JHPLT
Journal of Health Professions Learning and Technology

REVIEW

Vol. 1, No. 1 (Jul-Sep 2026) | e17
Published 14 September 2026
Published by Palmetto Innovative Education LLC, Mount Pleasant, SC | e-ISSN 3144-5454

Using Artificial Intelligence in Scholarly Review: Reconciling the Limits of Human Judgment with the Limits of the Machine

Debra E. Henninger-Borckardt, PhD[1]

[1] Medical University of South Carolina

[*] Correspondence: Debra E. Henninger-Borckardt, PhD, debra.henninger@gmail.com

ABSTRACT

Peer review remains the primary quality-control mechanism of scholarly publishing, yet the evidence based on human reviewer performance shows chronically low inter-rater reliability, documented bias, and a labor system straining under rising submission volumes. The emergence of large language models (LLMs) capable of generating manuscript feedback has prompted journals, including those in health professions education, to ask whether artificial intelligence (AI) can serve as a reviewer, an aid to reviewers, or neither. This article synthesizes the empirical literature on human reviewer reliability and bias, catalogs the documented strengths and vulnerabilities of LLM-generated review, and evaluates the case for hybrid human-AI review systems. Across multiple large-scale studies, AI-generated feedback overlaps with human reviewer commentary at rates statistically comparable to the overlap between two independent human reviewers, and AI systems can process volume, apply criteria consistently, and surface omissions humans overlook. At the same time, AI reviewers hallucinate citations, are demonstrably manipulable through prompt injection embedded in manuscripts, and lack the contextual and ethical judgment that anchors legitimate peer review. The data available at present support neither wholesale automation nor blanket prohibition, but rather a structured hybrid model in which AI performs triage, consistency-checking, and completeness screening under human oversight, while accountability for the final editorial judgment remains with human reviewers and editors.

Keywords peer review; artificial intelligence; large language models; inter-rater reliability; hybrid review; research integrity; health professions education

Introduction

Peer review is the mechanism by which the scholarly community distinguishes credible claims from unsubstantiated ones prior to publication. Its legitimacy rests on an assumption that independent expert reviewers, evaluating the same manuscript, will converge on similar judgments about its validity and contribution. That assumption has never held up well empirically, and the strain on the system has only intensified as submission volume has grown faster than the supply of willing, qualified reviewers. Into this already-strained system has arrived a new candidate reviewer: the large language model (LLM). Since 2023, a substantial and rapidly growing body of empirical work has tested whether systems such as GPT-4 and Claude can generate manuscript feedback comparable to that of trained human reviewers, and journals across disciplines, including biomedical and health professions publishing, have had to formulate policy in the near-total absence of consensus guidance.

This article addresses four questions in sequence. First, what does the empirical literature actually show about the reliability and fairness of human peer review, independent of any comparison to AI? Second, what specific, documented failure modes affect AI-generated review? Third, what documented strengths does each party bring to the task? Fourth, and most directly relevant to editorial policy, does the accumulated evidence support integrating AI into the review pipeline in some structured form, and if so, what form is defensible given what is currently known.

The Documented Problems of Human Peer Review

The most rigorously established finding in the peer review literature is that independent human reviewers agree with one another far less than the system's public legitimacy implies. *Bornmann, Mutz, and Daniel (2010)* conducted the first meta-analysis of inter-rater reliability (IRR) in journal peer review, pooling 70 reliability coefficients drawn from 48 studies and 19,443 manuscripts. The mean intraclass correlation was .34 and the mean Cohen's kappa was .17, both of which fall well below the thresholds conventionally used to describe acceptable agreement. This finding is not an artifact of a single field or review format; a subsequent analysis of consensus in scientific review reported inter-rater reliabilities typically falling between .10 and .40 across disciplines and review settings, including grant panels at the National Science Foundation, where agreement ranged from .18 for chemical dynamics proposals to .37 for economics proposals.

Low agreement is not simply noise; a portion of it reflects documented, patterned bias. *Lee, Sugimoto, Zhang, and Cronin (2013)* reviewed the evidence on bias in peer review and catalogued effects tied to reviewer and author characteristics, including institutional prestige and geography, that are not supposed to bear on scientific merit. Reviewer disagreement is also substantively meaningful rather than purely random: *Fiske and Fogg (1990)* found that reviewers evaluating the same manuscript frequently raise entirely different criticisms of it, a pattern consistent with reviewers applying genuinely different standards of what counts as a flaw rather than making correlated, correctable errors.

The system is also expensive in a way that is easy to overlook because the cost is absorbed as unpaid labor. *Aczel, Szaszi, and Holcombe (2021)* estimated that reviewers worldwide spent more than 100 million hours on peer review in 2020, equivalent to over 15,000 years of labor, with the monetary value of U.S.-based reviewers' time alone exceeding 1.5 billion dollars for that year. This burden falls unevenly: qualified reviewers decline invitations at high rates, citing both a mismatch between the manuscript and their expertise and a simple lack of available time, and the pool of willing reviewers has not scaled with the growth in submissions. A quality-control system this dependent on uncompensated, unevenly distributed volunteer labor is structurally vulnerable, independent of any question about whether AI should be involved.

It is worth noting, following *Erosheva, Martinkov , and Lee (2021)* , that low IRR estimates should be interpreted carefully rather than treated as proof that peer review accomplishes nothing. Using data from the National Institutes of Health and the American Institute of Biological Sciences, they showed that IRR estimated from restricted, high-quality subsets of proposals will approach zero under many realistic scenarios even when the underlying review process is functioning reasonably, because agreement statistics are highly sensitive to how much genuine variability exists in the pool being rated. The honest conclusion from this literature is not that human peer review is worthless, but that it is measurably less reliable and less consistent than its institutional authority suggests, and that any comparison to AI-generated review must use this modest human baseline rather than an idealized one.

The Documented Problems of AI-Generated Peer Review

AI reviewers fail in ways that are different from, not simply smaller versions of, the ways human reviewers fail. The most consequential of these is fabrication. LLMs can generate confident, well-formatted critique that includes citations, statistics, or methodological claims that do not correspond to anything in the manuscript, a phenomenon documented extensively in the broader hallucination literature and specifically flagged as a risk for scholarly review because reviewers and editors may not independently verify AI-generated claims before acting on them.

A second and more troubling category of failure is that AI review systems can be deliberately manipulated by the very manuscripts they are asked to evaluate. Investigative reporting and controlled studies converged on this finding in 2025 and 2026: hidden instructions embedded in manuscript text, invisible to human readers but processed by the AI system, can steer an LLM-generated review toward unjustifiably favorable judgments. One analysis reported that such hidden prompts achieved success rates above 98 percent across different models and remained more than 94 percent effective even after the injected text was paraphrased, with review scores inflated from roughly 5.3 to roughly 8.0 on a standard scale. A companion survey of 194 authors, reviewers, and editors found that 80 percent supported greater transparency requirements around AI use in review, and that some respondents, frustrated by inconsistent AI-generated feedback, had themselves considered using prompt injection to secure more favorable reviews (Zhu et al., 2025; Tang & Li, 2026). A controlled study using 1,441 papers from ICLR 2023 and NeurIPS 2022 similarly found that field-specific injected instructions successfully manipulated an LLM reviewer's ratings and stated weaknesses, even though generic, non-targeted prompts produced only minor effects (Zhu, Xiong, Ma, Lu, Liu, & Li, 2025).

AI reviewers also carry systematic biases distinct from, though not necessarily less severe than, the affiliation and geography effects documented in human reviewers. The literature describes an authority bias in which models favor manuscripts associated with prestigious authors or institutions, a length bias favoring longer submissions independent of quality, and a documented tendency to inflate ratings for lower-quality papers specifically, even as agreement with human judgment improves on stronger submissions (Zhu et al., 2025). Beyond bias, there is a rigor problem: LLM-generated critique has been characterized as comparatively shallow, more likely to restate limitations the authors themselves already disclosed than to identify unstated methodological weaknesses, which is arguably the more valuable function of a skilled reviewer.

Finally, there is a confidentiality problem specific to how these tools are used in practice. Uploading an unpublished manuscript to a commercial AI system to generate a review may expose confidential, unpublished research to a third party outside the control of the author, the journal, or the reviewer, which is why a substantial majority of surveyed journal policies prohibit submitting manuscript content to generative AI tools, and why bodies including the International Committee of Medical Journal Editors and the U.S. National Institutes of Health currently restrict or prohibit AI use in review altogether (see the policy analysis by *the inevitable future of peer review* , 2025, and Wang & Gong, 2026, on journal AI policies).

What Human Reviewers Still Do Better

The case for retaining human reviewers does not rest on nostalgia; it rests on specific, documented capacities that current AI systems lack. In the most direct controlled comparison available, *Farber (2025)* had ten human experts and Claude 3 independently review the same ten unpublished humanities and social science manuscripts. The AI system was markedly faster and more internally consistent across manuscripts, but human reviewers demonstrated a clearly superior ability to situate a manuscript within the relevant literature and to engage with its scholarly context, while the AI system's engagement with existing literature scored substantially lower on the same rubric. The study's own conclusion was not that either party should be replaced, but that a hybrid workflow, with AI augmented by human review rather than the reverse, was likely to outperform either approach alone.

Human reviewers also retain an edge in exactly the failures where AI review is weakest: catching what an author did not say, weighing whether a stated limitation is fatal to the paper's conclusions, and applying ethical judgment about research conduct, participant welfare, or clinical risk that has no clean textual signature for a model to detect. Even the most favorable large-scale study of AI review feedback, *Liang et al. (2024)* , which compared GPT-4-generated comments to human reviewer comments across 3,096 papers from 15 Nature-family journals and 1,709 International Conference on Learning Representations submissions, found that of 111 substantive comments made by human reviewers but missed entirely by the AI system, only six were rated as major concerns, but the AI missed them nonetheless. Human accountability is also structurally load-bearing in a way AI accountability is not: when a human reviewer signs a recommendation, an identifiable person with professional standing and disciplinary consequence has taken responsibility for that judgment, which remains true of no AI system currently in use.

What AI Reviewers Demonstrably Do Well

The evidence for AI's contribution is stronger, and more surprising, than casual skepticism about chatbot-written reviews would predict. In the largest empirical comparison to date, *Liang et al. (2024)* found that the overlap between GPT-4-generated feedback and human reviewer feedback, averaged across the Nature-family and ICLR samples, was statistically comparable to the overlap between two independent human reviewers evaluating the same manuscript (roughly 31 to 39 percent AI-human overlap versus roughly 29 to 35 percent human-human overlap). Because human-human agreement is itself the low baseline documented above, this is a more meaningful finding than it first appears: on this specific overlap metric, the AI system's disagreement with a human reviewer was not measurably larger than the disagreement one human reviewer would have with another. The overlap was highest, at nearly 44 percent, for manuscripts that were ultimately rejected, meaning the AI system was most aligned with human judgment precisely where its feedback mattered most for identifying a paper's weaknesses.

Researcher perception data from the same study reinforces this pattern in practice rather than in retrospective comparison alone. In a prospective survey of 308 researchers from 110 U.S. institutions who received AI-generated feedback on their own submitted manuscripts, 57.4 percent rated the feedback helpful or very helpful, and 82.4 percent rated it more beneficial than the feedback they had received from at least some of their human reviewers on the same or comparable manuscripts. Authors specifically noted instances where the AI system raised a substantive concern, such as a need for additional data visualization or an unaddressed data privacy issue, that their human reviewers had not raised at all.

AI systems also perform well on structured, high-volume literature tasks that are closely related to reviewing, if not identical to it. *Khraisha, Put, Kappenberg, Warraitch, and Hadfield (2024)* evaluated GPT-4 against human reviewers on systematic-review screening and data extraction across peer-reviewed and grey literature in multiple languages and

found accuracy approaching human performance for data extraction once chance agreement was properly controlled for, though performance on screening was more variable and, once corrected for dataset imbalance, ranged from negligible to moderate depending on the review stage and language. The consistent finding across this literature is that AI performs best on well-defined, checklist-like components of evaluation and worst on holistic, judgment-dependent components, which maps closely onto where a hybrid system would want to route each task.

AI as a Simulated Human Reviewer: What the Comparative Data Actually Show

The preceding sections make clear that the question "can AI review as well as a human?" is underspecified, because it depends entirely on which human baseline is used for comparison. If the implicit comparison is to an idealized, perfectly reliable expert reviewer, AI review fails badly, as does the actual human peer review system. If the comparison is to the empirically documented human reviewer, whose agreement with a second human reviewer is itself weak, patterned by non-scientific factors, and inconsistent across the same manuscript, the AI-human overlap data reported by Liang et al. (2024) indicate that a well-constructed LLM system can approximate a second human reviewer on at least one meaningful, quantifiable dimension: the degree to which its comments track the same issues a trained human would flag. This is a narrower and more defensible claim than "AI can replace reviewers," and it is the claim the data actually support.

This framing also clarifies why the Farber (2025) finding and the Liang et al. (2024) finding are not in tension with one another. AI can simultaneously match human-level overlap on issue identification while remaining measurably weaker on literature engagement and contextual depth, because these are different competencies. An AI system trained to simulate the output pattern of a human reviewer, meaning the kind and specificity of comments a reviewer typically produces, can succeed at that narrower simulation task without possessing the underlying judgment that produced comparable output in the human case. This distinction matters enormously for editorial policy, because it means AI-generated review should be evaluated, and trusted, task by task rather than accepted or rejected as a single undifferentiated capability.

Toward a Hybrid Review System: Does the Evidence Warrant It

Taken together, the empirical record supports neither the position that AI should replace human peer reviewers nor the position that AI has no legitimate role in the review pipeline. It supports a narrower, structured claim: given documented human unreliability on one side and documented AI manipulability and shallowness on the other, a hybrid model that assigns each party the tasks it performs best is more defensible than either pure-human or pure-AI review, provided the specific vulnerabilities of AI review are actively managed rather than assumed away.

This is also the direction the field's own literature, not merely the AI literature, is converging on. The Committee on Publication Ethics has stated plainly that AI tools cannot substitute for human judgment and expertise in review, while at the same time acknowledging that a hybrid approach combining AI tools with human judgment is a reasonable response to the current and future strain on the review system (*Shah & Jawaid* , 2025). *Farber (2025)* reached the same conclusion from a controlled comparison rather than a policy statement, proposing specifically that AI handle initial technical screening while human reviewers concentrate on interpretive, contextual evaluation. Some journals have already begun piloting structured versions of this division of labor; NEJM AI's experimental "Fast Track" review process issues a decision within one week of submission based on an editor's evaluation combined with two AI-generated reviews, though commentary on this model has cautioned that fully replacing a second or third human perspective with AI risks losing the benefit of genuinely differing human viewpoints on manuscripts that require methodological expertise spanning more than one AI-covered domain.

A defensible hybrid model, consistent with the evidence reviewed here, would route work as follows. AI systems are well suited to completeness and consistency screening (are required elements, disclosures, and statistical reporting present), to flagging candidate issues for a human reviewer's attention at higher volume than a human alone could manage, and to structured literature and data-extraction tasks of the kind evaluated by *Khraisha et al. (2024)* . Human reviewers should retain sole responsibility for judgments requiring contextual disciplinary knowledge, assessment of a manuscript's contribution relative to the live scholarly conversation, and any ethical or clinical-risk judgment, precisely the functions *Farber (2025)* and Liang et al. (2024) found AI performed weakest on or did not perform at all. Editorial decision authority should remain exclusively human. And because AI review is uniquely vulnerable to manipulation through content embedded in the manuscript itself, any AI use in review requires disclosure to authors and editors, restrictions against uploading confidential manuscript content to uncontrolled third-party systems, and technical safeguards against the prompt-injection vectors documented by *Zhu et al. (2025)* and related work, none of which are optional refinements but rather preconditions for the model's basic integrity.

Implications for Health Professions Education Journals

The case for a carefully bounded hybrid model is arguably stronger, not weaker, in health professions education publishing specifically. Reviewer pools in specialized health professions subfields are smaller than in large biomedical or computer science fields, which means the reviewer-fatigue and expertise-mismatch problems documented by *Aczel, Szaszi, and Holcombe (2021)* and by the broader peer review burden literature are likely to be felt more acutely here, not less. At the same time, the subject matter of health professions education, curriculum design, learner assessment, clinical simulation, and patient-facing training interventions, routinely involves judgment calls about learner welfare, assessment validity, and clinical realism that fall squarely into the category of contextual and ethical judgment where the evidence reviewed above shows AI performing weakest. A journal in this field adopting AI-assisted review should therefore weight the human-retained categories heavily: any manuscript touching learner assessment validity, simulated or real patient safety, or vulnerable-population research should not have its substantive evaluation delegated to an AI system, even as AI-assisted completeness and consistency screening could reasonably ease the burden on an already-strained reviewer pool for the majority of routine methodological and reporting checks.

Limitations of the Current Evidence Base

This synthesis should be read with several caveats appropriate to a genuinely young literature. Most of the largest empirical comparisons, including *Liang et al. (2024)* , were conducted on general scientific and computer science venues rather than health professions education journals specifically, and the single controlled human-versus-AI comparison discussed here (*Farber, 2025*) examined humanities and social science manuscripts with a sample of ten papers, which is informative but too small to generalize confidently. The prompt-injection findings are recent, drawn substantially from 2025 and 2026 preprints and early publications, and the specific attack success rates reported should be expected to change as both models and journal safeguards evolve. No study identified in this review has yet directly tested a hybrid AI-human workflow, of the kind proposed here, against a purely human review baseline in a health professions education context, using publication quality or downstream citation impact as the outcome. That study, effectively a pilot of the model proposed in this article, is the logical next empirical step, and journals considering adopting any form of AI-assisted review would be well served to treat their own adoption as a monitored pilot rather than a settled policy.

Conclusion

Human peer review is not the gold standard its institutional authority implies; its own reliability literature documents chronic disagreement, patterned bias, and an unsustainable labor burden. AI-generated review is not a ready replacement; its own literature documents hallucination, demonstrated susceptibility to manipulation embedded in the very manuscripts it evaluates, and a shallower, less contextually grounded form of critique than trained human reviewers produce. The comparative data available at present, most directly the finding that AI-human comment overlap is statistically similar to human-human comment overlap, support a specific and bounded conclusion: AI can credibly simulate a portion of what a human reviewer does, particularly issue identification and completeness screening, without yet approximating what a human reviewer is for, which includes contextual judgment, ethical reasoning, and accountable disciplinary expertise. A hybrid system that assigns each party the tasks the evidence says it performs well, with disclosure, confidentiality safeguards, and defenses against manuscript-embedded manipulation built in from the start, is the model the current data actually justify. Neither uncritical adoption nor blanket prohibition is.

References

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. Research Integrity and Peer Review, 6, 14. https://doi.org/10.1186/s41073-021-00118-2

Bornmann, L., Mutz, R., & Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. PLOS ONE, 5(12), e14331. https://doi.org/10.1371/journal.pone.0014331

Erosheva, E. A., Martinkov , P., & Lee, C. J. (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. Journal of the Royal Statistical Society Series A: Statistics in Society, 184(3), 904-919. https://doi.org/10.1111/rssa.12681

Farber, S. (2025). Comparing human and AI expertise in the academic peer review process: Towards a hybrid approach. Higher Education Research & Development, 44(4), 871-885. https://doi.org/10.1080/07294360.2024.2445575

Fiske, D. W., & Fogg, L. (1990). But the reviewers are making different criticisms of my paper! Diversity and uniqueness in reviewer comments. American Psychologist, 45(5), 591-598. https://doi.org/10.1037/0003-066X.45.5.591

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Research Integrity and Peer Review, 8, 4. https://doi.org/10.1186/s41073-023-00133-5

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., & Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. Research Synthesis Methods, 15(4), 616-626. https://doi.org/10.1002/jrsm.1715

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. Journal of the American Society for Information Science and Technology, 64(1), 2-17. https://doi.org/10.1002/asi.22784

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A

large-scale empirical analysis. NEJM AI, 1(8). https://doi.org/10.1056/AIoa2400196

Tang, S., & Li, Z. (2026). Prompt injection in manuscripts: Exploiting loopholes or crossing ethical lines? Research Integrity and Peer Review, 11(1). https://doi.org/10.1186/s41073-025-00187-7

Shah, F., & Jawaid, S. (2025). The inevitable future of peer review: Human and AI integrated peer review system. Pakistan Journal of Medical Sciences, 41(4). https://pmc.ncbi.nlm.nih.gov/articles/PMC12022572/

Wang, Z., & Gong, M. (2026). A cross-disciplinary analysis of AI policies in academic peer review. Learned Publishing, 39(1), e2035. https://doi.org/10.1002/leap.2035

Zhu, C., Xiong, J., Ma, R., Lu, Z., Liu, Y., & Li, L. (2025). When your reviewer is an LLM: Biases, divergence, and prompt injection risks in peer review (arXiv:2509.09912). arXiv. https://arxiv.org/abs/2509.09912

Ethics approval

No IRB approval required, as this is a review.

Competing interests

The author works for Figment Learning Labs and this journal is associated with that company.

Use of artificial intelligence

The authors used generative artificial intelligence to draft portions of the text of this manuscript. All AI-generated drafting was subsequently revised by the authors, who reviewed the final text and take full responsibility for its accuracy and integrity.

How to cite this article

Henninger-Borckardt, D. (2026). Using Artificial Intelligence in Scholarly Review: Reconciling the Limits of Human Judgment with the Limits of the Machine. Journal of Health Professions Learning and Technology, 1(1), e17.

Open access

This is an open access article distributed under the terms of the CC BY 4.0 licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Permanent link	https://jhplt.org/pages/article.php?id=25
Publisher	Palmetto Innovative Education LLC, Mount Pleasant, SC
e-ISSN	3144-5454