Can AI Replace the Peer Reviewer (Yet)?

Da Huang, Avner Kalay, James Schallheim

Discussion by **Alejandro Lopez-Lira**, University of Florida

22nd Annual WashU CFAR Finance Conference · Session 6: AI and Measurement in Finance · November 2026

What the paper does

2,591 reviewed papers
Utah Winter Finance Conference, 2017 to 2026

→

2 human reviewers
and 10 AI models score each paper

→

Compare scores
with each other and with later citations

Compression

AI rarely gives a 1 or a 5 (humans: 11.5%, 7.9%)

Agreement

AI agrees with a human about as much as two humans agree (0.19)

Citations

Humans predict better at the top and bottom, AI in the middle

No conference in 2021. Tables 2 and 6, Figure 3. Top and bottom: mean human score at most 1.5 or at least 4.0. Middle: 73% of papers.

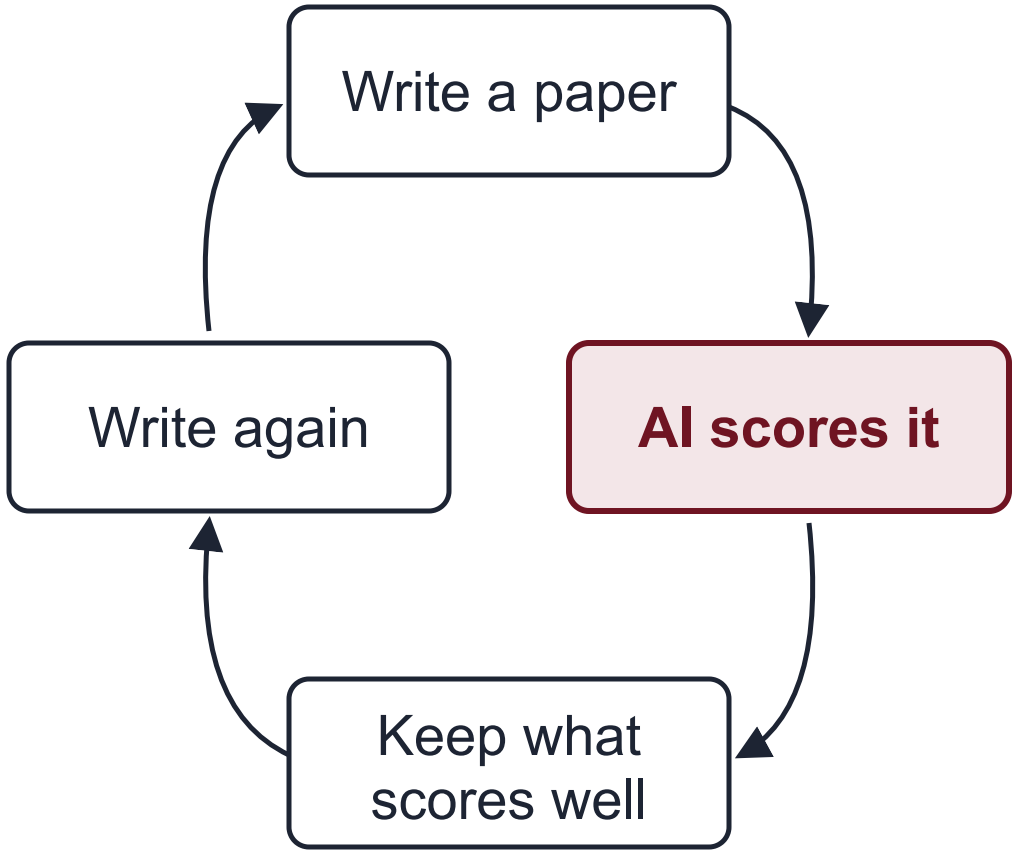
I like this paper a lot

The question	Reruns	Real human data
AI makes papers cheap to write, but review cannot keep up	Same papers and prompt, rerun on every new model	Committee scores, 2017 to 2026, before and after ChatGPT

Rising paper supply: Section 1. Reruns, CanAIReview.com: Section 2.4. Committee scores: Section 2.1. ChatGPT launched November 2022.

Why I care

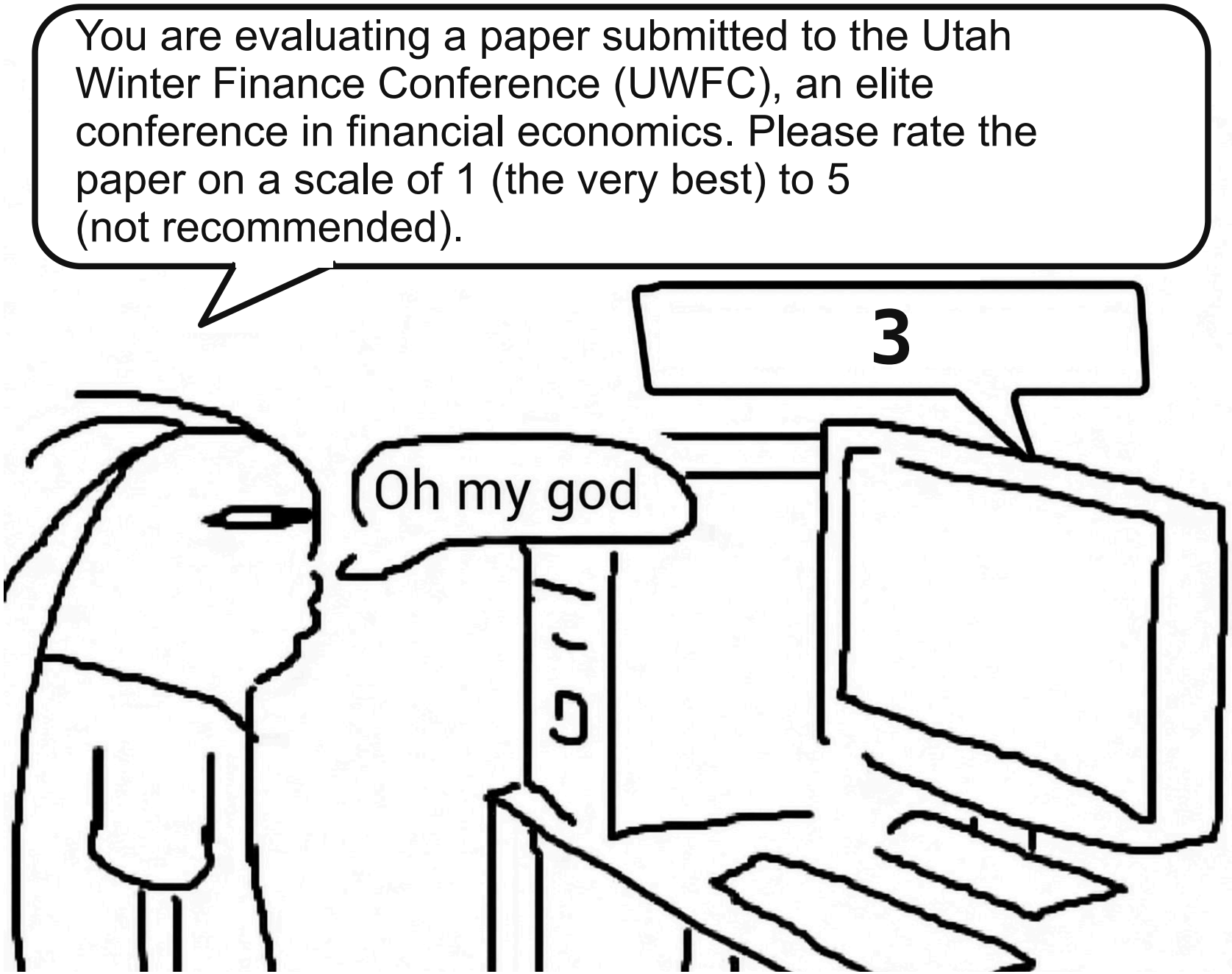
If AI can tell a good paper, we can train it to write one



- Writing drafts is already cheap
- Scoring papers well is the hard part
- This paper tests AI scoring

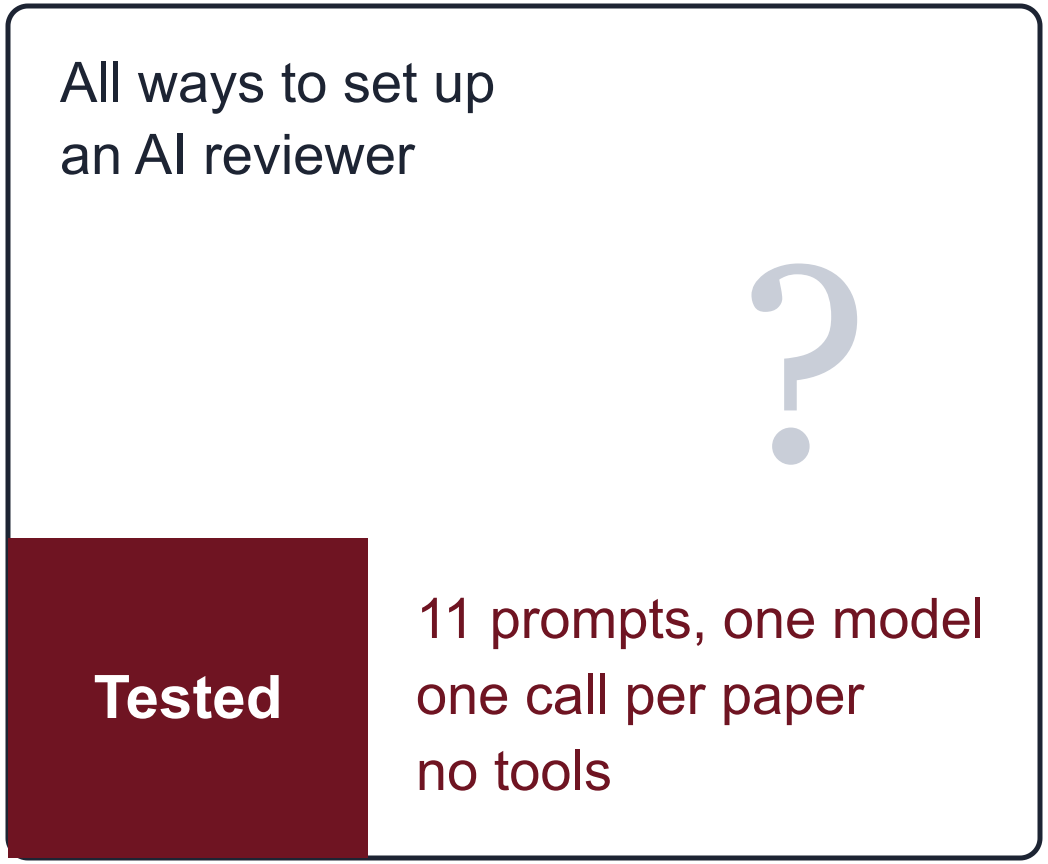
> ZeroPaper already does this
> github.com/alejandroll10/zeropaper

Lower bounds on detected ability



Opening of the prompt given to every model, abridged (Appendix C2; rubric follows). AI rarely gives a 1 or a 5 (Table 2).

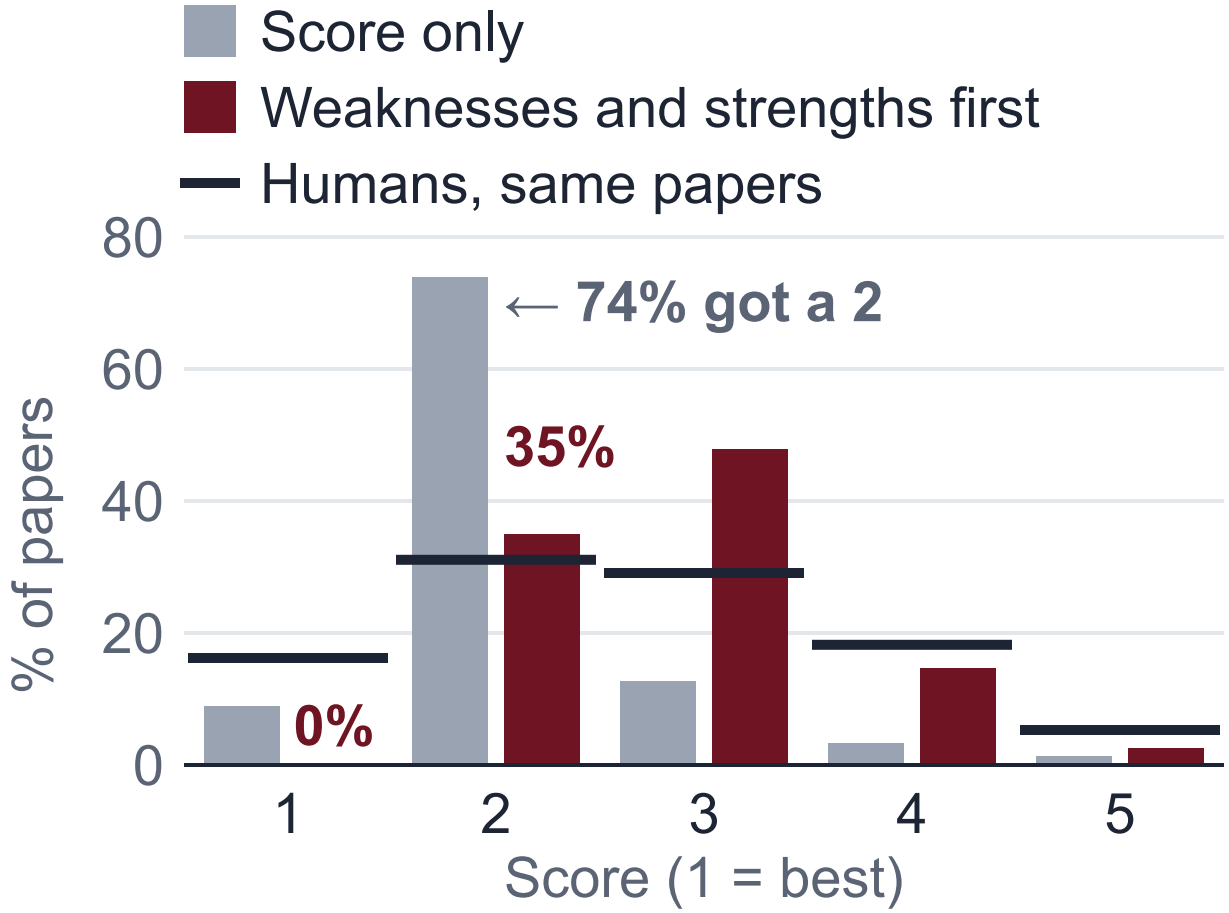
Failing to elicit a skill does not prove it is absent



- One success proves the skill exists
- A study can try only a few setups
- We learn what researchers could elicit

Prompt search: Opus 4.7 (main study: ten models). Sections 2.2 and 2.3, Table 3. Same logic as Remark 1 in Lopez-Lira, Tang, Zhu (2025).

One small change moved the results a lot



- Model lists weaknesses and strengths first
- Mismatch with humans falls from 0.85 to 0.45
- Untried prompts could do better

Opus 4.7, same instructions, same 157 papers. Mismatch = sum of gaps in score shares (L1), 0.85 to 0.45. Table 3.

Memorization

A memorization test gives only a lower bound

Recognized: 11%

Not recognized: memorized or not?

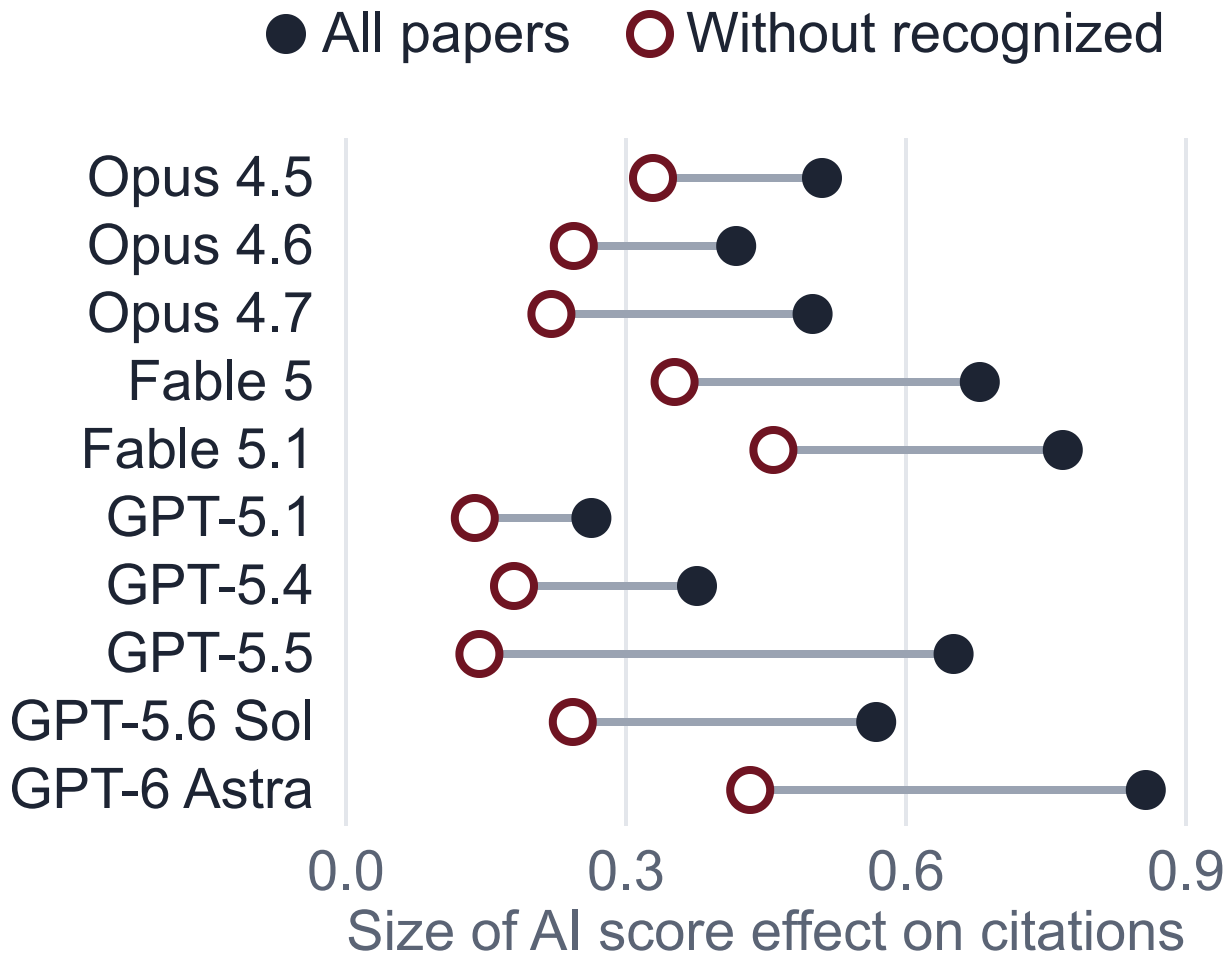


All 2,488 probed papers

- Model says it remembers and names the title
- Removing author names does not stop recognition
- Few papers appeared after the models' training

Opus 4.7 probe: Section 4.2. Post-cutoff tests: Section 4.1. Lopez-Lira, Tang, Zhu (2025), arXiv:2504.14765.

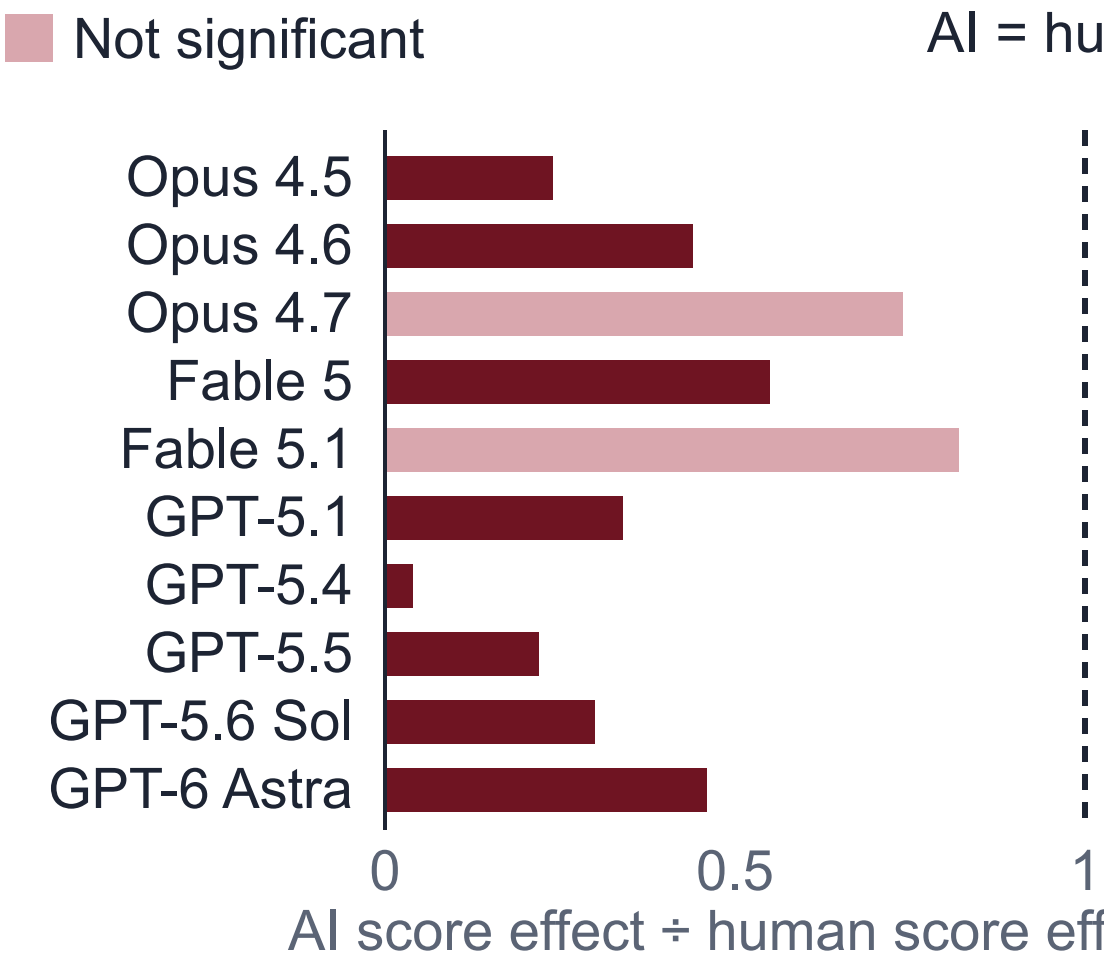
The papers AI recognizes are the famous, highly cited ones



- Dropping them halves AI's link in the middle
- Significant at 5% in 9 models, then in 3
- So the test cannot separate memory from fame

Opus 4.7 recognizes 11%, mostly top-journal papers. Middle: mean human score 1.5 to 4. Average AI effect 0.56 to 0.27. Table 10.

Even if AI memorized outcomes, human scores predict citations better at the top



- Human edge in all ten, significant in eight
- Memorization could only inflate AI
- So the human edge is understated

Absolute effects on future citations. Top: mean human score at most 1.5. Light bars: not significant at 10%. Opus 4.5: AI sign wrong. Table 6.

The right benchmark

There may be no single right score

Common view: 1/3 of the weight

Own taste: 2/3



Weight in one human reviewer's score

- Two reviewers correlate at only 0.19
- The paper reads disagreement as taste
- So the target depends on whose taste

Weight on the shared view is 0.33. Its share of score variance is 0.19, the reviewer correlation. Section 3.2, Welch (2014); Table 2.

Each result uses a different benchmark

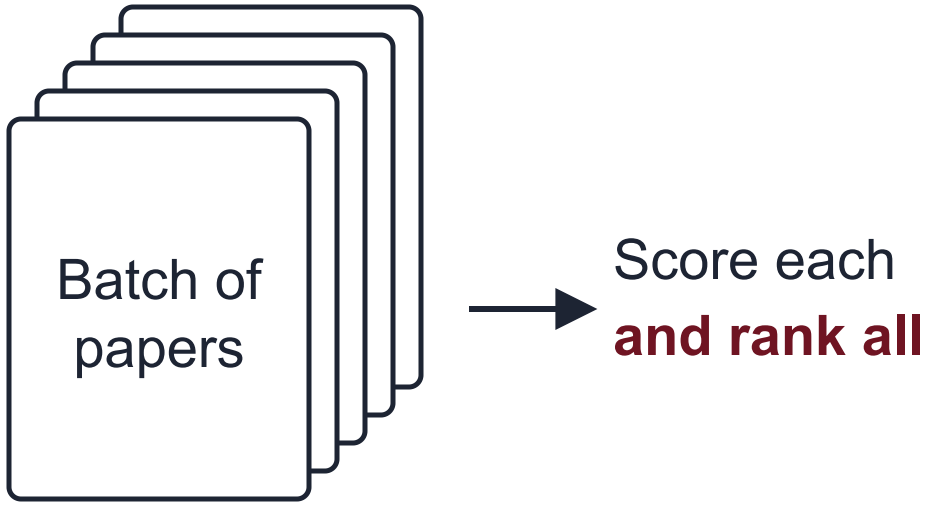
Result	AI is compared with	Concern
Score spread	Spread of human scores	Single reviews are noisier than consensus
Agreement	One human reviewer	At most 0.43 with one human
Program choice	Ten program papers a year	Committee picks reflect its own taste
Citation test	Future citations	A proxy, for AI and humans alike

0.43 is the square root of 0.189. Footnote 2: citations are "admittedly not a conference reviewer's stated task," but the comparison is fair. Tables 2, 4, 5, 6.

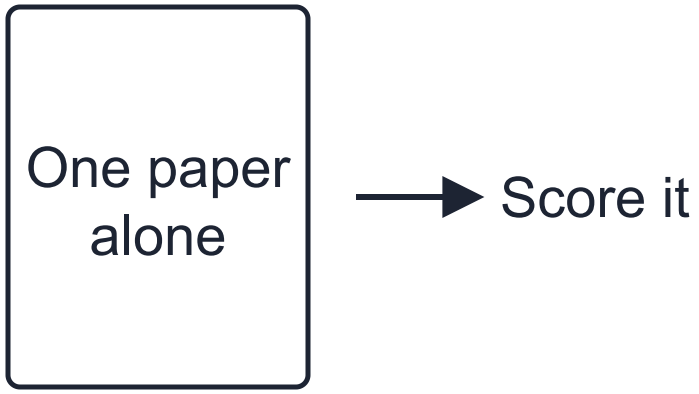
Suggestions

Give the AI the reviewers' task

Human reviewer



AI in this paper



- Few 1s and 5s may come from seeing one paper
- Ask the AI to rank a batch
- Compare with reviewers' own rankings

Reviewers rank all papers they receive, and the paper does not use these rankings (Section 2.1). Each AI review is a single isolated call (Section 2.2).

Two changes to the citation test

Which score better predicts a paper's later citations?

	The paper	Try
AI score	One AI review	**Average of several**
AI compared with	Mean of two reviewers	**Each reviewer's record**

Tables 5 and 6. Section 4.3 reruns check consistency only. The paper does not use reviewer identity.

Conclusion

In this test, AI cannot yet replace reviewers, but it can complement them.

- Results are lower bounds on AI's ability
- Memorization cannot explain the human edge

Lower bounds: Section 2.3, Table 3. Memorization and the human edge: Table 6, Section 4.1.

For reproducibility, report the settings

Reported	Not reported
Full prompt and output fields	Thinking on or off, effort level
One call per paper, no tools	Temperature, seed, output limit
Model names and release dates	Exact model version strings
Fixed protocol for new models	Text extraction tool and code

Prompt and output fields: Appendix C2. Protocol: Sections 2.2 and 2.4. None of the not-reported items appears in the paper.