

Peer Review

Review of: "Are AI Peer Reviewers More Demanding Than Humans? A Cross-Disciplinary Study of AI-Based Manuscript Evaluation"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The study asks whether AI tools evaluate manuscripts more strictly than human reviewers. The author took 24 already published papers, eight per discipline: toxicology, Earth/environmental sciences, and food science and technology (Abstract; Methods, "Study design and AI tools"). Each was the first experimental article of the April 2026 issue (May 2026 for Toxicology) of eight toxicology journals, the first article of the June issues of eight environmental journals, or chosen by the same criterion from eight food-science journals (Methods, three subsections). Three roles were assigned: a desk editor (DeepSeek, Mistral, Kimi) deciding "Continue to Review" or "Reject without Review"; five reviewers (Gemini, Copilot, ChatGPT, Qwen, Grok) choosing Accept, Minor revisions, Major revisions, or Rejection; and two meta-reviewers (Claude and Perplexity) ranking the five reports and giving a final recommendation. Each tool was queried once with identical prompts and asked to treat the papers as newly submitted. Results appear in Tables 1-3 (one per discipline) and Table 4 (counts per tool and discipline).

As I read it, the author reports that "Major revisions" predominated although all papers had been accepted, and that only Gemini recommended acceptance. Each tool kept a similar profile across disciplines, with Copilot strictest and Gemini most permissive. Stricter recommendations did not mean more useful reports, since ChatGPT and Qwen reports were ranked most useful (Abstract; Results and Discussion). The Conclusions go further, stating that AI tools "tend to apply more demanding and less permissive evaluative criteria than those typically observed in human peer review", and could be "potential alternatives to human reviewers" (Conclusions, paragraphs 1-2). The author describes the work

as a proof-of-concept, hypothesis-generating exploratory investigation (Introduction, last paragraph; Results and Discussion, the Limitations paragraph and the one after it).

2. Whether the conclusion holds on the data

I'd like to start with what I found valuable. I recounted Table 4 from the per-paper rows of Tables 1-3, and every cell matches, for all five reviewer tools and both meta-reviewers (by my count, as rendered on Qeios). Because the per-paper tables are given in full, any reader can repeat this check. The design is simple and transparent, and the Limitations paragraph in Results and Discussion names the small sample, the published versions, the single queries, and the AI meta-reviewers.

What the data show is this. Across all 120 reviewer recommendations (5 tools x 24 papers), Table 4 gives Accept 3, Minor 49, Major 68, and $3 + 49 + 68 = 120$. Major is $68/120 = 57\%$. No tool recommended Rejection for any paper (Tables 1-3). Per discipline, out of 40 recommendations each, I count toxicology 0 Accept / 16 Minor / 24 Major, environmental 2 / 17 / 21, and food 1 / 16 / 23 (Table 4). So the environmental set differs from the others by one Minor and two or three Major recommendations. Would that be enough to call it "a somewhat less stringent profile" (Results and Discussion, paragraph 2)? Copilot gave Major for 22 of 24 papers ($8 + 7 + 7$, Table 4), and all 3 Accept recommendations came from Gemini.

The central claim is where I would add the most. The only human data point is that every paper was accepted. No human reviewer reports or first-round human recommendations were collected, and the Methods say the comparison "is indirect and should not be interpreted as a controlled head-to-head assessment". Results and Discussion, paragraph 2, says the frequency of critical recommendations "cannot be interpreted as evidence that AI systems outperform human reviewers". Conclusions paragraph 1 then says this "strongly suggests" AI tools apply "more demanding and less permissive evaluative criteria than those typically observed in human peer review", and Conclusions paragraph 4 says testing on already accepted papers "can reinforce the conclusion". These readings pull in different directions. A Major revision is also a normal first-round human outcome, so could the data separate "stricter than humans" from "behaves like a first-round reviewer"? It would help to say which of the two is intended.

Two smaller claims are only partly supported by the tables. The Abstract says "ChatGPT and Qwen produced the most detailed and constructive comments". By my count, as rendered on Qeios (Tables 1-3, "Best report" columns), Claude put ChatGPT first 11 times and Qwen 11 times (of 24). Perplexity put ChatGPT first 9 times, Copilot 9 times, Qwen 5 times, and Grok once, so for Perplexity, Copilot ties

ChatGPT. The two meta-reviewers chose the same best report in 11 of 24 papers (4 toxicology, 3 environmental, 4 food), though they agreed on the final recommendation in 22 of 24.

Second, Gemini's reports "were consistently ranked lowest" (Results and Discussion, after Table 4). By my count, as rendered on Qeios, Perplexity ranked Gemini last in 21 of 24 papers, but Claude did so in 15 of 24, and in none of the 8 toxicology papers (Table 1, Claude column), where Claude placed Grok last 6 times. Would it be more accurate to qualify "consistently" by meta-reviewer and discipline? The tool-specific profiles rest on 8 papers per discipline and no statistical test (Methods: "no formal sample size calculation"; Limitations: "does not include formal statistical analysis"), so I would present them as observations in this sample.

3. Reproducibility

The paper reports the versions only in part (Methods, "Study design and AI tools"). Six tools have a version or mode named: DeepSeek (expert mode), Mistral (Le Chat), Kimi (K2.6 Instant), Qwen 3.6, Claude Opus 4.8, and Perplexity Pro. Gemini, Copilot, ChatGPT, and Grok appear only as "the free versions", without a model name or version.

I searched the text and did not find the dates on which the tools were queried. Without dates and model names for the free versions, a reader cannot tell which model produced each report. The prompts are described in paraphrase with a few quoted fragments (Methods, paragraph on prompts), and the full prompt texts are not given. The desk prompt asked for a 4–5 line justification and to "avoid fakes and hallucinations", but the complete wording is not available to a reader who wants to rerun it.

The desk prompt refers to "the attached papers". I did not find whether the tools received the typeset journal PDF, with the journal name and dates visible, or a stripped text. I also did not find whether papers were sent one per chat or several at once. On single runs, the Methods state that "All AI tools were queried once using identical prompts", and the Limitations paragraph says the analysis "does not account for potential variability arising from stochastic outputs or prompt sensitivity". Conclusions paragraph 3 calls the tools "reproducible evaluative agents". With one run per tool and paper, this run-to-run reproducibility was not measured.

Data Availability says the data are "available from the corresponding author upon reasonable request". I would suggest depositing in a public repository the full prompts, the raw 120 reviewer reports, the 48 meta-reviews, and the 72 desk decisions, with the query dates and the model shown in each interface. It

would also help to repeat each query at least a few times on a subset to show how often the recommendation changes.

4. What specifically to fix

1. In Results and Discussion, paragraph 2, the sentence says that environmental sciences had "the only cases in which 'Accept in its present form' recommendations were issued". As rendered on Qeios, by my count, Table 3 (Journal of Agricultural and Food Chemistry, Gemini) shows ACCEPT, and Table 4 lists 1 Accept for Gemini in food sciences. Could the sentence be checked and corrected?
2. Objective (a) in the last paragraph of the Introduction is to evaluate AI tools for initial manuscript screening, but the desk decisions of DeepSeek, Mistral, and Kimi are not reported in any table. It would help to add a table of the 72 decisions (3 tools x 24 papers) with the short justifications and to state the desk outcome for each set. The environmental and food subsections say the papers were "carried forward", and the toxicology set is covered only by "As in the previous stages" in the food subsection; the food wording "so that the complete set could be evaluated" also leaves open whether some papers were forwarded despite a reject decision.
3. Conclusions, paragraphs 1, 2, and 5 use "strongly suggests", "potential alternatives to human reviewers", and "AI-driven peer review", which go beyond a design that Methods itself calls indirect. I would align the Conclusions and the Abstract with the cautious wording already present in Results and Discussion, paragraph 2.
4. The "neutrality" advantage in Conclusions, paragraph 2, is not tested by any data in the study, while the Introduction mentions that AI reviewers have "their own algorithmic biases". I would either mark it clearly as an opinion or remove it.
5. In Conclusions, paragraph 3, "reproducible evaluative agents" is a strong phrase. With one query per tool and paper, the study shows stable profiles across papers, but not reproducibility of the same review. I would rephrase it or rerun a subset of the papers to support it.
6. The Abstract and Results state that "ChatGPT and Qwen produced the most detailed and constructive comments" and that Gemini "consistently ranked lowest". As rendered on Qeios, by my count from Tables 1-3, Perplexity ranked Copilot first 9 times, as often as ChatGPT (9), and Qwen 5 times, and Claude never ranked Gemini last in Table 1. Do these counts agree with the author's own tally? It would help to report first-place and last-place counts per meta-reviewer in a small table so readers can check these statements.

7. The inclusion rules differ between the two sets: toxicology excluded reviews (Methods, "Toxicology papers"), while the environmental set includes a systematic review and meta-analysis (Bonell et al., The Lancet Planetary Health). I would explain this or replace the paper. It would also help to give the journal selection rule, that is, why these eight journals were chosen in each field.
8. A simple agreement statistic would strengthen the analysis. One example is the agreement between the two meta-reviewers (by my count, the same final recommendation in 22 of 24 papers, the same best report in 11 of 24). Another is a test of tool versus discipline effect on the 120 recommendations in Table 4.
9. A few small items remain. Table 2 has the row "Environmental Health Perspective", while Methods says "Environmental Health Perspectives". Reference 19 (Laue et al.) is dated 2025, while the environmental set is described as June issues. I would state the coding rule for "border" verdicts in Tables 2-3. Reference 33 contains the misspelling "Inhibior".

5. Recommendation

Major revisions. The data are honest and internally consistent (Table 4 matches Tables 1-3 by my count), and the question is timely. The changes I would ask for are mainly in reporting and framing, but two of them go to the core of the paper: the claims about humans in the Conclusions and the Abstract need to match the indirect design, and objective (a) has no reported desk-screening results. I would also ask to share the prompts and the raw reports, and to check the Accept sentence in Results and Discussion against Table 3. With these changes, the paper would be a useful baseline for the prospective head-to-head comparison by publishers on newly submitted manuscripts that the author proposes (Results and Discussion, last paragraph; Conclusions, last paragraph).

Declarations

Potential competing interests: No potential competing interests to declare.