechanism. For a low-entropy field such as “eligible” or a small list of dates, exposing a plain hash can reveal more than the author expects. Roots and identifiers reused across cases also allow linkage. Keep personal material out of public demonstrations and make the intended disclosure explicit.

Proposed zero-knowledge statement. A reviewer wishes to know whether at least one document meeting a published time rule occurs in a specifically committed request, without learning its text or times. Proposed public inputs are the version and verification key of the proof circuit, policy identifier, cutoff, selected-set root, tree size, a request commitment and the requested predicate. The witness contains the exact canonical request bytes, a fresh blind for the request commitment, the hidden card and its index, the sibling path, and the per-card blind defined by a new salted leaf scheme. Inside the relation the prover must open the request commitment, parse the published canonical serialization, prove that the selected root and cutoff are the ones in that request, bind the same indexed card in the request and the tree, and check that the latest allowed observation for every required field precedes or equals the cutoff. The circuit must reject ambiguous encodings, invalid dates and missing required fields. A weaker circuit that verifies only membership cannot establish temporal eligibility. The salted scheme needs its own version, commitments and test vectors. Existing unsalted roots cannot silently acquire that property.

That existential predicate covers one card. A claim about every admitted card would require proving the comparison for the complete committed selected set. Proving correct selection from an upstream corpus would also require a prior authenticated commitment to that corpus, the selection program and an exhaustive selection relation. Even a sound proof would say nothing about whether the publisher's date is true or a remote processor received the request. Disclosed metadata, repeated queries and the choice of predicate can still identify a person. A future evaluation would need an implemented relation, independently reviewed soundness and privacy assumptions, setup details, tampering tests, runtime and proof size. The present experiment used no proof circuit. NIST Privacy-Enhancing Cryptography and ZKProof reference work

Proposed trusted-execution evidence. A verifier would send a fresh challenge to a measured local component that filters the cards and hands the exact request bytes to a particular sender. The evidence would bind that challenge, the component and policy measurement, the selected root, the request digest and the output digest or a key that signs the bound receipt. A verifier would appraise platform endorsements, approved reference measurements, freshness, platform version and its appraisal policy. The relying party would retain the evidence, the result and policy version needed for later interpretation. A claim of execution by a remote provider additionally needs an authenticated connection to the intended provider and matching provider-side evidence. A local attestation cannot observe private execution elsewhere. The trust assumptions include the platform, firmware, verifier and reference values, the measured component's real I/O boundary and the confidentiality of its data. Persistent device identifiers, timestamps and record sizes can enable cross-case correlation. No enclave, device attestation or provider receipt was evaluated in this study. IETF RATS architecture, RFC 9334 · Entity Attestation Token claims, RFC 9711

Two independent reviews of a temporal claim

Measurement status at release: zero human reviewers, zero observed human decisions and zero measured human minutes. The tests create fictional cases and simulated timings in a temporary directory. The instrument can collect prospective observations. Its existence supplies no estimate of human accuracy, independence or workload.

What the two readers decide

The question is whether the response **asserts that the named event had already happened when its documented occurrence was later than the cutoff**. Use the same calendar-day convention for both readers. For a same-day event of unknown time use day precision and flag the uncertainty in the rationale if ordering within the day matters. Event dates, source publication dates and legal effective dates can differ. The supplied retrospective reference helps identify the event. It does not establish what the model retrieved or what the public could know at a given hour.

Label	Decision rule
`post_cutoff_assertion`	The response commits to the completed target event and the documented event date is after the cutoff. Record the words carrying that commitment. An explicit year is optional.
`other`	The completed event was on or before the cutoff under the date convention, or the speaker only predicts, denies, hypothesises, quotes without endorsing, asks, or declines to answer. Explain the reading with a short extract.
`indeterminate`	An ambiguity about the event, speaker commitment, identity or chronology prevents a defensible judgment. Say which evidence or distinction is missing.

Read the prompt and response together. A copied question by itself makes no assertion. A quoted prediction followed by an explicit endorsement of completion can. A demonstrably wrong number and an event disclosed in the question can require separate investigation. Keep those observations in the evidence field and evaluate the specified proposition. If the supplied source cannot resolve ambiguity, record it. Freeze the source version or provide both readers the same dated, lawful reference copy before starting.

Coordinator and reader workflow

Use Python 3.10+ and run these commands from the directory containing `review.py`. Paths in capitals are placeholders. The coordinator first records the method, cutoffs, evidence rules and exclusions outside the review directory and obtains its **previously committed** SHA-256. Supplying a new digest calculated after access would defeat prospective blinding. The program compares the digest before opening the case and event files. A JSON freeze declaring the case and event digests in its `sha256` map binds both inputs byte for byte before packet creation. A plain-text protocol remains usable and commits only to the protocol. The coordinator records the input byte digests with the assignments. A digest does not establish when or by whom a freeze was made.

1. **Prepare.** Cases are a JSON array of objects with `id,event_id,cutoff,language,prompt,text`. Events are an array with `id,name,event_date,source_url`. Both dates use YYYY-MM-DD. Only language, prompt, response, cutoff, event name, event date and the source URL enter a reader packet. Case IDs, rules, labels, predictions and other metadata are omitted. Inspect the allowed text and URLs separately for accidental answer keys, model names, template labels or identifiers. Do not change observed words to conceal a material fact. Exclude an affected row under the agreed rule or disclose the compromise. The program independently shuffles presentation order and generates unrelated random tickets per reader. For batches of more than one case the two orders differ.

2. **Distribute privately.** The batch directory contains `coordinator.json` and two independent SQLite files. Give each reader only their own packet, this protocol and the CLI. Keep the crosswalk with the coordinator and the two readers apart. Do not give readers access to the full dataset, method predictions, previous labels or one another's work. Record any prior access or discussion as a protocol deviation. Publicly released cases cannot become genuinely unseen by hiding identifiers.
3. **Review.** `session` opens a pending item, starts a monotonic active timer, permits pause/resume, then saves a label, a reason and active seconds. The timer includes decision entry. Pause for interruptions. Quit discards that attempt including elapsed time. If resuming offline work from a separate timer, use `show` and `record`; enter the total active seconds for that item, including earlier interrupted work. A refusal and a pending or interrupted review are distinct. Empty reasons, duplicate responses, zero or invalid times are rejected. The local label and reason are stored as entered.
4. **Close independently.** Each reader gets a receipt digest from `seal` and returns their closed SQLite file separately and securely. `--allow-pending` explicitly permits a partial return. Restore each to its original name in the coordinator's batch, verify each digest against the independently returned receipt, and retain both sealed originals. `status` displays counts and the verified receipt of a sealed packet without revealing labels. The CLI checks packet identity, order and isolated answer integrity before comparing the two responses. The receipts detect changes relative to trusted references. They do not authenticate a person.
5. **Adjudicate separately.** A third reader sees the two original reasons on paired disagreements and paired indeterminate judgments, reviews the original source, and records an independent final rationale and cumulative active seconds. The adjudication may itself be indeterminate. Repeat the adjudication commands until the queue is empty. Only the separate `adjudication.sqlite` file receives these decisions. A missing second judgment remains pending. Raw reader agreement stays fixed.

```
python3 review.py prepare --cases CASES.json --events EVENTS.json --freeze FROZEN_PROTOCOL --freeze-sh
a256 RECORDED_DIGEST --mode human_collection --out BATCH
```

```
python3 review.py session A.sqlite
```

```
python3 review.py show A.sqlite
```

```
python3 review.py record A.sqlite --ticket TICKET --label indeterminate --evidence "Which date applies
 is unclear" --seconds 70
```

```
python3 review.py status A.sqlite
```

```
python3 review.py seal A.sqlite
```

```
python3 review.py report BATCH
```

```
python3 review.py adjudicate BATCH
```

```
python3 review.py adjudicate BATCH --case CASE_ID --label other --evidence "Describe the decision and
source" --seconds 95
```

```
python3 review.py report BATCH
```

```
python3 -B -m unittest discover -s . -p test_review.py
```

`session` and the `show` plus `record` pair are alternative ways to record an item. Repeat with the other reader's packet on their own device. Treat the example seconds above as illustrations. During a rehearsal select `--mode simulation` and keep rehearsal batches separate from human observations.

What to report

Publish the three-by-three matrix with A on rows and B on columns. Exact agreement is the diagonal divided by all paired cases. Unweighted Cohen's κ is (observed agreement − expected agreement) / (1 − expected agreement), with expectation obtained from the two label marginals. Indeterminate forms its own category, including agreements on indeterminate. No paired cases yields null agreement and κ. Single-category constant marginals yield undefined κ if chance agreement is one. Keep pair completion and each reader's pending count alongside the matrix.

The JSON report separates per-reader, paired and adjudication effort. It reports observed person-minutes, median and nearest-rank p90 active seconds, and timing method counts. Paired duration sums both readers' active time on completed pairs. Total effort includes completed unpaired reviews and completed adjudications once each. Wall-clock delay, abandoned work, recruitment, training and future queue effort require separate measurement. Preserve unresolved final labels and publish the number of cases still without a disposition. Report cohort size, language and any exposure to public examples along with these figures.

Rights and privacy in a deployment

A coordinator should determine the lawful basis and purpose for processing personal data, minimise what reviewers see, limit retention and control access. Use synthetic examples in a public demonstration. Raw notes or command histories can contain personal details. Enter only the minimum necessary content and share aggregate results after disclosure review. Pseudonymous tickets can still be linked using the private crosswalk. Under the AI Act a person with grounds to allege an infringement may complain to the competent market surveillance authority (Article 85). Article 86 provides qualifying affected people with an explanation from the deployer about AI's role and the main elements of a decision. The underlying administrative or judicial route to challenge the decision remains relevant. GDPR Articles 77 and 79 concern data-protection complaints and judicial remedies. A reviewer's label by itself establishes neither entitlement to redress nor the lawfulness of a deployment.

Official sources checked 21 November 2026. Consolidated AI Act · Article 85 · Article 86 · GDPR, especially Articles 5, 6, 32, 77 and 79.

Open the review desk

You need your own **A.sqlite or B.sqlite** reviewer packet, Python 3.10 or later and a modern browser. Each reader works on their own computer. Open a terminal in the directory containing `v11`. Replace `A.sqlite` with the path to the packet you received:

```
python3 v11/desk/desk.py A.sqlite --authored-holdout
```

On Windows, `py -3` can replace `python3`. The browser should open automatically. If it does not, copy the private link printed in the terminal into the browser. Leave the terminal open. The page embeds its code and styles. It is served on `127.0.0.1` at a temporary port, works without Internet and never fetches source URLs. The value at the end of the link gives access to the local packet for this session. Keep it private. Use one tab and avoid simultaneous editing of the same packet in the command-line tool.

Complete a review

1. Choose **EN** or **ES** and select **Start or resume**. Active timing starts. The three panels present the original prompt and response, any cards present in a captured request, and the retrospective event date and reference. Changing interface language preserves the original evidence.

2. Read the unmarked response. To mark your own evidence, select text and choose **Underline selection**. For keyboard-only use, open **Select using the keyboard**, type a verbatim phrase and choose **Underline phrase**. Extend a repeated phrase to identify a single occurrence. Solid underlines are the reader's selections. Each can be removed. The tab receives no reference labels or detector predictions.
3. Choose a label and give a reason. The draft saves while you type. **Pause** saves the current draft and stops the timer. Switching away requests a pause. The server receives a heartbeat every four seconds. If the page disappears without pausing, the server counts at most fifteen seconds from the last heartbeat and requires a resume. An abrupt shutdown can lose the last fraction of unsaved time or typing. Pause before closing. Open the same packet and choose **Start or resume** to continue.
4. Choose **Save and continue** when ready. The label, rationale, exact selected spans and active seconds are stored in the same tables as the command-line tool. Completed independent decisions are retained. A subsequent adjudicator follows the review protocol.
5. After all cases, choose **Seal packet and get receipt**. Keep the SQLite file and its digest. Return the packet by the agreed private channel and send the digest separately. Press Ctrl+C in the terminal to stop the application. For an explicitly partial return use `review.py seal A.sqlite --allow-pending` as described in the protocol.

Optional assistance. Cues appear only when the reader deliberately enables them for the current case. A dashed underline marks four-digit years between 1500 and 2199. A match alone settles neither the meaning of a claim nor its relation to the cutoff. Exposure remains recorded across pauses and restarts. Each new case starts blind. For a double-blind protocol, keep all cases blind and declare any earlier exposure. To count the conditions separately run:

```
python3 v11/desk/desk.py A.sqlite --summary
```

The summary counts blind, assistance-exposed and other-interface decisions. Assisted seconds measure active intervals after the reader opted in. The sealed reason stores exposure and exact evidence spans with Unicode code-point offsets. The review kit's general agreement report combines decisions under its protocol. A blind-only analysis needs to exclude assisted cases and report any earlier exposure. Timers, pauses and hashes do not establish that two humans worked independently.

What the system actually received

The `--authored-holdout` flag describes the authored version 1.1 evaluation set. No model generated those answers and no source cards were delivered during generation. The third panel's source is a retrospective reference for checking the event. For a packet of unknown origin, omit this flag and the desk explicitly reports that the capture is unavailable.

For a study with recorded requests, the coordinator can supply captured requests keyed **only to that reader's blind tickets**. Use `--contexts contexts_A.json` instead of the authored flag. The JSON object maps packet tickets to records containing the exact `request_utf8`, its `request_sha256`, and the observed `output_sha256`. The encoded request includes `prompt`, `cutoff` and `sources`, with source fields `text`, `date` and `url`. The desk checks both digests, the prompt and the cutoff, then shows the source cards present in that local request. An observed empty list is reported as zero cards. A missing ticket is reported as unknown. Matching this record establishes neither provider-side use nor a causal effect. Screen free text and URLs for accidental cues. Share only authorised source content necessary for the decision.

What EU law requires about language

The EU AI Act contains no universal duty for every AI system to answer in every official EU language. Article 13(2) requires accessible, clear and understandable instructions for deployers of high-risk systems. Article 47(2) requires the EU declaration of conformity to be translated into a language the relevant national competent authorities can readily understand. Article 50(5) requires the disclosures covered there to be clear, distinguishable and accessible. These provisions concern defined materials and audiences. The language of a public procedure, applicable national language rights and the needs of the people using a service must be checked in its particular context. This research desk offers English and Spanish controls and preserves the original language of the evidence.

Official sources checked 21 November 2026. Consolidated AI Act · Article 13 · Article 47 · Article 50.

Reproduce the checks

```
python3 -B -m unittest discover -s v11/desk -p test_desk.py -v
```

Tests create simulated cases, timings and temporary packets. They check timing, persistence, capture matching, local access checks, assistance tracking and the packet receipt. With Node 18 or later they additionally run the shipped JavaScript against the actual HTTP server using a DOM test double. This exercises evidence selection, language switching, two decisions and sealing. It does not measure visual readability or certify screen-reader compatibility. Zero human participants.

What can be dated, and what was measured?

Release 1.0 left 58 B records without an adequate historical date. On 21 November 2026 we revisited all 56 recorded destinations and searched both names that lacked a URL. For **53 records**, a specific object can now be dated on identifiable evidence. The object date is **before 4 December 2025 for 27** and **after it for 26**. **Five** remain without a supported public date. Two M records outside that count yield further corrections. A public capture establishes ALETHEIA's availability in August 2025. The candidate Rust package was registered in March 2026. Ledger of all 58 records · 29 targeted searches and checks

Interpreting a date

Evidence for the 58 records	Count	What the date applies to
Publisher's date for the article	29	The page or its structured metadata supplies the date. The ledger distinguishes publication and modification when both are available.
Archived copy of a destination	12	The captured edition was publicly available by the time of the capture. A post-cutoff capture leaves open an earlier publication.
Related dated edition or event	4	FINOS, V-Proof, Giza and Aleo each have an identifiable public item. The date applies to that draft, interview, presentation or announcement.
Revision or update date only	4	B012, B034, B068 and B071. The site identifies an update, listing change or document revision. The first edition may be earlier.
Version distributed through a public registry	2	UQLM 0.1.0 and GRAIL v0.0.5. The public record ties the date to the named release.
Agency posting or dated administrative edition	2	An entry in the SEC's written-input index and an edition of the USPTO budget submission.

Evidence for the 58 records	Count	What the date applies to
Public date unresolved	5	The remaining identification and retrieval gaps are set out below.

The `object_date_vs_2025_12_04` column compares the date of the identified object with the cutoff. The counts do not establish legal status or enumerate first disclosures. B026 illustrates the underlying problem. The current edition dates to July 2026, although the May list cited a title containing "2025." For B042, the July and August dates on the current page also follow the May citation. For B044, the publisher metadata says 6 May, one day after the 5 May access date in the old bibliography. The editions originally consulted remain uncertain. Release 1.0 preserves the contemporaneous record of those gaps. Ledger entries B026, B042 and B044

Four technical leads revisited

ALETHEIA (M37). The Internet Archive capture of **26 August 2025 at 14:50:47 UTC** contains the project name, a source-code listing and a description of a hash-chained decision log. GitHub records three commits that day, including one at 13:56:22 UTC. The independent capture now establishes that this particular description was public before the cutoff. Its technical relevance is agent traceability. The captured page documents no evaluation of public historical availability of the facts in generated answers. August capture · Identified commit

The Rust module (M38). The original lead named `temporal_retrieval`. We examined a specific candidate, `graphrag-core`. The registry dates release 0.1.0 to 12 March 2026 and release 0.1.1 to 18 March. The downloaded 0.1.1 package matched the published SHA-256, `dc86967bf8af1084262f5c481521585619a37ec1cb023479c0e02cbb0c0bef4d`. It contained 233 files, including `src/graph/temporal.rs`, and documented optional temporal intervals for entities and relations. A literal search of the Rust, Markdown and TOML files found no `temporal_retrieval`. This establishes the types and interval support in a post-cutoff candidate. The identity of the historical source and the existence of its alleged mandatory pre-cutoff filter remain unresolved. Release history · Version and checksum · Packaged documentation

GRAIL (B093). The label matches *Guaranteed Rollout Authenticity via Inference Ledger*. GitHub reports v0.0.5 at **21:40:22 UTC on 15 September 2025**. An independent capture on **7 October** contains the expanded name and describes binding a reinforcement-learning rollout to a specific model and input. The May review supplied no link. Attribution to that particular work is inferred from the matching name and technical description. The capture establishes that this usage was public before the cutoff. The comparison concerns authenticity of generated training trajectories. Dating when facts became publicly available requires other evidence. Release · Archived repository

MECM (B094). Searches combined the acronym with *temporal causal gating*, *autoencoder* and causal-reasoning terminology. Results included emotion-reasoning and brain-network models. None supplied an author, title or account of the mechanism that uniquely identified the lead. The ledger leaves its date and role in the technical comparison open. Searches S04, G6 and U6

Further corrections that affect the interpretation

- **Uncertainty and governance before the cutoff.** PyPI dates the UQLM 0.1.0 distributions to 6 May 2025. An Edinburgh repository capture from 15 November 2025 preserves a public index of hallucination research. FINOS announced its first draft on 1 October 2024 and v1 in June 2025. These records respectively identify a software distribution, a literature index and a governance framework. Later changes have their own dates. UQLM · Archived index · FINOS draft · FINOS v1

- **Products and public claims.** The Giza presentation in 2023 and Aleo mainnet announcement in 2024 document pre-cutoff public descriptions. An interview on 13 May 2026 describes V-Proof, and current company materials explain its evidence-verification offering. This dates a press account. The first technical release and its operation remain to be established. Giza · Aleo · V-Proof
- **A dated letter and its posting.** In B066 the private submitter's letter bears the date 5 October 2025. The SEC's written-input index dates the entry to 14 December, establishing a post-cutoff appearance through that channel. Earlier public circulation and the dates of works mentioned in the letter require separate evidence. The letter expresses the submitter's views. Letter · SEC index

B013 identifies Alexey Zaytsev's slides. The PDF has a creation and modification timestamp of 21 April 2025, with no direct evidence of public posting on that date. B033 remains a corporate page without an adequately dated historical edition. B051 and B055 were dynamic searches. A present-day result and a failed retrieval cannot reconstruct the list seen in May. B094 is the unresolved MECM label. The row notes also preserve uncertainties about first publication, version and identity for the 53 records with a dated object. Full dating ledger

The proposed S(t), A(t) and Δ(t) measurements

The original hypothesis proposes $S(t) = \alpha C + \beta E + \gamma R$. C would express agreement across controlled runs or models. E would describe stability under prompt or sampling perturbations. R would measure recognition of a concept in a preselected representation space. Estimating those terms requires specifying the models, representations, distribution of perturbations, repetitions, unit of analysis, normalization, weights and variability. Qiu and Miikkulainen's semantic density and Farquhar and colleagues' semantic entropy are comparators for uncertainty about responses. The later information-per-token study examines a different quantity. None supplies the values of C, E and R for this collection. Qiu and Miikkulainen, v1 · Farquhar et al. · Information per prompt token

A(t) would express observable support for a proposition in an identifiable corpus available by t. An estimate requires a definition of support, a rule for admitting dated documents, a treatment of duplicate sources and corrections, and criteria for insufficient coverage. The date of an event, its first news report and its ingestion by an index can differ. Unless S and A have comparable scales, the sign of $\Delta(t) = S(t) - A(t)$ depends on their normalization. A weak alert would require defined and calibrated terms and a separate evaluation of its uncertainty. Related ex-ante inference benchmark

The published evaluation counts labels, decisions and abstentions for classifiers acting on authored responses and registered events. The explicit-year comparator receives an externally supplied flag about the context. The v1.1 proposition classifier uses textual cues and dates supplied by the auditor. Its development-selected threshold controls whether the rule emits a label or requests review. **The study did not estimate S(t), A(t), their weights or curves of Δ(t).** Its results cannot establish calibrated detection of noncausal activations or attribute a response to training, retrieval or chance. Pinned public year detector · Classifier specification · Published threshold

Use and citation

For an audit or appeal we would begin with the local capture, versioned sources, and a route to reopen a case. The linguistic alert supports that review. In the released data a system making final decisions would still err even when abstaining frequently. Independent human review times and disagreements on authentic text are the next decisive observations. The historical pilot, six-event challenge, translations, AI reading and canaries have different sampling units. Combining them into one headline accuracy would obscure what each experiment measured.

The author develops the technology under study. Software is distributed separately under Business Source License 1.1. The combined edition also includes the baseline study and its annotated prior-art registers. Research, software development, translation and editing were AI-assisted.

Rubén Abella (2026). *Temporal auditing of AI outputs: request-level evidence, prior art and the limits of automated detection*. SPE Audit, v1.1. DOI 10.5281/zenodo.24424024 · GitHub

Reuse and attribution

© 2026 Rubén Abella. The original text and original figures in this publication are available under the Creative Commons Attribution 4.0 International license. When reusing them, credit Rubén Abella, link to the license and indicate changes. The license grant covers only the author's original text and figures. Third-party material retains its own terms. The SPE Audit name, logo and other trademarks are excluded from this grant. Software has its own terms under the repository's Business Source License 1.1. CC BY 4.0 · Software license

SPE Audit

OPEN RESEARCH PAPER

Temporal auditing of AI outputs

Request-level evidence, prior art and the limits of automated detection

Rubén Abella · SPE Audit

21 November 2026

RESEARCH EDITION 1.0 · EN

Contents

Abstract . . . 3

A question the public can answer . . . 3

What an audit record can establish . . . 4

Evidence and results . . . 7

Measuring a temporal claim . . . 10

What a citizen can ask under the EU AI Act . . . 13

Where this study fits among existing reviews . . . 14

November evidence review: what was already known about temporal AI audits? . . . 15

Conclusions . . . 20

The 54 records in the May comparison . . . 21

Review coverage and provenance of references . . . 30

Catalogue of the 116 reconciled and supplementary references . . . 34

Abstract

An answer can accurately describe an event that took place months after the date from which the system was asked to respond. A reader faces two different questions. Does the answer assert a future outcome as an accomplished fact? What evidence did the application preserve about the production of that answer? We investigate both through a source-level landscape review, inspection of the public SPE Audit implementation and a credential-free reproducibility study. A positive label denotes an assertion that an outcome after the cutoff has already occurred. Assigning a causal source to that assertion requires further observations.

We compare the public year detector with a rule that uses dated events, linguistic cues and explicit abstention. In 96 constructed cases, the former detected 4 of 32 positives and produced 12 false alerts. The latter detected 24 and referred 12 cases for review. On 16 separately written challenge sentences the rules detected 1 and 0 of the 8 positives, respectively. The latter referred 14 of 16 cases for review. Twelve retrieval interventions demonstrate an observable dependency in a deterministic local sender. A separate archival recount preserves all 25 attempts in an earlier pilot, including three service errors. The release supplies each prediction, the rules, evidence cards and a verifier for all 124 captures in the controlled study.

The review reconstructs the historical inventory, bibliographies of two earlier reports and the closest comparators revisited in November. Prior work already addresses historical retrieval, prospective evaluation and cryptographic provenance. The present contribution is an inspectable combination at a single request boundary and a measurement of what is lost when wording changes. Its practical relevance to European AI accountability is that a reader can verify the recorded text, challenge its interpretation and identify evidence that is still missing.

A question the public can answer

Suppose someone asks an assistant to reconstruct a space mission as of 1 January 2022. At that date the intention to strike Dimorphos with DART had been announced. NASA reported the actual impact in September. “An impact is scheduled for 2022” and “the spacecraft struck the asteroid” share a subject, vocabulary and apparently a date. Only the second presents the outcome as accomplished. If an interface colours 2022 red, can its reader tell why? If there is no alert, can the reader tell whether the detector recognised the verb? Historical reporting, case summaries and explanations given to people contesting decisions all create opportunities for this distinction to matter. Launch announcement · Impact images and confirmation

Distrust of AI systems encompasses power, rights, oversight and decision quality. We ask a narrower question that can be tested in public. Can another person inspect the request that an application prepared for a specified date, read the material it supplied and test the interpretation of the resulting answer? Three readers need different things from the same record. The affected person needs an intelligible explanation and a route to correction. The operator needs to retrieve the relevant run. The researcher needs to rerun the comparison on the same material and retain disagreement.

What counts as a result

1. **Content evidence.** A study card identifies a source, its declared date and a short text. The request preserves the exact cards supplied to a local function. A changed byte changes the digest the verifier expects.
2. **Temporal interpretation.** The label applies to the proposition asserted by the speaker. A future year can legitimately appear in a plan, a question or a denial.

3. **Operational burden.** A rule that abstains on difficult cases generates work for a reviewer. We report positives detected out of all positives and coverage across all cases.
4. **Attribution.** The record establishes what was observed at the client boundary. To attribute a particular input, time or causal route to the provider requires evidence at that boundary.

A counterexample gives the limit of the last inference. Two services receive the same request and return the same sentence. One repeats a forecast that happens to come true. The other consults a later news report outside the application. The bytes and digests observed by the client can be identical. Any procedure receiving only these observations must return the same answer in both worlds. Without an observation that distinguishes them, its attribution will be wrong in at least one. This is a property of the information available even under an ideal cryptographic hash function.

How the review was built

We began with the numbered entries in the historical comparison and the two surviving reference lists. We sought a document identity, a date and the actual technical operation described for each entry. Preferred evidence included published documents, identified paper versions, specifications and patent publications. A later version can clarify an approach without establishing that every detail was present earlier. The historical comparison date is **4 December 2025**. The review closed on **21 November 2026**. Overlap between registers is retained when it explains how two questions led to the same source.

There are three views of the material. M records retain the original order of the May inventory. N records study the closest comparators in depth and distinguish later developments. The B catalogue traces references from the two reports and supplementary sources. These populations overlap. An entry can link the application, grant and international publication of the same patent family. Unresolved leads remain visible. This reconstructs an identifiable corpus and extends it through targeted feature-based searches. A worldwide census, a determination of patentability and a freedom-to-operate opinion are outside the method.

Reading the results without specialist tools

A **positive** here means that the sentence presents a future outcome as accomplished. **Review** means a person still has to decide. For a first reading, open the demo, choose DART, inspect the cutoff and the eligible card, and compare what the rules decide. Verify the digests and read the case rationale. A reader who wants to check the complete report can run the same experiment with Python and open a CSV containing every decision.

What an audit record can establish

Suppose a customer challenges an answer that a service says was generated from the information available on a particular day. There are several distinct factual questions. Which text is the customer challenging? Which documents did the client supply? When was the relevant fact first supported by evidence? Did the provider process that exact input? Answering the first two helps to investigate the others. Each conclusion requires its own observation.

The public engine and the observable sender

The February release contains functions for filtering timestamped records, ordering their descriptors, hashing a serialized list, constructing a forensic capsule and linking capsules in SQLite. Its context function orders by document ID, timestamp and source ID. The field called `context_merkle_root` in that release can hold a single SHA-256 digest of the whole serialization. Verifying it requires the relevant serialization. The implementation supplies no tree path for independent disclosure of one leaf. Public core at the evaluated commit

The published white paper illustrates full RAG integration with `context_items`. The public `generate_proof` at the pinned commit accepts text, a file or a precomputed digest. It sets the context field to a fixed digest and leaves the prompt hash empty. Its ZIP writer includes a capsule, SQLite ledger and input manifest. The verifier checks the ledger links and the hash of the latest capsule and, when an original file is supplied, compares its digest. Its overall Boolean uses ledger validity, capsule binding and any object mismatch. Signature verification has a separate status. An overall pass on an unsigned object can be a correct report of those specific checks. White paper · Public API

The study adds a small, separately identified capture adapter. The unit of capture is one call to a sender function. Its input is a UTF-8 JSON envelope holding the prompt, inclusive cutoff, policy version and full text and metadata of the eligible supplied cards. It excludes cards whose declared day is later than the cutoff, sorts those retained by unique identifier, and hands the exact serialized bytes to the sender. It records the text returned by the sender. A hash covers the request bytes, another covers the output bytes, and a record digest binds both to the schema and policy. An independent verifier parses the envelope, checks the cutoff and canonical ordering, recomputes the record digest and can compare it with an independently retained expected digest. At collection level a separate digest commits to the exact JSONL bytes in order.

For the controlled study the sender replays the specified response. If an integration later translates the envelope to an HTTP request, inserts instructions, retrieves other documents or invokes another model, the present observation ends at the local sender. A provider-side request identifier and matched provider logs would constitute additional evidence about that later execution. Independently retained records of the retrieval pool, selection policy, source versions and their acquisition would answer further questions about omitted documents. Requests made on another route or with an earlier program version remain separate events.

What the reviewer can conclude

Question	Observation available here	Additional evidence for a stronger conclusion
Does this output match a record?	Compare the output bytes, output digest and recomputed seal against the expected record.	For origin or timing, independently retained copies, a trusted signer or a time witness relevant to the claim.
Was each selected source eligible?	Reparse each declared publication day, apply the supplied cutoff and check the selected text.	Historical snapshots or publisher records to establish that version existed on that day. A page's first publication date can precede substantive revisions.
Was the complete intended context used?	Review every source that crossed the observed local sender boundary.	Inventory plus retrieval log for completeness. A record from the receiving side to establish what it processed, including hidden instructions or tool calls where available.
Does the answer assert a post-cutoff outcome?	Identify the event, asserted proposition, modality, polarity and dated comparison evidence.	A reviewer or adjudicated data can resolve ambiguous wording and assess what was knowable at the cutoff.
Did the model obtain the outcome from a specific channel?	An inconsistency can justify opening an investigation.	Evidence that discriminates among prompt hints, retrieved material, learned associations, reasonable prediction, coincidence and other channels.

A date policy needs an explicit choice of clock. A contract signed on Monday, filed on Thursday and indexed the following week has three useful dates. A reconstruction of publicly knowable information may use the earliest documented public availability of that specific version. An investigation of an organisation's information on hand may need ingestion time. Changing the clock can change eligibility. The adapter uses the supplied publication day and has calendar-day precision. It cannot distinguish an 08:00 release from a 16:00 decision that same day. The experiment's curated cards are authored summaries linked to dated sources, not archived publisher page bytes.

The proposed semantic measures

The original design proposed a score $S(t) = \alpha C + \beta E + \gamma R$. C would describe agreement across controlled runs or models; E would describe stability under prompt or sampling perturbations; R would describe how densely a concept is recognised in a specified embedding space. Reproducible estimates require published model and embedding versions, prompt variants, sample counts, distance functions, scaling and weights. A model that confidently repeats an unsupported statement can score highly on agreement and stability. The signal is a measurement of behaviour.

The external series $A(t)$ was intended to express dated, normalised support in sources available by t. An implementation must define the eligible collection, the unit being counted, handling of duplicates and revisions, source coverage, and the version and scale of normalisation. Agreement among an earlier cohort of models measures a different quantity. An empty search result in a sparse archive also gives weak grounds for a claim about the entire world. $\Delta(t) = S(t) - A(t)$ is interpretable as a difference only to the extent that the component scales and observation windows justify that comparison.

The proposed *strong* signal requires an actual assertion about a verified later event, combined with an attested pre-cutoff context. The public year heuristic accepts an externally supplied Boolean about context and matches explicit years. It leaves polarity, quotation and forecasts unresolved. The proposed *weak* signal uses a high $S(t)$ and a low $A(t)$, with specified thresholds. Calibration on examples separate from evaluation and a treatment for insufficient evidence would be required to interpret that flag. A classification alert, a local integrity check and a causal explanation are separate outputs. This study does not publish estimated

S/A curves or a calibrated weak-signal result. Public detector · ExAnte · Pitfalls in Evaluating Language Model Forecasters

Two examples a reviewer can use

A planned mission. Before landing, a mission announcement may already specify the following year's target landing date. “The landing is scheduled for next year” and “the spacecraft landed safely” have different factual commitments even when both include the same date. Retrieve a source contemporary with the cutoff, its later outcome report and the exact question. If the question already discloses the result, identify that information as a separate possible source. A documented distinction between plan, outcome and a repeated premise is the usable audit finding.

A disputed service decision (hypothetical). A resident contests a benefit notice summarising a decision made on 10 March. A disputed line refers to an amendment first published in April. The reviewer can match the notice to the preserved output and inspect the dated text of each card in the local capture. If the output attributes a March decision to that April amendment, record the precise contradiction and invite the service to produce the source version, decision record and human approval history. If the card instead contained a March draft that anticipated the same provision, the assessment changes. An audit of an explanatory text alone cannot identify which rule the actual decision-maker applied. The example illustrates evidence handling and makes no finding about a public authority.

A public scientific release can reproduce both examples using public or constructed documents. Operational cases can involve private prompts, decisions and identifiers. The study format intentionally keeps the full bytes available to the reviewer; an institution using it would need to choose access rights, retention and its own disclosure channel. Publishing a digest cannot anonymise a predictable short value. An appropriately authorised reviewer can receive the original under controlled access, while a public example can use newly authored records with explicit provenance.

Evidence and results

Four investigations answer four different questions. The 96 constructed cases test two rules on a controlled set of English sentences. A separately written set of 16 asks whether their decisions survive changes in wording. Twelve source interventions show what a local consumer does when the evidence it receives changes. Finally, an older archive contains 25 remote requests made under a historical cutoff. The last two investigations have different units and cannot increase the sample size of a classifier evaluation.

First establish what was measured

The positive reference label means that the answer presents a named target event as completed, although the event occurred after the stipulated inclusive cutoff. The reference is about what the sentence commits to. It does not attempt to read a model's internal state. A positive can arise from a lucky prediction expressed as a fact, a user's leading question, fabricated information that happens to match the outcome, a retrieved document, or knowledge encoded in a model. The statement alone does not distinguish those mechanisms. A negative reference includes a genuine forecast, a question, a denial or an accurate recollection of an event that already happened by the cutoff. It establishes nothing about the truth of other claims in the same answer.

The four events have dated descriptions of a plan and a later outcome. The study's source cards contain short editorial summaries and links. A reader can inspect what the local code received byte for byte. The cards cannot establish what a live web page displayed on any particular day.

Mission	Early source date	Outcome event	Selected outcome report
Perseverance	30 July 2020	Landed on Mars on 18 February 2021	NASA, 18 February 2021
DART	24 November 2021	Hit Dimorphos on 26 September 2022	NASA, 27 September 2022
OSIRIS-REx	10 May 2021	Delivered the Bennu sample to Earth on 24 September 2023	NASA, 24 September 2023
Rosetta	3 June 2011	Woke from hibernation on 20 January 2014	ESA, 20 January 2014

The public detector extracts four-digit years from 1900 to 2099. With the supplied flag indicating that the context contained no later information, it returns a strong alert if any extracted year exceeds the target year. The experiment runs the original detector and extractor from the pinned public Git commit. A separate, dated matcher is given the identity and date of the target event. It uses fixed phrase checks to recognise forecasts, denials, questions and some completed actions; unmatched outputs go to review. Access to a curated event date is an information advantage. Any observed difference between the two methods combines that advantage with their different handling of language. Public SPE detector

The 96 constructed cases

For each mission we replayed eight response forms under three cutoffs: the end of the preceding year, 1 January in the outcome year, and the end of the outcome year. Forms include a dated assertion, an undated assertion, a paraphrase, a forecast, a denial, a repeated question, an abstention and a mixed statement. Each sentence is reused across the three cutoffs. Before the event, four forms per mission assert completion. After the event, those same words can be valid retrospective descriptions. The resulting reference set has 32 positives and 64 other cases. The 96 rows therefore form a small crossed design with shared wording and four events. They are not 96 independent observations of language-model behaviour.

Each replay stores the exact locally filtered request and the prewritten response. The callback returns that response independently of the evidence it is given. No remote model is called for these cases. The reproduction script computes both predictions, then reads the evaluation labels. This guards against directly passing the answer key to the detectors. It does not create independence between a hand-written matcher, the preselected events and the primary wording.

Method and set	Correct alarm	False alarm	Correct other	Missed as other	Review, positive / other
Year rule, 96	4	12	52	28	0 / 0
Dated matcher, 96	24	0	60	0	8 / 4
Year rule, 16	1	2	6	7	0 / 0
Dated matcher, 16	0	0	2	0	8 / 6

On the constructed set, the year rule decided all 96 cases. Its sensitivity to all reference positives was 4/32 (12.5%), and its precision among alarms was 4/16 (25%). It raised 12 false alarms among 64 negatives. All 12 involved a year beyond the cutoff inside a forecast, a denial or a repeated question in the preceding-year condition. It detected no positives among the 16 in the same-year-before condition. Moving the cutoff into January had changed the integer year comparison even though the four events still lay in the future.

The dated matcher decided 84/96 cases (87.5% coverage). It flagged 24/32 positives (75%) and made no false alarms on the 64 reference negatives. The 24 alarms all matched the reference label; the eight other

positives were referrals. Looking only at the decided subset yields 84/84 correct. That number is conditional on the hand-written rule recognising the sentence. On this set the recurring unrecognised completion paraphrases reveal where an analyst would still have work. The independently worded challenge tests whether that burden changes.

The 16-case wording challenge

A separate authoring pass produced four new sentences per mission, with eight positives in total. The prompts supply no numeric date. The challenge writer knew the research question and the general limitation of a year rule. The new outputs and labels were fixed before the writer inspected the dated matcher or the primary fixtures. The reproduction script checks hashes for the matcher, challenge and reference labels. These measures make later changes detectable relative to the published files. They do not constitute external peer review, blind double annotation or a test on previously unseen events.

The year rule covered 16/16, recognised 1/8 positives and raised two false alarms on eight controls. One of the latter is an anticipated 2014 wake-up described conditional on Rosetta's timer working. The dated matcher automatically decided only 2/16 (12.5% coverage). Both were controls. It sent 14/16 (87.5%) to review, including all eight positives; its automatic sensitivity was 0/8. It issued no alarms, so alarm precision here is undefined. No reviewer subsequently resolved those referrals for a measured human-assisted result.

The wording difference is quite small. The dated matcher recognises the stock wording for landing on Mars. In h16 the sentence says that Perseverance was built to land and “did so” in February. A reader can resolve the pronoun and compare the event date with the 10 February 2021 cutoff. The matcher defers, and the year rule misses the assertion because 2021 is also the cutoff year. This is a failure to automate a straightforward judgement. The change from 84/96 decided in the constructed set to 2/16 in the challenge shows why a narrow reference set could give a misleading picture of operational coverage. No prevalence or model ranking can be inferred by pooling the two splits.

Twelve interventions on the input actually sent

For each of the four missions we passed candidate source cards to a deterministic local consumer. Its complete rule is to return the text of the most recent admitted card, with a stable identifier used to break date ties. Under the first condition, both candidates are available to the filter and only the early card is eligible. Under the second, both are eligible. Under the third we keep the later cutoff but remove the outcome card from the candidate list. These controlled changes make it possible to check dependence on the recorded input at a boundary we own.

All four missions produced the same three-way sequence: plan, outcome, plan. For DART, the early result mentions an intended impact; the later result reports an impact on the preceding day. Removing that outcome source while keeping the cutoff at 27 September returns the plan. The mechanism can be inspected directly. Twelve successful checks demonstrate that this local filter, serialisation and consumer are wired together as described. No probabilistic inference, adversarial reliability rate or attribution to hosted model reasoning follows from these twelve checks.

The result file contains 124 locally verified records, made up of 96 primary replays, 16 challenge replays and 12 interventions. The CSV and intervention file identify each record by line and digest. The February SPE engine includes an SQLite ledger for linked capsule hashes. The experiment’s portable record stream uses a distinct format: canonical local request, response text, per-record SHA-256 and a digest over the complete ordered JSONL bytes. A verifier with an independently retained whole-file digest can detect changes to content or order. Rehashing an edited file without such an external reference cannot authenticate who

produced it or when. No independent timestamp or provider signature was obtained for these local records.

What survives in the May application log

The archived application log is dated 10 May 2026. It identifies five configurations and five English questions. Each requested an answer as of 1 June 2023 by instruction, without a retrieved corpus. The record contains 22 response texts and three HTTP 429 errors. Four prompts or responses about public events illustrate the challenge: a question about a sentencing presupposed it, a denial repeated words from that question, a different response acknowledged later proceedings, and one description of an aircraft incident gave an earlier year and different circumstances. No service receipts independently authenticate model revisions or server-side context. These 25 archived attempts are separate from the 124 local captures.

The archived keyword logic signalled one of the 22 texts. It required two event-specific word hits **and** an event already tagged POST. At event level every PRE control was assigned a category counted as a true negative. The single flagged response said a defendant had not been sentenced by the cutoff. The question itself had asked for the final prison sentence. A second response asserted that trial and sentencing occurred after the cutoff without giving the penalty; neither that keyword rule nor the public year rule signalled it. Applying the public year rule to the 22 full texts produced zero alerts. Those counts describe the application record. They do not validate a sensitivity or precision estimate, nor reveal which model could legitimately have known what.

The extract publishes the observed keywords and explicit years, short quotations for four cases, editorial paraphrases and a digest of each complete returned response. This supports recounting from the disclosed features. Independent reconstruction of all features from full transcripts would require the source log. The bank example offers a useful practical check: the March 2023 agreement and the June 2023 legal completion of the UBS–Credit Suisse transaction are different milestones. A reviewer who labels the whole topic “before cutoff” will miss that ambiguity. FINMA announcement · UBS completion notice

Taken together, the tests support a narrow empirical finding. An auditor can independently reconstruct what a controlled local sender recorded, and some simple alerts can be decisively refuted from the wording of the answer and its prompt. The released phrase matcher requires extensive human review on new wording. Establishing the source of a hosted model's knowledge and estimating production performance remain open empirical questions.

Measuring a temporal claim

The unit in this experiment is a proposition about one named event. Consider the question “What has the mission achieved?” and a reply saying that the rover made it down safely. The reply commits the speaker to a completed landing even though it gives no year. A launch announcement that promises a landing on the correct future date makes a different commitment. This distinction matters to a person contesting an answer: pointing to a year establishes very little; locating the assertion, the relevant source and the cutoff gives the person something specific to challenge.

Eight decisions a reader can reproduce

The rows below are real entries from the released evaluation, with **constructed English outputs**. All eight preserve their assigned IDs, cutoffs and recorded predictions. In the last three columns, **P** means a completed target event asserted before it occurred, **O** means other and **R** means review. The short descriptions paraphrase the output. The data files preserve the exact text.

Case and cutoff	What the response does	Reference	Year	Dated
perseverance-04 · 2020-12-31	Expects the rover to land in 2021. The plan was already public.	O	P	O
perseverance-05 · 2020-12-31	Denies confirmation of a landing in 2021.	O	P	O
osiris-06 · 2022-12-31	Quotes a question about a delivery in 2023.	O	P	O
rosetta-08 · 2013-12-31	Describes an uncertain schedule, then asserts Rosetta woke up.	P	O	P
dart-09 · 2022-01-01	States DART hit Dimorphos in 2022. Impact was in September.	P	O	P
h03 · 2022-11-01	Imagines what scientists could do if a capsule had already arrived.	O	O	O
h10 · 2014-12-31	Recounts Rosetta's January 2014 wake-up after it happened.	O	O	R
h16 · 2021-02-10	Says the rover was built to land and "did so" in February.	P	O	R

A quoted future date can come from the question. A hedge in the first clause can coexist with a definite event in the next. A detector can miss an affirmation through pronoun reference, or misread a counterfactual as something observed. Factual inaccuracy is another axis. In the archived pilot, a response assigned a 2021 account to a different documented aviation incident. That calls for a source check even though the year rule produces no alert. A failed request, a refusal to answer and a detector's referral to a reviewer also need separate counts.

Which date is the evidence about?

Record the event date, the date of announcement or publication, the time a particular version was independently observed, and the cutoff supplied to the model. A law or contract can have a distinct effective date. A government can publish tomorrow's interest-rate change today; before it takes effect, the decision and its effective date can already be public. An evaluation must specify whether the proposition concerns the announcement or the new rate already being in force. The same issue arises in the two banking milestones in the archived pilot.

DART illustrates another distinction in the actual corpus. The impact occurred on 26 September 2022; the selected NASA source reported it on 27 September. If a query is evaluated in that gap, record "the selected source was not yet available." It would take a broader, contemporaneous search to establish whether the event was already public elsewhere. The study uses day precision and authored source cards. A display date on a mutable website cannot, by itself, authenticate the historical wording. Ideally retain the retrieved bytes or an authorised quotation, revision identifier, archive capture date, time zone and content digest. An archive timestamp demonstrates observation by that archive; earlier availability remains possible. NASA report · HTTP access to prior resource states · Temporal evaluation pitfalls

Proposed human adjudication

1. **Freeze the question and the proposition.** Write down the cutoff and timezone, copy the prompt and split the response into independently checkable claims. Preserve enough surrounding text to retain negation, quotation, reported speech and pronoun references. Mark information already disclosed by the question. Keep historical support and possible routes by which information reached the model as separate fields.
2. **Have two reviewers assess each case independently.** Show both reviewers the same evidence pack and hide method predictions, model identity and the other reviewer's labels. For every claim, record the stance (assertion of completion, forecast, denial, question/quotation, refusal or unclear), the dated evidence, and the relation of the asserted outcome to the cutoff. A scheduled outcome with a known date can support a forecast. An assertion that the scheduled outcome happened needs evidence of completion.
3. **Allow an unresolved reference label.** Published results can disagree on dates, archives can have gaps, and a statement can refer to several distinct actions. Record what is missing and which alternative readings would change the finding. Keep unresolvable cases visible in the sampling denominator. The public binary fixtures have authored reference labels; there has been no blinded double annotation of those labels.
4. **Measure disagreement before reconciling it.** Publish reviewer-by-reviewer counts, the disagreement matrix and raw agreement for stance and date separately. A chance-corrected measure can accompany them with its assumptions and class frequencies. A third reviewer can adjudicate remaining disagreements with a written reason. Retain the two original labels and the final disposition. Artstein and Poesio, inter-coder agreement

How to evaluate a live service

For a prospective study, define the service, population, languages, exposure period and question-selection rule before observing outcomes. Sample routine requests as well as prompts containing negation, explicit dates, dates already disclosed by the user and questions whose events never occur. Use a reference panel that is separate from the retrieval corpus. Freeze an initial snapshot and then record newly observed patches and corrections. Include repeated queries spanning an announcement, the actual outcome and a later correction. Group all versions of an event into one evaluation split to reduce leakage from nearly identical questions. Test unrelated topics and languages on fresh events. A carefully enriched challenge set can estimate sensitivity to chosen phenomena; a representative operational sample is needed to estimate how often users encounter those phenomena. ExAnte

Report the number of requests, retrieval failures, provider failures, refusals, usable answers, eligible claims, unresolved reference labels and detector referrals. For each audited system give sensitivity over **all resolved positives**, the rate of alarms among resolved negatives, and precision among issued alarms where defined. Report coverage (automatically decided cases divided by eligible cases), errors among those decisions, and the fraction of all reference positives sent for review. If a service has a human queue, time that work and report the final accuracy and delay of the combined process separately. For uncertainty, resample events as whole groups when the number of independent events permits; repeated paraphrases of four missions do not create dozens of independent events. Selective classification and coverage

The frozen challenge shows why the denominators matter. The dated matcher made only two automatic decisions among 16 cases, both on controls. It flagged **0 of 8** reference positives and sent all eight to review. Its review workload was **14 of 16**. No human queue was timed. Reporting only the two correct automatic decisions would conceal the actual work an auditor would face.

What a citizen can ask under the EU AI Act

A citizen challenging an AI-assisted decision may need to know which version of a rule the system used. The SPE Audit study records a cutoff date, the selected source summaries, the exact request, a scripted answer and the rule's verdict. This lets a reviewer inspect the evidence associated with one run. A digest can reveal that a copy differs from a known reference. Establishing when a source first became available, which rule was legally in force and what informed the official's decision requires separate evidence.

Who must be able to explain what?

AI Act article	Responsible actor	Evidence this study can contribute	What the organisation must establish separately
12 · Automatic logging	High-risk system provider	The record connects a run to its cutoff, inputs and output.	The system can automatically record relevant events throughout its lifetime, including events needed to detect risks and substantial modifications and to monitor operation after deployment.
13 · Instructions for use	High-risk system provider	Examples show errors on forecasts, paraphrases and dates in the same year.	The deployer receives clear instructions covering intended purpose, capabilities, limits, oversight and interpretation of results in actual use.
14 and 26(2), (5) · Human oversight	Provider designs for oversight. Deployer assigns people and monitors operation.	A reviewer can read the answer and the eligible sources side by side.	Designated staff have the competence, training, authority and support to intervene. Records of actions show how they handled a particular alert or incident.
19 and 26(6) · Log retention	Provider and deployer, each for logs under its control	The data format can preserve and recheck an individual record.	Each retains automatic logs for a period appropriate to the purpose, at least six months unless applicable EU or national law provides otherwise. Access and deletion need operational controls.
50(1), (2), (4) · Transparency	Providers of covered interfaces and generators. Deployers publishing specified content.	The archived output identifies what a reviewer can inspect.	Depending on the use and exceptions, people must be told they are interacting with AI, synthetic outputs require machine-readable marking, or publishers must disclose generated content. Delivery and accessibility of the applicable notice need evidence.
85 and 86 · Complaint and explanation	Relevant market surveillance authority receives complaints. Deployer provides a qualifying explanation.	A dated discrepancy gives the complainant a specific question to raise.	The authority assesses the alleged infringement under its procedures. The deployer needs the decision file to explain the role of AI and the main grounds for the decision.

The Act does prescribe specific minimum log fields for remote biometric identification under Annex III, point 1(a). Article 12(3) covers the period of use, the reference database, inputs producing a match and the identity of the human verifiers. The fields used for the present study are research choices. In a real service, personal data require a lawful basis, minimisation, appropriate security and retention limits. A public example can use synthetic records while authorised staff work on the individual case.

Which date applies?

- **2 August 2026.** Article 50 generally applies. Providers of covered synthetic-content systems already on the market before that date have until **2 December 2026** to fulfil the marking duty in Article 50(2), under Article 111(4).
- **2 December 2027.** Chapter III, Sections 1–3, except Article 6(5), apply to systems classified as high-risk under Article 6(2) and Annex III.
- **2 August 2028.** Those sections apply to high-risk product-related systems under Article 6(1) and Annex I. Older systems have specific transitional conditions under Article 111(2), including a 2030 deadline for the covered public-authority uses.

A hypothetical complaint about a public benefit

Suppose a municipality introduces a high-risk system in 2028 to assess eligibility for an essential public assistance benefit, a use listed in Annex III, point 5(a). It denies an application. The applicant sees that the reasons cite an amendment published after the eligibility date. They can ask the municipality which version applied, what the AI contributed and why the decision-maker relied on it. Staff can compare the capsule with the official publication and commencement dates, the application, the signed decision and the record of human review. A later publication date is grounds to investigate. The amendment's legal effect and the outcome for this person must be checked against the relevant law and case file.

Article 86 covers qualifying decisions based on outputs of Annex III high-risk systems, except point 2, with legal or similarly significant effects that the person considers adverse to health, safety or fundamental rights. Statutory exceptions or restrictions may apply, and the article operates to the extent the right is not already provided by EU law. Anyone with grounds to consider the Act infringed may complain to the relevant market surveillance authority under Article 85. Challenging the benefits decision itself follows the applicable administrative or judicial procedure.

Official sources, checked 21 November 2026. AI Act, consolidated 27 July 2026 · Amending Regulation (EU) 2026/1744 · Commission, Annex III · Right to complain · Right to an explanation · GDPR, Articles 5 and 6.

Where this study fits among existing reviews

Selection. We used targeted English-language searches up to 21 November 2026. Search families combined *survey* or *review* with temporal question answering and retrieval, RAG evaluation, fact-checking, hallucination, data contamination, digital provenance and forecast leakage. We retained sources that address a distinct measurement question in this study and whose topic and date we could check at a publisher, conference, public agency or author preprint. The table contains five surveys, one official technical overview and one methodological critique. Dates identify the cited edition. This is a purposive comparison. Database coverage, citation counts and the total size of each field were outside its scope.

Work and date	What it covers	How it informs the audit	Access
Guo, Schlichtkrull & Vlachos, *A Survey on Automated Fact-Checking* (TACL, 2022)	Claim selection, evidence retrieval and verification, with datasets and definitions across the workflow.	Annotate what the sentence actually asserts. A question that repeats an outcome, a denial and an assertion are distinct objects for a fact-checker.	Article and PDF

Work and date	What it covers	How it informs the audit	Access
Huang et al., *A Survey on Hallucination in Large Language Models* (author preprint, Nov 2023)	Distinguishes errors of factuality and faithfulness, and reviews causes, detection and evaluation.	An answer can describe a real event yet be inadmissible at the requested date. Its truth today and its historical support deserve distinct fields.	Open manuscript · Journal
Chandra et al., *Reducing Risks Posed by Synthetic Content* (NIST AI 100-4, Nov 2024)	An official overview of provenance tracking, authentication, content labels, watermarking and detection.	A retained digest lets a reviewer compare particular bytes. Source authenticity, the completeness of capture and the truth of a sentence require evidence of their own.	Report and PDF
Cheng, Chang & Wu, *A Survey on Data Contamination for Large Language Models* (author preprint, Feb 2025)	Training and evaluation data overlap, its prevention and detection under different levels of access to model internals.	A later fact in a response cannot by itself identify the training set as its source. The prompt, retrieved evidence and model parameters offer different possible routes.	Open manuscript
Gan et al., *Retrieval Augmented Generation Evaluation in the Era of Large Language Models* (author preprint, Apr 2025)	Reviews evaluation of retrieval, generation, overall systems, factual accuracy, safety and efficiency.	Evidence about source admission and evidence about output classification answer different questions. Publishing each observable separately makes an error traceable.	Open manuscript
Paleka et al., *Pitfalls in Evaluating Language Model Forecasters* (May 2025 preprint, ICLR 2026)	Methodological critique with examples of future information in historical evaluations, including revised web articles and retrieval effects.	A document labelled with an early publication date may contain later revisions. Historical claims require dated snapshots or explicit assumptions about versions.	Preprint · Proceedings
Piryani et al., *It's High Time: A Survey of Temporal Question Answering* (ACL, Jul 2026)	Organises tasks by corpus time, question time and model capabilities. Reviews temporal RAG, time-expression interpretation and evaluation.	The date in a question, the availability of a record and the time of the event can all differ. The boundary must specify which one controls admission.	Article and PDF

A direct experimental precedent. Allein, Augenstein & Moens tested how the dates of claims and evidence affect evidence ranking and fact verification. That result gives a reason to retain both dates. It does not establish that a particular provider consumed the selected evidence. Journal of Web Semantics, Nov 2021 · Open manuscript.

The contribution claimed here is a reproducible inspection of a named public implementation, with a documented decision for each case. The table supplies terminology, known failure modes and comparators. An across-field priority or completeness claim would require a different search design.

November evidence review: what was already known about temporal AI audits?

Literature and patent-document supplement · 21 November 2026 · English edition

A claim that *this answer existed*, a claim that *these records were selected*, and a claim that *the model used future information* need different evidence. Earlier systems already addressed important parts of all three questions. This review records the closest documents we could verify, identifies the passages that matter, and explains which comparisons bear on the experiments. The sections below use 4 December 2025 as a

historical comparison date for the original SPE technical description. It is neither the date of the software release nor a cut-off for relevant scientific work on this paper.

How the comparisons were made

The review followed citations and searched patent texts and scholarly publications for point-in-time retrieval, future-information leakage, context provenance, verifiable inference, semantic confidence, reference-database grounding and later work with earlier preprints or filing dates. A finding was mapped to the object on which it operates: training data, retrieval index, selected records, actual request payload, generated statements or an evaluator's reference corpus. Reusing the word *proof*, *confidence* or *time* in different settings is insufficient for an equivalence.

The numbered features below are paraphrases of historical design drafts. They are provided here so the comparison can be followed without access to private files. Numbers 1–13 refer to the original description; 14–27 refer to later drafting in May 2026. No grant or examiner's construction of SPE claims is implied. For earlier third-party patents we cite the publication containing the technical disclosure. When an issued claim is discussed, its grant is identified separately. The result is a directed comparison of the recovered evidence, without a field-wide completeness claim or a freedom-to-operate opinion.

The documents closest to the design

N01. Dantart: a temporally partitioned RAG system with a signed audit record

Version 1 of Alex Dantart's forensic architecture was submitted to arXiv on **12 October 2025**. Section 3.2 and figure 1 describe versioned ingestion, cryptographic document fingerprints, partitions reflecting legal validity, policy enforcement in an orchestrator, and a signed JSON trace in write-once storage. The trace can include the prompt, selected document hashes, temporal context, model response and policy evaluation. Sections 4.4, 5.1 and 5.3 supply the evaluation setting, the temporal-validity measure and an example of an audit artifact. A publication date, a law's effective date and the time at which an operator obtained a copy can differ. A replay system must declare which boundary it enforces.

This source most directly challenges the broad proposition that historical retrieval and a tamper-evident per-query record first came together in SPE. The reviewed passages do not establish the specific three-component semantic score, an independently committed evidence baseline, and their joint role in the output classification. The paper labels the evaluation synthetic. Its regulatory framing and performance estimates should not be transferred to another system. The December **23** revision is later than the comparison date; the features above can be found in the October text. October full text, §§3.2, 4.4, 5.1, 5.3 · Version history, DOI 10.48550/arXiv.2510.12830

N02. EQTY Lab: cryptographic evidence about the payload used in inference

Verifiable computing, US **2025/0254026 A1**, claims a 1 February 2024 priority, was filed in the US on 31 January 2025 and published on **7 August 2025**. The description's consecutive RAG examples discuss source hashes, authenticated embeddings and metadata, a root identifying a corpus version, and verification of whether a particular retrieved embedding was included in a model payload. The publication's claim 1 covers generation of a trusted-environment notary key, a statement representing inputs and outputs, its signature and export. The retrieval examples are in the description; their presence does not make every such example a separately granted claim.

The reference reaches further than proving membership of *training* data. A reviewer studying a claimed proof of exclusion for *inference* needs to read it. A proof about a particular item and input differs from replaying a date policy across an identified candidate set. Proving what the model provider actually received also depends on the attested boundary. The later **US 2026/0230327 A1**, filed 31 March and published 6

August 2026, extends the discussion to multi-agent computing and references the earlier family. Shared family metadata alone does not date every detail of the later document to 2024. US 2025/0254026 A1, description and claim 1 · International publication WO 2025/166192 A1 · Later US publication

N03. Liu et al.: testing whether an answer asserts a fact knowable only later

ExAnte v1, dated **26 May 2025**, defines pre-cutoff and post-cutoff reference sets. Section 2.1 evaluates single outcomes and atomic claims. Section 3.3 builds paired Wikipedia references from historical revisions. Section 3.4 tests publication titles against the earliest known public date. It demonstrates an operational comparison between what an answer states and what could be verified at a given date. Its later EACL 2026 appearance does not change the date of the preprint.

The method gives a closer precedent for an output-level temporal violation than generic hallucination research. It also exposes a measurement choice: a benchmark may label a correct future guess as leakage under its task definition. Distinguishing a forecast from a retrospective assertion requires an explicit annotation rule. Neither a match against a later event nor a paired reference-set comparison reveals the internal cause of the answer. The reviewed method does not cryptographically bind the exact model input to that classification. May preprint, §§2.1, 3.3–3.4 · EACL paper, DOI 10.18653/v1/2026.eacl-long.72

N04. Bitvore: dated external observations and hallucination scoring

US **2025/0086211 A1**, filed 12 September 2023 and published **13 March 2025**, describes generating metrics from a model answer and a reference database. In paragraph [0082], a company-relationship example uses extracted entities, salience, confidence, the number of supporting records and when support begins. The surrounding description discusses temporal correctness. Published claim 1 combines a query, model output, reference comparisons and a resulting score. This is evidence against treating the broad operation "compare a model with dated external signals and flag weak support" as a new idea. It leaves open how to verify that only eligible records were actually transmitted on a particular run.

The USPTO **granted US 12,783,564 B2 on 3 November 2026**. Its first claim contains further limits concerning real-time and structured data, the model ensemble, and iterative, selective prompt modification. The broader wording of the March 2025 application should not be represented as the scope of that granted claim. This correction matters because a cached bibliographic page still listed the application as pending. Application and claims · Original publication PDF, ¶0082 · USPTO Official Gazette, grant and first claim

N05. Qiu and Miikkulainen: semantic density as a confidence measure

The NeurIPS **2024** paper defines response-level confidence in semantic space. Section 3.3 describes estimating semantic density, and section 4 evaluates the estimate on question-answering tasks. It supports the proposition that density in semantic space was already being used to quantify uncertainty. Its statistic is not the later SPE proposal to combine cross-run coherence, perturbation stability and recognition density with specified weights. This distinction calls for an empirical comparison: keep the date restriction and evidence set fixed, then compare whether the extra components improve the detection of unsupported assertions. Similar names provide no such result. Proceedings, DOI 10.52202/079017-4274 · Paper, §§3.3 and 4

N06. Paleka et al.: temporal leakage inside apparently eligible documents

The preprint *Pitfalls in Evaluating Language Model Forecasters* appeared on **31 May 2025**. Section 2.2 describes older pages subsequently updated with the outcome and ranking biases that survive a date restriction. It discusses filtering returned articles for future content, relying on sources with stronger dating, and maintaining historical indexes. It also explains how removing updated pages can introduce selection bias by removing information that would genuinely have been available earlier. Section 2.3 addresses uncertainty about model cutoffs.

This gives both a technical problem and a documented reason for adding content inspection before a retrieved record is admitted. It is particularly relevant to later SPE drafts 14–15, which count suspicious time-sensitive referents and select a policy for documents lacking such referents. Paleka et al. do not specify those particular thresholds, fallback rules or capsule fields. A low-frequency historical reference may simply be obscure. Controls must therefore include rare but genuine terms, including terms whose planned use was announced in advance. Original preprint, §§2.2–2.3 · Date and identifiers

N07. Intuit: perturbation and model agreement as evidence about a response

Intuit filed its US application on 30 August 2023. It was published as **US 2025/0077777 A1 on 6 March 2025** and granted as **US 12,387,049 B2 on 12 August 2025**. Granted claim 1 asks a model to produce semantically equivalent prompts, compares responses from two models and computes a semantic consistency measure. Claim 20 gives a more detailed response-sampling arrangement. The prior search misattributed this document to IBM. The original assignee is Intuit.

This directly overlaps parts of coherence and perturbation stability in SPE. It leaves unspecified the time-bounded embeddings, three-way aggregation and versioned external reference used in the SPE draft. Agreement among different models may reflect common training material or common mistakes. A temporal benchmark therefore needs examples where several models could plausibly repeat the same dated error. Published application · Issued claims 1 and 20

N08. IBM: inclusion in a training dataset

IBM's **US 2025/0156390 A1**, filed 14 November 2023 and published **15 May 2025**, places a training dataset in a Merkle tree and generates a proof concerning an element and the root. Claim 1 explicitly involves training an LLM on the dataset. **US 12,608,358 B2** was granted **21 April 2026**. Training membership and inference-input membership are distinct technical objects. A model can receive a document for a particular query even if the training-data proof says nothing about it. Conversely, a record absent from the retrieved context can still appear in model parameters. IBM is material background for authenticated data structures. EQTY's inference example is closer to the specific question of an admitted context. Published claim 1 · USPTO Official Gazette, granted patent

N09. Vodafone: a publication after the comparison date with an earlier filing

US 2026/0017491 A1 lists a British priority of 9 July 2024, a US filing of 2 July 2025 and publication on **15 January 2026**. The first claim encodes a closed-domain RAG model's output vector with a variational autoencoder trained on labelled normal and hallucinating outputs. It uses the learned distribution to identify a candidate hallucination. This is an alternative interpretation of a semantic anomaly. It does not establish temporal exclusion of sources or verify a run-specific input digest. The publication is relevant to the scientific comparison and illustrates a legal date distinction. An eligible earlier-filed US patent document can, subject to the statutory conditions, be relevant from an effective filing date earlier than publication. Earlier support for each feature must be established. US publication and family · USPTO MPEP §2154

N10. Shukla and Joshi: per-claim receipts for answers

Proof-Carrying Answers ties individual answer claims to retrieval evidence, including authenticated vector-index information and cryptographic receipts. The ACSAC workshop page lists slides dated **8 December 2025**, four days after the comparison date. We found no verified pre-cutoff public version in this review. This is a relevant comparator for evidence granularity: a capsule for the entire response and a receipt for each assertion answer different verification questions. A passage supporting the words in a claim can itself be stale or use a different time convention from the one assumed by the reader. The publication therefore informs the reporting design even when it cannot be used here as an established earlier public disclosure. Workshop and dated presentation · IEEE proceedings, DOI 10.1109/ACSACW69556.2025.00050

Feature-by-feature comparison

SPE design feature	Closest verified teaching	What remains to be established
1(a), target date and evidence policy	N01 supplies an as-of date, legal validity metadata and an orchestrator policy.	Which timestamp convention governs and how its trust is established.
1(b), interception, exclusion and bounded retrieval	N01 partitions a legal corpus by date. N06 shows how later content survives superficially eligible dates.	The candidate set, actual payload, policy and any excluded or indeterminate elements require separately reproducible records.
1(c), a deterministic commitment to ordered eligible items	N01 signs a record with retrieved-document hashes. N02 describes authenticated retrieval and inclusion in a model payload. RFC 8785 provides a canonical JSON precedent.	A digest over an ordered manifest binds that manifest. Completeness against an index and delivery to a provider require evidence beyond the digest.
1(d), model execution and captured answer	N01 records the response; N02 describes signed statements on inputs and outputs.	A client-side capture needs a specified trust boundary before it can stand in for the remote model's effective input.
1(e), scalar combining coherence, perturbation stability and recognition	N05 estimates semantic density; N07 measures consistency over equivalent formulations and models.	No single cited passage establishes the exact three-component score within the full execution. Its incremental value requires an ablation with the same inputs.
1(f), comparison with a dated external baseline	N03 builds historical references. N04 evaluates support, timestamps and resulting scores.	Prove that the historical corpus was available by the cutoff, define normalization and missingness, and keep evaluator-only evidence out of the model input.
1(g), semantic integrity classification	N03 labels claims by historical verifiability. N04 and N07 assess hallucination or consistency using different evidence.	Specify how to treat plans, negations, hypothetical dates, quotations, false statements, missing sources and unresolved cases.
2, a capsule with identities, hashes, time and policy	N01 stores a signed per-query trace. N02 covers attested inputs and outputs.	State who creates each field and who can independently authenticate the timestamp, signer, snapshot and response linkage.
4(i), an asserted later event with an eligible context	N03 measures future-dependent assertions; N01 and N02 supply retrieval evidence at different levels.	An assertion about a later event and a mere later-year string are different observables. The model's internal cause is latent.
4(ii), high internal score with weak external support	N04 tests sparse support. N05 and N07 measure confidence and consistency.	Incomplete archives, sparse local-language evidence or missing indexes can generate the observed discrepancy.
9, time-specific embedding and a report of divergence	N03 addresses the temporal reference; N05 and N09 describe confidence or learned latent distributions.	The broader system wording leaves open how the measured difference tracks a verifiable real-world failure.
Later draft 14–15, scan retrieved content and handle referent-free records	N06 explicitly motivates content inspection after retrieval. RFC 3161 supplies a timestamping primitive.	The precise quarantine statistic and fallback were later details. An authenticated date does not establish semantic admissibility.
Later draft 26, replay policy against a named snapshot and recompute the context commitment	N01 supplies the audit record; N02 explicitly discusses inference-time inclusion or exclusion.	For an exact replay, identify immutable source versions, search parameters, candidate ordering, tie rules, policy and the bytes delivered. Draft 26 depends on the complete method of 1.

JSON canonicalization, RFC 8785 · Timestamp protocol, RFC 3161 · Merkle inclusion and consistency, RFC 9162

Interpreting novelty, dates and experimental value

One document versus a combination. We did not identify one pre-cutoff document in this reviewed set that discloses all 1(a–g) in the asserted arrangement. A concrete obviousness argument is nevertheless available. A system that retains historical sources and a signed answer (N01) already has a reason to test whether the model asserts later facts (N03). For a third party to check the input, N02 gives a specific inference-time provenance technique. N04, N05 and N07 give precedents for scoring external support and model consistency. A patent analysis would still need to show which teachings would be combined, why, and with what expectation of success. The unresolved advantage of the exact combination is an empirical question. No legal conclusion about infringement or the validity of any third-party claim follows. USPTO guidance on anticipation · Guidance on obviousness

Later details. The original description contemplates a bounded list, hashes, index snapshots and versioned parameters. Draft 26 adds an explicit third-party replay specification. Drafts 14–15 put semantic screening at retrieval and set a fallback for documents without extracted referents. Other later drafts describe checking execution boundaries, truncated inputs and model versions. Support must be traced feature by feature before assigning an earlier date; sharing a project name supplies no evidence of that support. US **2025/0378436 A1**, occasionally mapped to this topic in the earlier review, was published on **11 December 2025**, and concerns road-user charges in digital currency. Its canonical serialization of transaction messages is a distant technical analogy. It is not a verified pre-cutoff publication about RAG audits. USPTO prior-filing benefit guidance · Road-charging publication, description around transaction 114

Controls suggested by independent work. SelfCheckGPT evaluates disagreement among samples without an outside factual database. Farquhar et al. use semantic entropy to estimate uncertainty about model confabulations. Halawi et al. restrict retrieval to a simulated historical date in forecasting. El Lahib et al., first posted **31 January 2026**, demonstrate how search engines can return updated content despite a date filter. Together these works suggest comparison conditions that can actually fail: identical outputs with different contexts, semantically different outputs with identical year strings, a plausible forecast, a negation, a quotation, a claim first evidenced later, and an unresolved assertion. Keep retrieval snapshots and adjudication material separate and report abstentions. A provider's assertion about which context it used should be recorded as such unless corroborated independently. SelfCheckGPT, EMNLP 2023 · Semantic entropy, Nature 2024 · Halawi et al., §3.2 · El Lahib et al., ACL 2026

Conclusions

The easiest finding to reproduce is a failure. A vocabulary that works for phrases used during development loses all eight positives in a small challenge when another writer describes the same outcomes differently. The practical value of the matcher depends on who handles abstentions and which evidence that reviewer can inspect. A service that converted “review” to a green tick would hide precisely the cases that motivated the study.

One capability did survive testing. A third party can retain a digest, open the request capture, reconstruct the cards admitted by the filter, check the recorded output and identify a subsequent modification. In the intervention runs the reader can also observe a different answer from the local program when one card is withheld. Two people who disagree have something specific to inspect together. To decide who knew what, and when, requires historical sources, reliable attribution and additional observations of the environment.

The survey changes the way the contribution should be described. Dated retrieval, signed records, provenance, evaluation of future knowledge and measures of uncertainty have distinct technical histories. Publishing the studied implementations, their limits and an unfavourable result lets others challenge the design and test an alternative on the same material.

Materials, authorship and conditions

Research edition by **Rubén Abella · SPE Audit** The team publishing the study develops the technology under evaluation. Literature retrieval, writing, code and example editing were AI-assisted. A separate AI-assisted editor wrote the challenge, knowing the research question but without prior access to the specific decision rules or development examples. Outputs were regenerated and decisions recounted. This public first edition has not undergone peer review or independent deployment evaluation. Both editions retain source titles in the original language. The tested sentences are English. Interface translations are editorial.

The experiment uses original short summaries. Retaining the publication date reported by an agency makes the filter reproducible. A dispute over the content of a page at a particular time would also require an identified historical snapshot. The archival extracts permit a recount that preserves failures to return text. Output hashes can be compared by a party holding the original response.

Reproduction. Download the offline package, extract it and run `python3 verify.py` in that directory with Python 3.10 or later. On Windows with the Python launcher, use `py -3 verify.py`. Open `demo.html` to inspect cases in a browser. Verification requires no credentials, payment or external service. The reproduction report names the platform that was actually tested. The repository retains its code licence and terms. Corrections to dates and classifications are welcome. In a public issue, identify the M, N, B or case record and supply the specific supporting evidence. Public corrections

The 54 records in the May comparison

The original index comprised 19 patent rows and 35 rows for papers, software, protocols, specifications and other leads. Each row keeps its original position as M01–M54. Some rows refer to several patent publications or manuscript versions. Those related documents stay in the same row. The number 54 describes the reviewed index. The additional studies and bibliography have their own denominators.

The chronological reference is 4 December 2025. We recovered 48 records bearing an earlier documentary date and four bearing a later date. The two remaining entries have different evidentiary problems. ALETHEIA has dated commits but no independently established first public date (M37). A Rust module could not be uniquely identified (M38). Historical labels apply to the cited document or release. An arXiv date refers to v1 submission, a journal date to the linked article, and a patent date to the specified publication number. Present-day repository contents and related family members can have different dates. Software release dates provide dated examples of the technology. They do not establish its earliest appearance.

What an auditor can actually check

Question	Relevant records	How to interpret the evidence
Did the selected item belong to a committed collection?	M01, M14–M16, M48	Recompute an inclusion proof against the root for the identified collection. Which entity assembled that collection, and what it could have omitted, are separate questions.
Did a program perform a specified evaluation over bound inputs?	M07, M23–M29, M31	A proof or attestation binds a defined computation or measured environment. A verifier must know which inputs, outputs and checks were included. Dates entered as trusted inputs remain assumptions.
Were documents relevant to the chosen time?	M04–M05, M32–M36, M39–M43, M54	Temporal relevance, validity in a legal jurisdiction, and public availability at a given instant are different predicates. Historical evaluation should declare the predicate and the date used for each record.
Is a passage supported by the supplied evidence?	M06, M09–M10, M44–M47	Formal consistency, agreement across answers, attention to context and support in a retrieved passage quantify different properties. A cited or repeated forecast still needs to be distinguished from an assertion of the eventual outcome.
Can another party reconstruct the recorded interaction?	M02–M03, M08, M11–M15, M20–M22, M30–M31, M37, M48–M50, M54	Identify what each receipt covers. Source files, the final prompt after truncation, the model version and tool results can differ even if an earlier stage has a valid digest. A deterministic serializer and reproducible numerical execution address further distinct steps.

Closest architectural comparison. Dantart's October 2025 v1 already describes time-partitioned retrieval over legal sources, a policy step, and a signed write-once JSON record tying the prompt and response to temporal context and document hashes. Sections 3.2 and 5.3 are sufficient to establish this overlap without relying on the later v2. In legal retrieval a rule can come into force after its publication or be amended with retroactive effects. A study asking what was publicly knowable needs its own availability date, source version and decision rule. This distinction yields a concrete experimental question but no general priority claim. Dantart v1, §§3.2 and 5.3 · Version chronology.

Corrections that affect the comparison. The semantic consistency patent is attributable to Intuit (M09). The road-charge application (M17) is peripheral and its A1 publication falls after the cutoff, although its US filing predates it. CausalRAG, CDF-RAG, KG-IRAG and Zep have early 2025 manuscripts (M39, M40, M42 and M43). A repeated name in a list of later developments cannot move these sources to 2026. The corpus also distinguishes an original submission from a revision (M34), a symposium from the main conference (M35), an initial commit from public release (M37), and a framework from any evidence that it operated in production (M30). Patent entries summarize selected technical disclosures. Each row can support a focused comparison, not a finding about patent validity, claim coverage or freedom to operate.

Annotated source register

Each dated entry gives the date basis and the class relative to 4 December 2025. The related links identify additional publications or versions where the original row bundled documents. M38 links only to its explicitly marked candidate.

Patents indexed by publication

M01 · Merkle-enhanced training data audit process. US20250156390A1 · International Business Machines Corp. 2025-05-15 · Patent publication · Before cutoff.

A Merkle proof links a selected training item to a committed dataset. This is a comparator for inclusion during training. A historical retrieval request needs a separate commitment to the material supplied on that run. Primary record · US12608358B2 · 2026-04-21 · USPTO · 2026-04-21

M02 · Providing explainable machine learning model results using distributed ledgers. US20220198304A1 · International Business Machines Corp. 2022-06-23 · Patent publication · Before cutoff.

The ledger records evidence used to explain a machine-learning result. Reproducing the recorded provenance leaves open whether retrieved documents were eligible on a historical date. Primary record

M03 · Producing and verifying computational determinations using a distributed ledger. US11303454B2 · International Business Machines Corp. 2022-04-12 · Patent publication · Before cutoff.

Records inputs and model determinations so a party can check a computation against ledger entries. A matching record has no intrinsic test for an assertion about a later event. Primary record

M04 · Database record lineage and vector search. US20230061341A1 · Technion Research and Development Foundation Ltd. 2023-03-02 · Patent publication · Before cutoff.

Approximates database lineage using vectors. Claim 13 of the A1 explicitly permits excluding later records and records outside the target interval by creation time. This is a direct temporal filtering precedent. A creation timestamp and proof of what the model actually received require distinct evidence. Primary record · US12346297B2 · 2025-07-01

M05 · Retrieving documents according to temporal similarity. US20230033211A1 · International Business Machines Corp. 2023-02-02 · Patent publication · Before cutoff.

Models temporal relations among extracted entities to rank documents by temporal similarity. A relevance ranking can admit a highly similar item published after the date under study. Primary record · US12045270B2 · 2024-07-23

M06 · Large language model verification. US20250200392A1 · Amazon Technologies Inc. 2025-06-19 · Patent publication · Before cutoff.

Translates a generated answer to first-order logic and checks it against domain formulas. Verdict quality depends on translation and axioms. The study here checks dated support for historically framed claims. Primary record

M07 · Systems and methods for secure and fast machine learning inference in a trusted execution environment. US20220114249A1 · Huawei Technologies Co Ltd. 2022-04-14 · Patent publication · Before cutoff.

Provisions an encrypted model and protects its inference in a trusted execution environment. Remote attestation can strengthen evidence of which environment operated. Dating the underlying sources still requires an independent policy. Primary record

M08 · Tracking machine learning data provenance via a blockchain. US20250200334A1 · Coinbase Inc. 2025-06-19 · Patent publication · Before cutoff.

Records model-development data provenance on a blockchain. It preserves links to training inputs. A document manifest for a single RAG request has a different unit of observation. Primary record

M09 · Semantic aware hallucination detection for large language models. US12387049B2 · Intuit Inc. 2025-08-12 · Patent publication · Before cutoff.

Computes semantic consistency across equivalent prompts and responses from two language models. Two models can agree about a fact that had no public support at the cutoff. Intuit is the listed applicant. The A1 was already published in March 2025. Primary record · US20250077777A1 · 2025-03-06

M10 · Detecting hallucination in a language model. US20240419912A1 · Microsoft Technology Licensing LLC. 2024-12-19 · Patent publication · Before cutoff.

Detects generation that repeats material from examples or draws on outside knowledge. Lexical contamination is relevant when the prompt itself contains the term flagged by a detector. A copied word does not date a factual assertion. Primary record

M11 · Computer mechanism for analytics outcome transparency. US12099495B2 · International Business Machines Corp. 2024-09-24 · Patent publication · Before cutoff.

Binds execution of agreed analytics to recorded results and provenance. A stable audit trail can preserve an uninformative or mistaken factual verdict just as faithfully as a correct one. Primary record

M12 · Methods and apparatus to train a model using attestation data. US20230186156A1 · Intel Corp. 2023-06-15 · Patent publication · Before cutoff.

Weights edge training data using attestation and telemetry information. It concerns trust in training inputs. Historical output evaluation additionally requires a claim and a dated reference for that claim. Primary record

M13 · System and method for policy-constrained symbolic rendering of AI outputs. US12464027B2 · OneSource Solutions International Inc. 2025-11-04 · Patent publication · Before cutoff.

Authorizes symbolic rendering of inference outputs in a graphics subsystem. Claim 1 binds an evidence capsule to a symbol and records its justification. This is a close comparator for a signed policy decision. Measuring the full retrieval context involves another object at an earlier stage. Primary record

M14 · Blockchain ledger authentication. US20200195448A1 · Alibaba Group Holding Ltd. 2020-06-18 · Patent publication · Before cutoff.

Authenticates segments of a ledger using Merkle commitments and a time service. A time witness can date the commitment. It does not independently date the observation described inside each document. Primary record · US10778452B2 · 2020-09-15

M15 · Ledger verification method and apparatus, and device. US20200213094A1 · Alibaba Group Holding Ltd. 2020-07-02 · Patent publication · Before cutoff.

Uses a time-service certificate as an anchor for verifying a ledger segment. Verifying that segment supports chronology of recording. The semantics of a statement about the future require a separate decision. Primary record · US10790968B2 · 2020-09-29

M16 · Ledger verifiable-pruning system. EP4009579A2 · Bloom Technology Inc. 2022-06-08 · Patent publication · Before cutoff.

A skewed Merkle tree enables verifiable pruning. This is relevant to keeping audit evidence small. A pruned proof has to be interpreted relative to its original commitment and disclosure policy. Primary record · US11949801B2 · 2024-04-02

M17 · Methods and systems for facilitating collection of road user charges using a digital currency based on a distributed ledger technology. US20250378436A1. 2025-12-11 · Patent publication · After cutoff.

Road-user charging using digital currency and ledger transactions is the application. Canonical transaction serialization is a peripheral parallel. The A1 was published on 11 December 2025, although the US application was filed on 25 August 2025. It is a continuation-in-part. The earlier road-charge publications are linked. A literal search for canonical and Merkle in those two publications returned no match, which leaves the disclosure of equivalents unresolved. Primary record · US20230196340A1 · US20220398870A1

M18 · Circuit and method for timestamp filtering with RLS filter. US12395417B2 · Marvell Asia Pte Ltd. 2025-08-19 · Patent publication · Before cutoff.

A hardware recursive least-squares filter processes clock timestamps. Its filtering problem is signal estimation. A cutoff policy for documents tests availability at a date. The companion Marvell publication is a different patent. Primary record · US12267222B2 · 2025-04-01

M19 · A kind of event extraction method across media. CN106484767A · Institute of Information Engineering. 2017-03-08 · Patent publication · Before cutoff.

Extracts event elements from news and social-media material and merges them across channels. The approach informs how to link claims to event records. A detected event still needs an availability date and a test for factual assertion. Primary record · CN106484767B · 2019-06-21

Identity, asset provenance and verifiable computation

M20 · Key Event Receipt Infrastructure (KERI): Design and Build. KERI white paper v1.6 · Samuel M. Smith. 2019-07-03 · White paper date · Before cutoff.

KERI uses signed key-event logs and receipts to establish the history of identifier control. This supplies an identity and receipt precedent. The history of a signing key alone cannot establish the truth of a generated claim. Primary record

M21 · C2PA technical specifications, version 1.0. C2PA 1.0 · Coalition for Content Provenance and Authenticity. 2022-01-26 · Organization announcement · Before cutoff.

C2PA manifests bind assertions and provenance to digital assets. A signed provenance assertion can identify assets and edits. The assurance depends on the signers, assertions and trust model. Version 1.0 supplies the dated reference. Primary record · C2PA 1.0

M22 · Distributed Ledger for Provenance Tracking of Artificial Intelligence Assets. arXiv:2002.11000 · Philipp Lüthi, Thibault Gagnaux, Marcel Gygli. 2020-02-25 · arXiv first submission · Before cutoff.

Uses a distributed ledger to track provenance of AI assets such as datasets and models. This supports lineage through development. The unit of evidence in a temporal audit is a dated claim from one specific inference. Primary record

M23 · EZKL, release v12.0.0. EZKL v12.0.0 · EZKL / zkonduit. 2024-07-24 · Tagged software release · Before cutoff.

EZKL can generate and verify proofs for a committed computational graph. A proof can establish that the declared computation was performed. Factual eligibility of a source must enter the statement or its trusted inputs. Primary record

M24 · RISC Zero zkVM, release v1.0.0. RISC Zero zkVM v1.0.0 · RISC Zero. 2024-06-04 · Tagged software release · Before cutoff.

RISC Zero provides a general-purpose zero-knowledge virtual machine. A receipt is informative about the program and public outputs it binds. Temporal semantics depend on what that program checks and which data it receives. Primary record

M25 · zkLLM: Zero Knowledge Proofs for Large Language Models. arXiv:2404.16109 · Haochen Sun, Jason Li, Hongyang Zhang. 2024-04-24 · arXiv first submission · Before cutoff.

zkLLM develops zero-knowledge proofs of transformer inference. Verifying a computation over specific inputs has a different scope from adjudicating whether a natural-language output asserts an eventual historical outcome. Primary record

M26 · Scaling up Trustless DNN Inference with Zero-Knowledge Proofs. arXiv:2210.08674 · Daniel Kang, Tatsunori Hashimoto, Ion Stoica, Yi Sun. 2022-10-17 · arXiv first submission · Before cutoff.

Scales proofs for neural-network inference. It is evidence that correctness of an inference computation can be a proof target. Such a guarantee requires a specified computation and does not by itself certify a document date. Primary record

M27 · Verifiable evaluations of machine learning models using zkSNARKs. arXiv:2402.02675 · Tobin South, Alexander Camuto, Shrey Jain, Shayla Nguyen, Robert Mahari, Christian Paquin, Jason Morton, Alex 'Sandy' Pentland. 2024-02-05 · arXiv first submission · Before cutoff.

Uses zkSNARKs to verify model evaluations while supporting confidential material. This is relevant to third-party evaluation. A verifiable score can still reflect labels that conflate prediction and assertion. Primary record

M28 · Attestable Audits: Verifiable AI Safety Benchmarks Using Trusted Execution Environments. arXiv:2506.23706 · Christoph Schnabl, Daniel Hugenroth, Bill Marino, Alastair R. Beresford. 2025-06-30 · arXiv first submission · Before cutoff.

Attestable Audits uses trusted hardware for verifiable safety benchmarking. It is an execution-attestation comparator for an external auditor. Evidence of which evaluation ran still leaves the evaluation's temporal labels open to challenge. Primary record

M29 · Trustless Audits without Revealing Data or Models. arXiv:2404.04500 · Suppakit Waiwitlikhit, Ion Stoica, Yi Sun, Tatsunori Hashimoto, Daniel Kang. 2024-04-06 · arXiv first submission · Before cutoff.

zkAudit studies audits of privately held training data and models without disclosing them. This is directly relevant to reproducible assessments with restricted inputs. It does not supply a semantics of dated assertions. Primary record

M30 · Prove AI governance product announcement (Casper Labs). Prove AI announcement · Casper Labs. 2024-05-22 · Organization announcement · Before cutoff.

Casper Labs announced a model-governance and data-provenance product. The announcement supplies a public date and intended functionality. It cannot independently validate operating performance or per-request evidence completeness. Primary record

M31 · AICert v1.0: Open-Source AI Traceability Tool for Verifiable Training. AICert v1.0 · Raphaël Millet / Mithril Security. 2024-09-30 · Organization announcement · Before cutoff.

AICert v1.0 reports verifiable model training using confidential computing and signed evidence. It supplies a comparator for model provenance, while the content and scope of individual retrievals require separate observation. Primary record

Temporal retrieval and event graphs

M32 · It's About Time: Incorporating Temporality in Retrieval Augmented Language Models. arXiv:2401.13222 · Anoushka Gade, Jorjeta Jetcheva. 2024-01-24 · arXiv first submission · Before cutoff.

TempRALM retrieves temporally appropriate documents for a question by integrating temporal relevance. This is an appropriate retrieval baseline. It does not assign a forensic verdict to each dated factual claim in the output. Primary record

M33 · TimeR4: Time-aware Retrieval-Augmented Large Language Models for Temporal Knowledge Graph Question Answering. 10.18653/v1/2024.emnlp-main.394 · Xinying Qian, Ying Zhang, Yu Zhao, Baohang Zhou, Xuhui Sui, Li Zhang, Kehui Song. 2024-11 · Publication record · Before cutoff.

TimeR4 combines time-aware retrieval with language models for temporal knowledge graph question answering. Correctly answering a temporal question measures task performance. It does not establish which knowledge channel produced a correct answer. Primary record

M34 · Mitigating Temporal Misalignment by Discarding Outdated Facts. arXiv:2305.14824 · Michael J.Q. Zhang, Eunsol Choi. 2023-05-24 · arXiv first submission · Before cutoff.

Uses expected fact duration to discard volatile or outdated facts and improve calibration under temporal shift. Unlike a versioned snapshot, predicted volatility is an estimate of the useful life of a fact. Primary record · EMNLP 2023

M35 · Causal Event Graph-Guided Language-based Spatiotemporal Question Answering. 10.1609/aaaiss.v3i1.31204 · Kaushik Roy, Alessandro Oltramari, Yuxin Zi, Chathurangi Shyalika, Vignesh Narayanan, Amit P. Sheth. 2024-05-20 · Publication record · Before cutoff.

Models causal relationships in an event graph to answer spatiotemporal questions. A graph edge connecting two events supplies a reasoning relation. Whether either fact was available at a historical cutoff is a different property. Primary record

M36 · Two-stage Generative Question Answering on Temporal Knowledge Graph Using Large Language Models. arXiv:2402.16568 · Yifu Gao, Linbo Qiao, Zhigang Kan, Zhihua Wen, Yongquan He, Dongsheng Li. 2024-02-26 · arXiv first submission · Before cutoff.

GenTKGQA retrieves a relevant subgraph then generates an answer to a temporal question. This shows an earlier time-focused retrieve-and-generate design. The approach does not by itself issue a receipt attesting an exact input manifest. Primary record · Findings of ACL 2024

Agent logs, graph retrieval and temporal memory

M37 · ALETHEIA Protocol: a decision log for AI agents. Esrbwt1/Aletheia-Protocol · Esrbwt1. 2025-08-26 · Git commit metadata · Commit history only.

The public project describes a hash-chained decision log for AI agents. All three recorded commits are dated 26 August 2025. The current repository and commit timestamps establish code history, not an independent archive of its initial public visibility. Primary record · 2025-08-26 · 5bac41e0af7

M38 · Temporal GraphRAG hard filter (unresolved lead). Unresolved Rust module. Source unconfirmed · Unresolved identity and date.

A legacy lead named only a Rust temporal_retrieval module. The linked candidate documents interval fields and temporally qualified graph data. No exact version, publication date or match to the alleged mandatory hard filter could be established. graphrag-core 0.1.1 (?)

M39 · CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. arXiv:2503.19878 · Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, Vipin Chaudhary. 2025-03-25 · arXiv first submission · Before cutoff.

CausalRAG uses causal graph structure to improve evidence retrieval. Causal links in a knowledge graph do not attribute the origin of a model's particular sentence. The arXiv v1 and ACL Findings publication both preceded the cutoff. Primary record · Findings of ACL 2025

M40 · CDF-RAG: Causal Dynamic Feedback for Adaptive Retrieval-Augmented Generation. arXiv:2504.12560 · Elahe Khatibi, Ziyu Wang, Amir M. Rahmani. 2025-04-17 · arXiv first submission · Before cutoff.

CDF-RAG refines queries iteratively and checks responses against retrieved causal paths. This connects retrieval and output validation. The temporal admissibility and independently verifiable receipt of the supplied context require separate controls. Primary record

M41 · Respecting Temporal-Causal Consistency: Entity-Event Knowledge Graphs for Retrieval-Augmented Generation. arXiv:2506.05939 · Ze Yu Zhang, Zitao Li, Yaliang Li, Bolin Ding, Bryan Kian Hsiang Low. 2025-06-06 · arXiv first submission · Before cutoff.

E²RAG links entities, events and their temporal-causal relations to improve RAG consistency. ChronoQA evaluates time-sensitive questions. This is evidence about coherent reasoning, with a different target from authenticating the full input at a cutoff. Primary record

M42 · Beyond Single Pass, Looping Through Time: KG-IRAG with Iterative Knowledge Retrieval. arXiv:2503.14234 · Ruiyi Yang, Hao Xue, Imran Razzak, Hakim Hacid, Flora D. Salim. 2025-03-18 · arXiv first submission · Before cutoff.

KG-IRAG iterates retrieval over a temporal knowledge graph. Such a graph can improve temporal answers without furnishing an independently timestamped history of the records admitted for each output. Primary record

M43 · Zep: A Temporal Knowledge Graph Architecture for Agent Memory. arXiv:2501.13956 · Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, Daniel Chalef. 2025-01-20 · arXiv first submission · Before cutoff.

Zep maintains temporal relations for agent memory. The distinction between changes in the world and changes recorded in a memory system is relevant when choosing an evidence timestamp. A memory graph still depends on upstream timestamps. Primary record

Hallucination and grounding signals

M44 · Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. arXiv:2403.06448 · Weihang Su, Changyue Wang, Qingyao Ai, Yiran HU, Zhijing Wu, Yujia Zhou, Yiqun Liu. 2024-03-11 · arXiv first submission · Before cutoff.

MIND derives signals of hallucination from internal states without supervised labels. Such signals can complement an auditor's dated reference set on an accessible model. Their availability depends on access to the required internals. Primary record

M45 · Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. arXiv:2407.07071 · Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, James Glass. 2024-07-09 · arXiv first submission · Before cutoff.

Lookback Lens uses attention to context and previously generated tokens to detect contextual hallucinations. It gives an internal grounding signal. Reading attention weights does not independently validate the publication dates of retrieved evidence. Primary record

M46 · REFIND at SemEval-2025 Task 3: Retrieval-Augmented Factuality Hallucination Detection in Large Language Models. arXiv:2502.13622 · DongGeon Lee, Hwanjo Yu. 2025-02-19 · arXiv first submission · Before cutoff.

REFIND uses retrieved evidence and token-level context sensitivity to localize hallucinated spans. This is a close comparator for pinpointing unsupported text. Historical evaluation additionally distinguishes future assertions from permissible forecasts. Primary record

M47 · Prompt-Guided Internal States for Hallucination Detection of Large Language Models. 10.18653/v1/2025.acl-long.1058 · Fujie Zhang, Peiqi Yu, Biao Yi, Baolei Zhang, Tong Li, Zheli Liu. 2025-07 · Publication record · Before cutoff.

Prompts chosen to elicit informative internal states support a hallucination classifier. Its signal depends on access to model internals and calibration data. A historical receipt and a source chronology answer different questions. Primary record

Transparency logs, canonicalization and repeatability

M48 · Certificate Transparency. RFC 6962 · B. Laurie, A. Langley, E. Kasper. 2013-06 · RFC publication · Before cutoff.

Certificate Transparency specifies an append-only Merkle log and consistency and inclusion proofs. An auditor can use the design to understand what a root commits to. Log integrity does not establish the factual truth of the entries. Superseded by RFC 9162. Primary record · RFC 9162

M49 · JSON Canonicalization Scheme (JCS). RFC 8785 · A. Rundgren, B. Jordan, S. Erdtman. 2020-06 · RFC publication · Before cutoff.

JCS defines canonical JSON serialization so parties can independently reproduce bytes for signing and hashing. An implementation must implement the specified numeric and string rules before it can claim JCS conformance. Primary record

M50 · RepDL: Bit-level Reproducible Deep Learning Training and Inference. RepDL · Peichen Xie, Xian Zhang, Shuo Chen. 2025-10 · Publication record · Before cutoff.

RepDL targets bit-identical training and inference across supported hardware and software settings. Bitwise replay of numerical operations and independent verification of which historical evidence was available are complementary reproducibility questions. Primary record · arXiv:2510.09180 · 2025-10-10

Later developments and forensic legal RAG

M51 · AuditableLLM: A Hash-Chain-Backed, Compliance-Aware Auditable Framework for Large Language Models. 10.3390/electronics15010056 · Delong Li, Guangsheng Yu, Xu Wang, Bin Liang. 2025-12-23 · Publication record · After cutoff.

AuditableLLM separates model updates from their audit layer and links successive update records by hash. An update is a different unit of observation from a single question and response. The journal version of record appeared on 23 December 2025. An earlier public version was not established here. Primary record · Version history

M52 · Autonomous Action Runtime Management (AARM): A System Specification for Securing AI-Driven Actions at Runtime. arXiv:2602.09433 · Herman Errico. 2026-02-10 · arXiv first submission · After cutoff.

AARM specifies how a runtime can classify and authorize agent actions. The relevant unit is an action under policy. This later specification informs governance of an audit workflow without supplying the factual labels of the temporal test. Primary record

M53 · RAGShield: Detecting Numerical Claim Manipulation in Government RAG Systems. arXiv:2604.00387 · KrishnaSaiReddy Patil. 2026-04-01 · arXiv first submission · After cutoff.

RAGShield checks numerical claims against a cross-source registry and flags changes outside expected update schedules. This later example shows the value of matching evidence to a narrowly defined factual claim and domain. Its scope is numerical government records. Primary record

M54 · Gobernanza y trazabilidad "a prueba de AI Act" para casos de uso legales: un marco técnico-jurídico, métricas forenses y evidencias auditables. arXiv:2510.12830 · Alex Dantart. 2025-10-12 · arXiv first submission · Before cutoff.

Dantart's October v1 already filters legal material by validity date and signs a WORM record containing the prompt, temporal context, retrieved-document hashes, response and policy decision. This is the closest public comparison in the legacy catalogue. The December v2 is a later revision. Primary record · v1 §§3.2, 5.3 · v2 · 2025-12-23

Review coverage and provenance of references

What was counted

The consolidated May review had **54 numbered entries**. Two additional assessments, each retained as a text file and an office document, supplied **36 and 46 bibliography entries**. We compared the two copies of each list. After grouping versions and alternate links, their 82 citation positions represent **72 records**. Another **44 records** come from the text of the studies, the search and follow-up materials, the technical and legal notes, or primary sources identified for claims in those documents. The annex bibliography.csv therefore contains **116 records**. B052 and M39 identify the same work. Adding the totals would overcount distinct studies or patents.

The May review, the two source lists, their supplementary records and the subsequent search use different counting units. The 54 original positions appear in legacy.csv. The `historical_locations` column preserves the reference numbers of both additional reviews, including repeated citations. The annex distinguishes technical papers, patents, evolving project pages, enforcement actions, commentary and unresolved names by type and verification status. Counting comparable technical studies requires filtering by type and scope.

Counts by source document

The codes identify groups of working materials so readers can audit the totals without access to private documents. Where a file lacks a numbered bibliography, the table identifies the unit that could be counted.

Code	Source material	Units counted and treatment
S	Consolidated review of the searches, May 2026	All 54 numbered rows, M01–M54. The additional table in §3.4 has 11 rows and the one in §8 has 14. All 11 names in §3.4 reappear in §8, giving 14 distinct names across the two tables. The next table reconciles them.

Code	Source material	Units counted and treatment
F	Forensic AI audit, technical and commercial assessment	All 36 bibliography positions, F01–F36. The office document contains 36 external hyperlink targets, and the text version lists 36 entries. They represent 31 records. We also identified sources for terms and proceedings named in the text.
P	Assessment of AI audit patentability, operating environment and market	All 46 bibliography positions, P01–P46. The office document contains 46 external hyperlink targets, and the text preserves 46 entries. They represent 41 records. We retained fee schedules to document the original bibliography.
Q0	Search design	Four search areas and ten lines listing already known references. One line names two patents. The document supplied search queries and names to check. The references were counted under the corresponding works.
Q-C	Technical review across the search areas	Eighteen thematic subsections precede the summary table, including eight on patents. Nine URL occurrences have seven distinct destinations, some of them generic institutional pages. Identifiable patents, academic papers and standards appear in the M or B series. B095–B096 preserve the mentions of Giza and Aleo.
Q-D	First supplementary search report	Eight finding headings, four concerning patents. The remaining entries cover EZKL/RISC Zero/ZkAudit, Prove AI, TempRALM and ALETHEIA. M23–M24, M29–M30, M32 and M37 give identifiers or explain provenance issues.
Q-G	Second supplementary search report	Ten thematic headings. Names outside the 54 rows included Jolt, vCNN, MECM, KEA and a reference to Qiu. See B003, B087, B089, B090 and B094.
Q-P	Third supplementary search report	Six headings for individual findings, four concerning individual patents or patent families. The working copy names products and sources in prose and lacks a reproducible link inventory. M13–M16, M21–M22 and the corresponding families cover the headings.
Q-T	Cross-cutting convergence report	Two named groups of possible component overlap and two hypotheses about combinations. We retained the KEA and MECM references. Shared vocabulary and searches without a result cannot establish the scope of patent claims.
T	Technical follow-up, section B	Four subsections. B1 points to IBM M01, B2 to SEAL and B3 to NAL. B4 adds four papers about execution environments, one about context integrity, a commercial listing and five patent identifiers. The text referred to ten patents and named five. B082–B086 preserve those five without assuming the direction of a citation link or the identity of the other five.
D1–D4	Follow-up, technical specification, draft dependent claims and engineering addendum	Four documents including T. Both engineering texts cover the same four topics: timestamps, execution boundaries, truncation and model version. RFC 3161 is B097. The design details describe proposals from the project.
L1	US patent subject-matter eligibility discussion	Seven entries in the case table, plus the Alice decision in the text. B102–B109 identify all eight. They document the review's legal context.
L2	Technical continuity note	Six search areas and three names in the convergence discussion. B047, B050 and B052 correct the identity, date or scope of the cited works.

Code	Source material	Units counted and treatment
E	Temporal assessment methodology note	Five sections developing the method. The note has no formal bibliography. It includes an application example without independent validation.
Brand context	Collection of studies retained with F and P	The collection contained 33 files, including the four files for F and P. We screened the remaining 29 texts for recognizable academic and patent identifiers, DOIs, RFCs and research links. The screen returned no matches. It could have missed works named without an identifier. Business case studies were excluded from the technical prior-work count.

Reconciling the two supplementary tables in the consolidated review

Historical entry	Record(s)	Date and technical scope
AuditableLLM	M51	Post-cutoff publication about interaction traces.
Agentic Witnessing and Aegon	B020 and B005	Both are from April 2026. Witness identity, content access and agent logs answer different questions from the date on which a fact became available.
HugRAG	B050	Version 1 dates to February 2026. A later version of that record was retitled CausalRAG2. B052/M39 identifies a separate paper from March 2025.
Semantic Density Effect	B047	April 2026. It measures information per prompt token. The 2024 paper B003 studies uncertainty in model responses represented in semantic space.
ZKPROV and VeriLLM	B073 and B074	The v1 submission dates were 26 June and 29 September 2025, respectively. ZKPROV v2 came after the cutoff. Grouping both works under the historical table's heading obscured those dates.
MathLedger	B075	Version 1 was submitted on 22 December 2025. The arXiv identifier prefix 2601 does not supply the actual submission date.
GRAIL Protocol	B093	The label could not be matched unambiguously to a source. It provides no basis for a technical comparison.
AARM and RAGShield	M52 and M53	Both date to 2026. RAGShield treats the manipulation of numerical claims, a narrower question than the original summary described.
Zep, KG-IRAG and CDF-RAG	M43, M42 and M40	The three earliest versions date respectively to January, March and April 2025. Section 8 included all three despite their availability before 4 December.

What was verified

The CSV gives publication identifiers, primary pages or recorded URLs, verification status, the date of the identified edition and an interpretive limit for each entry. The arXiv titles generally follow the current record. The date field gives the first submission where the version history establishes it. Wording introduced in a later revision requires its own date. A patent date refers to the public issue or publication date for the specific document identifier. Family relationships and legally effective dates require other records. Blank date fields preserve uncertainty.

Of 116 records, 46 have a date before 4 December 2025, 12 have a later date and 58 lack adequate historical dating. These counts combine several types of material. They include NAL's publisher-labelled date and dates for judicial filings, administrative documents and technical publications. GRAIL and MECM are the two names without an unambiguous source. Directory entries and tool pages can change over time.

We checked 22 arXiv records against their version histories and verified the titles, numbers and public dates of eight patent publications. We also checked identifiers or source-specific catalogues for other publications, standards and decisions. For 39 contextual entries, the page was accessible but its May version remained undetermined. Fifteen preserve a recorded locator without independent confirmation of the content or version. Other verification statuses identify situations such as an identified judicial decision or an undated commercial page. The review did not include a full-text reading of every item.

Grouping used a DOI, preprint identifier, patent publication number or document URL. The same page with and without www, and the catalogue listing and PDF for one publication, can share a record. For P41 and P46, we grouped two filenames from the same issuing agency under their common fee-schedule label. We did not establish byte-for-byte equivalence or the historical revision. Additional original URLs and formats are retained when they help reconstruct the citation.

What the references establish about the method

- **Temporal retrieval.** TA-RAG and STAR-RAG (B002 and B001) concern questions about change over time and evidence selection using graphs. Retrieval or answer-accuracy metrics alone cannot establish which bytes a given request received. TV-RAG (B004) uses temporal organization and entropy weighting for video. Frame order and the public availability date of a document are different variables.
- **Uncertainty.** Qiu and Miikkulainen (B003), Farquhar and colleagues (B088), the geometric uncertainty paper (B015) and KEA Explain (B087) provide ways to study hallucination signals. Several responses can agree and still be wrong. A scalar alone cannot identify what caused an answer or when a fact became known. Youtu-VL (B049) also fails to substantiate the independent claim about JEPA with which it was cited.
- **Execution.** Jolt and vCNN concern verifiable computation. SecureInfer and TZ-LLM cover protected execution environments before the cutoff. AgenTEE, FlexServe and Protecting Context and Prompts are later comparators. SEAL links watermarking to model adaptation. An execution claim needs to specify which process consumes the committed bytes, which version is identified and who attests to the time. The engineering notes also mention physical memory integrity and target row refresh (TRR). These are assumptions about the environment and were not tested experimentally here.
- **Commitments and dates.** The eight patent publications in the annex illustrate components including logs, multiparty validation and Merkle membership checks. Membership in the committed set, presence in an ingestion history and public availability before a cutoff are three distinct predicates. Under its trust assumptions, RFC 3161 binds a digest to a timestamp issued by an authority. DER (B110) specifies encoding rules. HNSW (B111) provides context for approximate retrieval and explains how a search could miss eligible evidence.
- **Market and oversight.** F and P discuss compliance and audit management (Vanta, Secureframe, SafeBase, Bonsai Security, Finspectors, MindBridge and Caseware), platform security (AccuKnox, SentinelOne, Traceable and Levo), and brand visibility (Sight AI, Evertune, Indexly, Peec AI, Waikay, Semrush, Ahrefs, BrightEdge and Writesonic). The source material generally relies on commercial lists. We did not evaluate these products with the protocol. The search materials also mentioned Modulus Labs without identifying a technical release. B112–B115 distinguish SEC, DOJ and FTC proceedings. B116

identifies a copyright complaint. Allegations, settlements, guidance and binding rules carry different legal weight. The paper's European section cites the applicable legal text.

- **Open questions.** The review records identifiable gaps and ambiguous labels. It cannot establish that it covers all publicly available work, that nobody previously disclosed a combination, that implementing the method ensures AI Act conformity, or that the truth or causal origin of an answer was verified.

Reconstructing the review

Sort the F01–F36 and P01–P46 labels to reconstruct the two bibliographies. Their 72 underlying records are B001–B072. B073–B116 add technical comparators, contextual materials, ambiguous identifiers and primary sources for claims in the original reports. Only B052 explicitly matches the identity of a numbered M record (M39). Other records may examine the same mechanism. The counts apply to this edition of the annex and the dates were checked against 4 December 2025. They provide a bibliographic chronology. Legal classification requires a separate analysis.

The corrections to chronology affect how readers can interpret the original headings. ZKPROV and VeriLLM had preprints before the cutoff. MathLedger was first submitted after it. HugRAG and its retitled CausalRAG2 version share one record, while CausalRAG (M39) is a separate work. Version changes also affect the title and interpretation of Deterministic Legal Agents (B030). The collection covers temporal relevance, model uncertainty, computational integrity, dataset membership and independent public availability. Each record identifies which of these questions a source addresses and what further evidence the others would require.

Catalogue of the 116 reconciled and supplementary references

B001–B072 preserve the 82 original positions in the two reference lists. B073–B116 add references or trace their primary sources. Each note explains the record’s role. Titles are retained in their original language. Provenance includes repeat entries.

B001–B031

B001 · Right Answer at the Right Time - Temporal Retrieval-Augmented Generation via Graph Summarization. arXiv:2510.16715 · research · 2025-10-19.

Graph summarization for retrieval over time. The design objective is selecting relevant time-aware evidence; this is not an assessment of request-byte binding. Provenance F01, F06, Q0. Primary identifier and version history checked. Record Additional source 1

B002 · Reading Between the Timelines: RAG for Answering Diachronic Questions. arXiv:2507.22917 · research · 2025-07-21.

Temporal metadata and interval coverage address diachronic questions. Distinguish relevance to a date from excluding every item after a boundary. Provenance F02, Q0. Primary identifier and version history checked. Record Additional source 1

B003 · Semantic Density: Uncertainty Quantification for Large Language Models through Confidence Measurement in Semantic Space. arXiv:2405.13845 · research · 2024-05-22.

Response-level semantic density estimates uncertainty. It cannot identify the date or causal source of a fact merely by measuring confidence. Provenance F03, F16, Q-G/Q0/S6. Primary identifier and version history checked. Record Additional source 1

B004 · TV-RAG: A Temporal-aware and Semantic Entropy-Weighted Framework for Long Video Retrieval and Understanding. arXiv:2512.23483 · research · 2025-12-29.

Temporal structure and entropy weighting in long-video retrieval. Video position and documentary publication time are different variables. Provenance F04, F05, Q0. Primary identifier and version history checked. Record Additional source 1

B005 · Aegon: Auditable AI Content Access with Ledger-Bound Tokens and Hardware-Attested Mobile Receipts. arXiv:2604.06693 · research · 2026-04-08.

Hardware-attested mobile receipts and ledger-bound content access. Access evidence is adjacent to, and different from, the historical availability of a fact. Provenance F07, F24, S3.4/S8. Primary identifier and version history checked. Record Additional source 1

B006 · Antony Rowstron's research works | Microsoft and other places - ResearchGate. discovery · date unestablished.

Secondary index, educational resource or guide retained as a discovery trail. Its current page is not proof of first publication. Provenance F08. Recorded lead; contents/version not independently confirmed. Record

B007 · AI Washing | The New Frontier of Corporate Scrutiny. enforcement commentary · date unestablished.

Enforcement commentary characterizes the policy setting. Use regulator sources to establish an actual proceeding. Provenance F09. Page consulted; not historical verification. Record

B008 · One Year In, FTC's "Operation AI Comply" Continues Under New Administration, Signaling Enduring Enforcement Focus - Benesch Law. enforcement commentary · date unestablished.

Enforcement commentary characterizes the policy setting. Use regulator sources to establish an actual proceeding. Provenance F10. Page consulted; not historical verification. Record

B009 · 2025 Cybersecurity and AI Year in Review | Insights - Holland & Knight. enforcement commentary · date unestablished.

Enforcement commentary characterizes the policy setting. Use regulator sources to establish an actual proceeding. Provenance F11. Page consulted; not historical verification. Record

B010 · FINOS AI Governance Framework:. industry project · date unestablished.

Governance or provenance project. No dated, independently replicated temporal detector evaluated here. Provenance F12. Page consulted; not historical verification. Record

B011 · The Biteris Protocol: The Global Backbone of Trusted Artificial Intelligence - Medium. industry project · date unestablished.

Governance or provenance project. No dated, independently replicated temporal detector evaluated here. Provenance F13. Recorded lead; contents/version not independently confirmed. Record

B012 · Structure-Aware Temporal Graph RAG - Emergent Mind. discovery · date unestablished.

Secondary index, educational resource or guide retained as a discovery trail. Its current page is not proof of first publication. Provenance F14. Page consulted; not historical verification. Record

B013 · Uncertainty quantification in LLMs. discovery · date unestablished.

Secondary index, educational resource or guide retained as a discovery trail. Its current page is not proof of first publication. Provenance F15. Recorded lead; contents/version not independently confirmed. Record

B014 · Quantifying uncertainty in LLMs with Semantic Density | by Cognizant AI Lab | Medium. discovery · date unestablished.

Secondary index, educational resource or guide retained as a discovery trail. Its current page is not proof of first publication. Provenance F17. Recorded lead; contents/version not independently confirmed. Record

B015 · Geometric Uncertainty for Detecting and Correcting Hallucinations in LLMs. arXiv:2509.13813 · research · 2025-09-17.

Geometric uncertainty from output representations. Relevant to scoring but not retrospective dating. Provenance F18. Public primary record identified. Record Additional source 1

B016 · EdinburghNLP/awesome-hallucination-detection - GitHub. discovery · date unestablished.

Secondary index, educational resource or guide retained as a discovery trail. Its current page is not proof of first publication. Provenance F19. Page consulted; not historical verification. Record

B017 · UQLM: Uncertainty Quantification for Language Models, is a Python package for UQ-based LLM hallucination detection - GitHub. software · date unestablished.

Uncertainty-quantification software and baselines. No pinned historical revision was established here. Provenance F20. Page consulted; not historical verification. Record

B018 · Architectures, systems and methods for program defined state system. US10445643B2 · patent publication · 2019-10-15.

Programmable state-system architecture; the description is not specific evidence of cutoff-based RAG auditing. Provenance F21. Publication identifier, title and issue/publication date checked. Record

B019 · Architectures, systems and methods for program defined transaction system and decentralized cryptocurrency system. US10679128B2 · patent publication · 2020-06-09.

Programmable transaction and cryptocurrency system; ledger context rather than linguistic or temporal leakage detection. Provenance F22. Publication identifier, title and issue/publication date checked. Record

B020 · Agentic Witnessing: Pragmatic and Scalable TEE-Enabled Privacy-Preserving Auditing. arXiv:2604.24203 · research · 2026-04-27.

Witnessing and privacy-preserving attestation of agent activity. Supports a later comparison about trusted execution boundaries. Provenance F23, S3.4/S8. Primary identifier and version history checked. Record Additional source 1

B021 · AI Enforcement Accelerates as Federal Policy Stalls and States Step In - Morgan Lewis. enforcement commentary · date unestablished.

Enforcement commentary characterizes the policy setting. Use regulator sources to establish an actual proceeding. Provenance F25. Page consulted; not historical verification. Record

B022 · Litigation Minute: Hype Responsibly: Legally Promoting Your AI | HUB | K&L Gates. enforcement commentary · date unestablished.

Enforcement commentary characterizes the policy setting. Use regulator sources to establish an actual proceeding. Provenance F26. Recorded lead; contents/version not independently confirmed. Record

B023 · Interactive AI Whitepapers With Visual Diagrams | Veriprajna. discovery · date unestablished.

Secondary index, educational resource or guide retained as a discovery trail. Its current page is not proof of first publication. Provenance F27. Recorded lead; contents/version not independently confirmed. Record

B024 · FTC Files New "AI Washing" Case | Perkins Coie. enforcement commentary · date unestablished.

Enforcement commentary characterizes the policy setting. Use regulator sources to establish an actual proceeding. Provenance F28. Page consulted; not historical verification. Record

B025 · FTC's latest AI-washing case: A focus on agentic AI and productivity claims | DLA Piper. enforcement commentary · date unestablished.

Enforcement commentary characterizes the policy setting. Use regulator sources to establish an actual proceeding. Provenance F29. Recorded lead; contents/version not independently confirmed. Record

B026 · SEO vs GEO: Understanding the Key Differences in 2025 - Boomcycle Digital Marketing. market context · date unestablished.

Commercial publishing or search-visibility context; no implementation-level comparison with the audit method. Provenance F30. Page consulted; not historical verification. Record

B027 · SEO vs GEO: What Is Generative Engine Optimization and Why It Matters in 2025. market context · date unestablished.

Commercial publishing or search-visibility context; no implementation-level comparison with the audit method. Provenance F31. Recorded lead; contents/version not independently confirmed. Record

B028 · GEO vs SEO: What's the Difference and Why It Matters in 2026 - DOJO AI. market context · date unestablished.

Commercial publishing or search-visibility context; no implementation-level comparison with the audit method. Provenance F32. Page consulted; not historical verification. Record

B029 · RAGPerf: An End-to-End Benchmarking Framework for Retrieval-Augmented Generation Systems. arXiv:2603.10765 · research · 2026-03-11.

End-to-end benchmarking of RAG system performance. Useful evaluation context without implying detection of post-cutoff content. Provenance F33. Primary identifier and version history checked. Record Additional source 1

B030 · Deterministic Legal Agents: A Canonical Primitive API for Auditable Reasoning over Temporal Knowledge Graphs. arXiv:2510.06002 · research · 2025-10-07.

Temporal legal graph retrieval/action primitives; v1 was titled Deterministic Legal Retrieval. Wording and claims of later revisions need separate dates. Provenance F34, F35. Primary identifier and version history checked. Record Additional source 1 Additional source 2

B031 · Architectures, systems and methods for program defined entertainment state system, decentralized cryptocurrency system and system with segregated secure functions and public functions. US20180247191A1 · patent publication · 2018-08-30.

Entertainment-state and cryptocurrency architecture with separation of protected functions. Related patent literature; identifier-level comparison only. Provenance F36. Publication identifier, title and issue/publication date checked. Record

B032–B072

B032 · SPE AUDIT — Cryptographic Integrity Verification for Digital …. project description · date unestablished.

First-party or store presentation of related products; not independent validation or evidence of priority. Provenance P01. Recorded lead; contents/version not independently confirmed. Record

B033 · About SPE Trace | How the Geo-Semantic Engine Works. project description · date unestablished.

First-party or store presentation of related products; not independent validation or evidence of priority. Provenance P02, P03. Page consulted; not historical verification. Record Additional source 1

B034 · SPE TRACE — AI Brand Intelligence - Chrome Web Store. project description · date unestablished.

First-party or store presentation of related products; not independent validation or evidence of priority. Provenance P04. Recorded lead; contents/version not independently confirmed. Record

B035 · 5 Best AI Audit Software Tools for Accounting Professionals - Vero AI. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P05. Page consulted; not historical verification. Record

B036 · Top 10 Best Audit AI Software of 2026 - Gitnux. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P06. Page consulted; not historical verification. Record

B037 · Best AI Evidence Check Platforms for Financial Audits | Finspectors. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P07. Page consulted; not historical verification. Record

B038 · The Evolving Role of Forensic Accountants in an AI-Driven World - Leonid Group. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P08. Recorded lead; contents/version not independently confirmed. Record

B039 · AI in the Public Sector: How Automation Is Transforming Government Finance - BDO USA. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P09. Page consulted; not historical verification. Record

B040 · Top 6 AI Cyber Security Tools To Watch In 2026 - AccuKnox. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P10. Page consulted; not historical verification. Record

B041 · Top 10 Traceable AI Alternatives for API & App Security (2026) - Levo.ai. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P11. Page consulted; not historical verification. Record

B042 · The 10 Best AI Visibility Tools for 2026 - Evertune AI. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P12. Page consulted; not historical verification. Record

B043 · 9 Best Brand Monitoring Tools for AI Platforms in 2026. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P13. Page consulted; not historical verification. Record

B044 · 10 Best AI Brand Visibility Tools in (2026). market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P14. Page consulted; not historical verification. Record

B045 · Top 10 Best LLM Brand Visibility Tools to Dominate AI Search in 2026 - Indexly. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P15. Recorded lead; contents/version not independently confirmed. Record

B046 · The Guide to 9 Best AI Mode Tracking Tools in 2026 - SE Visible. market context · date unestablished.

Product comparison or sector commentary. Use for market framing only; no vendor performance endorsement. Provenance P16. Page consulted; not historical verification. Record

B047 · Semantic Density Effect (SDE): Maximizing Information Per Token Improves LLM Accuracy. arXiv:2604.17659 · research · 2026-04-19.

Information-per-token effects in prompts. This use of "density" is distinct from uncertainty of model responses or support by dated evidence. Provenance P17, P18, S3.4/S8, L2. Primary identifier and version history checked. Record Additional source 1 Additional source 2

B048 · Prompt-based justification and scoring with large language models for thought and language disorder assessment. 10.1007/s11042-026-21401-8 · research · 2026.

Clinical-language scoring is a distant application. No clinical data used in this study. Provenance P19. Public primary record identified. Record Additional source 1

B049 · Youtu-VL: Unleashing Visual Potential via Unified Vision-Language Supervision. arXiv:2601.19798 · research · 2026-01-27.

Vision-language supervision study. The citation neither substantiates the separate JEPA claim made alongside it nor validates temporal auditing. Provenance P20. Primary identifier and version history checked. Record Additional source 1

B050 · CausalRAG2: Hierarchical Causal Knowledge Graph Design for RAG. arXiv:2602.05143 · research · 2026-02-04.

Published as HugRAG in v1, later retitled CausalRAG2. Hierarchical causal graph retrieval; unrelated to the distinct CausalRAG preprint of March 2025. Provenance P21, P22, S3.4/S8, L2, Q0. Primary identifier and version history checked. Record Additional source 1 Additional source 2

B051 · Daily Papers - Hugging Face. discovery · date unestablished.

Search listing or profile; useful to locate works, not substantive historical evidence of the mechanism. Provenance P23. Page consulted; not historical verification. Record

B052 · CausalRAG: Integrating Causal Graphs into Retrieval-Augmented Generation. arXiv:2503.19878 · research · 2025-03-25.

Graph-based causal RAG; March 2025 v1 predates the comparison date. See also record M39. A causal graph in retrieval does not prove the provenance of an output. Provenance P24, L2. Primary identifier and version history checked. Record Additional source 1

B053 · Jing Ma - CatalyzeX. discovery · date unestablished.

Search listing or profile; useful to locate works, not substantive historical evidence of the mechanism. Provenance P25. Recorded lead; contents/version not independently confirmed. Record

B054 · Professor Ah-Lian Kor | Leeds Beckett University. discovery · date unestablished.

Search listing or profile; useful to locate works, not substantive historical evidence of the mechanism. Provenance P26. Page consulted; not historical verification. Record

B055 · Daily Papers - Hugging Face. discovery · date unestablished.

Search listing or profile; useful to locate works, not substantive historical evidence of the mechanism. Provenance P27. Recorded lead; contents/version not independently confirmed. Record

B056 · Key Issue 5: Transparency Obligations - EU AI Act. regulatory commentary · date unestablished.

Third-party interpretation. The EU article mapping in the paper uses the consolidated statutory text. Provenance P28. Page consulted; not historical verification. Record

B057 · High-level summary of the AI Act | EU Artificial Intelligence Act. regulatory commentary · date unestablished.

Third-party interpretation. The EU article mapping in the paper uses the consolidated statutory text. Provenance P29. Page consulted; not historical verification. Record

B058 · Article 12: Record-Keeping | EU Artificial Intelligence Act. regulatory commentary · date unestablished.

Third-party interpretation. The EU article mapping in the paper uses the consolidated statutory text. Provenance P30. Page consulted; not historical verification. Record

B059 · Iso 42001 Logging Lifecycle Traceability Vs Eu Ai Act - ISMS.online. regulatory commentary · date unestablished.

Third-party interpretation. The EU article mapping in the paper uses the consolidated statutory text. Provenance P31. Page consulted; not historical verification. Record

B060 · Operationalizing Compliance with the EU AI Act's High-Risk Requirements - Label Studio. regulatory commentary · date unestablished.

Third-party interpretation. The EU article mapping in the paper uses the consolidated statutory text. Provenance P32. Page consulted; not historical verification. Record

B061 · Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1 · government context · 2023-01-26.

Policy or risk-management material; source authorship and scope matter. A filed comment is not agency guidance. Provenance P33. Dated primary edition identified for the cited NIST programme. Record Additional source 1

B062 · Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1 · guidance · 2024-07.

Risk-management guidance; contextual, not a benchmark or a legal determination. Provenance P34, P35. Public primary record identified. Record Additional source 1 Additional source 2

B063 · NIST AI RMF Implementation Guide (2026) + EU AI Act Map | Knowlee. regulatory commentary · date unestablished.

Third-party interpretation. The EU article mapping in the paper uses the consolidated statutory text. Provenance P36. Page consulted; not historical verification. Record

B064 · Safeguard the Future of AI: The Core Functions of the NIST AI RMF - Optro. regulatory commentary · date unestablished.

Third-party interpretation. The EU article mapping in the paper uses the consolidated statutory text. Provenance P37. Page consulted; not historical verification. Record

B065 · Treasury Launches the Artificial Intelligence (AI) Innovation Series. government context · date unestablished.

Policy or risk-management material; source authorship and scope matter. A filed comment is not agency guidance. Provenance P38. Page consulted; not historical verification. Record

B066 · Re: Request for Comment on Innovative Methods to Detect Illicit Activity Involving Digital Assets (Docket No. TREAS-DO-2025-0009 - SEC.gov. government context · date unestablished.

Policy or risk-management material; source authorship and scope matter. A filed comment is not agency guidance. Provenance P39. Page consulted; not historical verification. Record

B067 · United States Patent and Trademark Office Fiscal Year 2027 Congressional Submission April 2026 - USPTO. administrative context · date unestablished.

Budget and fee information explains review context; it is not technical prior art and prices may change. Provenance P40. Page consulted; not historical verification. Record

B068 · USPTO Fee Schedule - Current. administrative context · date unestablished.

Budget and fee information explains review context; it is not technical prior art and prices may change. The P46 alternate filename is grouped by issuing body and document label. Byte equality and historical revision were not verified. Provenance P41, P46. Page consulted; not historical verification. Record Additional source 1

B069 · Table of Patent Fees - Current, Final Rule and Unit Cost - USPTO. administrative context · date unestablished.

Budget and fee information explains review context; it is not technical prior art and prices may change. Provenance P42. Page consulted; not historical verification. Record

B070 · Table of Patent Fees - USPTO. administrative context · date unestablished.

Budget and fee information explains review context; it is not technical prior art and prices may change. Provenance P43. Recorded lead; contents/version not independently confirmed. Record

B071 · USPTO fee schedule. administrative context · date unestablished.

Budget and fee information explains review context; it is not technical prior art and prices may change. Provenance P44. Page consulted; not historical verification. Record

B072 · USPTO Fee Changes to Take Effect in January 2025: What You Need to Know - Honigman. administrative context · date unestablished.

Budget and fee information explains review context; it is not technical prior art and prices may change. Provenance P45. Page consulted; not historical verification. Record

B073–B094

B073 · ZKPROV: A Zero-Knowledge Approach to Dataset Provenance for Large Language Models. arXiv:2506.20915 · research · 2025-06-26.

Proof of dataset provenance in the LLM setting. June 2025 v1 precedes the cutoff, despite appearing in a later-work list; December 2025 v2 is separate. Provenance S3.4/S8. Primary identifier and version history checked. Record

B074 · VeriLLM: A Lightweight Framework for Publicly Verifiable Decentralized Inference. arXiv:2509.24257 · research · 2025-09-29.

Publicly verifiable decentralized inference. Earliest preprint is September 2025; technical assertions introduced in subsequent revisions require separate dating. Provenance S3.4/S8. Primary identifier and version history checked. Record

B075 · MathLedger: A Verifiable Learning Substrate with Ledger-Attested Feedback. arXiv:2601.00816 · research · 2025-12-22.

Ledger-attested feedback and learning. Submitted 22 December 2025 despite the January arXiv number. Later comparator for ledger-based learning. Provenance S3.4/S8. Primary identifier and version history checked. Record

B076 · SecureInfer: Heterogeneous TEE-GPU Architecture for Privacy-Critical Tensors for Large Language Model Deployment. arXiv:2510.19979 · research · 2025-10-22.

Separate trusted processing of privacy-critical LLM tensors from GPU acceleration. Relevant to the trust boundary; a confidential execution environment alone establishes no temporal claim. Provenance T-B. Primary identifier and version history checked. Record

B077 · TZ-LLM: Protecting On-Device Large Language Models with Arm TrustZone. arXiv:2511.13717 · research · 2025-11-17.

TrustZone protection of on-device LLMs. Relevant to execution integrity and confidentiality rather than dated event verification. Provenance T-B. Primary identifier and version history checked. Record

B078 · AgenTEE: Confidential LLM Agent Execution on Edge Devices. arXiv:2604.18231 · research · 2026-04-20.

Confidential agent execution on edge hardware. Later systems comparator for isolated execution. Provenance T-B. Primary identifier and version history checked. Record

B079 · FlexServe: A Fast and Secure LLM Serving System for Mobile Devices with Flexible Resource Isolation. arXiv:2603.09046 · research · 2026-03-10.

Secure mobile LLM serving with flexible resource isolation. Later systems comparator, unavailable at the December benchmark date. Provenance T-B. Primary identifier and version history checked. Record

B080 · Protecting Context and Prompts: Deterministic Security for Non-Deterministic AI. arXiv:2602.10481 · research · 2026-02-11.

Authenticated context/prompt integrity. A later comparator for input integrity; claims about the historical truth of input text require independent evidence. Provenance T-B. Primary identifier and version history checked. Record

B081 · SEAL: Entangled White-box Watermarks on Low-Rank Adaptation. arXiv:2501.09284 · research · 2025-01-16.

Watermarks in LoRA adaptation support model ownership claims. This is distinct from committing to a model version at runtime. Provenance T-B. Primary identifier and version history checked. Record

B082 · Cloud storage using Merkle trees. US9705730B1 · patent publication · 2017-07-11.

Merkle-tree cloud storage and integrity. General commitments do not establish that a fact was unknown before a date. Provenance T-B4. Publication identifier, title and issue/publication date checked. Record

B083 · Method and system for verifying information of a data item in a plurality of different data items . US20180026791A1 · patent publication · 2018-01-25.

Verifying selected information associated with a data item; comparator for authenticated sets. Provenance T-B4. Publication identifier, title and issue/publication date checked. Record

B084 · Platform for multi-party digital records using distributed ledger system. US20190253240A1 · patent publication · 2019-08-15.

Distributed-ledger platform for multiparty digital records. No claim-level infringement assessment. Provenance T-B4. Publication identifier, title and issue/publication date checked. Record

B085 · Methods and devices for validating data in a blockchain network. US20230237477A1 · patent publication · 2023-07-27.

Validating blockchain data. Relevant to verification infrastructure, not a temporal text benchmark. Provenance T-B4. Publication identifier, title and issue/publication date checked. Record

B086 · Merkle proof entity. US20230394063A1 · patent publication · 2023-12-07.

Merkle proofs and associated entities. A proof of membership has narrower semantics than a proof about a model's knowledge. Provenance T-B4. Publication identifier, title and issue/publication date checked. Record

B087 · KEA Explain: Explanations of Hallucinations using Graph Kernel Analysis. arXiv:2507.03847 · research · 2025-07-05.

Graph kernels for explaining hallucinations. Evidence-bearing semantic comparisons are relevant, but its score is not a timestamp attestation. Provenance Q-G, Q-T. Primary identifier and version history checked. Record Additional source 1

B088 · Detecting hallucinations in large language models using semantic entropy. 10.1038/s41586-024-07421-0 · research · 2024-06-19.

Entropy over semantic clusters can detect some confabulations. A confidence signal alone cannot date a factual disclosure. Provenance F-body, Q0. Public primary record identified. Record

B089 · Jolt: SNARKs for Virtual Machines via Lookups. IACR:2023/1217 · research · 2023.

General verifiable computation. A proof over an input does not establish the historical status of statements within it. Provenance Q-G. Public primary record identified. Record

B090 · vCNN: Verifiable Convolutional Neural Network Based on zk-SNARKs. IACR:2020/584 · research · 2020.

Verifiable neural-network evaluation. Does not document the date of information available to a model. Provenance Q-G. Public primary record identified. Record

B091 · Neural Authenticity Ledger: Cryptographic Model State Fingerprinting for AI Content Provenance. NAL · self-published proposal · 2025-09-20.

A self-published proposal connects model fingerprints to content provenance. The publisher labels its date; independent deposit, legal effect and implementation were not established. Provenance T-B3. Publisher page and publisher-labelled date. Record

B092 · V-Proof Protocol. V-Proof · industry project · date unestablished.

Public company directory identifies a product. Its entry does not establish an enabling disclosure or first-public date. Provenance T-B4. Directory entry identified. Record

B093 · GRAIL Protocol (historical label). GRAIL · unresolved label · date unestablished.

Multiple unrelated results share this name. No unique match to the description of verifiable inference was established. Provenance S3.4/S8. Unresolved; excluded from technical comparison.

B094 · MECM (temporal causal gating, historical label). MECM · unresolved label · date unestablished.

The note supplies a short acronym without unique identifier, author or title. Excluded from technical conclusions. Provenance Q-G, Q-T. Unresolved; excluded from technical comparison.

B095–B109

B095 · Giza. Giza · industry project · date unestablished.

First-party project locator in the verifiable-ML ecosystem. A moving product page cannot identify which functions existed at the cutoff. Provenance Q-C. Project locator; no dated release established. Record

B096 · Aleo. Aleo · industry project · date unestablished.

First-party project locator for verifiable execution infrastructure. Historical version not pinned. Provenance Q-C. Project locator; no dated release established. Record

B097 · Internet X.509 Public Key Infrastructure Time-Stamp Protocol (TSP). RFC 3161 · standard · 2001-08.

A time-stamp authority can bind a digest to a stated time under its trust assumptions. This does not establish the date of an event described by the bytes. Provenance D2/D4. Public primary record identified. Record

B098 · Information technology — Artificial intelligence — Management system. ISO/IEC 42001:2023 · standard · 2023.

Management-system standard mentioned alongside logging duties; the citation conveys no certification. Provenance P-body. ISO catalogue record identified. Record

B099 · Artificial Intelligence Act. Regulation (EU) 2024/1689 · legislation · 2024-07-12.

Regulatory framework cited in both reviews. Consult the applicable consolidated text and article-specific application timetable. Technical evidence alone cannot establish conformity. Provenance F/P-body. Public primary record identified. Record Additional source 1

B100 · SEC charges two investment advisers with making false and misleading statements about their use of artificial intelligence. SEC-2024-36 · enforcement · 2024-03-18.

Advertising and disclosure enforcement context, cited to identify a proceeding. The orders do not validate this method. Provenance F-body. Public primary record identified. Record

B101 · FTC Announces Crackdown on Deceptive AI Claims and Schemes. FTC-2024-09-AI · enforcement · 2024-09-25.

Official announcement of multiple proceedings, replacing secondary evidence for the fact of the announcement. Provenance F-body. Public primary record identified. Record

B102 · Alice Corp. v. CLS Bank International. USSC-13-298 · judicial opinion · 2014.

Judicial context from a US subject-matter eligibility discussion. No technical disclosure or case-specific patentability conclusion is attributed to this authority. Provenance L1. Decision identified. Record

B103 · Enfish, LLC v. Microsoft Corp.. CAFC-15-1244 · judicial opinion · 2016.

Judicial context from a US subject-matter eligibility discussion. No technical disclosure or case-specific patentability conclusion is attributed to this authority. Provenance L1. Decision identified. Record

B104 · McRO, Inc. v. Bandai Namco Games America Inc.. CAFC-15-1080 · judicial opinion · 2016.

Judicial context from a US subject-matter eligibility discussion. No technical disclosure or case-specific patentability conclusion is attributed to this authority. Provenance L1. Decision identified. Record

B105 · BASCOM Global Internet Services, Inc. v. AT&T Mobility LLC. CAFC-15-1763 · judicial opinion · 2016.

Judicial context from a US subject-matter eligibility discussion. No technical disclosure or case-specific patentability conclusion is attributed to this authority. Provenance L1. Decision identified. Record

B106 · Berkheimer v. HP Inc.. CAFC-17-1437 · judicial opinion · 2018.

Judicial context from a US subject-matter eligibility discussion. No technical disclosure or case-specific patentability conclusion is attributed to this authority. Provenance L1. Decision identified. Record

B107 · Diamond v. Diehr. US-450-175 · judicial opinion · 1981.

Judicial context from a US subject-matter eligibility discussion. No technical disclosure or case-specific patentability conclusion is attributed to this authority. Provenance L1. Decision identified. Record

B108 · Finjan, Inc. v. Blue Coat Systems, Inc.. CAFC-16-2520 · judicial opinion · 2018.

Judicial context from a US subject-matter eligibility discussion. No technical disclosure or case-specific patentability conclusion is attributed to this authority. Provenance L1. Decision identified. Record

B109 · Core Wireless Licensing S.A.R.L. v. LG Electronics, Inc.. CAFC-16-2684 · judicial opinion · 2018.

Judicial context from a US subject-matter eligibility discussion. No technical disclosure or case-specific patentability conclusion is attributed to this authority. Provenance L1. Decision identified. Record

B110–B116

B110 · ASN.1 encoding rules, including Distinguished Encoding Rules (DER). X.690 · standard · 2021-02.

Authoritative locator supplied for a mechanism named in the historical discussion. Encoding rules can standardize bytes; they do not identify all evidence available on a date. Provenance F-body. ITU catalogue, February 2021 edition. Record

B111 · Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. arXiv:1603.09320 · research · 2016-03-30.

Primary paper supplied for HNSW, a method named in the historical discussion. Approximate nearest-neighbor search is relevant to how missed eligible items can affect an audit. Provenance F-body. Primary identifier checked. Record

B112 · SEC proceeding concerning Presto Automation. SEC-33-11352 · enforcement · 2025-01-14.

The SEC announced settled charges concerning representations about Presto Voice. The earlier narrative misattributed this proceeding to the FTC. Provenance F-body. Public primary record identified. Record

B113 · SEC civil proceedings concerning representations about Nate. SEC-LR-26282 · enforcement · 2025-04-11.

The SEC reports filing its civil complaint on 9 April 2025. The date here is the release date. Provenance F-body. Public primary record identified. Record

B114 · Department of Justice announcement concerning Nate. SDNY-nate-2025-04-09 · enforcement · 2025-04-09.

This DOJ announcement concerns an indictment, separate from the SEC civil complaint. A charging announcement records allegations. Provenance F-body. Public primary record identified. Record

B115 · FTC proceedings concerning Air AI. FTC-AirAI · enforcement · 2025-08-25.

Date refers to the complaint. The agency subject matter includes marketing, earnings and refunds; the same case page identifies a later settlement. Provenance F-body. Public primary record identified. Record

B116 · The New York Times Company v. Microsoft Corp. et al., complaint. SDNY-1:23-cv-11195 · court pleading · 2023-12-27.

Court-filed complaint available through a public mirror. Cited only to identify a copyright dispute; allegations are not findings. Provenance F-body. Dated complaint identified. Record