https://dx.doi.org/10.21037/jmai-2026-0071
2026-05

Can AI help with peer review in transplantation research?

Marc Raynaud, Charles Brunet[1,2] and Alexandre Loupy
Université de Paris Cité, INSERM, PARCC, Paris Institute for Transplantation and Organ Regeneration, Paris, France
Artefact Research Center, Paris, France

Comment on: Shen SM, Wang Z, Paul K, et al. Evaluation of Large Language Models for Peer Review in Transplantation Research: Algorithm Validation Study. JMIR AI 2026;5:e84322.

Keywords: *Peer review; large language model-assisted peer review (LLM-assisted peer review); transplantation*

Peer review in medicine is under increasing pressure. Submission volumes continue to rise, while reviewers and editors face increasing workload, delays, inconsistency, and well-described risks of bias [1, 2]. Used well, large language models (LLMs) could help pre-structure reviews, summarize manuscripts, flag reporting omissions or internal inconsistencies, and support more systematic methodological scrutiny, allowing human reviewers to focus on novelty, clinical relevance, and judgment [3]. Major medical journals are already discussing or piloting such hybrid models: JAMA has explicitly outlined a human- in-the-loop vision for artificial intelligence (AI)-assisted editorial assessment [4], and NEJM AI has described "Human+AI" review workflows as part of accelerated evaluation processes [5]. Recent work also suggests that carefully designed LLM feedback can improve the specificity and actionability of peer review at scale [6]. Transplantation is an especially important domain in which to study this question. Peer review in transplantation often requires simultaneous appraisal of immunology, histopathology, organ disease, donor and recipient characteristics, etc. At the same time, the field is evolving rapidly. AI is already being used across transplantation for organ allocation [7, 8], rejection monitoring [9] and outcome prediction [10], while xenotransplantation is moving from experimental promise toward early clinical reality [11]. AI-assisted review could be thus particularly valuable, as a structured reader capable of improving consistency and surfacing methodological issues that deserve closer human attention. Against this background, the study by Shen and colleagues is timely [12]. The authors address an important and underexplored use case: whether LLMs can support peer review in transplantation research, and whether they exhibit affiliation bias when evaluating manuscripts. Their decision to test multiple open-weight models and prompting strategies, and to examine fairness alongside accuracy, is a genuine strength. In particular, the focus on affiliation bias is valuable, because one of the most appealing arguments for AI-assisted review is the possibility of reducing at least some forms of human bias. Encouragingly, no clear affiliation effect was observed. Their overall conclusion is appropriately cautious: the open-weight models they studied are not accurate enough to replace human reviewers, and any use in peer review still requires close human oversight. At the same time, the study should probably be interpreted as an early benchmarking exercise rather than as a decisive test of AI-assisted peer review in transplantation. A major limitation, openly acknowledged by the authors, is the choice of ground truth: the benchmark target was journal quartile rather than editor-rated review quality or manuscript-specific editorial usefulness. This is a pragmatic choice, and the authors explain why they adopted it, but journal quartile is at the very best an indirect proxy for the quality judgments that matter in real editorial workflows. A more rigorous benchmark would compare AI-generated reviews against structured scoring by experienced editors or domain experts. A further concern is that the evaluation framework raises questions about generalizability. The optimal prompt- temperature configuration was selected in a pilot

Corresponding author(s): alexandre.loupy@inserm.fr
Artefact Research Center, Paris, France

phase using Llama 3.3 on a subset of papers and then applied across all models. That design is understandable from a practical standpoint, but it assumes that hyperparameters identified on one architecture transfer meaningfully to others, which is far from established. More broadly, given the pace of model development, any benchmark of open- weight LLMs now has a short half-life. Shen and colleagues tested several models likely influential at the time, but the field is evolving so quickly that periodic re-benchmarking against newer frontier open-weight systems is likely to be necessary if such studies are to inform editorial practice. Finally, the apparent advantage of retrieval-augmented generation (RAG) deserves careful interpretation. In the study, RAG was the only setup able to ingest full papers, whereas the other prompting methods were limited to abstracts because of token constraints. That makes the comparison realistic in one sense, but not fully symmetric in another. The better performance of RAG may therefore reflect both the value of retrieval and the simple fact that it had access to much richer input. Relatedly, because retrieval accuracy is foundational to the validity of a RAG-based review pipeline, more implementation detail would have been helpful, particularly regarding how the system ensured that the retrieved document corresponded to the target paper being evaluated. Several broader issues also deserve mention. LLMs are too often benchmarked against other LLMs rather than against human reviewers, who are themselves far from unbiased; head-to-head comparisons should be the standard. It is also useful to distinguish three roles LLMs may play in peer review: literature search, editing assistance, and access to specialized domain knowledge. Key risks include over-reliance for nuanced judgments, bias inherited from non-comprehensive training data, though recent frontier models show markedly lower hallucination rates in scientific contexts. In specialized settings, a circularity whereby models trained on prior peer-reviewed literature may reinforce its dominant assumptions. The next generation of studies should therefore move closer to the editorial endpoint that matters most: decision support. What is now needed are prospective evaluations Journal of Medical Artificial Intelligence, 2026 in which journal editors receive AI-generated reviews, or AI-augmented reviewer reports, and then score them for usefulness, correctness, specificity, fairness, and impact on editorial confidence and time to decision. Such work should compare multiple deployment modes: AI as a pre-review triage tool, AI as a structured methodological checker, AI as a meta-review assistant synthesizing human reports, and AI as a parallel reviewer whose report is visible to editors but not determinative. The central question is whether it helps editors make better and faster decisions. Building on such evidence, journal editors may then be in a position to propose practical guidelines for integrating LLMs into the peer-review process.

A. Acknowledgments

Large language models were used to improve scientific writing.

B. Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, Journal of Medical Artificial Intelligence. The article has undergone external peer review.

Peer Review File: Available at https://jmai.amegroups.com/article/view/10.21037/jmai-2026-0071/prf.

Funding: None.

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://jmai.amegroups.com/article/view/10.21037/jmai-2026-0071/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Open Access Statement: This is an Open Access article distributed in accordance with the

Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license).

References

[1] Alina Tvina, Ryan Spellecy, and Anna Palatnik. “Bias in the peer review process: can we do better?” In: *Obstetrics & Gynecology* 133.6 (2019), pp. 1081–1083.

[2] Samir Haffar, Fateh Bazerbachi, and M Hassan Murad. “Peer review bias: a critical review”. In: *Mayo Clinic Proceedings*. Vol. 94. 4. Elsevier. 2019, pp. 670–676.

[3] Nitya Thakkar et al. “A large-scale randomized study of large language model feedback in peer review”. In: *Nature Machine Intelligence* 8.3 (2026), pp. 326–336.

[4] *JAMA Network Peer Review Academy*. 2026.

[5] Arjun K Manrai et al. *Accelerating science with human+ AI review*. 2025.

[6] Qiyao Wei et al. “The ai imperative: Scaling high-quality peer review in machine learning”. In: *arXiv e-prints* (2025), arXiv–2506.

[7] Rintaro Yanagawa et al. “Development and validation of a machine-learning model to reduce futile procurements in donations after circulatory death in liver transplantation in the USA: a multicentre study”. In: *The Lancet Digital Health* 7.10 (2025).

[8] Bima J Hasjim et al. “The AI agent in the room: informing objective decision making at the transplant selection committee”. In: *medRxiv* (2024), pp. 2024–12.

[9] Alexandre Loupy et al. “Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study”. In: *Bmj* 366 (2019).

[10] Daniel Yoo et al. “A machine learning-driven virtual biopsy system for kidney transplant patients”. In: *Nature communications* 15.1 (2024), p. 554.

[11] Xiaowei Hu et al. “Xenotransplantation: from proof of concept to clinical reality”. In: *Swiss Medical Weekly* 155 (2025), p. 4945.

[12] Selena Ming Shen et al. “Evaluation of large language models for peer review in transplantation research: Algorithm validation study”. In: *JMIR AI* 5 (2026), e84322.