AutoDesign: Meta-Harness Optimization for Long-Horizon Agentic Design

Yaxin Luo*, Haobin Jiang*, Jialv Zou, Xu Huang, Wenhao Yan, Haodong Li, Zhengrong Yue, Jing Li, Xiaofu Chen, Xiaohan Zhao, Jiacheng Liu, Jiacheng Cui, Zhiqiang Shen†, Xiaotong Li†

Meituan · MBZUAI · HUST · Peking University · Tsinghua University · CUHK · SJTU

Project autodesign.designanything.ai Code github.com/Yaxin9Luo/AutoDesign Demo designanything.ai

Overview

Outline

Problem

Multimodal design is a long-horizon agentic process

Multimodal sources

Papers, figures, tables, code, and human guidance

→

Agentic process

Extract, plan, author, render, critique, and revise

→

Human-facing artifact

A poster, slide deck, webpage, or conference video

- The source is long and heterogeneous; evidence must remain traceable.
- The output is rendered, so layout and readability are part of correctness.
- Failures appear over many actions, not in a single model response.

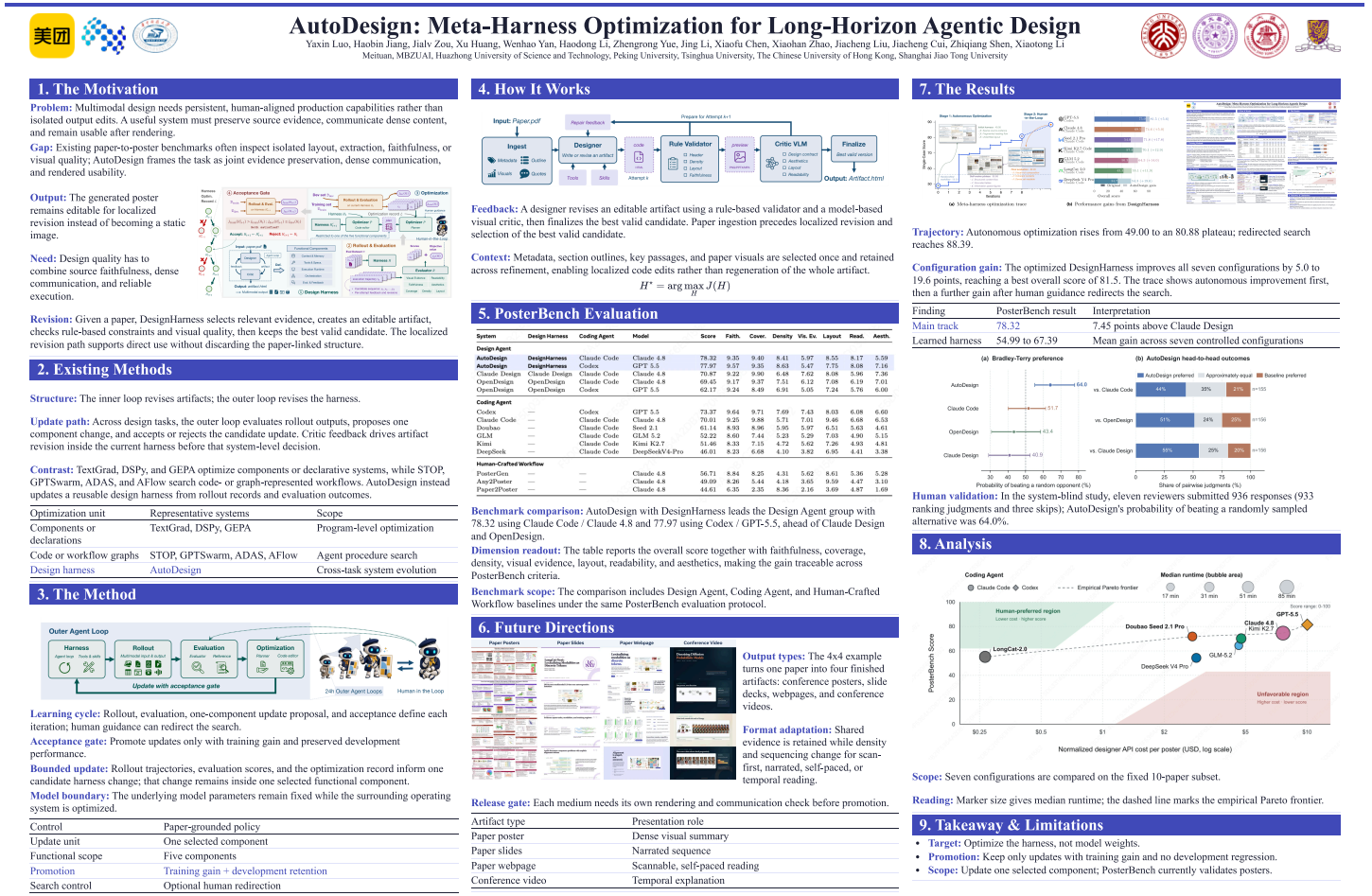
Problem

Paper-to-poster couples evidence, rendering, and revision

Paper-to-poster is a compact stress test for the full system.

- **Source-grounded:** claims, numbers, and visuals must stay tied to the paper.
- **Render-aware:** typography, density, and hierarchy must work at poster scale.
- **Locally revisable:** a detected failure should not force full regeneration.

Figure 2 was generated from the AutoDesign paper itself in approximately 40 minutes with negligible human intervention.



AutoDesign: Meta-Harness Optimization for Long-Horizon Agentic Design

Yaxin Luo, Haobin Jiang, Jialv Zou, Xu Huang, Wenhao Yan, Haodong Li, Zhengrong Yue, Jing Li, Xiaofu Chen, Xiaohan Zhao, Jiacheng Liu, Jiacheng Cui, Zhiqiang Shen, Xiaotong Li

Meituan, MBZUAI, Huazhong University of Science and Technology, Peking University, Tsinghua University, The Chinese University of Hong Kong, Shanghai Jiao Tong University

1. The Motivation

Problem: Multimodal design needs persistent, human-aligned production capabilities rather than isolated output edits. A useful system must preserve source evidence, communicate dense content, and remain usable after rendering.

Gap: Existing paper-to-poster benchmarks often inspect isolated layout, extraction, faithfulness, or visual quality; AutoDesign frames the task as joint evidence preservation, dense communication, and rendered usability.

Output: The generated poster remains editable for localized revision instead of becoming a static image.

Need: Design quality has to combine source faithfulness, dense communication, and reliable execution.

Revision: Given a paper, DesignHarness selects relevant evidence, creates an editable artifact, checks rule-based constraints and visual quality, then keeps the best valid candidate. The localized revision path supports direct use without discarding the paper-linked structure.

2. Existing Methods

Structure: The inner loop revises artifacts; the outer loop revises the harness.

Update path: Across design tasks, the outer loop evaluates rollout outputs, proposes one component change, and accepts or rejects the candidate update. Critic feedback drives artifact revision inside the current harness before that system-level decision.

Contrast: TextGrad, DSPy, and GEPA optimize components or declarative systems, while STOP, GPTSwarm, ADAS, and AFlow search code- or graph-represented workflows. AutoDesign instead updates a reusable design harness from rollout records and evaluation outcomes.

Optimization unit	Representative systems	Scope
Components or declarations	TextGrad, DSPy, GEPA	Program-level optimization
Code or workflow graphs	STOP, GPTSwarm, ADAS, AFlow	Agent procedure search
Design harness	AutoDesign	Cross-task system evolution

3. The Method

Learning cycle: Rollout, evaluation, one-component update proposal, and acceptance define each iteration; human guidance can redirect the search.

Acceptance gate: Promote updates only with training gain and preserved development performance.

Bounded update: Rollout trajectories, evaluation scores, and the optimization record inform one candidate harness change; that change remains inside one selected functional component.

Model boundary: The underlying model parameters remain fixed while the surrounding operating system is optimized.

Control	Paper-grounded policy
Update unit	One selected component
Functional scope	Five components
Promotion	Training gain + development retention
Search control	Optional human redirection

4. How It Works

Feedback: A designer revises an editable artifact using a rule-based validator and a model-based visual critic, then finalizes the best valid candidate. Paper ingestion precedes localized revision and selection of the best valid candidate.

Context: Metadata, section outlines, key passages, and paper visuals are selected once and retained across refinement, enabling localized code edits rather than regeneration of the whole artifact.

$$H^* = \arg\max_H J(H)$$

5. PosterBench Evaluation

System	Design Harness	Coding Agent	Model	Score	Faith.	Cover.	Density	Vis. Ev.	Layout	Read.	Aesth.
Design Agent											
AutoDesign	**DesignHarness**	Claude Code	Claude 4.8	78.32	9.35	9.40	8.41	5.97	8.55	8.17	5.59
AutoDesign	**DesignHarness**	Codex	GPT 5.5	77.97	9.57	9.35	8.63	5.47	7.75	8.08	7.16
Claude Design	Claude Design	Claude Code	Claude 4.8	70.87	9.22	9.90	6.48	7.62	8.08	5.96	7.36
OpenDesign	OpenDesign	Claude Code	Claude 4.8	69.45	9.17	9.37	7.51	6.12	7.08	6.19	7.01
OpenDesign	OpenDesign	Codex	GPT 5.5	62.17	9.24	8.49	6.91	5.05	7.24	5.76	6.00
Coding Agent											
Codex	—	Codex	GPT 5.5	73.37	9.64	9.71	7.69	7.43	8.03	6.08	6.60
Claude Code	—	Claude Code	Claude 4.8	70.01	9.25	9.88	5.71	7.01	9.46	6.68	6.53
Doubao	—	Claude Code	Seed 2.1	61.14	8.93	8.96	5.95	5.97	6.51	5.63	4.61
GLM	—	Claude Code	GLM 5.2	52.22	8.60	7.44	5.23	5.29	7.03	4.90	5.15
Kimi	—	Claude Code	Kimi K2.7	51.46	8.33	7.15	4.72	5.62	7.26	4.93	4.81
DeepSeek	—	Claude Code	DeepSeekV4-Pro	46.01	8.23	6.68	4.10	3.82	6.95	4.41	3.38
Human-Crafted Workflow											
PosterGen	—	—	Claude 4.8	56.71	8.84	8.25	4.31	5.62	8.61	5.36	5.28
Any2Poster	—	—	Claude 4.8	49.09	8.26	5.44	4.18	3.65	9.59	4.47	3.10
Paper2Poster	—	—	Claude 4.8	44.61	6.35	2.35	8.36	2.16	3.69	4.87	1.69

Benchmark comparison: AutoDesign with DesignHarness leads the Design Agent group with 78.32 using Claude Code / Claude 4.8 and 77.97 using Codex / GPT-5.5, ahead of Claude Design and OpenDesign.

Dimension readout: The table reports the overall score together with faithfulness, coverage, density, visual evidence, layout, readability, and aesthetics, making the gain traceable across PosterBench criteria.

Benchmark scope: The comparison includes Design Agent, Coding Agent, and Human-Crafted Workflow baselines under the same PosterBench evaluation protocol.

6. Future Directions

Output types: The 4x4 example turns one paper into four finished artifacts: conference posters, slide decks, webpages, and conference videos.

Format adaptation: Shared evidence is retained while density and sequencing change for scan-first, narrated, self-paced, or temporal reading.

Release gate: Each medium needs its own rendering and communication check before promotion.

Artifact type	Presentation role
Paper poster	Dense visual summary
Paper slides	Narrated sequence
Paper webpage	Scannable, self-paced reading
Conference video	Temporal explanation

7. The Results

Trajectory: Autonomous optimization rises from 49.00 to an 80.88 plateau; redirected search reaches 88.39.

Configuration gain: The optimized DesignHarness improves all seven configurations by 5.0 to 19.6 points, reaching a best overall score of 81.5. The trace shows autonomous improvement first, then a further gain after human guidance redirects the search.

Finding	PosterBench result	Interpretation
Main track	78.32	7.45 points above Claude Design
Learned harness	54.99 to 67.39	Mean gain across seven controlled configurations

Human validation: In the system-blind study, eleven reviewers submitted 936 responses (933 ranking judgments and three skips); AutoDesign's probability of beating a randomly sampled alternative was 64.0%.

8. Analysis

Scope: Seven configurations are compared on the fixed 10-paper subset.

Reading: Marker size gives median runtime; the dashed line marks the empirical Pareto frontier.

9. Takeaway & Limitations

- Target: Optimize the harness, not model weights.
- Promotion: Keep only updates with training gain and no development regression.
- Scope: Update one selected component; PosterBench currently validates posters.

Paper Figure 2 · Self-authored artifact example

Problem

Artifact-level repair does not create a reusable design capability

Response-level feedback

Repairs the current artifact.

- Fixes a clipped figure or missing section
- Improves one execution trajectory
- Ends when that output is delivered

→

System-level learning

Changes the reusable production system.

- Finds failures that recur across tasks
- Updates one harness component
- Carries the improvement into later runs

AutoDesign moves feedback from an isolated output into persistent design capability.

Contributions

1 **AutoDesign**
A meta-harness optimization framework for human-aligned multimodal design.

2 **PosterBench**
A 100-paper, seven-dimensional, source-grounded evaluation protocol with controlled tracks and blind human validation.

3 **DesignHarness**
An executable paper-to-poster system evolved through rollout evidence and bounded harness updates.

4 **Long-horizon production**
Editable, ready-to-use posters generated within 40 minutes for less than $3 under the reported configurations.

Overview

Outline

Meta-harness optimization leaves model weights fixed

Design harness

$$y \sim H(\pi_\theta, x, c)$$

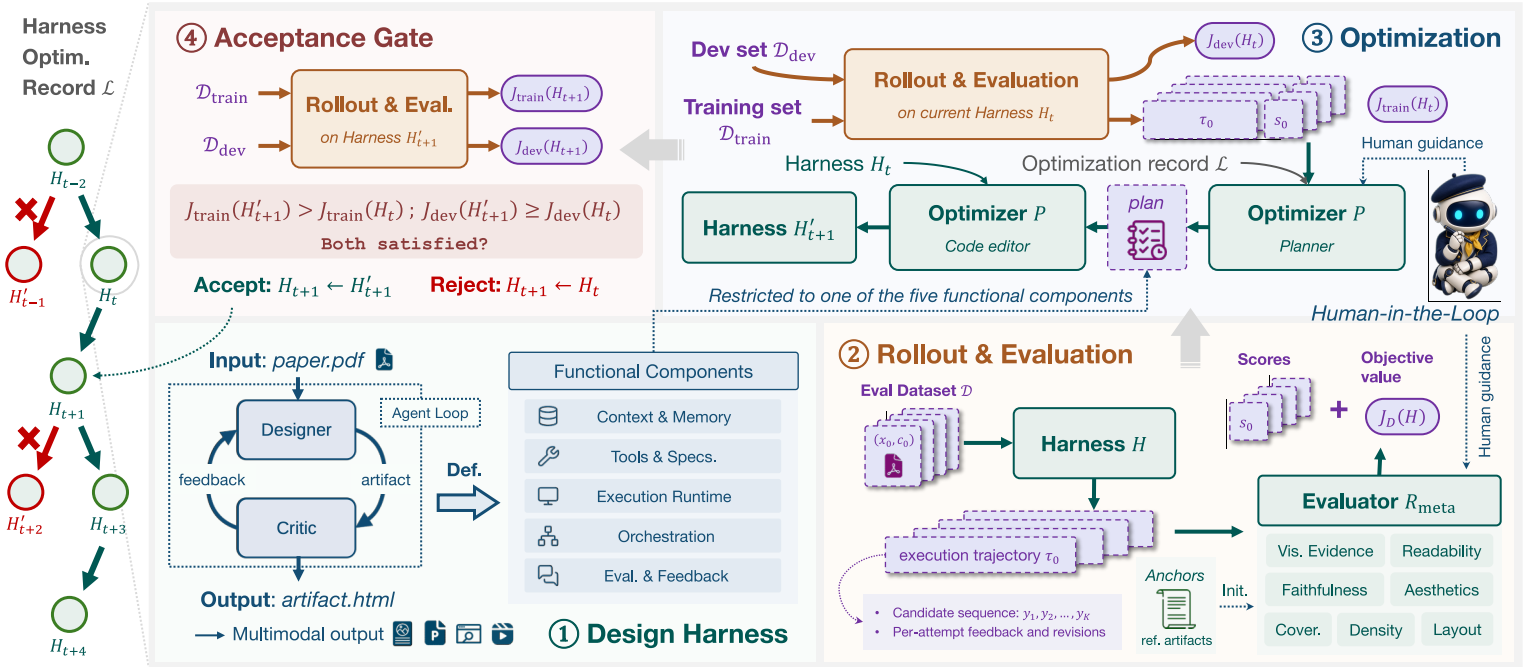
Transforms a multimodal source into an artifact through an execution trajectory.

Meta-harness objective

$$H^\star = \arg\max_H J(H)$$

Improves the implementation of the design harness across tasks.

> The parameters θ of the underlying model remain fixed.



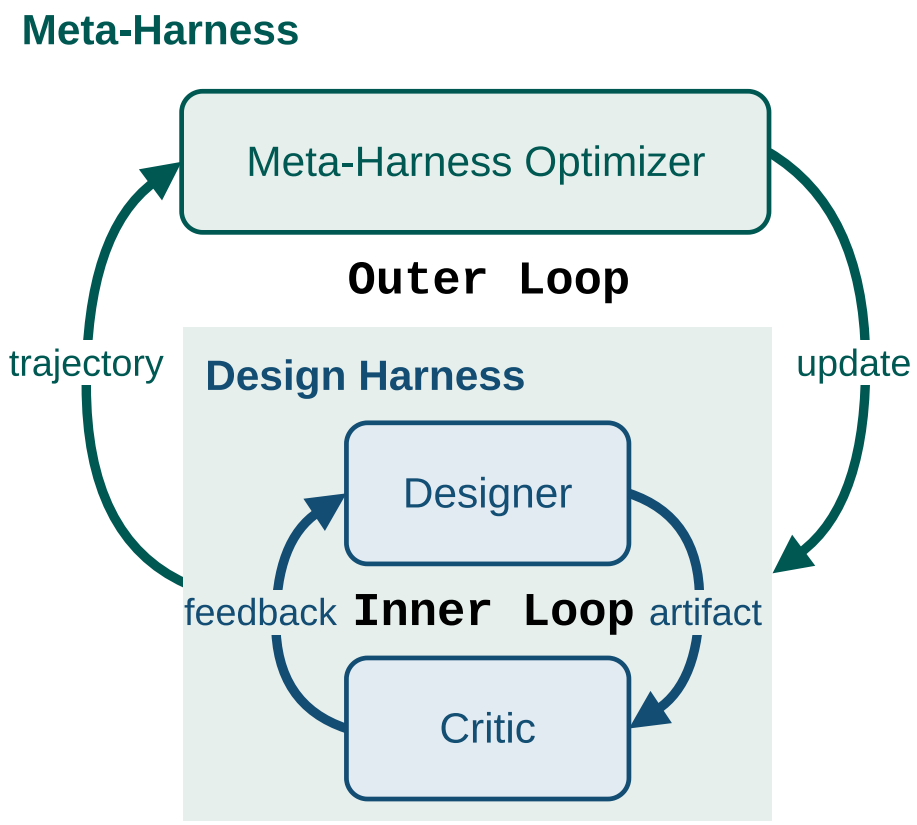
Paper Figure 3 · Design harness, rollout evaluation, optimization, and acceptance

A design harness has five functional components

Component	Responsibility
Context and Memory	Source management, prompts, skills, reusable assets, and persistent state
Tools and Specifications	Authoring tools and editable artifact rules for layout, typography, and provenance
Execution Runtime	Workspace support for authoring, rendering, validation, and export
Orchestration	Routing, attempt budgets, loop control, candidate selection, fallback, and finalization
Evaluation and Feedback	Rule-based validation, model-based critique, and localized repair feedback

Each outer-loop iteration may update exactly one of these five functional components.

Nested loops separate artifact repair from harness learning



Paper Figure 4 · Two nested feedback loops

Inner loop

Revises one artifact under a fixed harness H.

$$y_k = M_{design}(y_{k-1}, f_{k-1}; x, c)$$

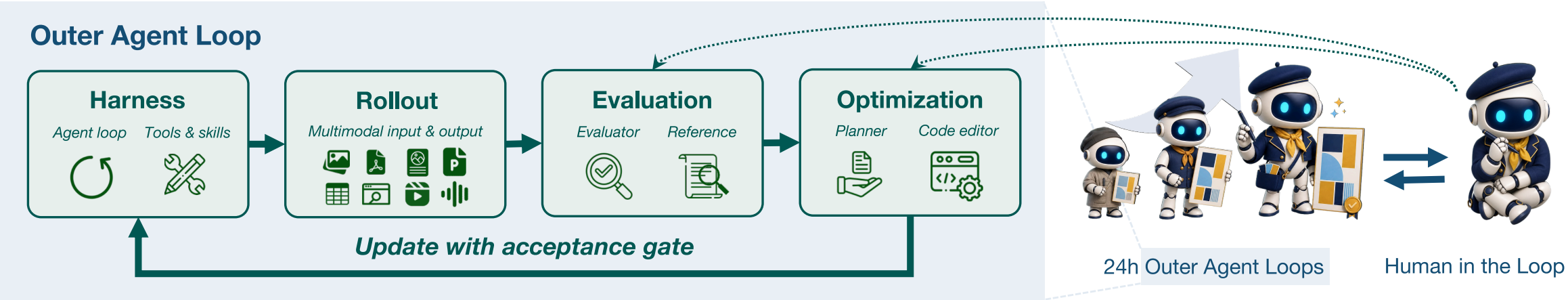
Outer loop

Uses trajectories across tasks to update the harness.

$$H'_{t+1} = P(H_t, \tau_t, s_t, L)$$

Method

Each outer-loop iteration follows four stages



Paper Figure 5 · One outer-loop iteration

1
Rollout
Run the current harness on training tasks and retain trajectories.

2
Evaluation
Score outputs with the fixed optimization-time evaluator.

3
Update proposal
A planner and code editor modify one harness component.

4
Acceptance
Promote only when training improves and development does not regress.

Method

The acceptance gate protects development performance

1	Run H_t on the training tasks and collect trajectories τ_t and scores s_t.
2	The optimizer proposes H'_{t+1} from the active harness, evidence, and optimization record.
3	Evaluate the candidate independently on training and development tasks.
4	Accept only when training improves and development does not decline.
5	Otherwise retain H_t; append the decision and checkpoint to the optimization record.

$$\text{Accept}(H'_{t+1}) \Leftrightarrow$$

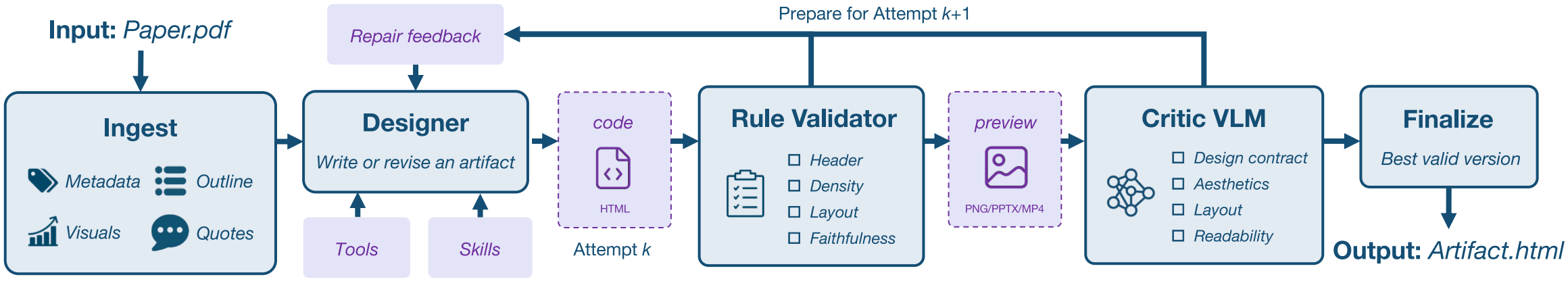
$$J_{\text{train}}(H'_{t+1}) > J_{\text{train}}(H_t)$$

$$\wedge\, J_{\text{dev}}(H'_{t+1}) \geq J_{\text{dev}}(H_t)$$

- Development results are hidden from the optimizer.
- Rejected candidates do not replace the active harness.
- One selected functional component changes per iteration.

Method

DesignHarness produces an editable, validated artifact



Paper Figure 6 · Optimized paper-to-poster production loop

Ingest

Metadata, sections, key passages, figures, and source locations

Author

A coding-agent designer creates or revises editable HTML

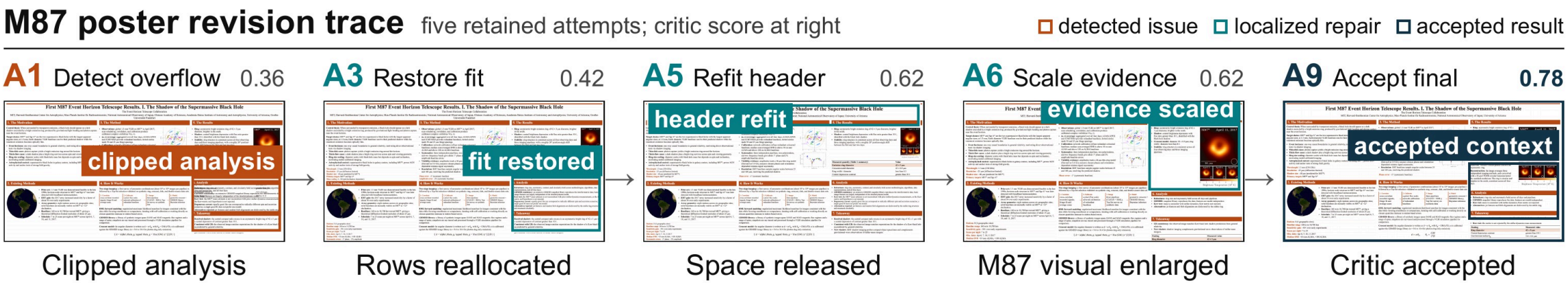
Validate

Rule checks and a visual critic return localized feedback

Finalize

The best valid candidate becomes the self-contained output

Localized feedback enables localized repair



Paper Figure 11 · Five selected attempts from one poster run

A1 → A3

A clipped analysis lane is repaired by reallocating rows.

A5/A6

Header refitting and evidence scaling improve hierarchy without rebuilding valid regions.

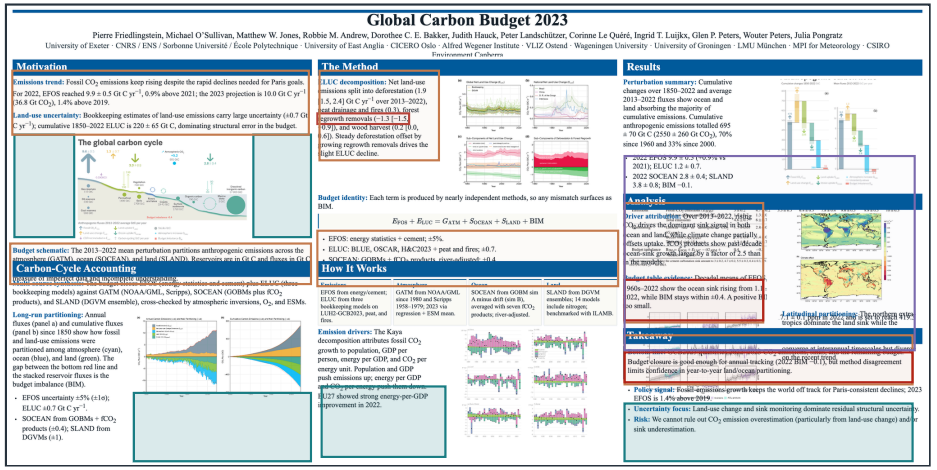
A9

The final attempt preserves repaired context and is accepted at 0.78.

Outline

PosterBench measures communication and rendered usability

(a) Localized audit trace



Section-local boxes expose unused capacity inside otherwise dense panels.

OCR tokens are matched against source-PDF numerals before scoring faithfulness.

Spatial checks retain overlaps and cropped text as repairable evidence.

The source-conditioned VLM judges evidence use and readability at poster scale.

(b) Evaluation dimensions

Each dimension consumes a traceable subset of rendered, source, and VLM evidence.

Faithfulness
OCR + source

Coverage
brief + sections

Density
spatial audit

Visual Evidence
visual rubric

Layout
overlap audit

Readability
clipping audit

Aesthetics
VLM rubric

Protected gate
RENDER INTEGRITY
hard constraint

Main Track

100 papers across five disciplines

PosterBench-mini

Shared 10-paper subset for controlled studies

Dimensions

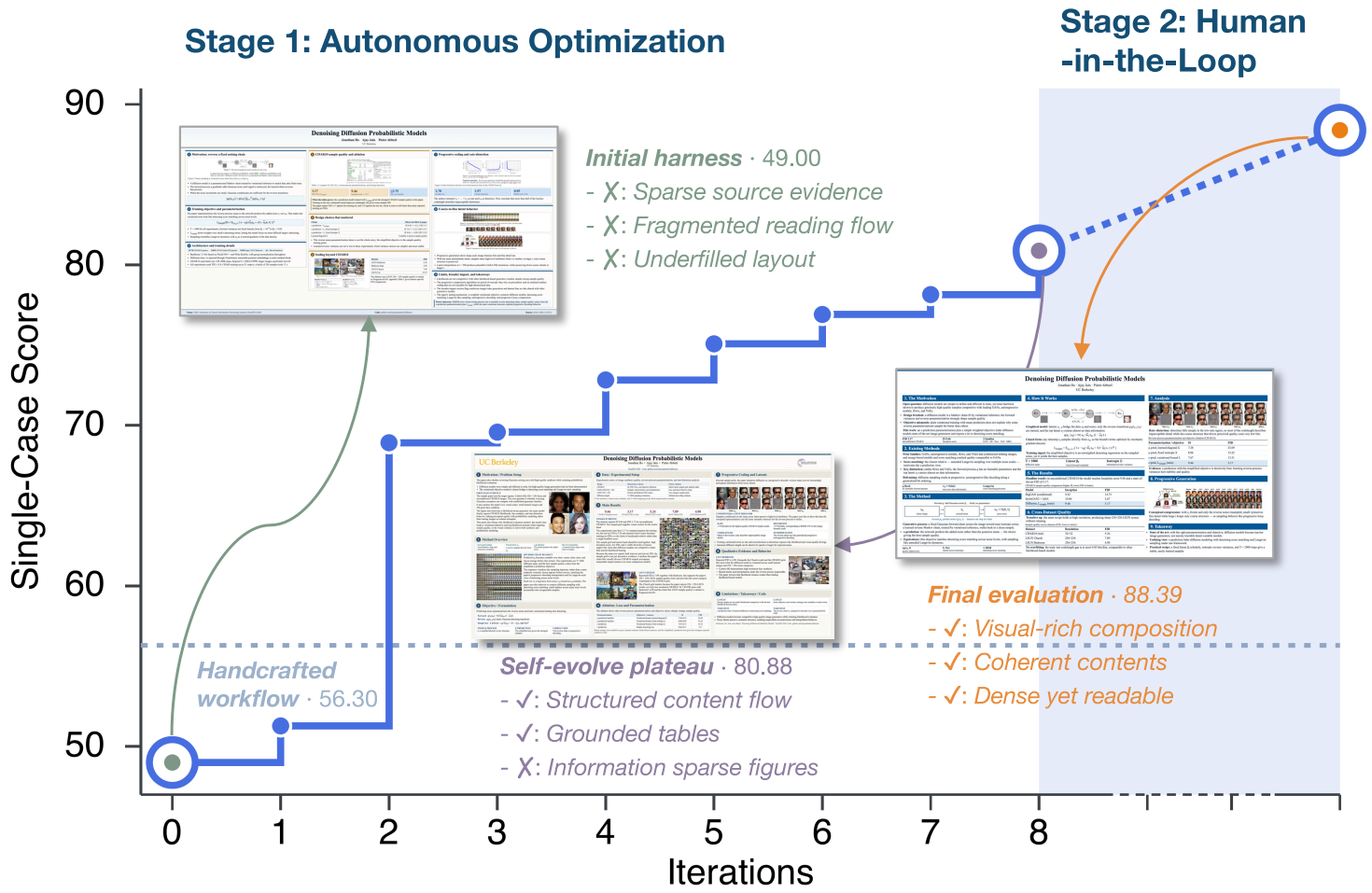
Faithfulness, Coverage, Density, Visual Evidence, Layout, Readability, Aesthetics

Checks

Programmatic audits, rubric-guided VLM judgments, and protected render-integrity gates

Paper Figure 7 · Traceable audits and seven evaluation dimensions

Optimization raises harness quality from 49.00 to 88.39



Paper Figure 1a · Representative optimization trace

49.00

Initial harness

Sparse evidence, fragmented reading flow, underfilled layout

80.88

Autonomous plateau

Structured content and grounded tables, but information-sparse figures

88.39

Final evaluation

Human guidance redirects the search after the autonomous plateau

Evaluation

AutoDesign leads the 100-paper Main Track

PosterBench Main Track · selected rows from Table 1

System	Design harness	Coding agent	Model	Score
AutoDesign	DesignHarness	Claude Code	Claude 4.8	**78.32**
AutoDesign	DesignHarness	Codex	GPT-5.5	**77.97**
Claude Design	Claude Design	Claude Code	Claude 4.8	**70.87**
OpenDesign	OpenDesign	Claude Code	Claude 4.8	**69.45**
Codex	—	Codex	GPT-5.5	**73.37**
Claude Code	—	Claude Code	Claude 4.8	**70.01**
PosterGen	—	—	Claude 4.8	**56.71**

Paper Table 1 · 100-paper complete-system comparison

Main Track · 100 papers

78.32

AutoDesign

Claude Code + Claude 4.8

70.87

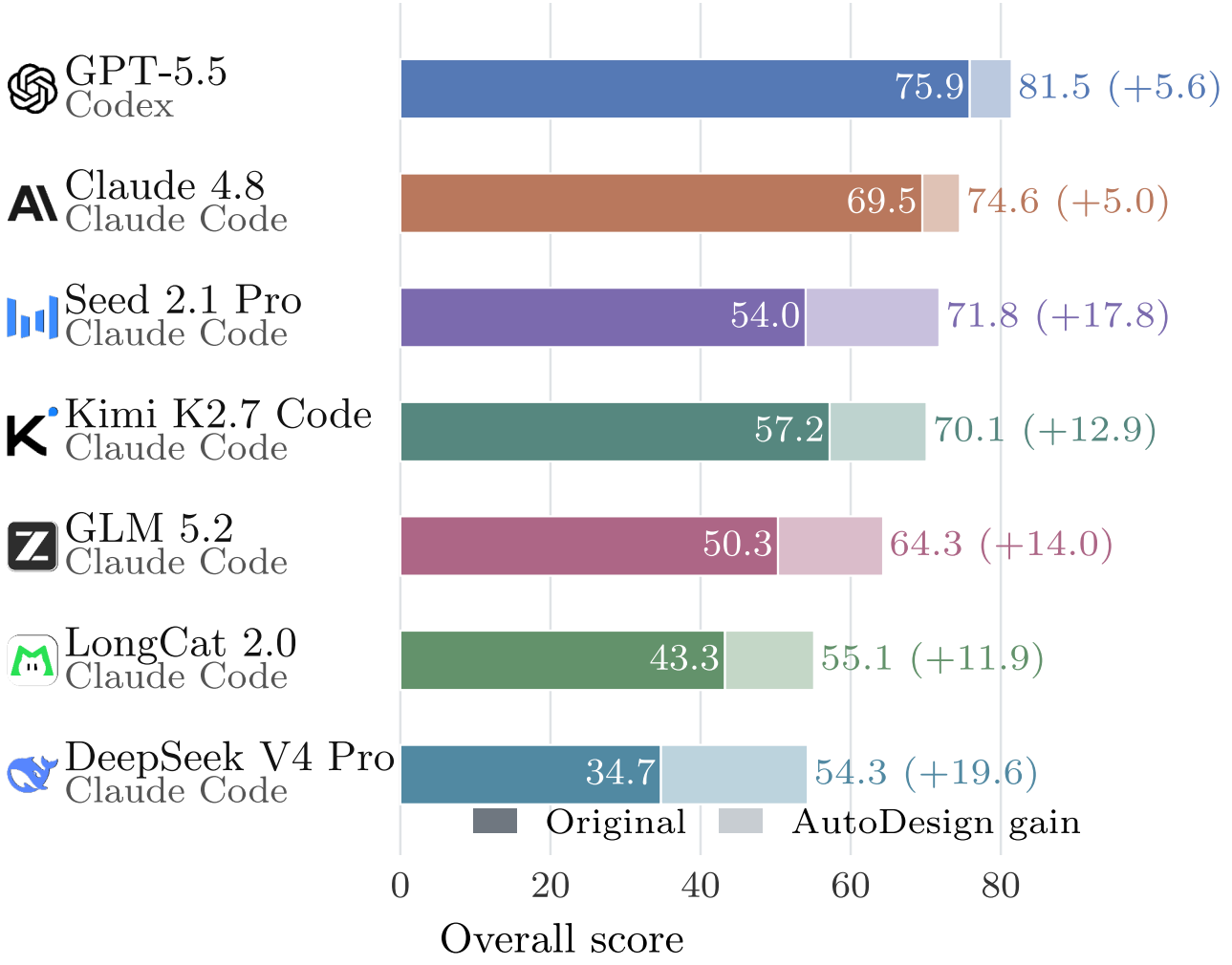
Claude Design

Matched commercial baseline

+7.45 PosterBench points

Evaluation

DesignHarness improves all seven fixed configurations



Paper Figure 1b · Before and after DesignHarness attachment

PosterBench-mini · 10 papers · 7 fixed configurations

7 / 7

configurations improve

The coding agent and model stay fixed within each comparison.

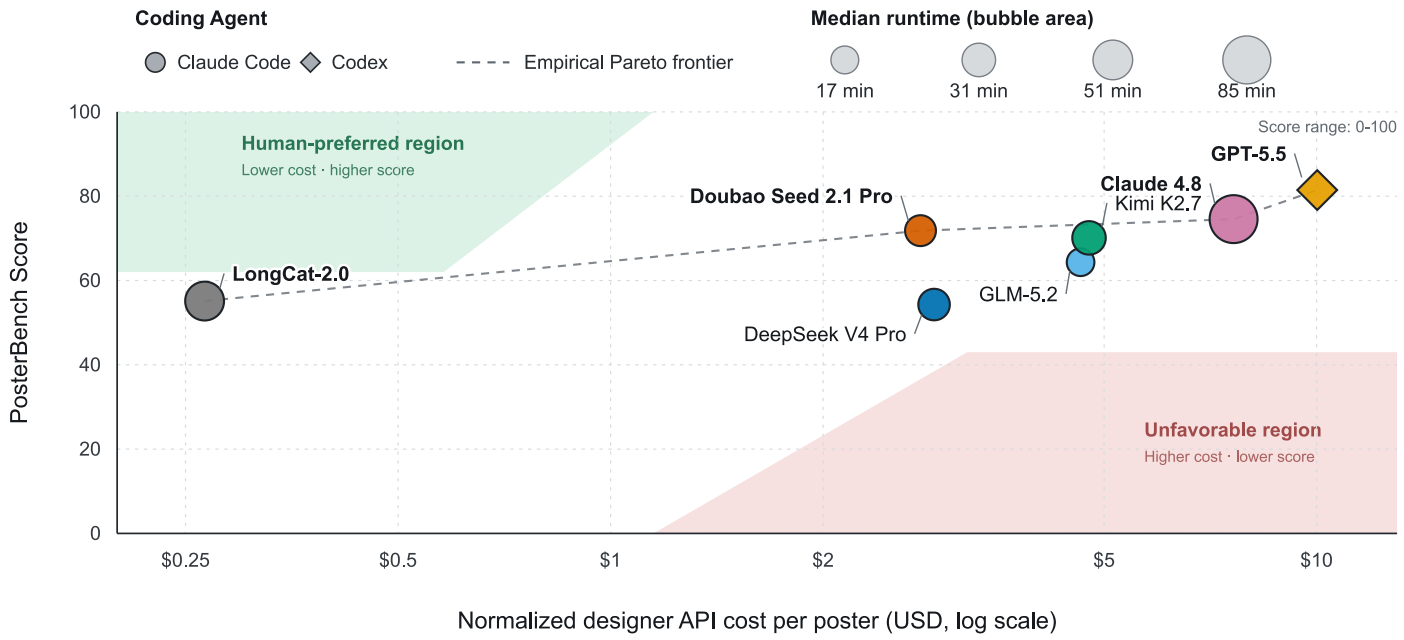
54.99 → 67.39

mean PosterBench Score

Average increase: +12.40 points

Gain range: +5.01 to +19.56

Model configurations trace a cost–performance frontier



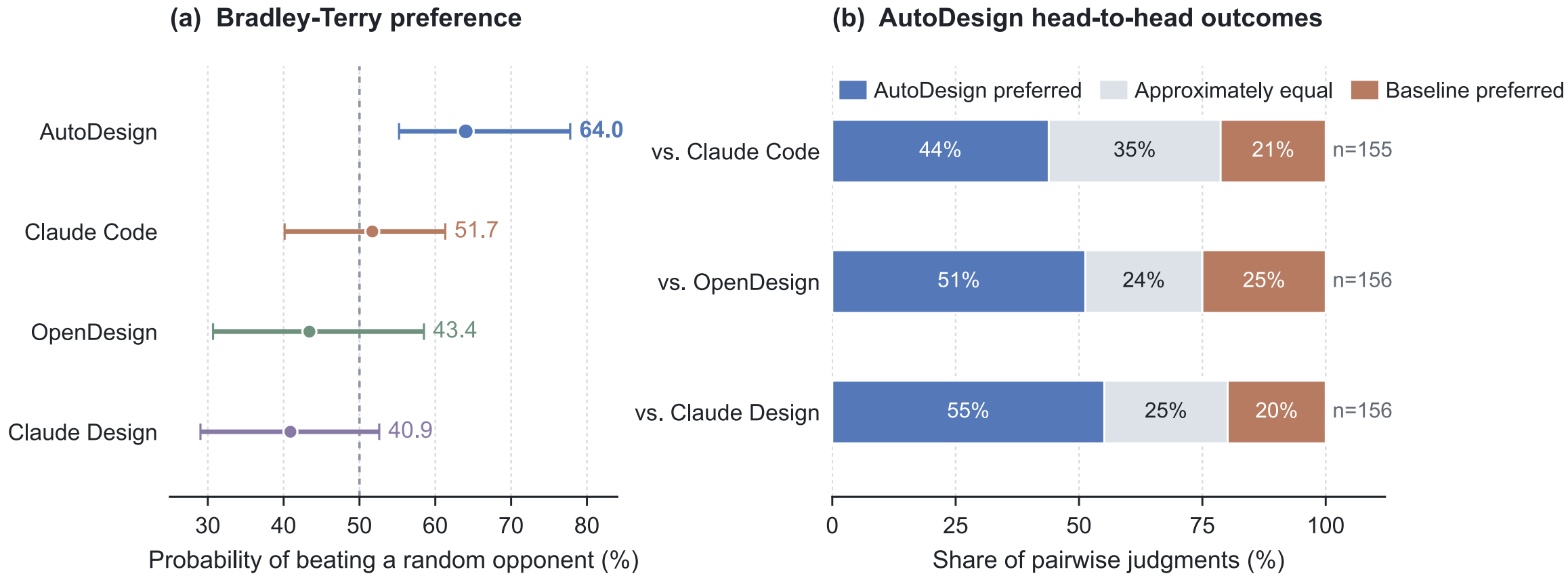
Paper Figure 8 · Marker area denotes median runtime

normalized designer-only API cost proxy · PosterBench-mini

- **LongCat 2.0:** 55.13 at approximately $0.27 per poster.
- **Doubao Seed 2.1 Pro:** 71.83 at $2.75.
- **GPT-5.5:** highest score, 81.46 at $10.02.

The x-axis is a designer-only API proxy under the paper's pricing assumptions, not an all-in deployment estimate.

AutoDesign has the highest system-blind human preference



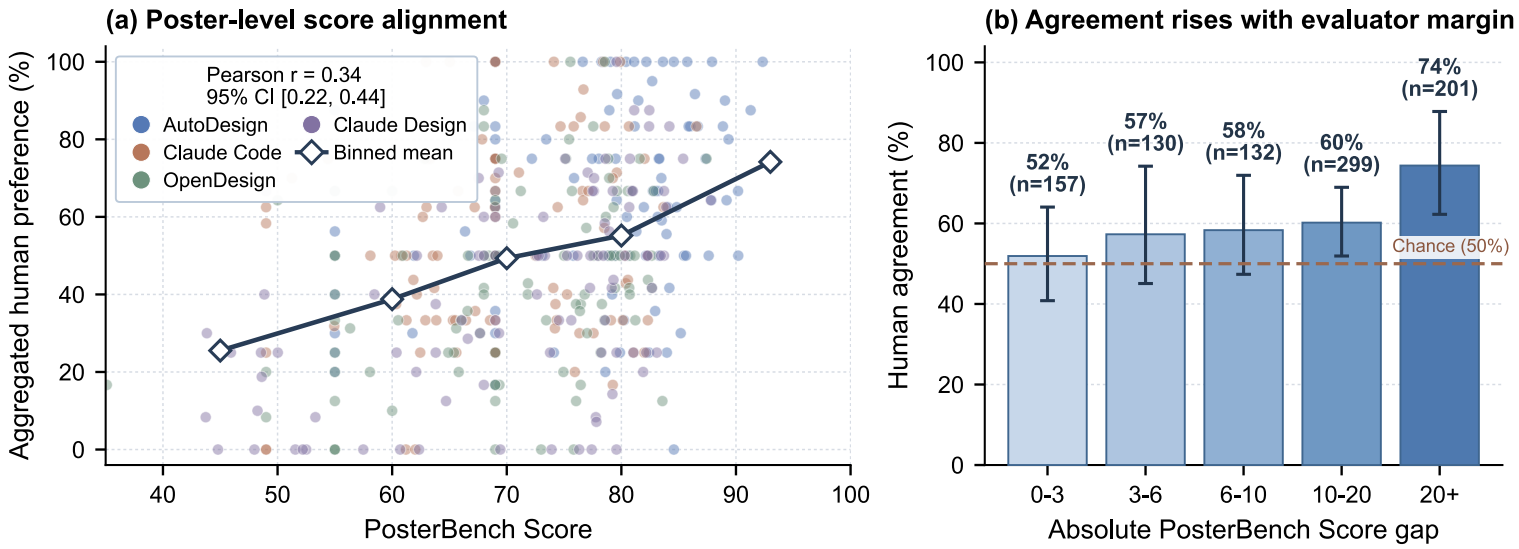
Paper Figure 9 · Bradley–Terry estimates and head-to-head outcomes

11 reviewers
Volunteer, fully system-blind evaluation

936 responses
933 ranking judgments and 3 skips

64.0%
Highest Bradley–Terry estimate; 95% interval 55.2–77.8%

Larger PosterBench margins better predict human choices



Paper Figure 10 · Association and agreement by score margin

r = 0.34

Positive but imperfect poster-level association; 95% interval [0.22, 0.44].

51.9%

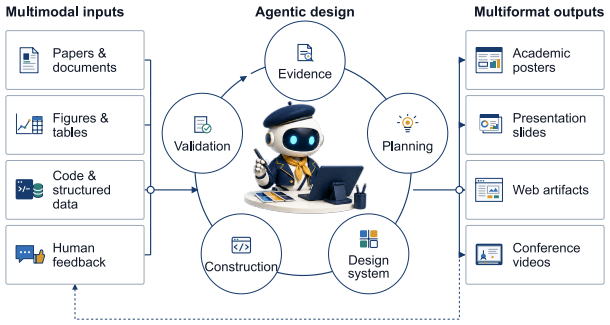
Human agreement for 0–3 point PosterBench gaps.

74.4%

Human agreement when the gap is at least 20 points.

PosterBench is an informative comparison signal, not a substitute for immediate human preference.

Posters are validated; other media remain pilots



Paper Figure 12 · Long-term system direction

Paper Figure 13 · Current cross-medium pilot artifacts

Validated now

Paper-to-poster generation under the frozen PosterBench protocol

Demonstrated as pilots

Paper slides, webpages, and conference videos

Required next

Medium-specific data, evaluator, rendering gate, and communication objective

Conclusions

1 **Optimize the harness, not the model weights.** Recurrent design failures become persistent system updates.

2 **Validate the complete artifact.** PosterBench combines source grounding, rendered usability, controlled comparisons, and blind human evidence.

3 **Keep outputs editable and claims scoped.** Posters are formally evaluated; other media require their own evidence contracts.

Project autodesign.designanything.ai **Code** github.com/Yaxin9Luo/AutoDesign **Demo** designanything.ai

Questions?