Qeios

Peer Review

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

Iliano Fasolino[1]

1. Department of Computer Science, University of Milan, Italy

Compared to v1, this is a real step forward. Extending the same protocol from toxicology alone to environmental and food science papers, and summarizing tool behavior in Table 4, makes the exploratory design more interesting than a single-field pilot. In places, the author also labels the work as hypothesis-generating, which fits what the design can support. The match between the design and the claim in the title and abstract has still not closed.

The title asks whether AI peer reviewers are “more demanding than humans.” The experiment does not answer that question. The AI tools reviewed final published versions of papers that had already been accepted. The human side of the comparison is the fact of prior acceptance, not human reviews of the same text. The Methods section itself says the comparison is indirect and should not be read as a controlled head-to-head. Expanding from eight to twenty-four papers strengthens tool-profile stability across fields; it does not fix the human-comparison design. That stability claim is useful. It is not yet evidence that AI is tougher than human reviewers on the manuscripts those humans actually saw.

Severity of recommendation and usefulness of the report come apart, and with three fields in view, the pattern is easier to see. Copilot is often harsh; Gemini is often lenient and alone issues Accept recommendations; ChatGPT and Qwen are repeatedly ranked more useful by the meta-reviewers even when they are not the softest graders. Table 4 shows the severity profiles; the ranking columns in Tables 1–3, and the prose after Table 4, show usefulness coming apart from severity across disciplines. For editors thinking about AI-assisted workflows, “which tool writes a usable report” matters at least as

doi.org/10.32388/MYNCCD

much as “which tool says Major.” I would rewrite the abstract and opening so this is the headline, and demote the human-comparison claim to a carefully hedged secondary observation.

The meta-review step still uses Claude and Perplexity to rank the five AI reports and to issue overall recommendations. Sarah Fankhauser’s review of v1 already flagged the circularity and the lack of prompt-sensitivity checks; both still apply here, and they matter more now that “stable tool profiles” are presented as a cross-disciplinary result. Those rankings are AI judgments of AI judgments. Without a human-coded sample of report quality, I cannot tell how much of the usefulness story is about report quality and how much is about preference patterns among the meta-reviewers. I ask the author to human-code usefulness on a subset of reports and compare those scores to the Claude and Perplexity ranks.

The Conclusions still outrun the Methods. Phrases such as “potential alternatives to human reviewers,” “neutrality,” and “AI-driven peer review” do not follow from single-shot prompts on twenty-four published papers, with tools chosen from the author’s personal experience. The neutrality claim is the hardest to sustain on this evidence. Commercial models carry training and policy biases; tool selection was not systematic; and interpersonal bias among human reviewers is not measured here, so it cannot be shown to have been reduced. I accept that AI might help with some relational problems in specialized fields. That is an editorial hypothesis, not a finding of this study. The prospective publisher study sketched at the end is exactly the right next experiment. It should stay as future work, not as present warrant for calling AI a viable substitute.

A few reporting gaps keep the paper thinner than the design suggests. There are still no simple inter-tool agreement statistics on the Accept/Minor/Major/Reject calls, even though five reviewers scored the same twenty-four papers. There are no repeated draws under the same prompts, so stochasticity and prompt sensitivity remain unquantified. The actual review texts are not deposited; “available upon request” is weak for a study whose substance is the content of those reports. Tables of verdicts alone leave readers unable to judge whether “Major” meant a serious methodological attack or a generic request for more discussion. Depositing prompts, model versions where known, and the raw recommendations (even if anonymized by paper) would make the work far more usable.

The three-stage editorial emulation is unevenly reported relative to v1’s clearer desk note. DeepSeek, Mistral, and Kimi are introduced as desk editors, yet the Results barely show what they decided. Methods note that papers were carried forward to review; readers still need a clear desk-screen table or a frank

statement that desk outcomes were uniform and therefore not informative. Either report those outcomes properly or drop the claim of a three-stage editorial simulation.

Declarations

Potential competing interests: No potential competing interests to declare.