

PEER REVIEW · AI AGENTS

Agent-assisted peer review



A technical and critical analysis of Sakana AI's Multi-Layered Review (MLR): what the 73% headline really measures

Aridio Silva · Independent Researcher · @aridiosilva

ORCID 0009-0008-2411-6995 · 13-October-2026

Roadmap: the structure of the paper

1

Introduction

2

Claim check

3

Foundations

4

State of the art

5

The benchmark

6

The MLR system

7

Experimental protocol

8

Results

9

Independent evidence

10

Discussion and
credibility

11

Limitations

12

Gaps

13

Outlook

14

Requirements

15

Conclusion

Executive summary



73.43%

synthetic contradictions
in the main claim (4
reviews)



40.95%

across all benchmark
distances (4 reviews)



16.11%

real retraction reason,
exact match



0

precision (false-alarm)
measurements in the
paper

Why it matters

Catching errors is the costliest peer-review task;
MLR targets exactly that.

Positive implications

- ◆ Strong on errors planted in the central claim
- ◆ Low cost: about US\$ 0.47 per review

Negative implications

- Drops to 16–26% on real errors
- False alarms were not measured

The problem: peer review under pressure

- More submissions, lower-quality reviews (Teo et al., 2026)
- ▲ NeurIPS: about half of the papers accepted by one committee would be rejected by another
- ◆ LLMs already in use: NEJM AI study, ICLR 2025 trial (20k reviews) and the AAAI-26 pilot (22,977 submissions)
- ▶ Gap: evaluating by imitation of human reviews, not by finding errors

Why it matters

Catching errors is the costliest review task, and where capacity is scarcest.

Positive implications

- ◆ Scale error screening
- ◆ Free up human reviewer time

Negative implications

- Risk of delegating without reading
- Inflated scores under indiscriminate use

Four research questions

Q1

What does the headline claim, and what does the paper actually measure?

Q2

How were the benchmark, system and protocol built?

Q3

Which results hold, and which depend on favorable conditions?

Q4

Which gaps and threats to validity remain?

Why it matters

Clear questions separate fact from headline.

Positive implications

- ◆ Verification traceable to the source
- ◆ Focus on evidence, not opinion

Negative implications

- Scope limited to the paper and open sources
- Part of the appendix was not read

The headline versus what the paper measures

Aspect	Headline	Paper (Teo et al., 2026)
Who detects	"3 agents"	Only the Review Agent (Claude Sonnet 4)
How many reviews	Implied: one	Four; one hit is enough
Error type	"Errors"	Synthetic contradictions (GPT-4.1)
Severity	"Central claims"	Distance 0 in the knowledge graph
Dataset	"Papers"	257 machine-learning papers
Benchmark total	Not cited	40.95% (4 reviews) and 28.32% (1 review)
Real errors	Not cited	26.07% (similar) and 16.11% (exact)

Why it matters

The headline blends configurations and hides the slice measured.

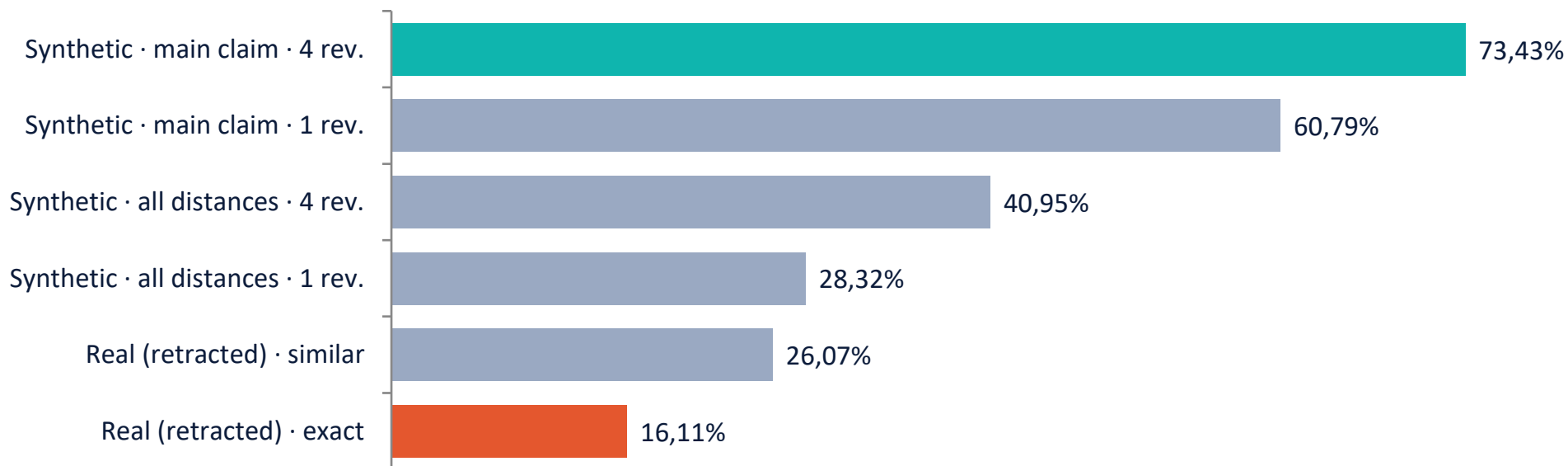
Positive implications

- ◆ 73.43% is real and reported in the paper
- ◆ Result is verifiable

Negative implications

- A hasty reading overstates the system
- Press repeats the headline

73% falls to 16% when errors are real



MLR hit rate by error type and configuration (%). Source: Teo et al. (2026), Tables 3 and 14.

Why it matters

Same system, six measurements: the number depends on error type and ensemble.

Positive implications

- ◆ Synthetic: 73.43% on the main claim
- ◆ A single review already hits 60.79%

Negative implications

- Exact real error: only 16.11%
- Ensemble multiplies cost by 4

Method and verification legend

- ◆ Confirmed in the paper: figure read in the paper text or an excerpt of it
- ▲ Secondary source: press or blog coverage, to be checked
- ✦ Own inference: calculation or interpretation by the author of this analysis
- Read: main text, appendices A.1 to A.4 and excerpts of A.5, A.6 and B.1
- Not read: OpenReview (TMLR), original Substack, Figure 28, A.5, A.7, B.2 to B.4

Why it matters

Tracing each number to its source avoids repeating headlines.

Positive implications

- ◆ Rigor on sources and dates
- ◆ Limits stated openly

Negative implications

- OpenReview blocked: TMLR reviews unread
- Original Substack not located

Core concepts

● Assistance vs replacement

AI supports the human reviewer; it does not replace them

● Graph and distance

Distance to the main claim measures severity

● LLM as judge

A model decides whether the error was found

● Precision and recall

Recall: errors found; precision: correct flags

● Synthetic contradiction

Logical error inserted in a controlled way

● Understand before critiquing

Structured reading before the review (Keshav)

● Detection in ≥ 1 of k

Counts if any review hits (like pass@k)

● Prompt injection

Hidden instruction that tries to manipulate the reviewer

Why it matters

A shared vocabulary avoids confusing recall with precision.

Positive implications

- ◆ Critical reading of figures
- ◆ Basis for comparing systems

Negative implications

- New terms raise the barrier
- Definitions vary across papers

Four phases, from 2007 to 2026

1. First tests

2007–2023

Keshav (2007)
Yuan et al. (2022)
Liu & Shah (2023)
Robertson (2023)

2. Scale and agents

2024

Liang et al. (NEJM AI)
MARG
AgentReview
AI Scientist
Reviewer2
WithdrarXiv
Skarlinski et al.
Ye et al.

3. Verification

2025

DeepReview
ReviewAgents
GAR
SPOT
Zhang & Abernethy
FLAWS
Thakkar et al.
Hidden Prompts

4. Deployment

2026

AAAI-26 pilot
ReviewGrounder
PRISM
ReviewBench
PeerReview Bench
MLR (Sakana AI)
AI Scientist (Nature)

Why it matters

Knowing the trajectory shows what MLR actually adds.

Positive implications

- ◆ Mature field: from tests to pilots at scale
- ◆ Reuse of ideas (Three-Pass, graphs)

Negative implications

- Fast pace makes replication hard
- Part of the literature lacks verified authorship

From exploratory test to deployment

- 2022–2023: models generate feedback and detect inserted errors
- ▶ 2024: usefulness measured at scale, multi-agent systems and the AI Scientist
- ▲ 2025: deep reasoning, error benchmarks and hidden text in manuscripts
- ◆ 2026: deployment (AAAI-26), expert evaluation and MLR
- The original AI Scientist reviewer extracts PDF text without reading figures and tables

Why it matters

MLR answers known limits: it sends the PDF as text and image.

Positive implications

- ◆ Evidence accumulating since 2022
- ◆ Real pilot on 22,977 submissions

Negative implications

- Perceived usefulness $\neq$ rate of flaws found
- Inflated scores and undeclared use

AI review systems

System	Approach	Available evidence
MLR (Teo et al., 2026)	3-pass chain + auxiliary agents (Claude)	Section 8 of this work
AI Scientist Reviewer	Ensemble + meta-review + reflection	Reads PDF text, no figures
AgentReview	Simulates reviewers, authors and AC	Baseline in the paper
MARG	Lead agent + experts	No direct comparison
GAR	Paper graph + 3 to 6 reviewers	Acceptance prediction
Reviewer2	Aspect prompts + generation	Described in the paper
ReviewAgents	Reasoning chains; trained agents	Ablation on number of agents
DeepReview	Structured reasoning	DeepReview-13K / Bench
ReviewGrounder	Rubrics + tools	ReviewBench
Reviewer3.com	Commercial multi-agent	Compared in ReviewBench
Stanford · OpenAIRewiew · CMU	Agentic review platforms	P/R/F1 on PeerReview Bench

Why it matters

The field grew in generation, evaluation and error verification.

Positive implications

- ◆ Many approaches to compare
- ◆ Open platforms already available

Negative implications

- Few direct comparisons among them
- Models change, results age

Error-detection benchmarks

Benchmark	Error type	Scale	Evaluation
Contradiction Benchmark (Teo)	Synthetic, graph with severity	1,164 in 257 ML papers	LLM judge (o3)
WithdrarXiv-Check	Real, retractions	211 papers	LLM judge (o3)
FLAWS (Xi et al.)	Synthetic, invalidates claims	713 paper-error pairs	Automatic
SPOT (Son et al.)	Real, automatic verification	Not verified	Not verified
PeerReview Bench (Kim et al.)	Real, expert-validated items	78 papers	AI meta-reviewer
ContraDetect (Skarlinski)	Contradictions across papers	Not verified	Experts

Why it matters

The right benchmark defines what 'good' means for an automatic reviewer.

Positive implications

- ◆ Synthetic errors give an unambiguous target
- ◆ Scales without manual annotation

Negative implications

- Synthetic can be easier than real
- Tasks are not comparable with each other

Positioning of MLR

- Only benchmark on the list that grades error severity by graph distance
- ▶ Other 2026 works evaluate by rubrics, dimensions or process
- ▲ Study with Claude agents on data analysis: good on implementation errors, weak on domain critique
- ✦ AI and humans find largely different problems: a complementarity thesis

Why it matters

Positioning MLR avoids comparing different tasks as if they were the same.

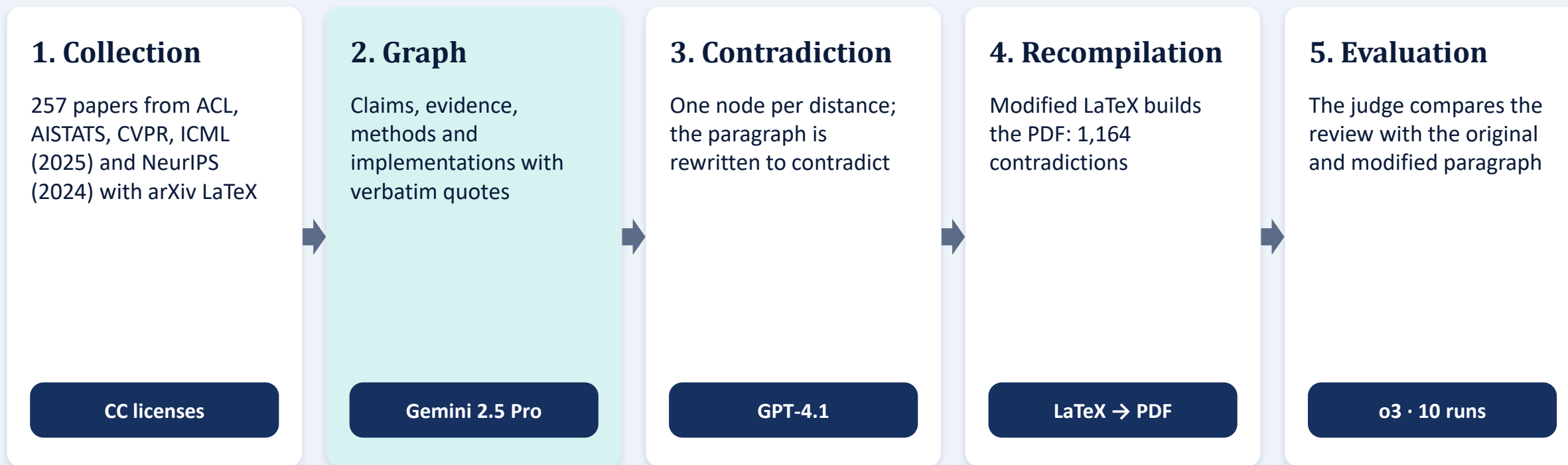
Positive implications

- ◆ Complements the human reviewer
- ◆ Focus on verification, not imitation

Negative implications

- Covers ML; other domains untested
- 'Jagged' profile across error types

How the benchmark is built



Why it matters

Each contradiction comes from a graph node and has known severity.

Positive implications

- ◆ Unambiguous, scalable target
- ◆ Severity graded by distance

Negative implications

- Not every LaTeX recompiles
- LLM-made errors can sound artificial

Knowledge graph and distance

Node type	Subtypes / definition
Claim	Main, secondary, tertiary
Evidence	High, medium, low
Methodology	New technique or approach
Implementation	Experimental details
Mathematical Theory	New, proven theorems

Distance 0 = main claim. Result: 1,164 contradictions (distances 0 to 8); analysis on 0 to 6. The 'Contradicts' relation appears in only 0.4% of the graph.

Why it matters

Distance turns 'how serious is the error' into a number.

Positive implications

- ◆ Compares errors of equal weight
- ◆ Reveals the fall by severity

Negative implications

- Severity is assumed, not measured
- Few samples beyond d6

The judge (o3) and its validation

Judge	Outlines (%)	Reviews (%)
o3 (chosen)	97.42	99.90
o4-mini	94.70	99.52
GPT-4.1	92.63	100.00
GPT-5.2	88.77	96.35
Gemini 2.5 Pro	72.40	93.11
Gemini 3 Pro	99.87	99.92



86.8%

judge sensitivity on 50 verified hits



≈13%

of true hits may be missed: rates are conservative

Accuracy on papers with no inserted error (26 papers, 130 contradictions). Source: Teo et al. (2026), Table 11.

Why it matters

A poorly calibrated judge distorts every benchmark result.

Positive implications

- ◆ Judge validated on error-free papers
- ◆ 10 repetitions reduce noise

Negative implications

- Judge from a different family than the systems
- Cross-family bias not discussed

Benchmark limitations

- ▲ LLM-generated contradictions can sound artificial
- Audit of 50 cases: 34% incoherent with neighboring sentences and 8% obviously AI-made
- The authors see this as making the task easier, not harder
- ✦ An adversarial system could compare the edited PDF with the original on arXiv

Why it matters

Shows how far the 73% can generalize to natural errors.

Positive implications

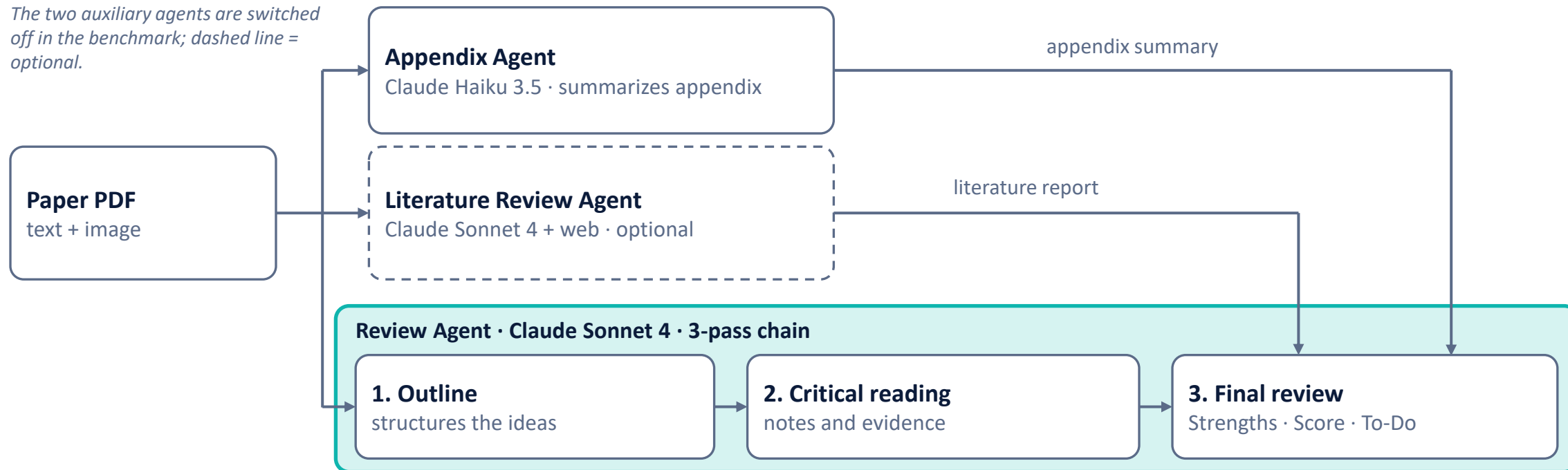
- ◆ Limitations acknowledged by the authors
- ◆ Manual audit of the dataset itself

Negative implications

- Real-world recall tends to be lower
- Possible leakage by searching arXiv

MLR architecture and workflow

The two auxiliary agents are switched off in the benchmark; dashed line = optional.



Why it matters

Only the Review Agent was measured on the benchmark; the other two are auxiliary.

Positive implications

- ◆ PDF as text and image preserves formulas
- ◆ Summarized appendix avoids already-answered critiques

Negative implications

- Main text limited to 10 pages
- Appendix summary by a smaller model loses detail

Roles, models and artifacts

Component	Model	Function
Appendix Agent	Claude Haiku 3.5	Summarizes the appendix; context management
Literature Review Agent (optional)	Claude Sonnet 4 + web search	Assesses novelty and covers the knowledge cutoff
Review Agent	Claude Sonnet 4	3-prompt chain over the main text

Artifacts produced

Paper outline

Critical notes

Final review + To-Do

Appendix summary

Literature report

Why it matters

Off-the-shelf commercial models, no fine-tuning: easy to reproduce and swap.

Positive implications

- ◆ No GPU or training
- ◆ Five artifacts useful to the author

Negative implications

- Dependence on third-party APIs
- Model versions age

The three-pass chain

1

1. Outline

Structures the paper's core ideas at a high level



2

2. Critical reading

Notes with evidence, assumptions, weaknesses and gaps; corrects pass 1



3

3. Final review

Strengths, Weaknesses, Questions, Recommendation, Score (1–10) and To-Do

Why it matters

Inspired by Keshav's Three-Pass: understand before critiquing.

Positive implications

- ◆ Strong gain on errors far from the claim
- ◆ Auditable intermediate notes

Negative implications

- Only the score rubric is detailed in the prompt
- Higher cost than a single prompt

How the multi-agents operate

- Sequential coordination with no debate between agents
- ▶ Two auxiliary agents feed inputs to a single reviewer agent
- ▲ No author or area-chair roles, fine-tuning or extra reasoning
- ◆ Own inference: 'multi-agent' is generous; the analytical part is a prompt chain
- ◆ Single-prompt variant: cost about 1/3, drop of about 3.5 points overall

Why it matters

Shows what really drives the result: the model and the chain, not an 'orchestra'.

Positive implications

- ◆ Simple to implement and debug
- ◆ Cheaper variant available

Negative implications

- 'Multi-agent' label can mislead
- No debate to reduce false alarms

MLR cost per review

Component	Input (tokens)	Output (tokens)	Cost (US\$)
MLR (appendix + review)	189,062	3,913	0.42 + 0.05
Appendix Agent	66,521	1,011	0.05 + <0.01
Review Agent	122,541	2,902	0.37 + 0.04
Literature Agent (optional)	312,767	2,032	0.94 + 0.03



≈ **US\$ 0.47**

per review, without the literature agent



≈ **US\$ 1.9**

for 4 reviews (own calculation)

Source: Teo et al. (2026), Table 9. Prices in US\$ per million tokens: Sonnet 3/15; Haiku 0.8/4.

Why it matters

Low cost makes a second read of every manuscript feasible.

Positive implications

- ◆ About US\$ 0.47 per paper
- ◆ Half the tokens of AI Reviewer

Negative implications

- The 4-review ensemble multiplies cost ~4×
- Literature adds ~US\$ 0.97

Evaluation map: what was tested where

Evaluation	Sample	MLR configuration	Metric
Contradiction Benchmark	1,164 contradictions, 257 papers	Review Agent only	Hit rate; o3 × 10
WithdrarXiv-Check	211 of 245 papers (≤ 30 pages)	Review Agent only	Retraction reason; o3
Review focus	Expert Review Dataset	Appendix + Review (web)	KL vs humans
Conference submissions	ICML 2025 · ICLR 2025 · NeurIPS 2024	Appendix + Review (web)	Score correlation
Manipulation	50 rejected papers	Review Agent only	Score shift
User study	38 sessions	Three agents	Likert; acceptance

Why it matters

The three-agent system was only evaluated as a whole in the user study.

Positive implications

- ◆ Several complementary evaluations
- ◆ Configurations declared

Negative implications

- 73% comes from the single agent, not the three
- Full system has no detection rate

Baselines and an open question

- LLM-Review: GPT-4.1, first 10 pages, text truncated at 6,500 tokens
- ▶ AI Reviewer: o4-mini, 5-review ensemble, meta-review and up to 5 reflection rounds
- ▲ AgentReview: GPT-4o, 3 reviews, first phase only
- ✦ MLR measured with 4 reviews, between the baselines' ensembles of 3 and 5
- Open: how does the paper handle pages beyond the 10th in WithdrarXiv (papers up to 30 pages)?

Why it matters

Baselines differing in model and design make the comparison partly confounded.

Positive implications

- ◆ Known, accessible baselines
- ◆ Ablation with Claude in LLM-Review

Negative implications

- Deep errors may fall beyond 10 pages
- Ensembles of different sizes

Detection by distance (Table 14 of the paper)

Hit rate (%)	d0	d1	d2	d3	d4	d5	d6	Total
MLR, 4 reviews	73.43	47.17	31.67	27.41	21.46	15.21	38.89	40.95
MLR, 1 review	60.79	32.43	19.56	14.47	12.85	4.17	16.67	28.32
MLR-Combined (subset)	56.67	31.42	18.03	9.89	3.77	1.43	0.00	24.81
LLM-Review + Claude	35.40	22.75	7.77	7.68	6.99	8.33	0.00	16.43
LLM-Review (GPT-4.1)	14.56	7.89	3.63	3.03	2.20	2.29	0.00	6.39
AgentReview	14.81	6.33	3.82	2.98	1.30	0.00	0.00	5.95
AI Reviewer	11.17	7.73	6.89	3.73	1.71	2.08	0.00	6.50

d7 and d8: 0.00 for all, except AI Reviewer (14.00 at d7). d6 has only 18 samples.

Why it matters

The 73% is the most favorable slice: the benchmark total is 40.95%.

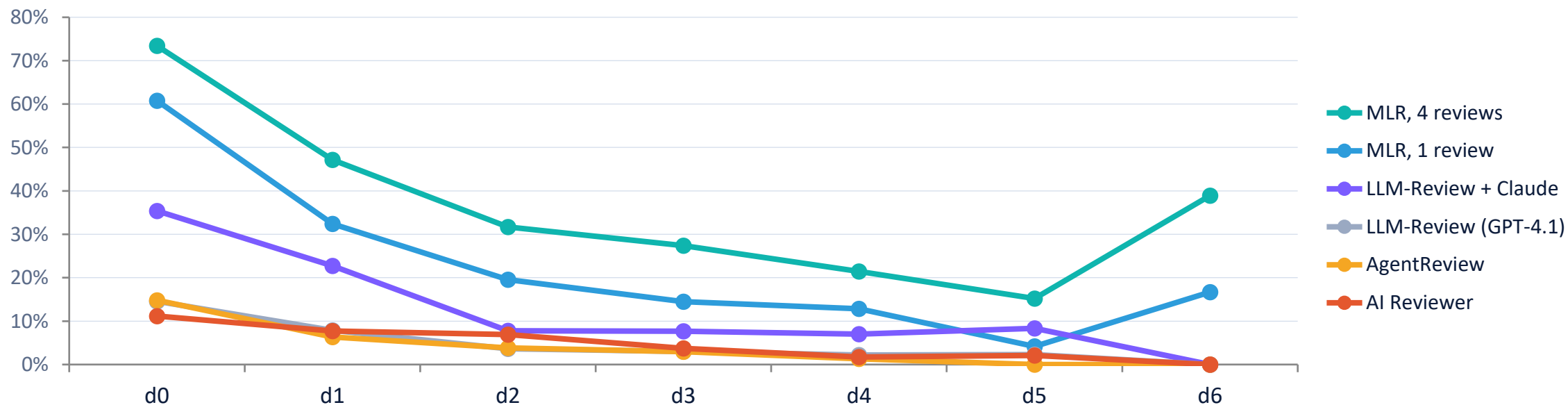
Positive implications

- ◆ ≈ 6.3× the best baseline overall
- ◆ Monotone fall calibrates severity

Negative implications

- Errors 4 steps away: only 21.46%
- Few samples at d5 and d6

Detection falls from 73% to 21% from d0 to d4



Hit rate (%) by distance to the main claim. d6 has 18 samples. Source: Teo et al. (2026), Table 14.

Why it matters

The ensemble lifts the curve but not its shape: distant errors escape more.

Positive implications

- ◆ MLR beats the GPT baselines at every distance
- ◆ 4 reviews add ~12 points

Negative implications

- Peripheral errors have low recall
- d6 is sampling noise (18 items)

Four readings of the numbers

- The 73% is the most favorable slice: 40.95% overall; about 6.3× the best baseline
- ▶ Uncertainty: with ~250 items at d0, the 95% interval is ± 5 points
- ▲ Ensemble: $P_4 = 1 - (1 - 0.6079)^4 \approx 0.976$ if independent; observed 0.7343: systematically hard items
- ✦ Judge: true recall $\lesssim 73.43 \div 0.868 \approx 84.6\%$, a rough upper bound

Why it matters

Own calculations show what the ensemble and the judge do to the measurement.

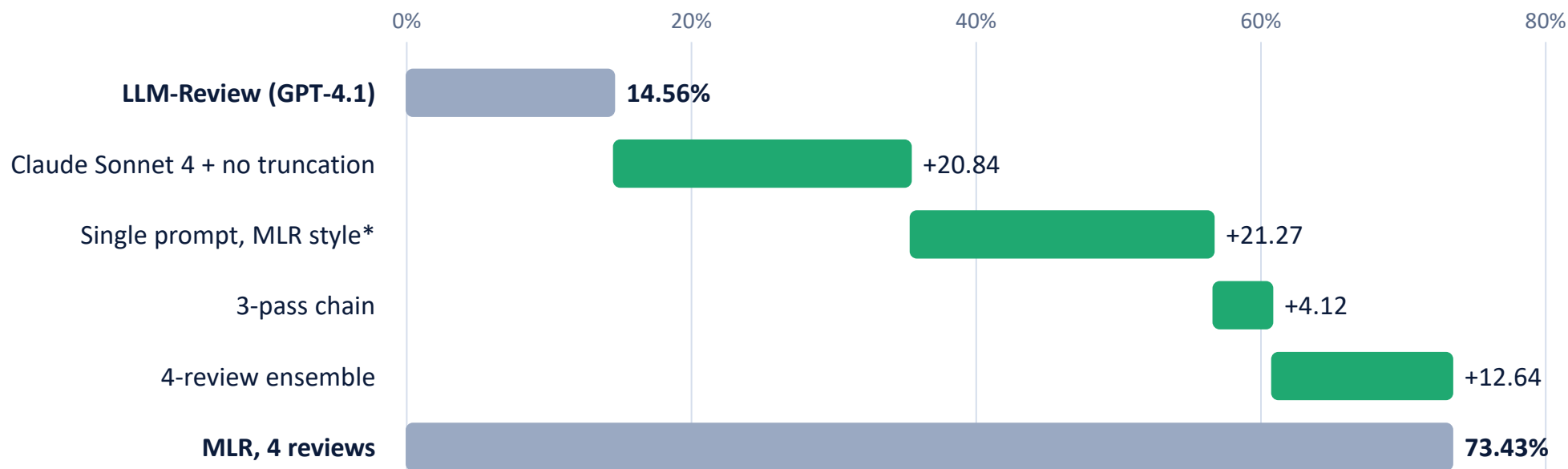
Positive implications

- ◆ Recall may be underestimated
- ◆ Ensemble helps even with hard items

Negative implications

- Diminishing returns from the ensemble
- Upper bound assumes uniform sensitivity

Of the 59 points gained at d0, the chain accounts for 4



* MLR-Combined: 497-item subset. The model effect is confounded with removing truncation. Own calculation on Teo et al. (2026), Table 14. Gray = level; green = gain (percentage points).

Why it matters

Separating model, truncation, chain and ensemble shows where the gain comes from.

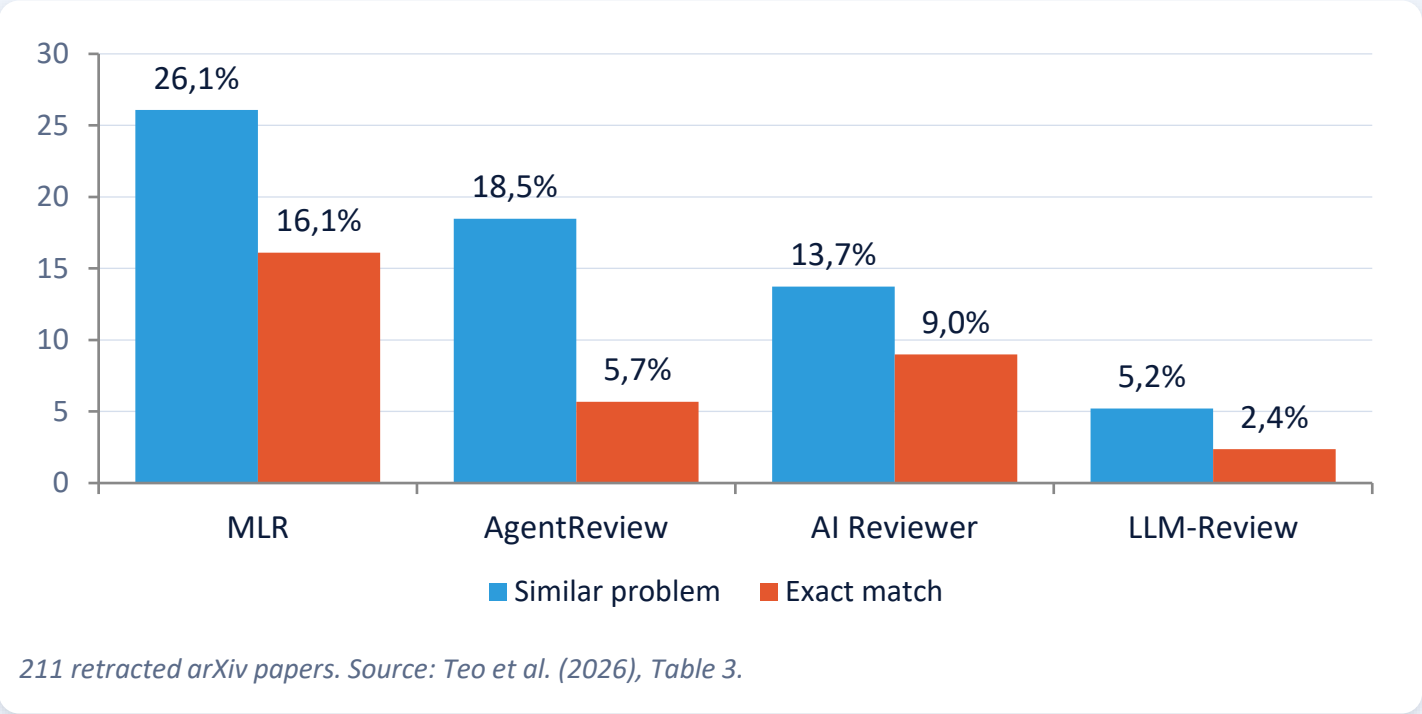
Positive implications

- ◆ Chain matters on distant errors (d4: 12.85 vs 3.77)
- ◆ Single prompt costs ~1/3

Negative implications

- Model effect confounded with truncation
- MLR-Combined on a subset

Real retracted papers (WithdrarXiv-Check)



System	Similar (%)	Exact (%)
MLR	26.07	16.11
AgentReview	18.48	5.69
AI Reviewer	13.74	9.00
LLM-Review	5.21	2.37

Why it matters

Real errors are the acid test: MLR's lead shrinks to 5–8 points.

Positive implications

- ◆ MLR still leads all four systems
- ◆ Both 'similar' and 'exact' criteria reported

Negative implications

- 'Similar' was defined after seeing results
- Domain outside ML and the 10-page cap

Review focus versus humans

System (KL; lower = more similar)	Strengths	Weaknesses	Average
MLR	0.316	0.164	0.240
LLM-Review	0.172	0.072	0.121
AI Reviewer	0.210	0.306	0.266
AgentReview	0.299	0.058	0.178
Claude Sonnet 4.5 (simple prompt)	0.125	0.069	0.097

MLR emphasizes validity and experiments; humans emphasize clarity and novelty.

Why it matters

Shows whether AI sees what humans see, or something different.

Positive implications

- ◆ A perspective complementary to humans'
- ◆ Focus on validity and experiments

Negative implications

- Divergence may simply be weaker alignment
- Hypotheses not separated by the paper

Alignment with human scores

Pearson	MLR	LLM-Review	AI Reviewer	AgentReview	Human reference
NeurIPS 2024	0.451	0.358	0.328	0.167	0.781
ICML 2025	0.429	0.169	0.439	0.006	0.684
ICLR 2025	0.586	-0.013	0.538	0.195	0.742

At ICML the AI Reviewer wins (Pearson 0.439; Spearman 0.416 vs 0.429 and 0.333 for MLR). BERTScore among humans: only 12–18.

Why it matters

Shows how closely AI scores follow human judgment.

Positive implications

- ◆ MLR leads at 2 of 3 venues
- ◆ Moderate to good alignment (0.43–0.59)

Negative implications

- Below the human reference (0.68–0.78)
- High scores for rejected papers

Manipulation by text injection

System	Score before	After	Shift	Shift (own score)
MLR	7.30 ± 1.10	8.00 ± 2.88	0.70 ± 2.46	0.26 ± 1.85
LLM-Review	6.88 ± 1.32	9.30 ± 1.47	2.42 ± 1.47	n/d
AI Reviewer	7.16 ± 1.05	7.88 ± 0.62	0.72 ± 0.98	1.32 ± 1.09
AgentReview	5.51 ± 0.69	7.19 ± 0.98	1.68 ± 1.09	0.43 ± 0.69

50 rejected papers; text in tiny white font after the conclusion. MLR detected the manipulation in 8 of 50.

Why it matters

A reviewer that reads the author's text has a real attack surface.

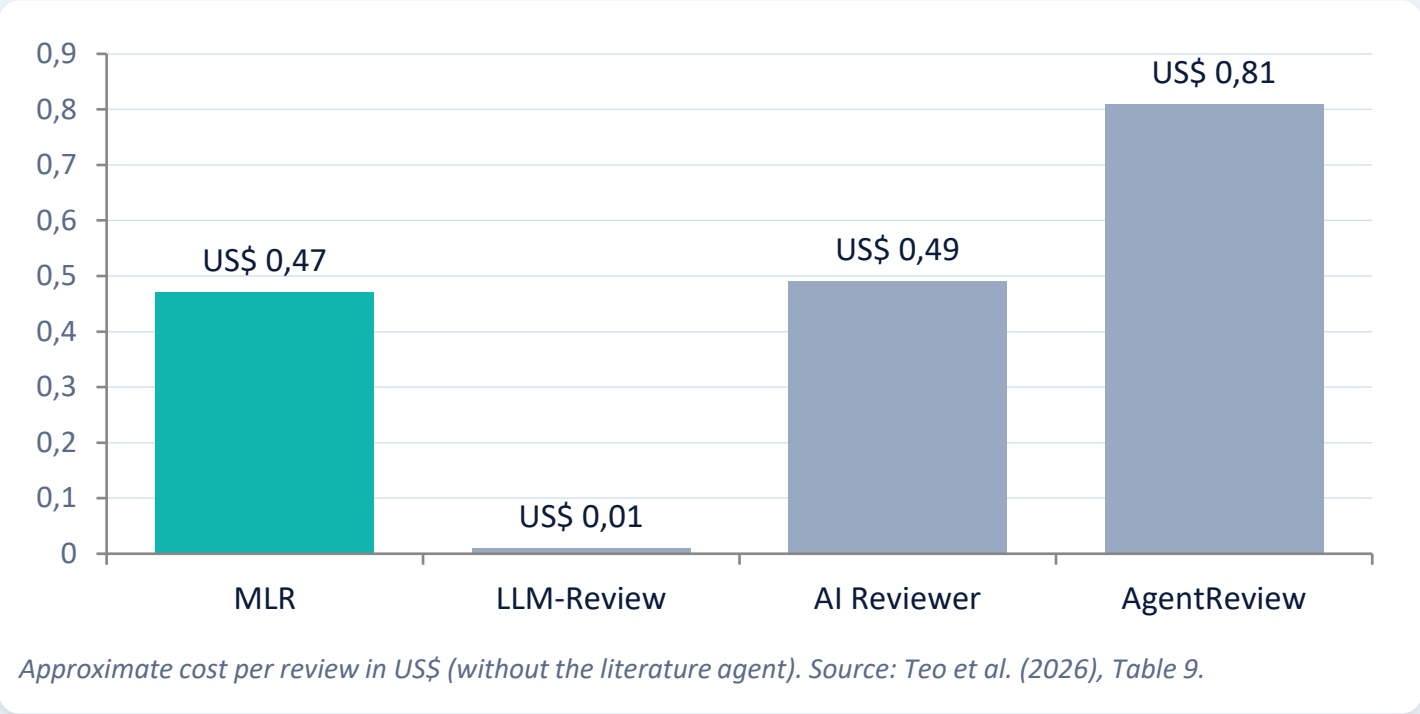
Positive implications

- ◆ MLR had the smallest average shift
- ◆ Image input helps detection

Negative implications

- Every system was influenced
- High standard deviation for MLR (2.46)

Cost compared per review



System	Input	Output
MLR	189,062	3,913
LLM-Review	6,517	740
AI Reviewer	403,654	12,448
AgentReview	310,964	3,404

Why it matters

Cost and performance must be read together to decide adoption.

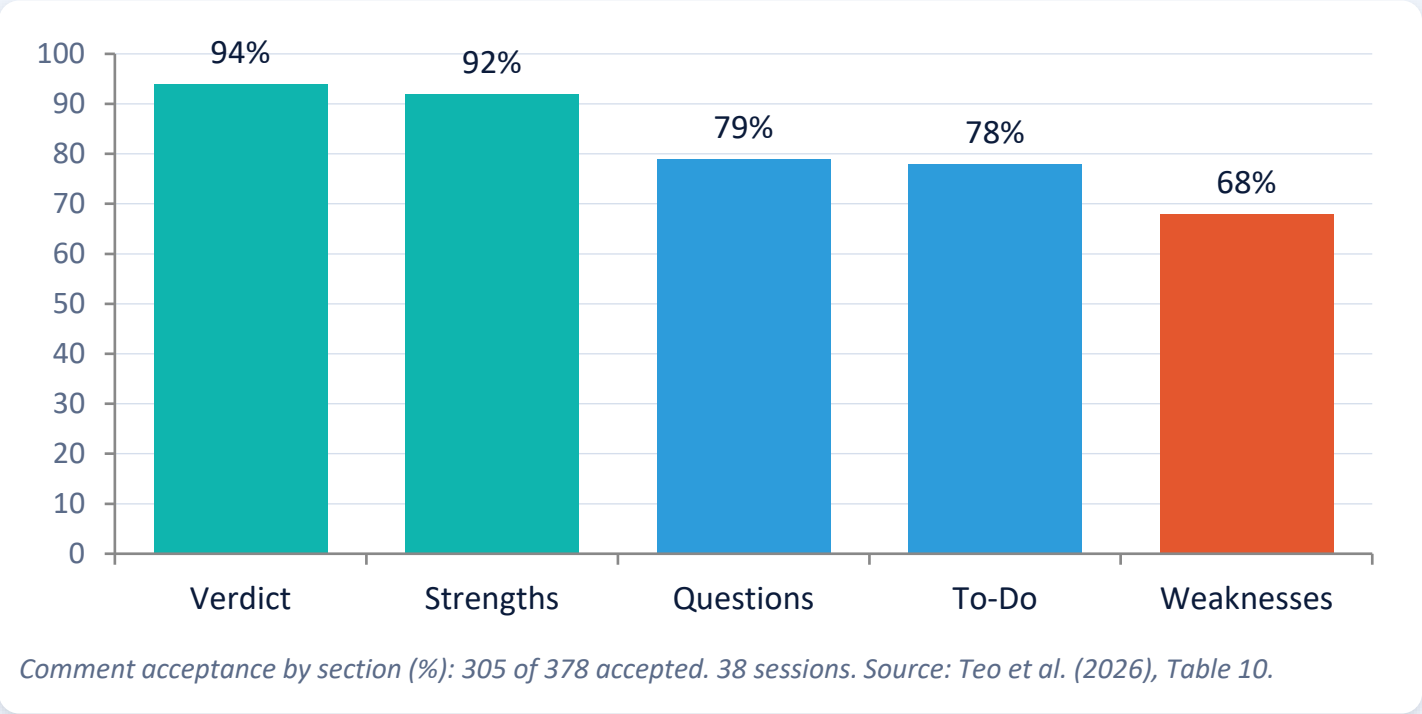
Positive implications

- ◆ MLR costs less than AgentReview
- ◆ Half the tokens of AI Reviewer

Negative implications

- LLM-Review is 47× cheaper
- Sonnet's per-token price is higher

User study: 81% acceptance



Metric (1–5)	Mean
Perceived benefit	4.25
Manuscript improvement	4.04
Reuse intention	3.74
Usefulness	3.70
Criticality (3 = balanced)	2.88

Why it matters

The only test of the full three-agent system, but qualitative.

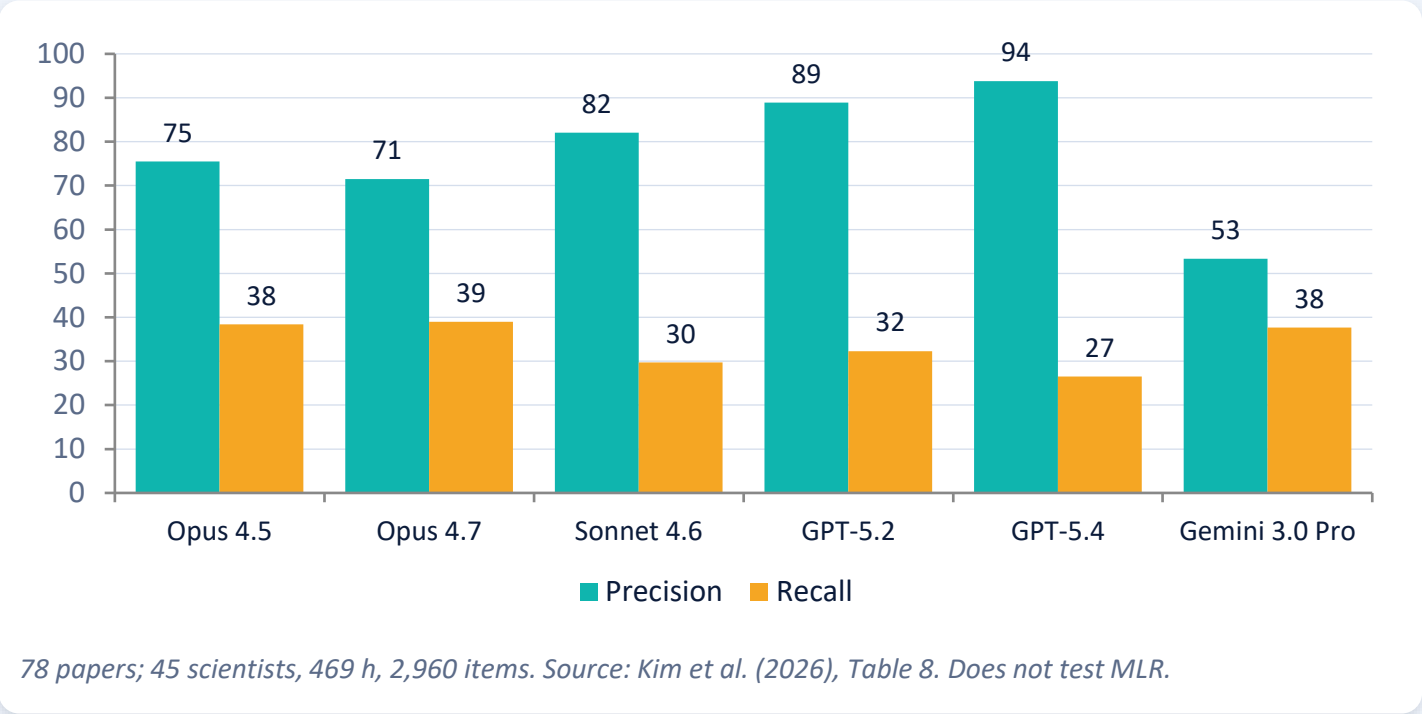
Positive implications

- ◆ High acceptance on strengths and recommendation
- ◆ Authors report manuscript improvement

Negative implications

- Weaknesses, which carry the errors: 68%
- No control group; self-selection

PeerReview Bench: real precision and recall



Reviewer	P	R	F1
Claude Opus 4.5	75.49	38.39	50.89
Claude Opus 4.7	71.47	39.00	50.46
Claude Sonnet 4.6	82.03	29.74	43.65
GPT-5.2	88.92	32.28	47.37
GPT-5.4	93.81	26.55	41.38
Gemini 3.0 Pro	53.35	37.65	44.14

Why it matters

Gives the order of magnitude of precision and recall that Sakana's paper does not measure.

Positive implications

- ◆ Frontier AI reviewers are useful
- ◆ High precision in some models

Negative implications

- Recall of 27–39% of validated critiques
- Study does not test MLR

Quality, overlap and failure modes

Indicator	Result
Item correctness	AI 82–86% vs top human 92.3%
Fully positive items	GPT-5.2 60.0% vs top human 48.2%
Overlap among reviewers	AI-AI ≈ 21% vs human-human ≈ 3%
AI items with no human match	26%, of which 81.8% correct

Failure modes (comment count)

- Misaligned field norms (54+46)
- Misreading: 'paper does not say X' (37)
- Redundancy among AI reviewers (28)
- Vague or verbose comments (24)

Strengths (comment count)

- ◆ Statistical rigor (45)
- ◆ Code inspection (28)
- ◆ Domain depth (27)
- ◆ Internal consistency (15)

Why it matters

An AI-only panel would reduce the diversity of review perspectives.

Positive implications

- ◆ AI finds 26% of items humans miss
- ◆ Strong on statistics and code

Negative implications

- 14–18% of items incorrect
- Redundancy among AI reviewers

Claims and evidence matrix

Claim	Strength	Caveat
73% on central claims	High (synthetic)	Ensemble and favorable slice; total 40.95%
"Three agents"	Low	The benchmark uses only the Review Agent
Better than baselines	High	Different models and truncation; judge from another family
Design adds beyond the model	Moderate	Model confounded with truncation; subset
Aligns with human scores	Moderate	0.43–0.59 vs 0.68–0.78 human reference
Detects real errors	Low to moderate	16.11% (exact) and 26.07% (similar)
More robust to injection	Low	8 of 50; high standard deviation
Benefits researchers	Suggestive	38 sessions, no control; weaknesses 68%
Low cost	High	4-review ensemble multiplies cost ~4×
Precision of flags	None	Not measured: the central gap

Why it matters

Each claim gets an evidence strength, not just a number.

Positive implications

- ◆ Separates what holds from what does not
- ◆ Makes the judgment explicit

Negative implications

- Strength rating is own judgment
- May change with new data

Strengths, cautions and the headline

- ◆ Strengths: TMLR, open preprint, model ablation, validated judge, limitations discussed
- Caution: vendor authorship, own benchmark, data 'on request'
- ▲ Caution: precision not measured; judge from another family; secondary press repeats the headline
- Precise wording: one Claude Sonnet 4 agent, 4 reviews, detected 73.43% of synthetic contradictions in the main claim
- ✦ On real errors: 16–26%; false alarms not measured

Why it matters

Credibility is high for what is measured and low for what the headline suggests.

Positive implications

- ◆ Result reproducible in principle
- ◆ Peer-reviewed publication

Negative implications

- Vendor authorship
- Data and code only on request

Limitations acknowledged by the authors

- ▲ Synthetic errors may be easier than genuine ones
- On real errors the lead over baselines is less pronounced
- ▶ Analysis limited to ML; no comparison with proprietary systems
- ✦ All systems, including MLR, are vulnerable to manipulation
- Deep analysis only on the main text up to 10 pages; code and data on request

Why it matters

Stating limits is part of the result's credibility.

Positive implications

- ◆ Transparency from the authors
- ◆ Facilitates focused replications

Negative implications

- Independent replication depends on the authors
- Scope restricted to ML

Threats to validity (own analysis)

Type	Threat	Likely effect
Construct	Synthetic ≠ real error; 'at least 1 of k' favors sensitivity	Overstates real-world recall
Construct	Precision not measured	Unknown cost of triaging false alarms
Internal	Model and truncation confounded; baselines with different models	Uncertain gain attribution
Internal	'Similar' criterion defined after seeing results	Possible optimism (mitigated)
External	ML, 10 pages, 2025 Sonnet 4	Limited generalization
Statistical	~250 items at d0; 18 at d6; judge sensitivity 86.8%	Wide intervals; recall possibly underestimated

Why it matters

Shows where the result may be biased and in which direction.

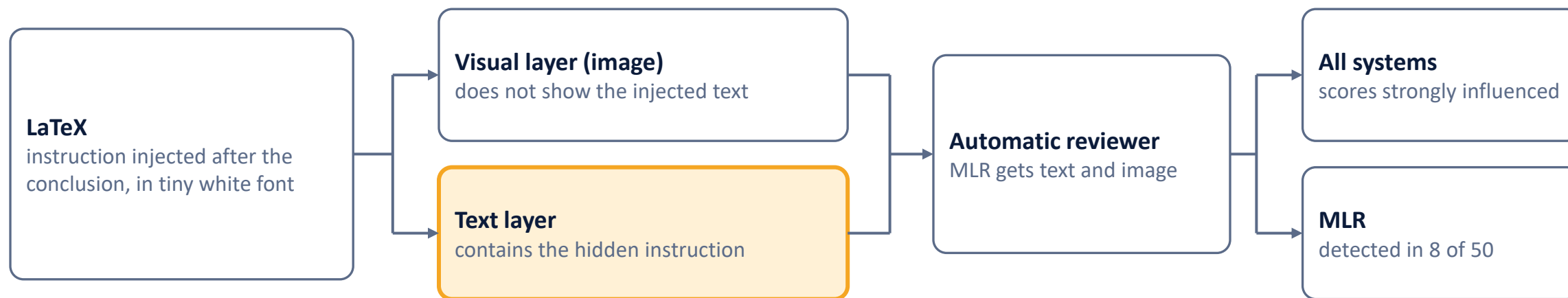
Positive implications

- ◆ Makes the biases explicit
- ◆ Points to what to measure next

Negative implications

- Several biases push recall upward
- Few biases push it downward

Prompt injection: hidden text in the PDF



Authors' hypothesis: MLR's design and text-plus-image input explain the partial detection.

Source: Teo et al. (2026), Section 4.5 and Appendix A.6.

Why it matters

The instruction exists in the PDF text layer, but not in the page image.

Positive implications

- ◆ Text-plus-image input helps detection
- ◆ Possible defense: compare the two layers

Negative implications

- 18 real manuscripts with hidden prompts (Jul/2025)
- Explicit detection in only 8 of 50

Gaps: closed and remaining

Gap	Status	Note
Table 14, ensembles, chain ablation, A.6, AAAI-26 source	Closed	Transcribed and checked in the paper
MLR precision	Open	External evidence only sizes other reviewers
Exact prompt text	Partial	Only output formats of passes 2 and 3
Figure 28 (1 to 6 reviews)	Open	Curve values not recovered
Appendices B.2 to B.4; A.5; A.7	Open	Only the variant's summary result
Pages beyond the 10th in WithdrarXiv	Open	Handling not clarified
TMLR reviews (OpenReview)	Open	Access blocked
Original Substack post	Open	Not located

Why it matters

Saying what is unknown is part of the result.

Positive implications

- ◆ Five gaps closed with the paper
- ◆ Objective list of what is missing

Negative implications

- Precision remains unanswered
- Full prompts still unavailable

Outlook for improvement

Proposed by the authors

- ▶ Benchmark with more natural errors (retractions)
- ▶ Domains beyond ML
- ▶ Robustness to manipulation
- ▶ Compare with proprietary systems
- ▶ Literature agent with stronger retrieval (PaperQA)
- ▶ Swap the model to cut cost

Proposals of this analysis (own inference)

- Measure precision and false alarms
- Report single-review recall with CIs
- Judges from several families
- Injection defenses (text vs image)
- Code execution and reproducibility
- Long documents, figures and tables
- Re-evaluate with newer models
- Isolate the chain effect
- Calibrate severity by field

Why it matters

The agenda separates what the authors propose from what this analysis adds.

Positive implications

- ◆ Clear path to close the precision gap
- ◆ Practical defenses against injection

Negative implications

- Items from the analysis are own inference
- Require data and human annotation

Responsible use and governance

- Treat MLR as support for the human reviewer, with disclosure of use
- ▶ Human oversight and compliance with each venue's rules
- ▲ AI-only panels reduce diversity of perspectives (Kim et al., 2026)
- ◆ Author-side risk: polishing manuscripts against automated checks
- Reviewer-side risk: delegating the review without reading the paper

Why it matters

Without governance, an efficiency gain becomes a scientific-integrity risk.

Positive implications

- ◆ ICLR, NeurIPS, ICML and AAAI policies already exist
- ◆ Human in the loop preserves accountability

Negative implications

- Undeclared use and rule-breaking
- Generous scores and prompt injection

Reproduction and deployment requirements

Requirement	Detail
Models	Claude Sonnet 4 and Haiku 3.5; for the benchmark: GPT-4.1, Gemini 2.5 Pro and o3
Data	LaTeX of arXiv papers under CC licenses; LaTeX compilation
Prompts	Figures 12, 13, 17 and 18 to 22 of the paper's appendix
Input	PDF with text and image; main text up to 10 pages; ML
Cost	≈ US\$ 0.47 per review; ≈ US\$ 1.9 with 4 reviews; +≈ US\$ 0.97 with literature
Availability	Data and code on request from the authors
Governance	Disclosure of use, human oversight, venue rules

Why it matters

Anyone adopting it should measure recall and precision on their own real errors.

Positive implications

- ◆ No GPU or training
- ◆ Predictable cost per paper

Negative implications

- Exact replication needs the paper's versions
- Data and code only on request

Conclusion

MLR is solid at finding contradictions planted in the central claim, but the 73% headline describes the best case.

What holds

- ◆ 73.43% at d0 (4 reviews, synthetic)
- ◆ Lead over baselines
- ◆ Cost $\approx$ US\$ 0.47 per review

What does not hold

- 'Three agents': it was only one
- 40.95% overall; 16–26% on real errors
- False alarms: not measured

How to use it

- ▲ Second read, with a human in the loop
- ▲ Measure precision on your own data
- ▲ Defend against prompt injection

Why it matters: The main gap to close is measuring false alarms.

News sources

Primary source

Teo, Yamada, Kotyan, Imajuku, Clanuwat (2026). Beyond Imitation: A Framework and Benchmark for LLM-Assisted Peer Review. TMLR (07/2026); arXiv:2610.11087 (8 Oct 2026). arxiv.org/abs/2610.11087 · openreview.net/forum?id=7iX2Z2bPFB

Secondary coverage (to be checked)

- ▲ MarkTechPost (10 Oct 2026). Sakana AI's LLM Peer Review System Catches 73% of Core-Claim Errors. marktechpost.com
- ▲ FourWeekMBA (2026). Sakana AI Reviewer Catches 73% of Planted Main-Claim Errors. fourweekmba.com
- ▲ Beri (12 Oct 2026). Sakana's AI Reviewer Caught 73% of Planted Errors, 16% of Real Ones. beri.net
- ▲ Remio (2026). Sakana AI Peer Review Catches 73% of Core Errors, but Real Papers Expose the Limit. remio.ai
- ▲ Tech AI Magazine (2026). Sakana AI's Review System Achieves 73% Error Detection Rate. techaimag.com

The original Substack post cited in the claim was not located; verification relies on the paper and the sources above.

References (1/3)

- Anthropic (2024). Claude 3 model card.
- Anthropic (2025). System card: Claude Opus 4 & Claude Sonnet 4.
- Avaliação independente do AI Scientist (2025). Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future? arXiv:2502.14297.
- Benchmark centrado em processo (2026). Beyond Final Decisions: A Process-Centric Benchmark for Transparent AI-Assisted Peer Review. arXiv:2609.05947.
- Beri, R. (2026). Sakana's AI Reviewer Caught 73% of Planted Errors, 16% of Real Ones. beri.net (secondary).
- Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; Wortman Vaughan, J. (2023). The NeurIPS 2021 consistency experiment. arXiv:2306.03262.
- Biswas, J. et al. (2026). AI-assisted peer review at scale: The AAI-26 AI review pilot. arXiv:2604.13940.
- Bougie, N.; Watanabe, N. (2025). Generative reviewer agents: Scalable simulacra of peer review. EMNLP 2025 Industry, 98–116.
- Chiang, C.-H.; Lee, H. (2023). Can large language models be an alternative to human evaluations? ACL 2023.
- Cortes, C.; Lawrence, N. D. (2021). Inconsistency in conference peer review: revisiting the 2014 NeurIPS experiment. arXiv:2109.09774.
- D'Arcy, M.; Hope, T.; Birnbaum, L.; Downey, D. (2024). MARG: Multi-agent review generation for scientific papers. arXiv:2401.04259.
- FourWeekMBA (2026). Sakana AI Reviewer Catches 73% of Planted Main-Claim Errors (secondary).
- Gao, X. et al. (2025). ReviewAgents: Bridging the gap between human and AI-generated paper reviews. arXiv:2503.08506.
- Gao, Z.; Brantley, K.; Joachims, T. (2024). Reviewer2: Optimizing review generation through prompt generation. arXiv:2402.10886.
- Hidden Prompts (2025). Hidden prompts in manuscripts exploit AI-assisted peer review. arXiv:2507.06185.

References (2/3)

- Jin, Y. et al. (2024). AgentReview: Exploring peer review dynamics with LLM agents. EMNLP 2024, 1208–1226.
- Keshav, S. (2007). How to read a paper.
- Kim, S. et al. (2026). On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. arXiv:2605.20668.
- Lála, J. et al. (2023). PaperQA: Retrieval-augmented generative agent for scientific research. arXiv:2312.07559.
- Liang, W. et al. (2024a). Can large language models provide useful feedback on research papers? NEJM AI 1(8), A10a2400196.
- Liu, R.; Shah, N. B. (2023). ReviewerGPT? An exploratory study on using LLMs for paper reviewing. arXiv:2306.00622.
- Lu, C. et al. (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. arXiv:2408.06292.
- Lu, C. et al. (2026). Towards end-to-end automation of AI research. Nature 651(8107), 914–919.
- MarkTechPost (2026). Sakana AI's LLM Peer Review System Catches 73% of Core-Claim Errors. 10 Oct 2026 (secondary).
- POMP (2026). Jagged AI in Scientific Peer Review: Evidence from POMP Data Analysis. arXiv:2605.07855.
- PRISM (2026). PRISM: A Multi-Dimensional Benchmark for Evaluating LLM Peer Reviewers. arXiv:2605.26730.
- R&D World (2025). Sakana AI claims first fully AI-generated paper passes peer review at top AI conference (secondary).
- Rao, D. et al. (2024). WithdrarXiv: A large-scale dataset for retraction study. arXiv:2412.03775.
- ReviewBench (2026). An Extensible Framework for Benchmarking Human and AI Manuscript Review. bioRxiv 10.64898/2026.04.17.719279.
- ReviewGrounder (2026). Improving Review Substantiveness with Rubric-Guided, Tool-Integrated Agents. arXiv:2604.14261.

References (3/3)

- Robertson, Z. (2023). GPT-4 is slightly helpful for peer-review assistance. arXiv:2307.05492.
- Russo, G. et al. (2025). The AI review lottery. Proc. ACM HCI 9(7), 1–28.
- Shin, H. et al. (2025). Mind the blind spots: A focus-level evaluation framework for LLM reviews. EMNLP 2025, 35618–35644.
- Skarlinski, M. D. et al. (2024). Language agents achieve superhuman synthesis of scientific knowledge. arXiv:2409.13740.
- Son, G. et al. (2025). SPOT: A benchmark for automated verification of scientific research. arXiv:2505.11855.
- Teo, R. S. Y.; Yamada, Y.; Kotyan, S.; Imajuku, Y.; Clanuwat, T. (2026). Beyond Imitation: A Framework and Benchmark for LLM-Assisted Peer Review. TMLR 07/2026; arXiv:2610.11087.
- Thakkar, N. et al. (2025). Can LLM feedback enhance review quality? A randomized study of 20k reviews at ICLR 2025. arXiv:2504.09737.
- Wallis, S. (2013). Binomial confidence intervals and contingency tests. J. Quantitative Linguistics 20(3), 178–208.
- Xi, S.; Rao, V.; Payan, J.; Shah, N. B. (2025). FLAWS: A benchmark for error identification and localization in scientific papers. arXiv:2511.21843.
- Ye, R. et al. (2024). Are we there yet? Revealing the risks of utilizing LLMs in scholarly peer review. arXiv:2412.01708.
- Yuan, W.; Liu, P.; Neubig, G. (2022). Can we automate scientific reviewing? JAIR 75, 171–212.
- Zhang, T. M.; Abernethy, N. F. (2025). Reviewing scientific papers for critical problems with reasoning LLMs. arXiv:2505.23824.
- Zhu, M.; Weng, Y.; Yang, L.; Zhang, Y. (2025). DeepReview: Improving LLM-based paper review with human-like deep thinking process. ACL 2025, 29330–29355.

Glossary (1/2)



MLR

Multi-Layered Review: Sakana AI's layered review system



Distance (d)

Shortest graph path to a main claim; d0 is the central one



Synthetic contradiction

Logical error inserted in a controlled way by an LLM



Ensemble of k reviews

k independent reviews; counts if any one hits



Recall

Fraction of known errors the reviewer finds



Precision

Fraction of the reviewer's flags that are correct



LLM judge

Model that decides whether the review detected the error



KL (Kullback-Leibler)

Divergence between distributions; compares reviewer focus



Prompt injection

Hidden instruction in the document that tries to manipulate the reviewer

Glossary (2/2)

WithdrarXiv-Check

Papers withdrawn from arXiv with retraction comments

PeerReview Bench

Scientist-validated benchmark, with precision and recall

Three-Pass Approach

Keshav's three-pass reading method (2007)

Knowledge graph

Network of the paper's claims, evidence, methods and implementations

Appendix Agent

Agent (Haiku 3.5) that summarizes the paper's appendix

Review Agent

Agent (Sonnet 4) that runs the 3-pass chain

Truncation

Cutting the input text due to a context limit

Wilson interval

Confidence interval for proportions in small samples

Retraction

Formal withdrawal of a paper due to error or misconduct

Verification note and copyright

Primary sources read

Main text and appendices A.1 to A.4 of the paper, excerpts of A.5, A.6 and B.1 (Teo et al., 2026), and the main text of Kim et al. (2026).

Secondary sources

MarkTechPost, FourWeekMBA and Beri: used only for cross-checking.

Not verified

Original Substack, TMLR reviews on OpenReview, Figure 28 and Appendices A.5, A.7, B.2 to B.4 in detail.

Own calculations

Flagged on the slides as 'own calculation' or 'own inference'.

© Aridio Silva · @aridiosilva · 13-October-2026

Aridio Silva, Independent Researcher · ORCID 0009-0008-2411-6995 · Text, tables, charts, diagrams and images in this infographic are authored by Aridio Silva.