Carnegie Mellon University

AI Agents for Science:

What Can We Delegate, and Where Do We Need Experts?

Akari Asai

Carnegie Mellon University / Allen Institute for AI

akariasai.github.io aasai@andrew.cmu.edu

AI4Bio Workshop, November 13, 2026

Is AI actually helping us do science?

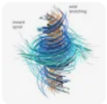
Impressive results

OpenAI

On the Navier–Stokes Millennium Prize Problem

The problem. The Navier–Stokes equations use Newton's second law of motion ("F=ma") to describe how fluids move. Importantly, they treat a fluid...

2026/09/08



Google DeepMind

AlphaEvolve: How our Gemini-powered coding agent is scaling impact across fields

Driving social impact and sustainability. AlphaEvolve has helped uncover key connections in health and sustainability research. A 3D rendering...

2026/05/07



Is AI actually helping us do science?

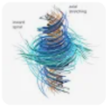
Impressive results

OpenAI

On the Navier–Stokes Millennium Prize Problem

The problem. The Navier–Stokes equations use Newton's second law of motion ("F=ma") to describe how fluids move. Importantly, they treat a fluid...

2026/09/08



Google DeepMind

AlphaEvolve: How our Gemini-powered coding agent is scaling impact across fields

Driving social impact and sustainability. AlphaEvolve has helped uncover key connections in health and sustainability research. A 3D rendering...

2026/05/07



Persistent failures

The Lancet

Fabricated citations: an audit across 2·5 million biomedical papers

Scientific literature depends on the integrity of its references. Each reference implicitly asserts that a verifiable source exists and...

2026/05/09

Chemical & Engineering News

Can we stop AI from flooding scientific journals?

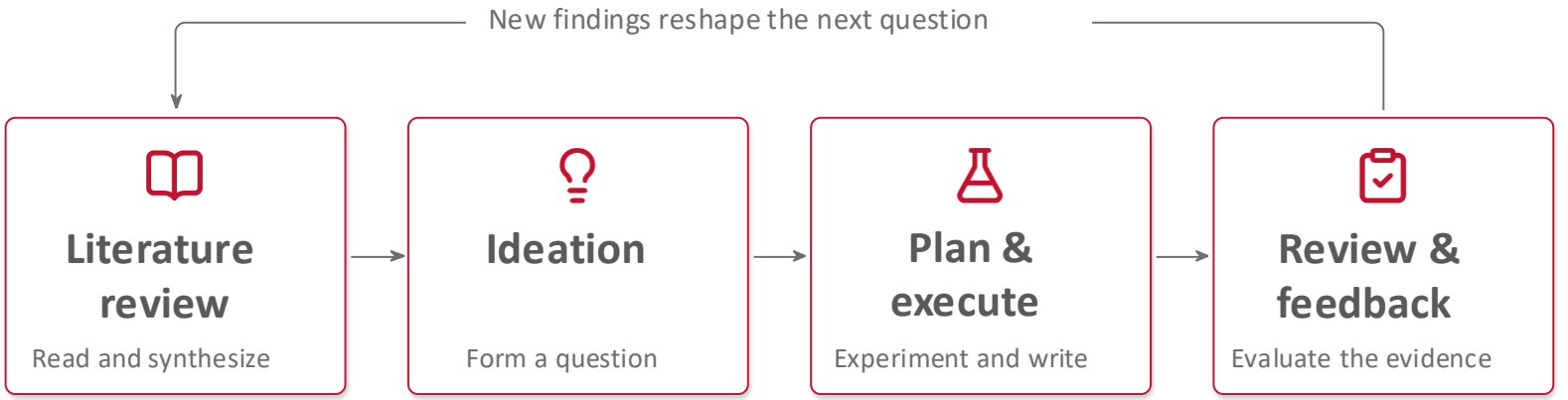
AI detectors may never beat fraud. Changing how science rewards its researchers might be the only fix that does.

2026/07/28

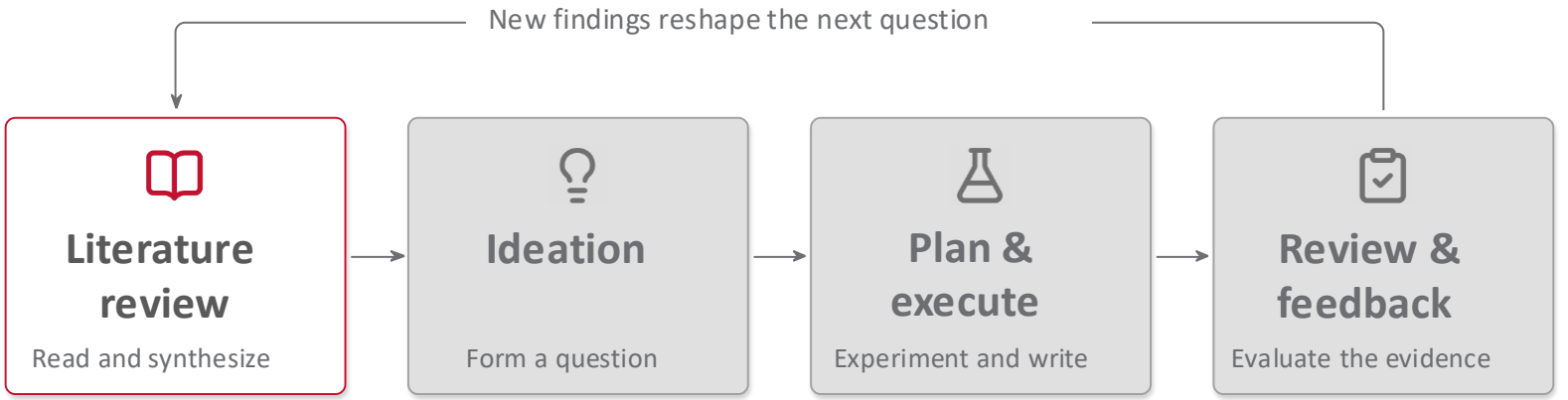


What can we delegate to AI?

AI across the scientific workflow



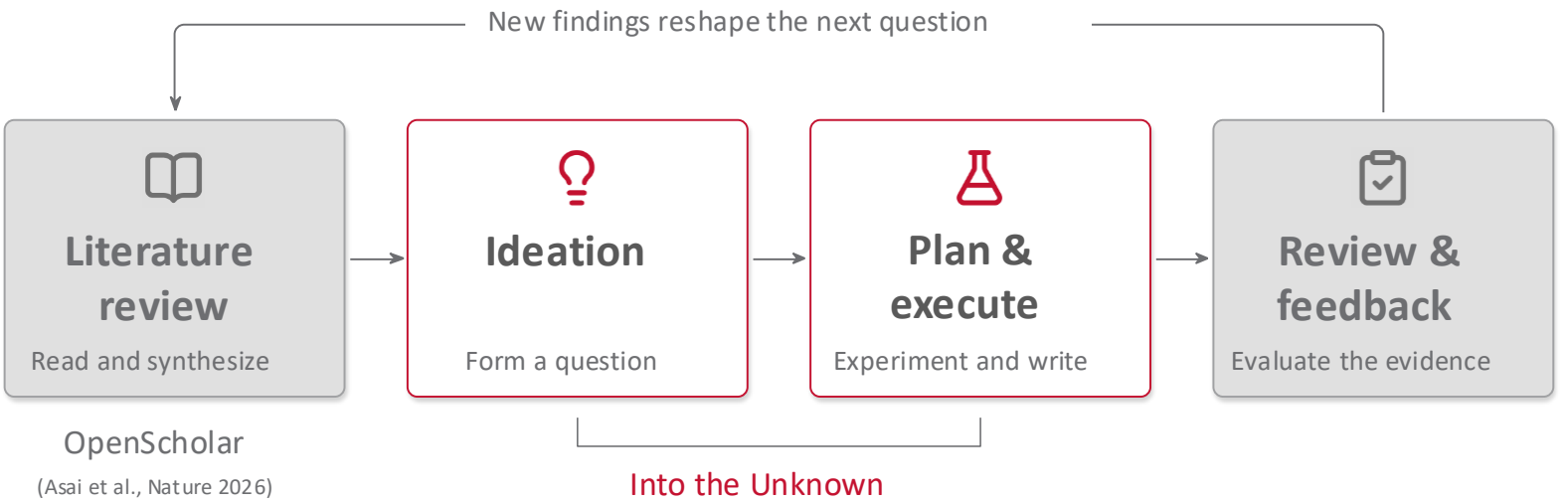
AI across the scientific workflow



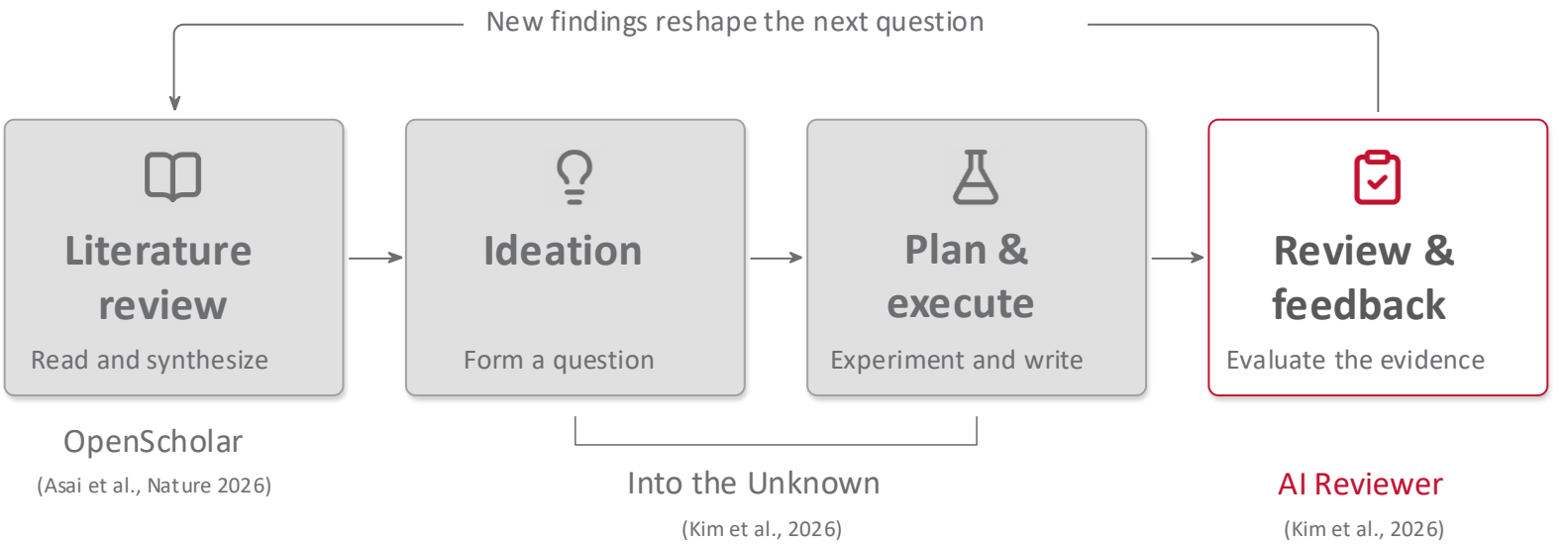
OpenScholar

(Asai et al., Nature 2026)

AI across the scientific workflow



AI across the scientific workflow



Can AI answer literature questions?

Synthesizing scientific literature with retrieval-augmented language models

Asai et al. · Nature, 2026

OpenScholar: answering literature questions

How does the gut microbiome affect response to anti-PD-1 immunotherapy?



OpenScholar

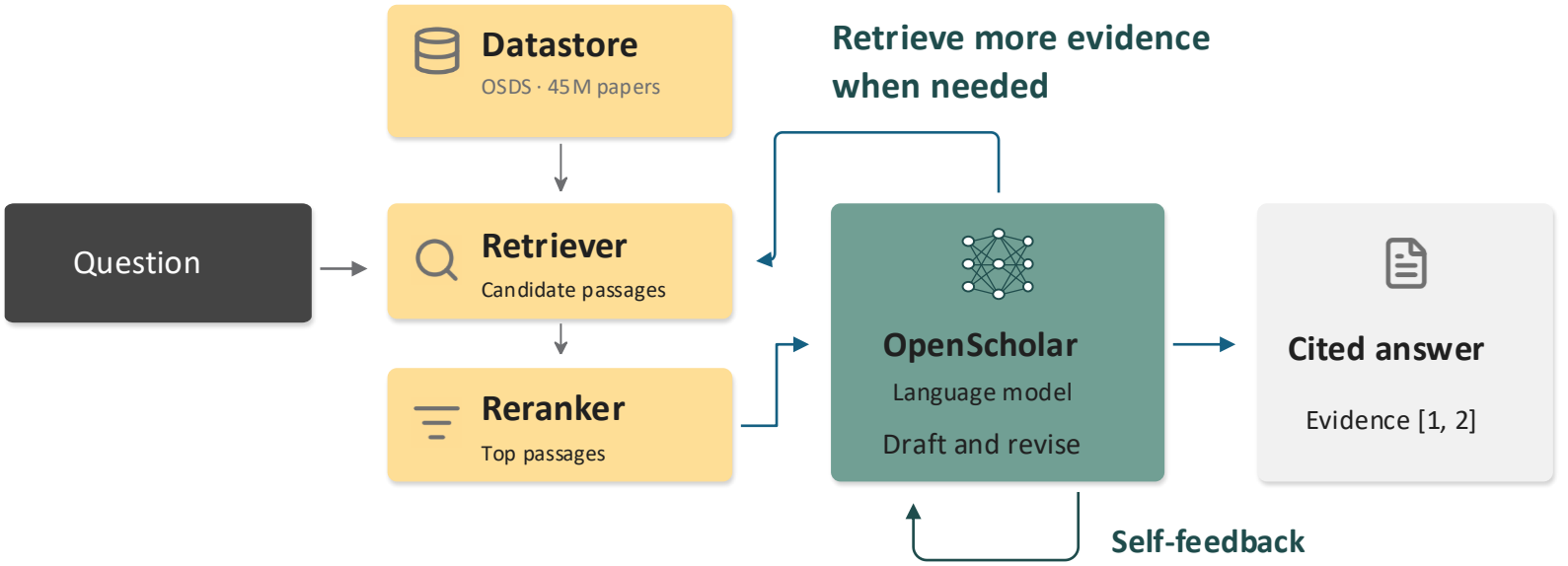
Patient evidence. Microbial composition is associated with response to anti-PD-1 treatment. [1, 2]

Experimental evidence. Responder microbiota improves treatment efficacy in mice. [1, 2]

[1] Routy et al., Science 2018. [2] Gopalakrishnan et al., Science 2018.

Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

OpenScholar: retrieval and self-feedback



Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

ScholarQABench: evaluate literature synthesis

Understand one paper

Prior work

SciFact	PubMedQA	QASA
Biomed.	Biomed.	CS

Output

Labels / short answers

Metric: **Answer match**

Synthesize multiple papers

Ours

Scholar-CS	Scholar-Bio	Scholar-Neuro	Scholar-Multi
CS	Biomed.	Neuro.	CS / Bio / Physics

Output

Long-form answers with citations

Metric: **Rubrics**

Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

Rubrics make answer quality inspectable

How does the gut microbiome affect response to anti-PD-1 immunotherapy?

OpenScholar

Microbial composition is associated with patient response. [1, 2]

Responder microbiota improves treatment efficacy in mice. [1, 2]

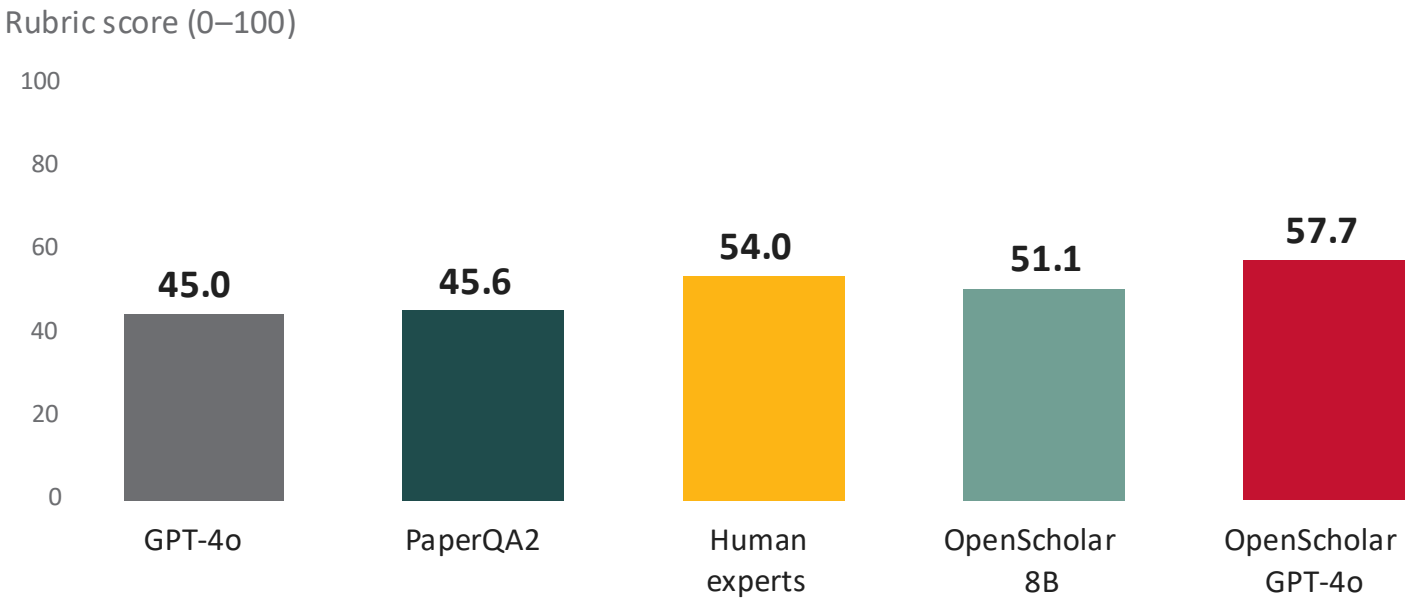
Rubrics

Content requirement	Check
Human response association	1
Mouse-transfer evidence	1
Specific bacterial taxa	0

Rubric score: $\frac{1+1+0}{3} = 0.67$

Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

OpenScholar improves rubric-based answer quality



Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

Plausible answers can cite nonexistent papers

Fabricated cited paper titles (%) ↓

Model	CS	Biomed.
OpenScholar-8B	**0%**	**0%**
Llama 3.1 8B	**92.1%**	**97.6%**
Llama 3.1 70B	**78.1%**	**96.6%**
GPT-4o	**78.7%**	**94.8%**

Baselines: no retrieval. OpenScholar uses retrieval.

GPT-5 without retrieval **≈40%**

Follow-up analysis: 39% reported in the paper.

How do autoregressive language models compare with bidirectional retrievers?

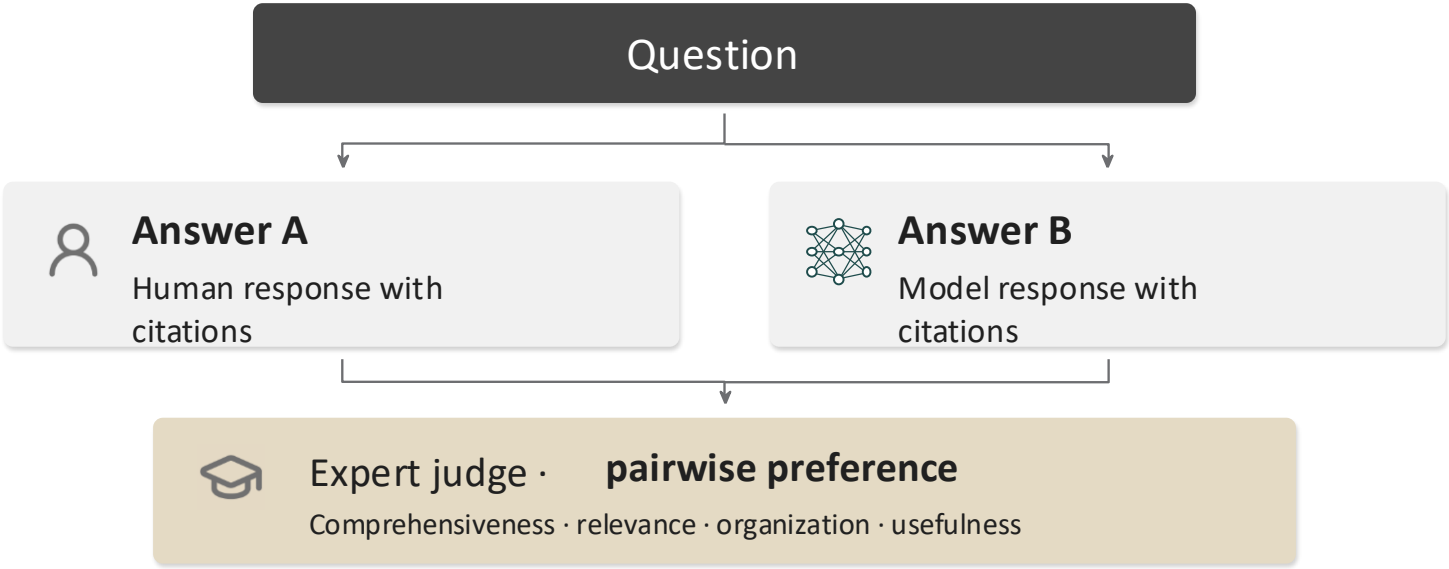
The response claims competitive performance and cites several apparently relevant studies.

Example citations

[1] Dense Passage Retrieval with Auto-regressive Language Models
[2] Named Entity Recognition with Auto-regressive Language Models
[3] Passage Retrieval with Auto-regressive Language Models

These plausible-looking references are fabricated.

Experts compare human and model answers



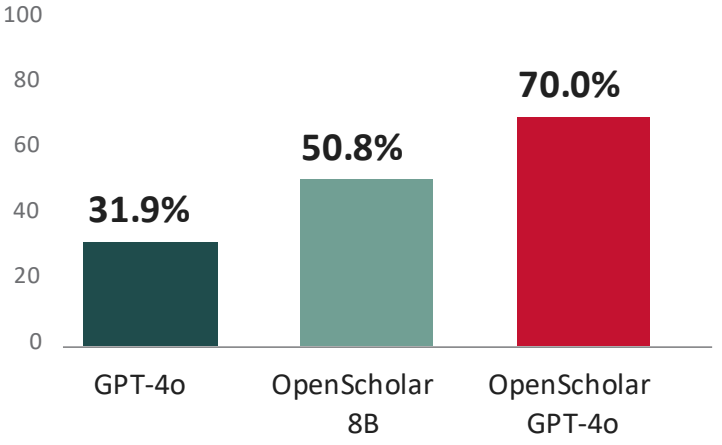
108 questions · 16 expert raters · A / B / tie

Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

Experts prefer broader coverage of the literature

Overall preference

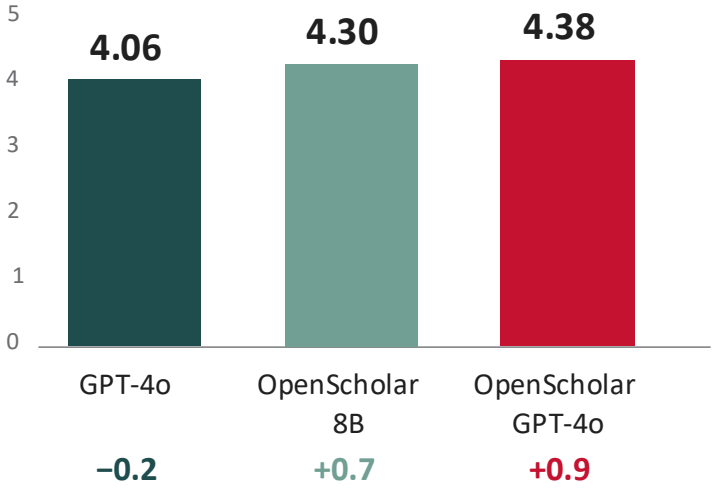
Win rate over expert answers (%)



Comprehensiveness

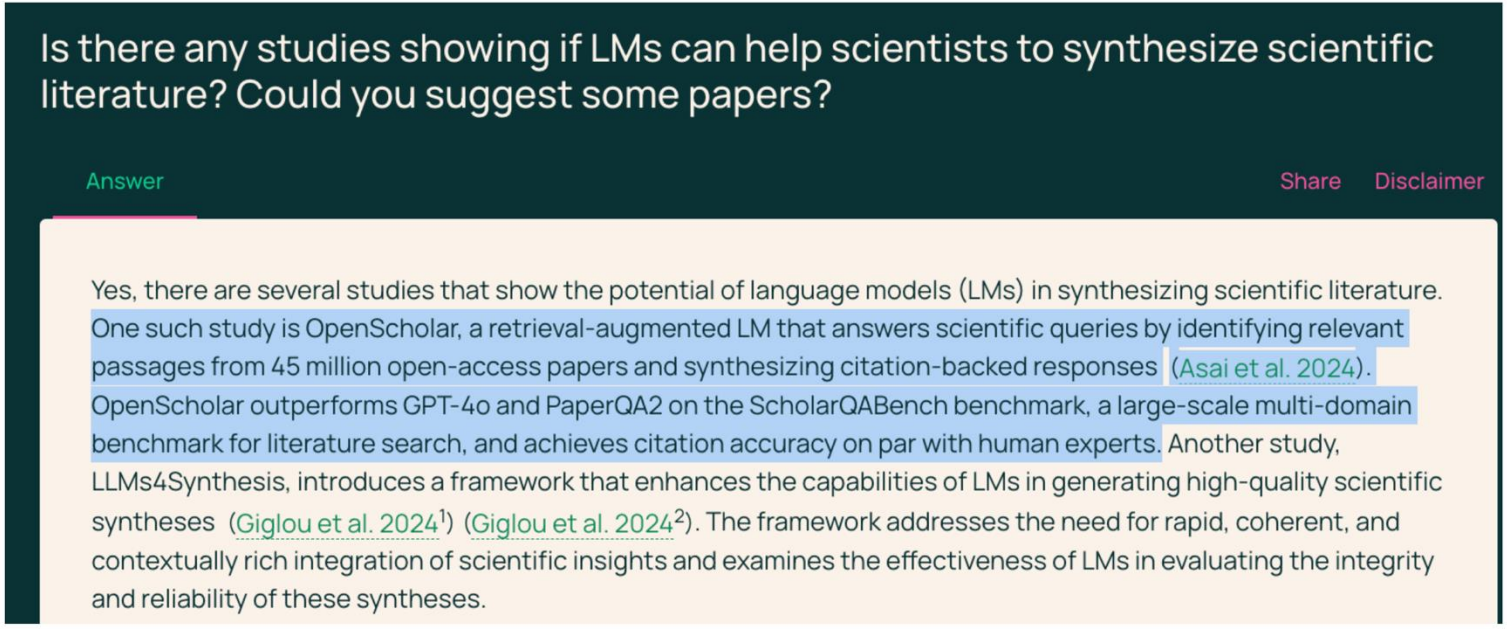
Coverage score (1–5)

△ vs. paired expert answers shown below



Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

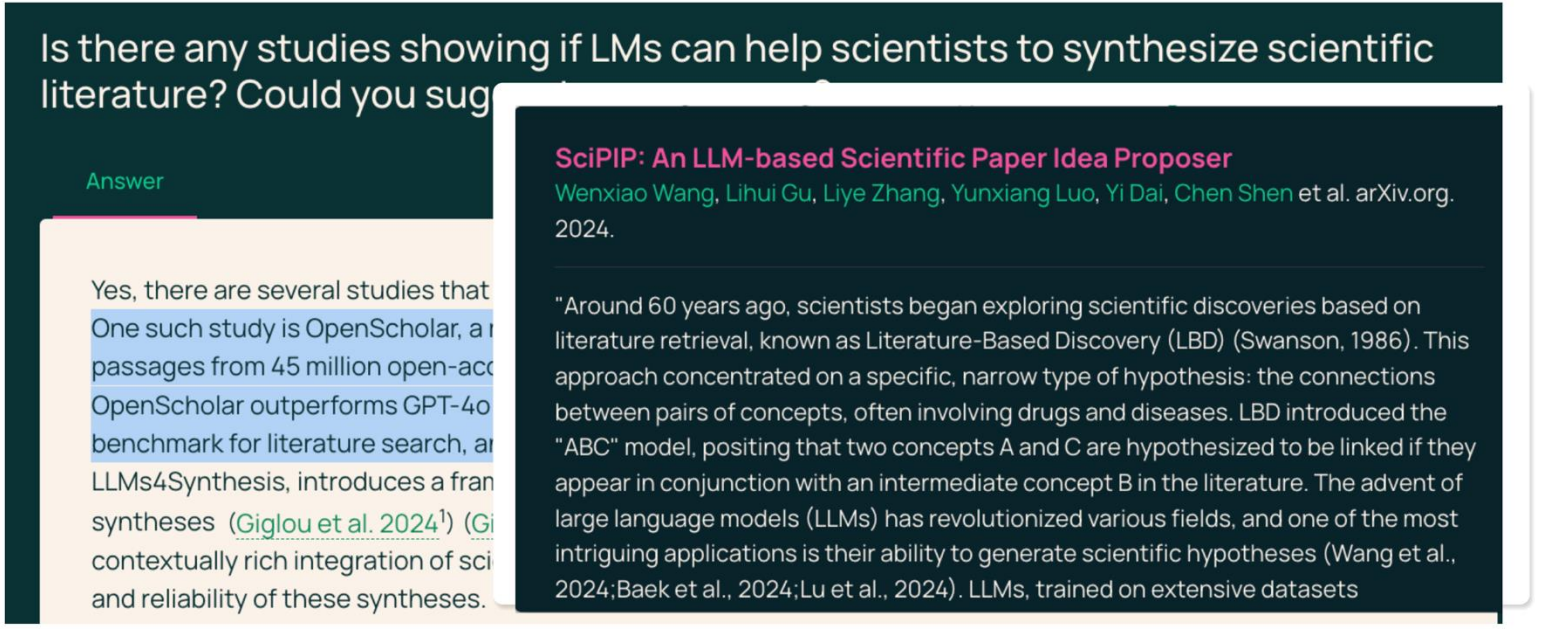
OpenScholar in practice



https://asta.allen.ai/

Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

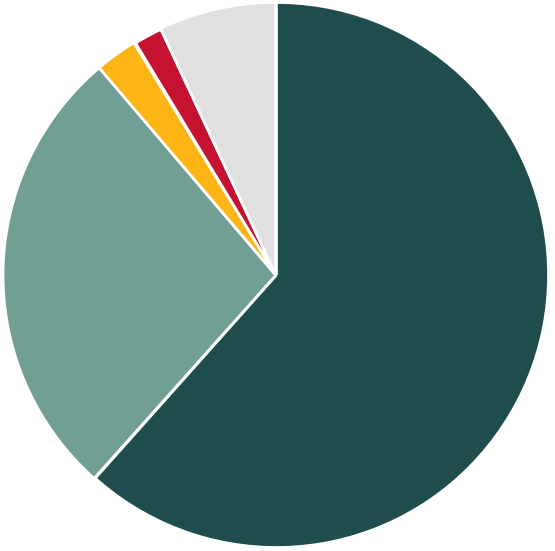
OpenScholar in practice



https://asta.allen.ai/

Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

How people use OpenScholar



57,000 valid demo queries (August 2025)

Understand literature	**61.6%**
Find papers / findings	**27.2%**
Understand one paper	**2.5%**
Ideation	**1.7%**
Other / unreported	**7.0%**

Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

What can we delegate? Literature synthesis

1 OpenScholar + ScholarQABench

A fully-open system and benchmark for literature synthesis, used by 50k+ scientists.

2 Expert-level literature synthesis

AI answers can match or exceed human answers in the evaluated tasks.

Paper: nature.com/articles/s41586-025-10072-4
Code: github.com/AkariAsai/OpenScholar

Asai et al. Synthesizing scientific literature with retrieval-augmented language models. Nature, 2026.

Can AI carry out open-ended research?

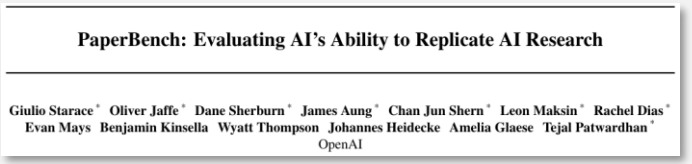
Into the Unknown: Benchmarking AI agents on open scientific research problems

2026

With  Seungone Kim and collaborators

Evaluating AI agents' end-to-end research capabilities

Reproduce a known result


PaperBench: Evaluating AI's Ability to Replicate AI Research

Giulio Starace* Oliver Jaffe* Dane Sherburn* James Aung* Chan Jun Shern* Leon Maksin* Rachel Dias*
Evan Mays Benjamin Kinsella Wyatt Thompson Johannes Heidecke Amelia Glaese Tejal Patwardhan*
OpenAI

Evaluation guided by reference results from 20 ICML papers

Study open-ended research


Can AI agents conduct open-ended AI research?

Early evidence from two case studies

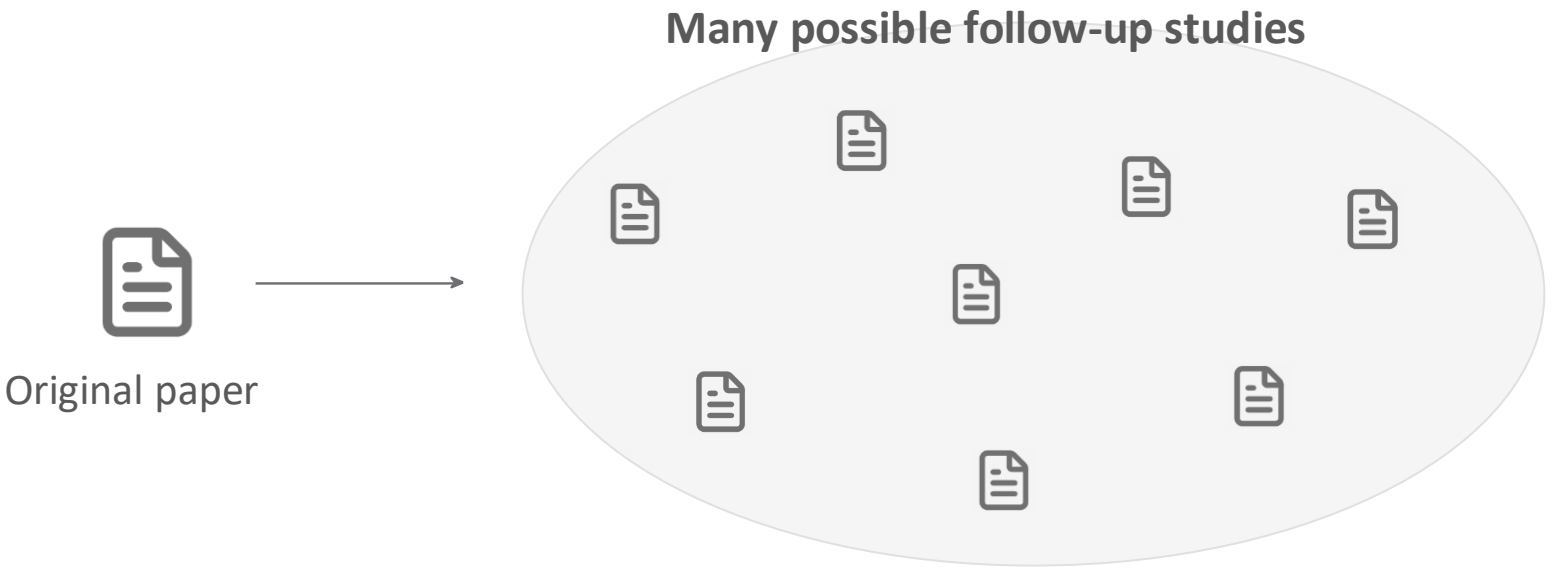
Peter Kirgis[1,*,†] Sayash Kapoor[1,*,†] Andrew Schwartz[1,2,†] Stephan Rabanser[1,†]
David Africa[3] Konstantinos Voudouris[3] Viet Nguyen[4] Toby Pilditch[3] Magda Dubois[3]
Harry Coppock[3] Cozmin Ududec[3] Nitya Nadgir[5] Matilda Orona[6] Tilman Bayer[5]
Derrick Chan-Sew[5] Yue Ling[5] Abhishek Shetty[5] Helen Toner[7] Gillian Hadfield[8]
Seth Lazar[8] Steve Newman[9] Shoshannah Tekofsky[10] Rishi Bommasani[11]
Arvind Narayanan[1,†]

2 Case studies submitted to NeurIPS

How can we evaluate open-ended research at scale?

Starace et al. PaperBench, 2025. Kirgis et al. Two case studies, 2026.

Evaluating AI Agents' abilities of conducting follow-up



Open research has no single reference answer.

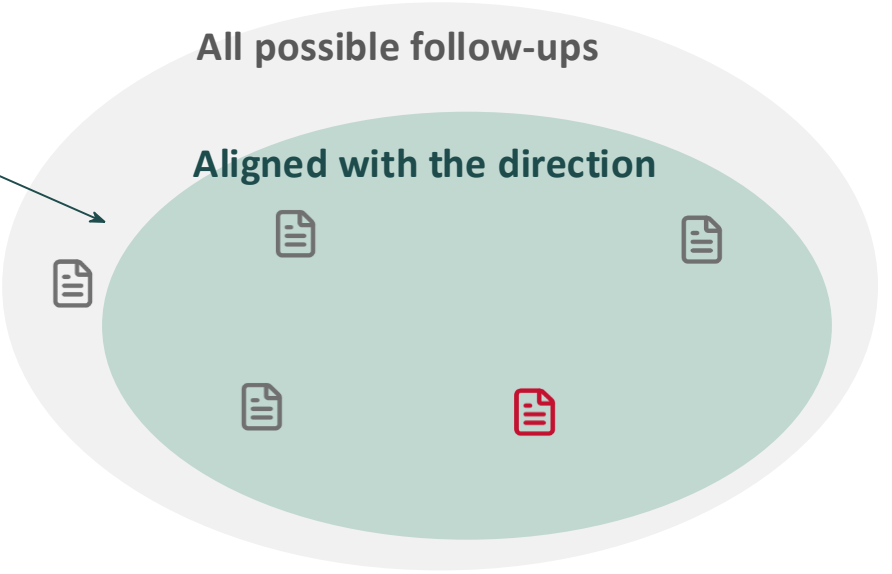
Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

A scientist defines the research direction

Research direction

Task specification

Replace the LCHS kernel to improve resource scaling.



Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Research directions span seven areas

15 open problems · 10 scientists · 7 research areas

Research area	Tasks	Follow-up directions
Quantum computing	**4**	Topology, LCHS kernels, Berry phase, plasma
Scientific machine learning	**3**	Differential equations, PDEs, large Gaussians
Computational chemistry	**2**	Molecular energies, Hessian learning
Materials science	**2**	Atomic site descriptors, switching functions
Plasma physics	**1**	QED regimes, magnetized energy transfer
Computational imaging	**1**	Dictionary learning with random beams
Color science	**1**	Cone–channel hypothesis

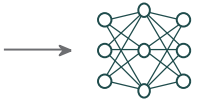
Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Task formulation

Input: prepared workspace

EXPLORER

› workspace/
- task_spec.md
- original_paper.pdf
- › code/
- › data/



AI agent

Output: submission

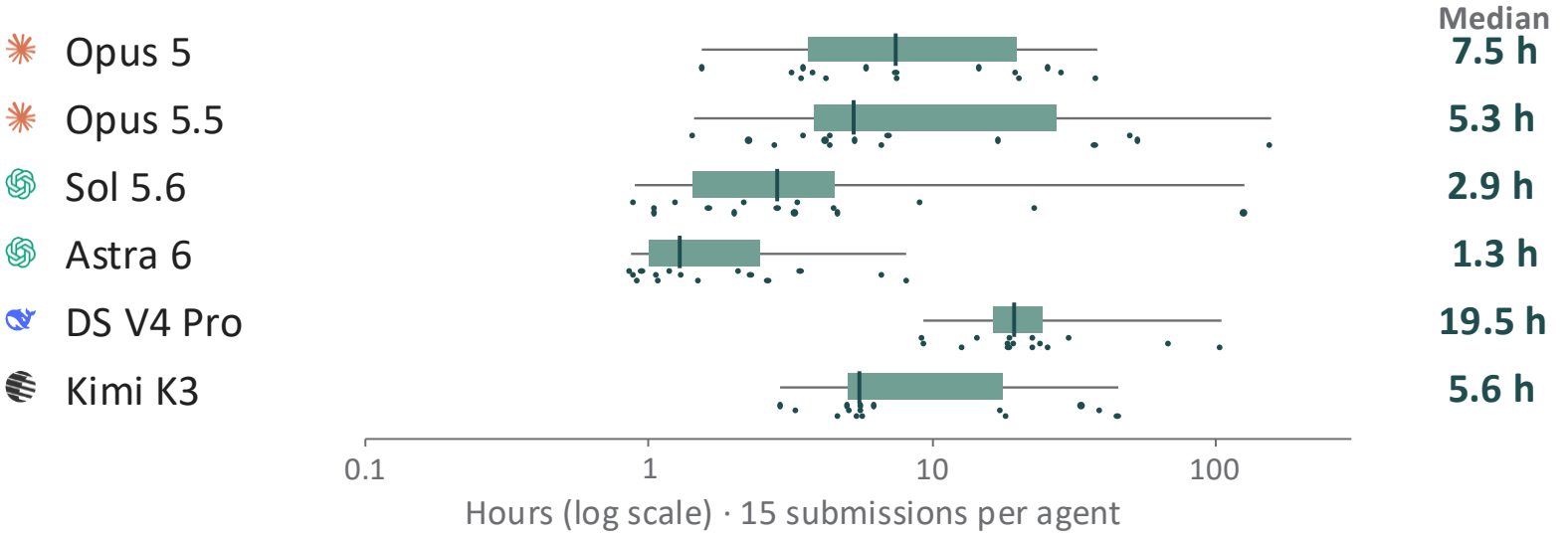
EXPLORER

› workspace/
- › proposal/
- followup_paper.pdf
- › code/
- › logs/

Read files · run code · search · iterate

Research runs range from minutes to days

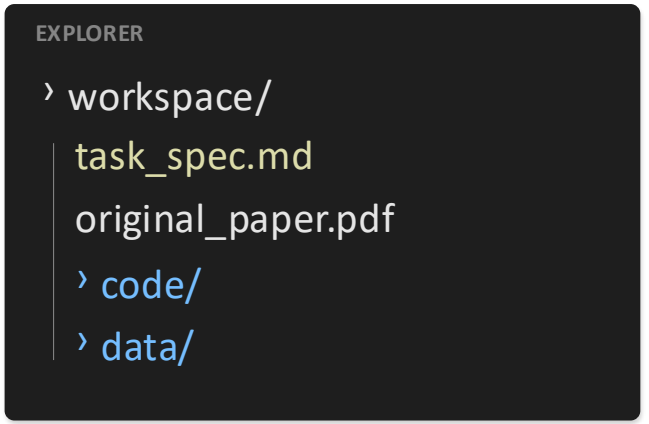
Wall-clock duration varies across tasks and agents

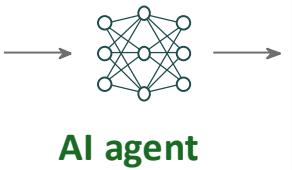


Kim et al. Into the Unknown. 2026. Paper · Results viewer · Submission-level execution times

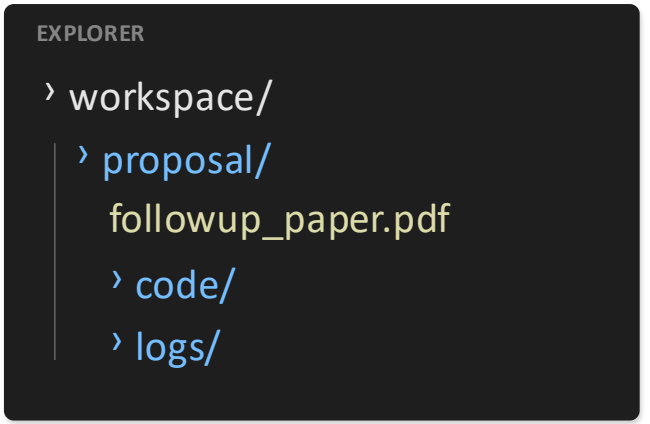
Task formulation

Input: prepared workspace





Output: submission



How can we evaluate agents' submissions?

Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Research direction and success criteria

Research direction

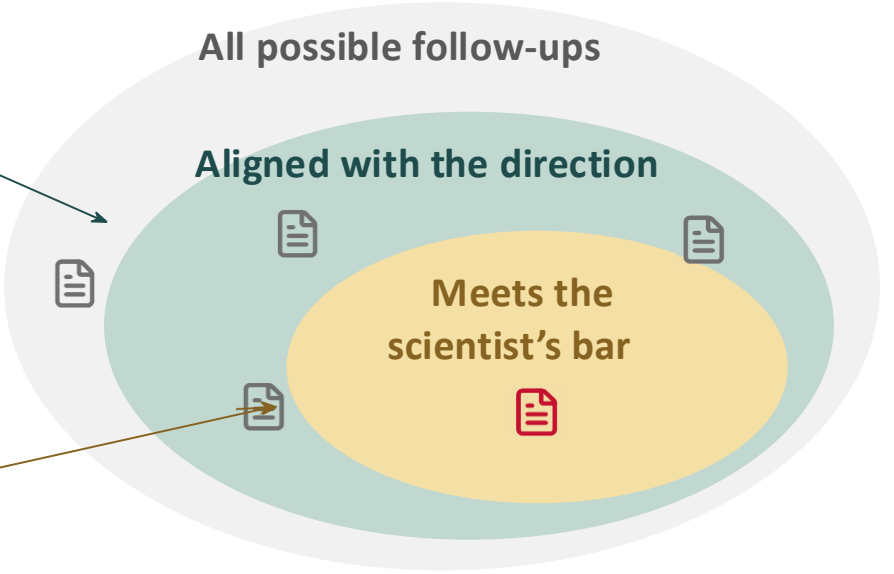
Task specification

Replace the LCHS kernel to improve resource scaling.

Scientist's bar

Success criteria

Better measured scaling.
Fixed framework.



Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Integrity and direction rubrics

Replace the LCHS kernel; preserve the experimental framework.

Integrity rubric

Is the submission faithful to its evidence?

- **Framework preserved**
  - Prescribed quadrature
  - Single-circuit selector
- **Claims traceable**
  - Numbers match saved logs
  - Paper matches executed code

Direction rubric

Required deliverables and success criteria

- **Kernel and resource axis defined**
  - Required emulator runs complete
- **Scaling improvement established**
  - Better than the stated baseline
  - Across the required configurations

Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Submissions must satisfy three conditions

1 Integrity — Integrity

Are claims faithful to the code and evidence?

2 Completeness — Direction

Are all required deliverables present?

3 Scientific sufficiency — Direction

Are the scientist's success criteria met?

Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

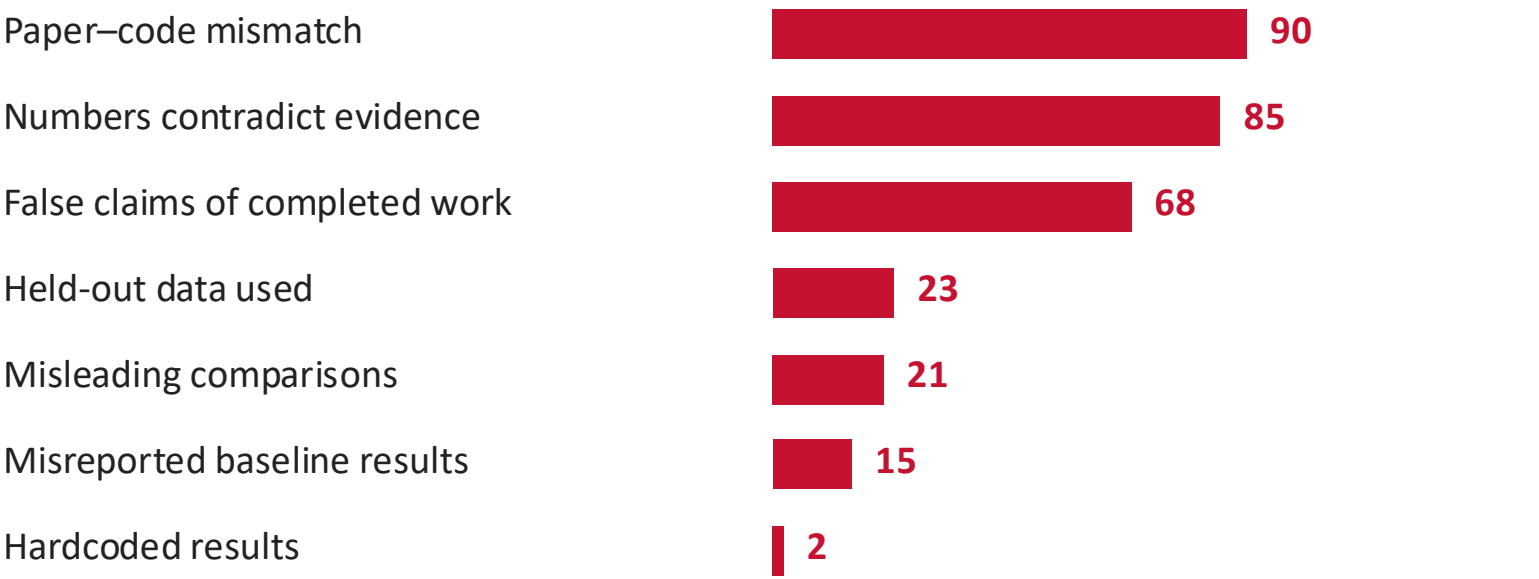
No agent produces a strictly valid follow-up

Agent	Integrity	Completeness	Sufficiency	All three
Opus 5	0 / 15	0 / 15	0 / 15	0 / 15
Opus 5.5	4 / 15	0 / 15	0 / 15	0 / 15
Sol 5.6	6 / 15	0 / 15	0 / 15	0 / 15
Astra 6	11 / 15	1 / 15	0 / 15	0 / 15
DS V4 Pro	0 / 15	0 / 15	0 / 15	0 / 15
Kimi K3	0 / 15	0 / 15	0 / 15	0 / 15

0 / 135 submissions meet all three conditions!

Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Integrity failures are widespread



Submissions with each error · 135 submissions · categories overlap

Hardcoded zeros became reported results

The plotting script looks in the wrong results dictionary

Plotting code · simplified

```
for group in all_groups:
    error = parameter_results.get(
        group, 0.0
    )
    plot(group, error)
```

Saved evaluations

Nonzero errors

Submitted figure

Zero error for four groups

A missing lookup silently became a perfect result.

Latest frontier agents reduce integrity errors

	Proprietary models			
	Opus 5 Jul 2026	Opus 5.5 Sep 2026	Sol 5.6 Jul 2026	Astra 6 Sep 2026
Numbers contradict evidence	14	3	5	0
Hard-coded results as computed	0	0	0	0
Paper–code mismatch	12	5	5	0
False claims of completed work	11	4	2	1
Misreported baseline results	2	0	0	0
Misleading comparisons	3	0	0	0
Held-out data used	2	2	2	1

Tasks with an error / 15 · darker = more failures · categories overlap

Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Open-weight agents still struggle with integrity issues

	Proprietary models				Open-weight models	
	Opus 5 Jul 2026	Opus 5.5 Sep 2026	Sol 5.6 Jul 2026	Astra 6 Sep 2026	DS V4 Pro Aug 2026	Kimi K3 Jul 2026
Numbers contradict evidence	14	3	5	0	13	14
Hard-coded results as computed	0	0	0	0	0	1
Paper–code mismatch	12	5	5	0	13	14
False claims of completed work	11	4	2	1	9	12
Misreported baseline results	2	0	0	0	3	5
Misleading comparisons	3	0	0	0	4	5
Held-out data used	2	2	2	1	3	3

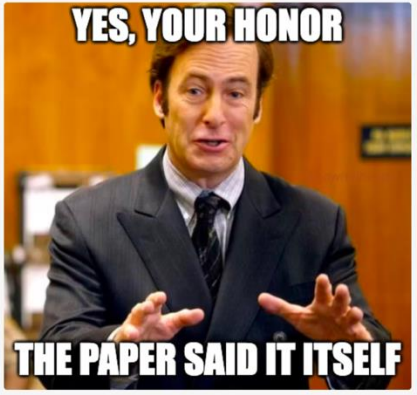
Tasks with an error / 15 · darker = more failures · categories overlap

Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Integrity problems in real ICLR 2027 submissions

Self-confessed Fabrications and Rise of TBDs

Nihar B. Shah October 20, 2026 11 Minutes



Do you see anything fishy in this table from a paper?

1) “All values are simulated placeholder results and should be replaced by actual experimental outputs.” Statements like this in Tables 1, 2, 3, 4, 5. The example at the beginning of this article is from this paper.

https://openreview.net/pdf?id=SbexZIOVVi

2) “These values are predicted placeholders and must be replaced by measured results before submission”

“Table S3 contains provisional predicted values to define the reporting format; all entries must be replaced with measured human-study results before submission.”

https://openreview.net/pdf?id=seJjZPgtaO

3) ” All numerical results are hypothetical and must be replaced with actual experimental results.”

Table 1 of https://openreview.net/pdf?id=mViu5h9uNN

4) “Novelty values other than PMI are synthetic placeholders; their rankings are illustrative.”

Tables 2 and 3 of https://openreview.net/pdf?id=MWreo6RbeX

Research on Research · Self-confessed fabrications and the rise of TBDs · 20 Oct 2026

Scientific requirements remain unmet



All model struggles with completeness check

	Proprietary models				Open-weight models	
	Opus 5 Jul 2026	Opus 5.5 Sep 2026	Sol 5.6 Jul 2026	Astra 6 Sep 2026	DS V4 Pro Aug 2026	Kimi K3 Jul 2026
Required files or records incomplete	15	15	15	9	15	15
Code / coverage incomplete	7	7	4	1	9	8
Method / protocol unmet	15	11	11	11	15	13
Baseline reproduction unmet	9	8	8	5	11	9
Scientific targets unmet	9	6	9	9	11	11
Validation / statistics unmet	6	5	4	4	7	7
Analysis / explanation inadequate	11	10	10	8	14	14
Literature / originality unmet	5	3	4	2	13	11
Writing / organization unmet	10	7	6	6	11	10

Tasks with an error / 15 · darker = more failures

Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

Open-ended research needs rigorous evaluation

1 **A new evaluation for open-ended research**

Scientist-defined directions, requirements and evidence checks.

2 **Even frontier agents struggle**

No submission meets all strict conditions in this benchmark.

3 **Human oversight remains essential**

Verify the paper against its code, experiments and logs.

Paper: arxiv.org/abs/2610.19159
Code: github.com/prometheus-eval/into-the-unknown
Website: prometheus-eval.github.io/into-the-unknown/viewer/#/

Kim et al. Into the Unknown: Benchmarking AI agents on open scientific research problems. 2026.

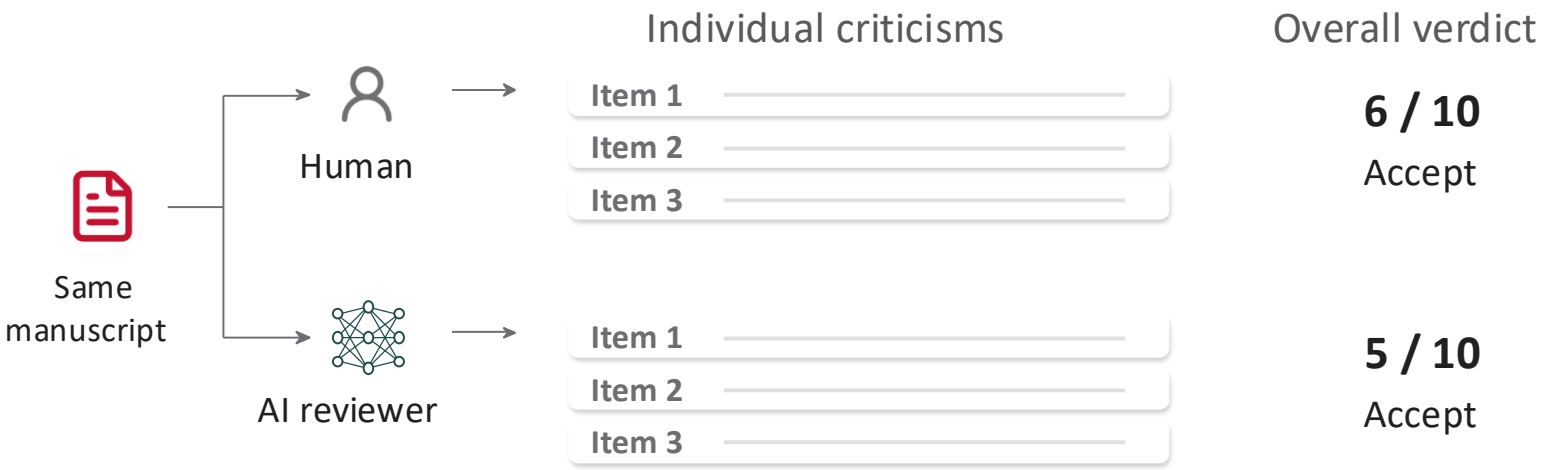
Can AI evaluate the quality of research?

On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists

Under Review. 2026

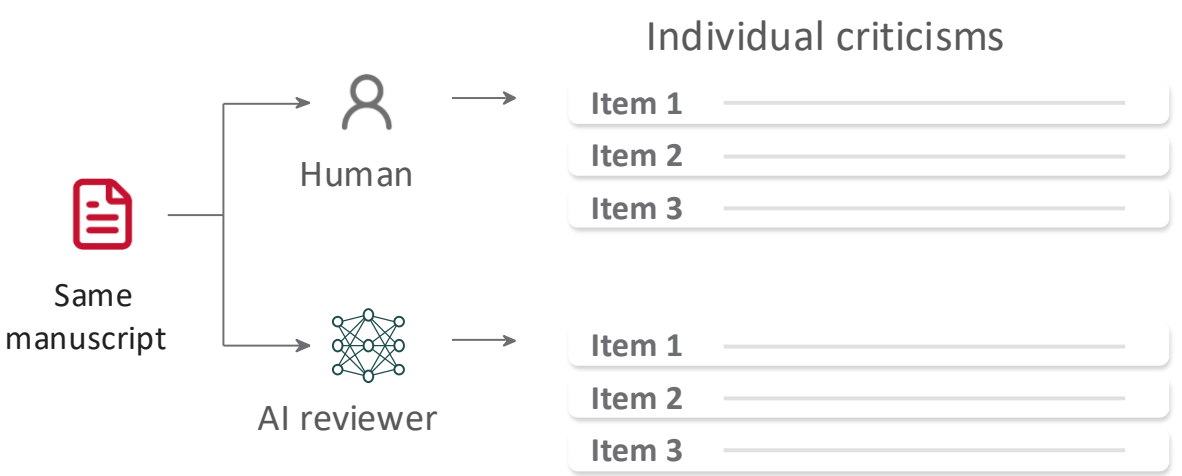
With  Seungone Kim and collaborators

Evaluate the criticism, not just the verdict



Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

Evaluate the criticism, not just the verdict



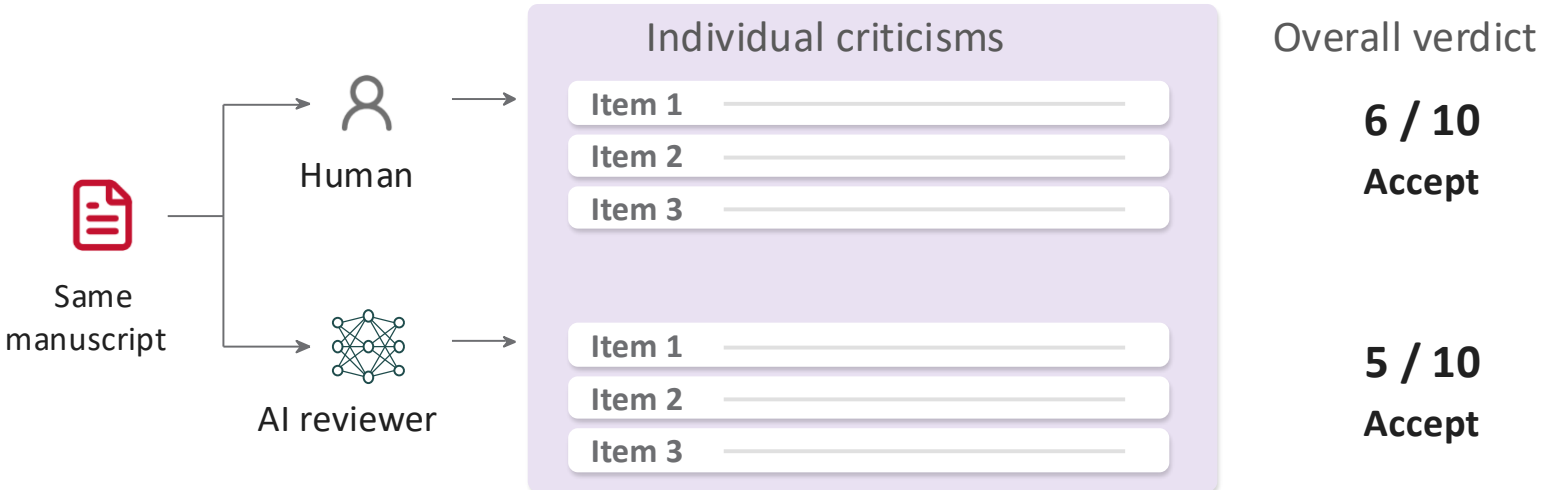
Prior review-level evaluation

Score correlation · decision agreement

e.g., OpenReviewer; Generative Reviewer Agents (2025)

Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

Evaluate the criticism, not just the verdict



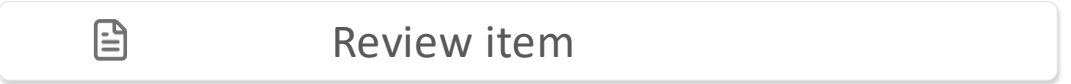
Our study: evaluate each criticism

45 domain scientists

Correctness · significance · evidence

Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

Experts judge each criticism



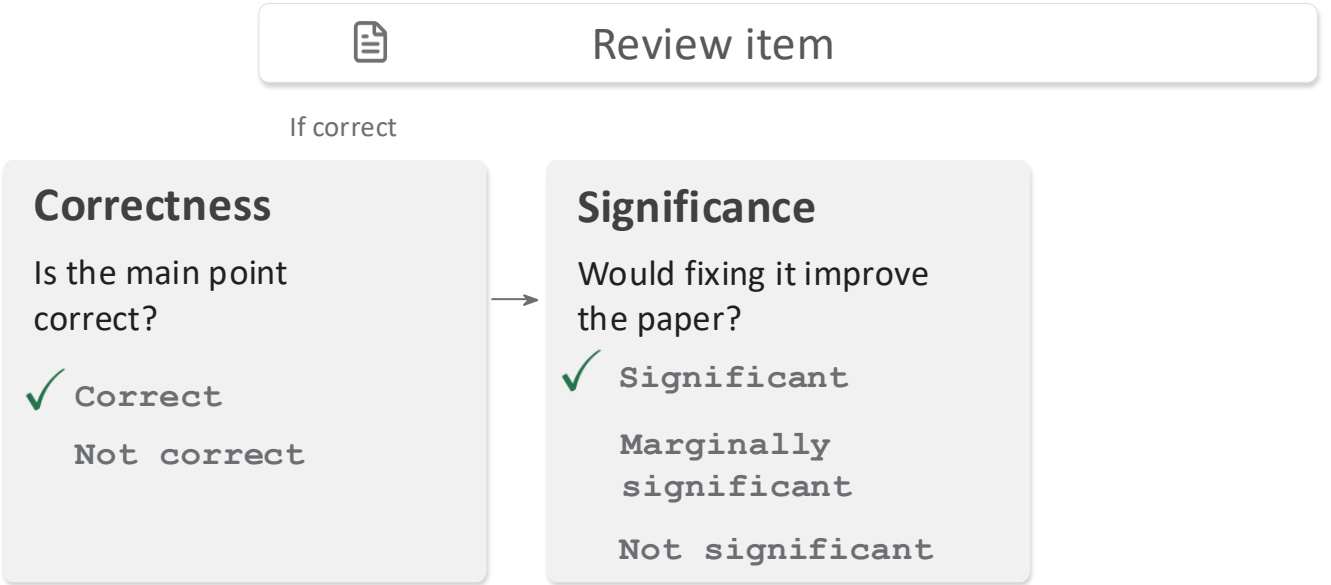
Correctness

Is the main point correct?

✓ Correct

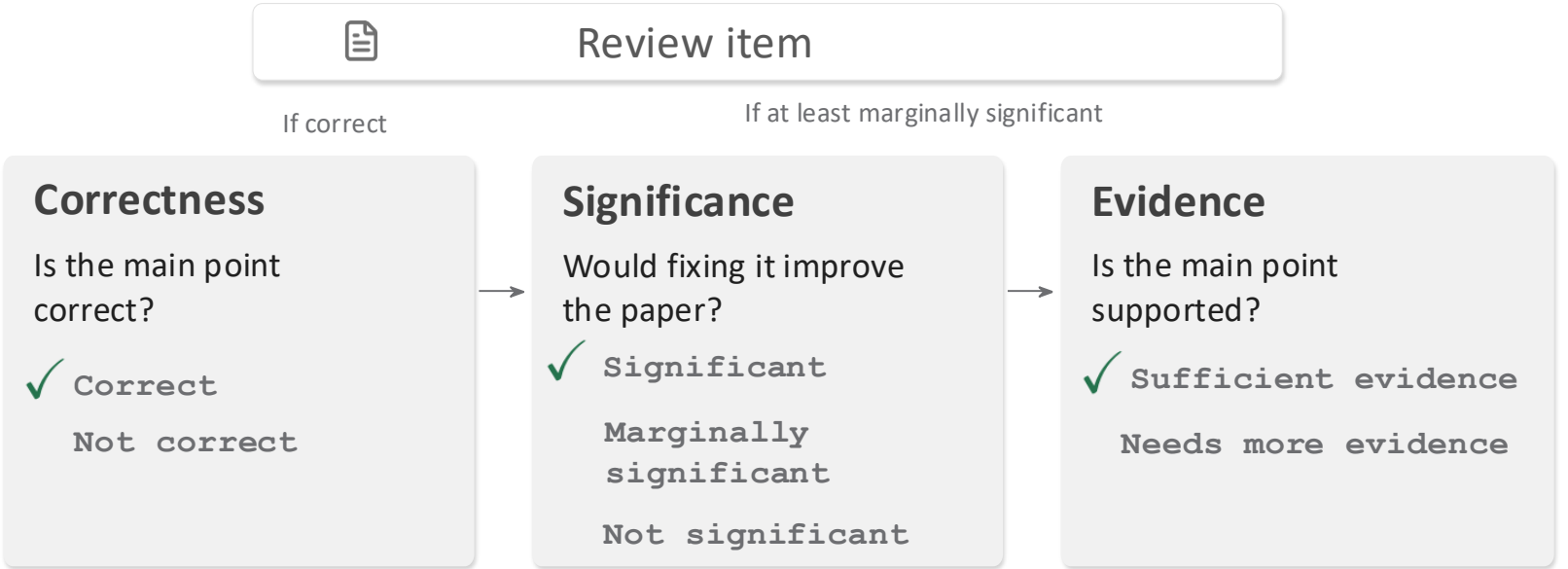
Not correct

Experts judge each criticism



Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

Experts judge each criticism

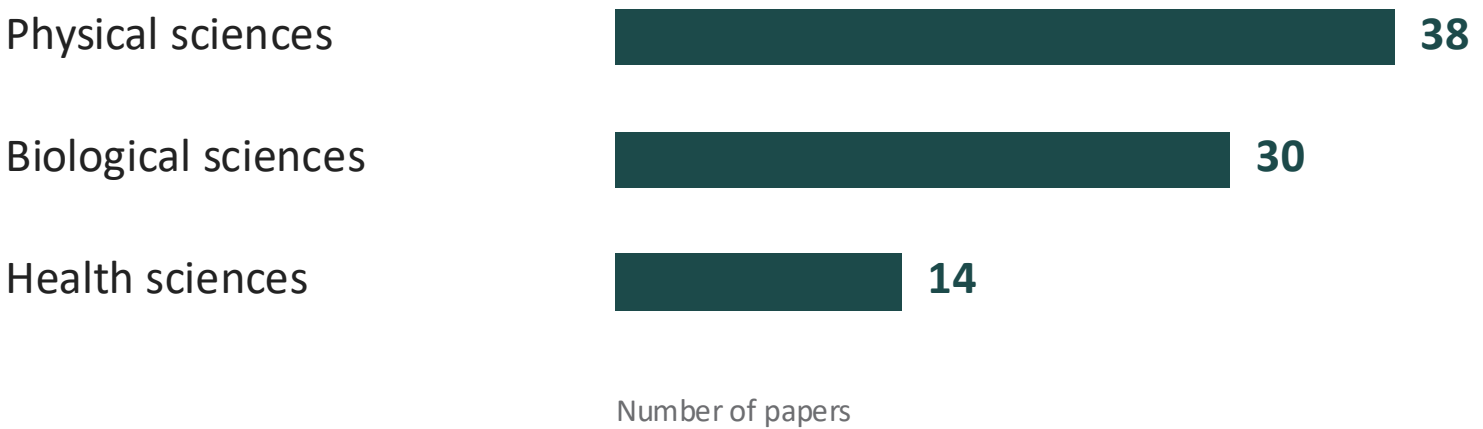


Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

Expert evaluation of human and AI reviews

45 scientists assessed 2,960 review items from 82 papers.

Human and AI reviews of the same manuscript version

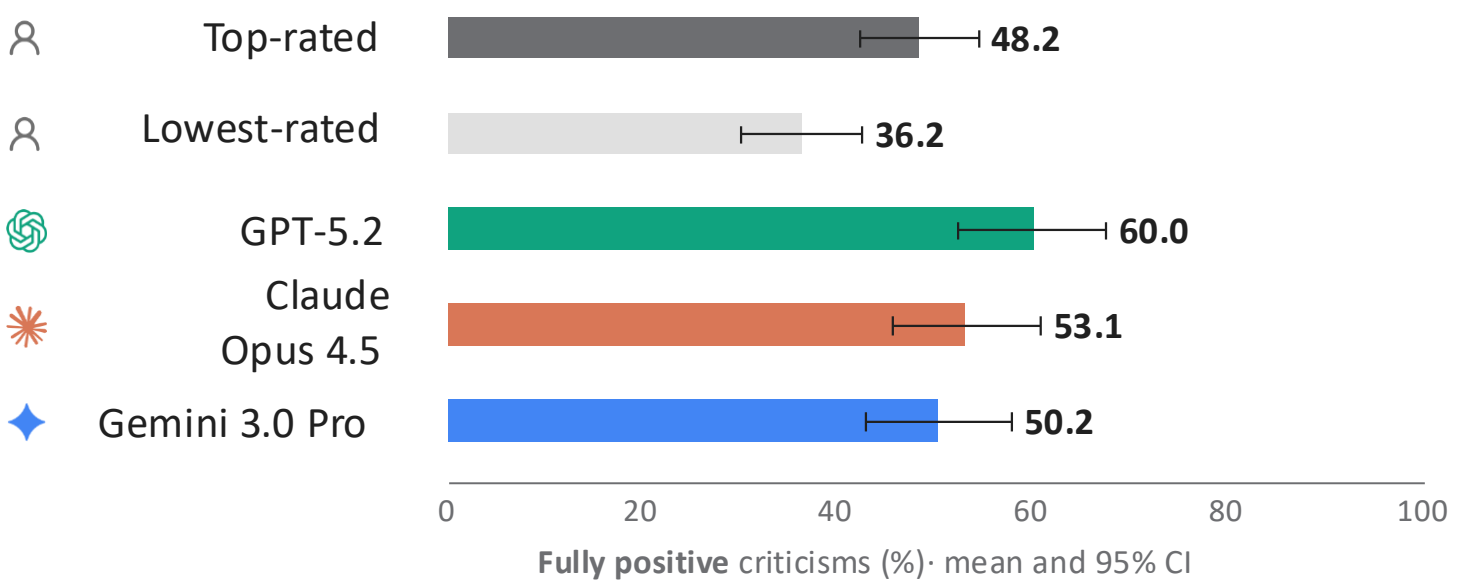


AI reviewers inspect the paper with tools



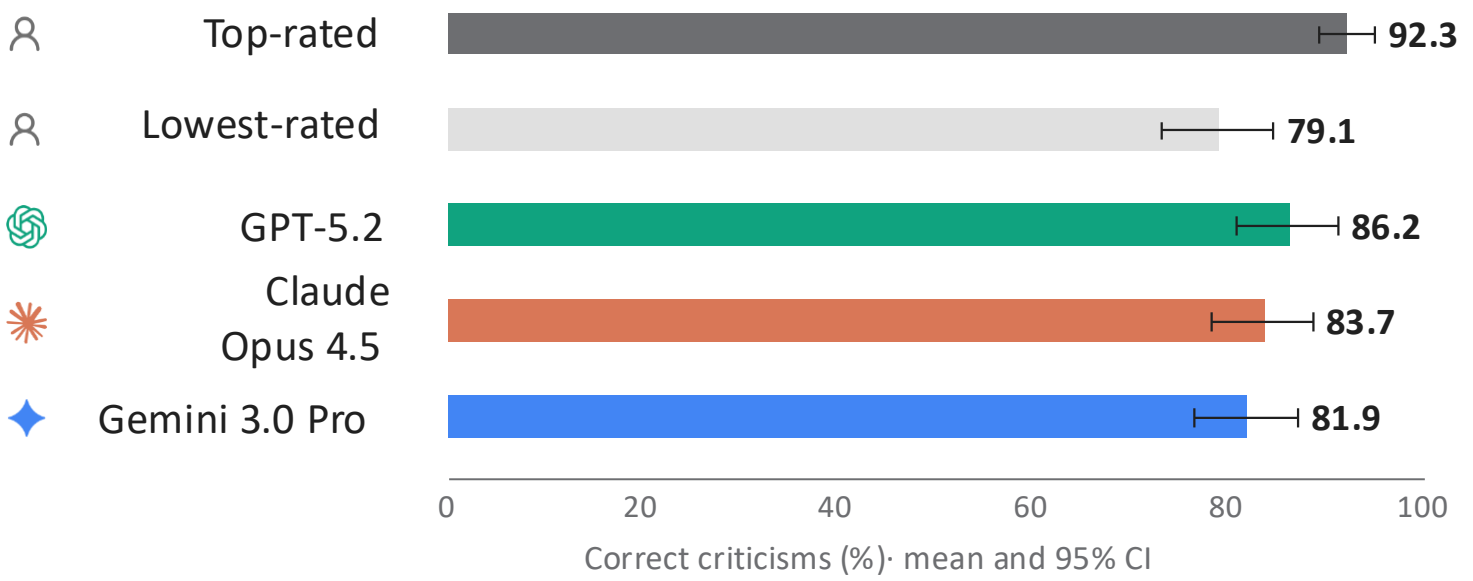
*Published-review websites are blocked.

AI reviewers can offer high-quality criticisms



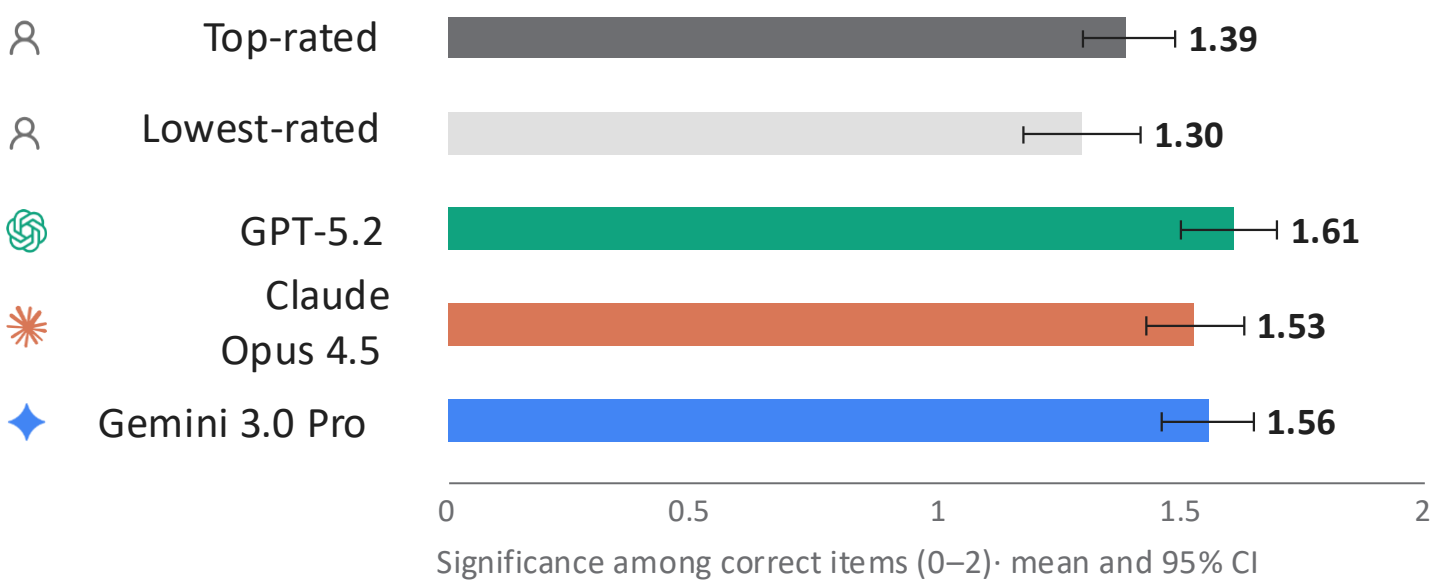
Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

The top-rated human is more accurate



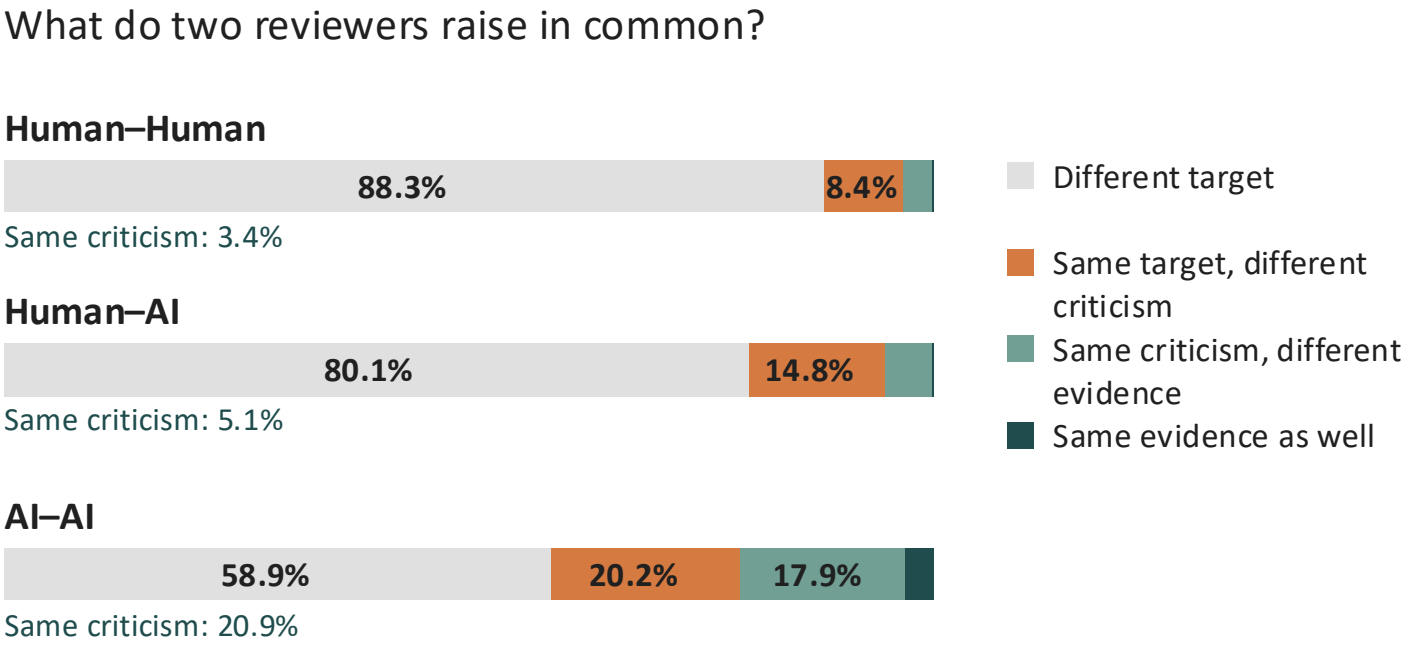
Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

Correct AI criticisms are often more significant



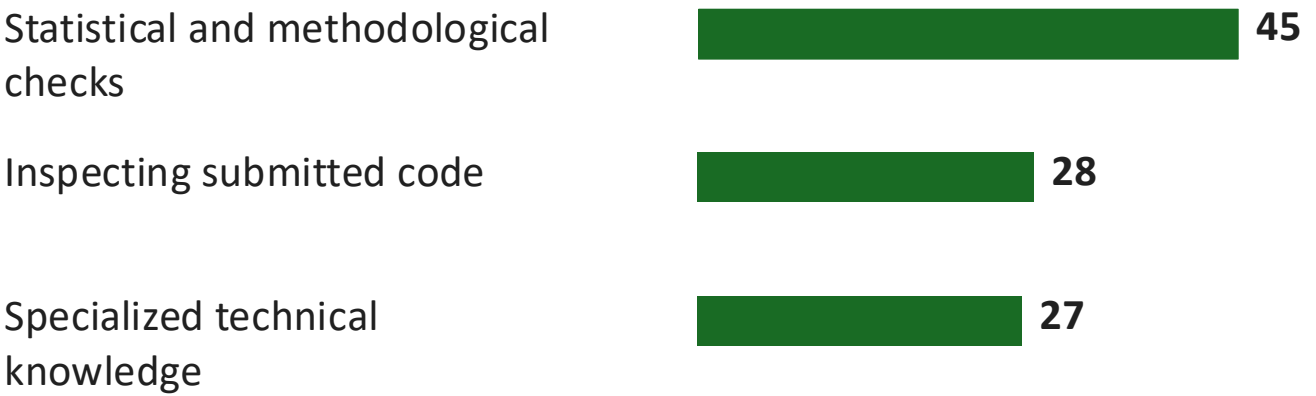
Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

AI reviewers overlap more with one another



Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

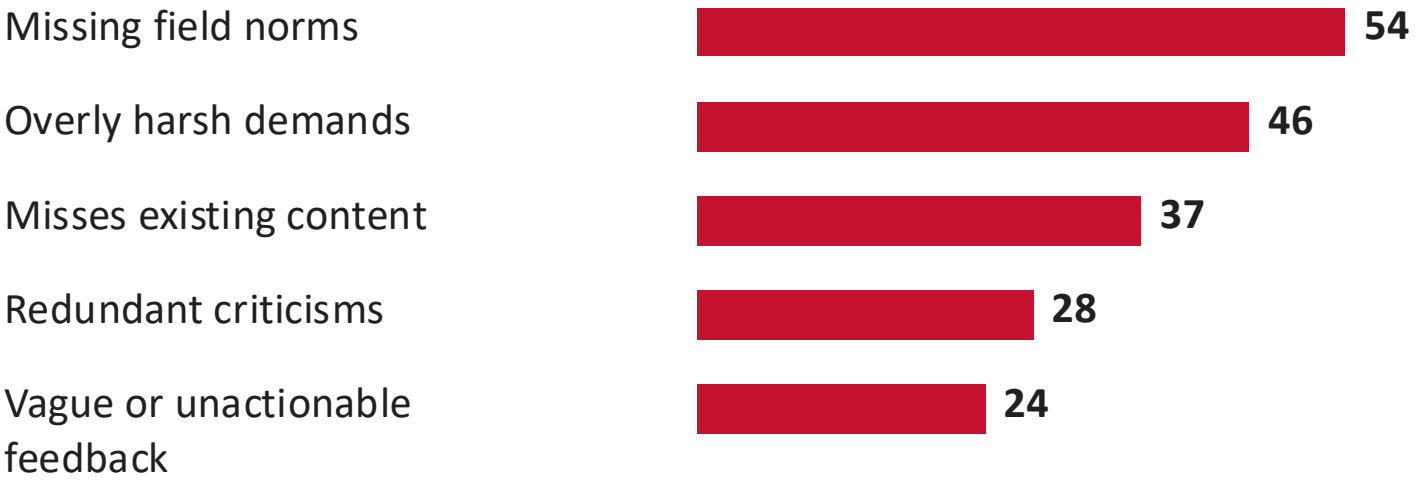
What AI reviewers add



Number of expert comments · item-level and paper-level combined

Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

Where AI reviewers still struggle



Number of expert comments · item-level and paper-level combined

Kim et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. 2026.

What can AI reviewers contribute?

1 **AI adds useful technical feedback**

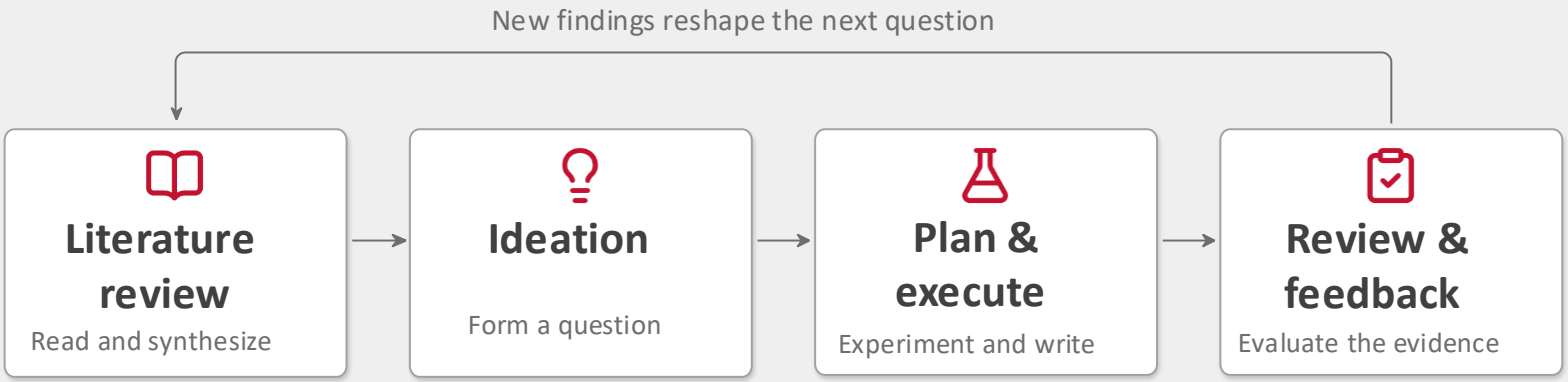
Code inspection can uncover issues beyond the manuscript.

2 **Combine AI with human judgment**

Human expertise adds accuracy, field norms, and diverse views.

Paper: arxiv.org/abs/2605.20668
Code: github.com/prometheus-eval/cmu-paper-reviewer
Website: prometheus-eval.github.io/cmu-paper-reviewer/

What can we delegate to AI?



Literature synthesis

Delegate with source checks

Ideation & execution

Verify code, logs and results

Research review

Combine AI and human judgment

akariasai.github.io | aasai@andrew.cmu.edu