Research Article

Are AI Tools Tougher Reviewers Than Toxicologists? Artificial Intelligence as Editor and Peer Reviewer of Scientific Manuscripts in Toxicology

Jose L. Domingo[1]

1. School of Medicine, Laboratory of Toxicology and Environmental Health, Universitat Rovira i Virgili, Spain

The peer-review system is under growing strain owing to the exponential increase in manuscript submissions and the progressive decline in the availability of qualified reviewers. This preliminary study examined whether artificial intelligence (AI) tools can reliably perform two key editorial functions, desk decision-making and manuscript review, using eight recently published experimental papers in toxicology as test cases. Three AI tools were tasked with acting as Editors making desk decisions (Continue to Review or Reject without Review), while five further AI tools (reviewed the same papers as if they were submitted manuscripts, recommending Accept, Minor revisions, Major revisions, or Reject. Two additional AI tools subsequently evaluated the five reviewer reports and issued overall recommendations. All eight papers were recommended for external review by the three desk-decision AI tools, in full agreement with the original editorial decisions. Notably, none of the eight papers evaluated by the AI tools received an "Accept in its present form" recommendation from any of the five AI reviewers. Most received Major revision recommendations. These observations suggest that, within this exploratory setting, AI tools tend to issue more conservative recommendations than those implied by the final editorial outcomes. Among the tools acting as reviewers, ChatGPT and Copilot produced the reports most consistently rated as useful by the meta-reviewer AI tools, while Grok performed least well. These findings, derived from a limited and non-equivalent comparison framework, suggest that AI tools may hold potential as decision-support instruments in editorial workflows, warranting validation in larger and methodologically controlled studies.

Corresponding author: Jose L. Domingo joseluis.domingo@urv.cat

Introduction

Throughout the present century, the number of changes affecting the entire scientific publication system has been substantial. These changes have impacted all sectors involved in the process, from authors to publishing companies, including reviewers, editors, and journal subscribers. The changes have been particularly profound with respect to the most prominent publishing companies in the academic sector (Springer, Elsevier, Taylor & Francis, Wiley, etc.), which have progressively shifted from publishing most articles under a subscription model to disseminating them under Open Access arrangements, in some cases at very high costs. Although a considerable number of scientific journals still maintain the "hybrid" model (at no cost to authors), a large proportion now exclusively accept manuscripts for publication under the Open Access modality. Under this paradigm, hundreds of new journals spanning all disciplines have emerged over the past 2-3 years, whose scientific quality is more than questionable. Scientists receive dozens of spam e-mails daily inviting them to submit to these new journals, which raise serious concerns regarding their scientific rigor, publication quality, and international standing, given that most of them are not indexed in the most prestigious international databases such as Web of Science or Scopus in general, or PubMed/MEDLINE or Embase in particular. This implies that the curricular value for authors of publishing in such journals is very limited, if it exists at all. The lack of rigor characteristic of many of these journals is often reflected in their promotional e-mails, which invite submissions to journals bearing no relation whatsoever to the areas of expertise of the recipients, which is a suggestive indicator of how they operate.

Focusing specifically on those publishers held in high regard by international scientific communities, a critically important problem has emerged in a large proportion of their journals with respect to recruiting experienced reviewers to evaluate the constantly and increasingly growing number of manuscripts monthly received. My extensive experience as Editor-in-Chief of international journals has enabled me to identify this serious and highly relevant issue. For decades, the peer-review process has been, and continues to be, the key of manuscript acceptance or rejection in scientific journals. It is practically futile to invite the leading specialists in each field to review specific manuscripts, for which their expertise would be essential. Similar results are obtained when inviting reviewers at the next level of seniority, such that in many cases, after receiving numerous declinations or no response whatsoever, Editors end up inviting colleagues with limited scientific experience (often due to their early career stage), who tend to be more willing to accept review invitations. Furthermore, given the various ethical issues identified over time in certain journals regarding reviewer suggestions by authors, the difficulties faced by Editors in identifying the most appropriate reviewers for the manuscripts they handle are considerable. It is not unusual to invite 20 or 30 reviewers before finding at least two who accept, generally with an enormous delay in manuscript evaluation, a delay that has been increasing year after year. It is far from uncommon for authors to receive an initial decision six months or more after manuscript submission. This

situation is paradoxical in an era in which everything is conducted online, which should theoretically reduce the time required for any action or decision.

The above difficulties are well documented. Thus, available survey data indicate that peer review typically requires multiple days to complete and that time constraints constitute one of the most frequently cited challenges among researchers[1]. The escalating volume of submissions is a key driver. For example, Elsevier receives more than one million manuscript submissions annually across its 2,900 journals, of which only around 30% are ultimately published[2]. This structural imbalance between submission volume and reviewer availability has prompted growing interest in technology-assisted editorial workflows, and particularly, in the potential of AI tools to support, or partially automate, various stages of the peer-review process[3][2].

At present, a large proportion of Editors are overwhelmed by the substantial challenges involved in processing the large volumes of manuscripts they receive and in deciding whether to send them for external review, or to reject them directly. This pressure generates numerous opportunities for error, as it is impossible to thoroughly read all submitted material, meaning that initial decisions carry a considerable degree of arbitrariness, along with probable biases. For manuscripts deemed suitable for external review, the long process of finding and securing the agreement of the most appropriate reviewers begins. It should be noted that a certain number of potential reviewers have decided not to review for journals that are published exclusively under the Open Access model, unless they receive some form of compensation from the publishers. The principle of “you scratch my back, and I’ll scratch yours” has ceased to operate in the context of Open Access publications[4].

The problem described above is serious and demands imaginative solutions. In recent years, and particularly since 2024, the integration of Artificial Intelligence (AI) tools into the scientific publication process has grown substantially. Nowadays, most journals with a degree of prestige, include now in their author guidelines the requirement to disclose the use of AI in manuscript preparation, a requirement that, while entirely reasonable, may have a certain aspirational quality in terms of enforcement. What role, then, could AI tools play in the editorial evaluation process of manuscripts? This constitutes the primary goal of the present manuscript, whose specific aims are: 1) to examine how AI tools can assist Editors in their desk decisions, 2) to assess whether AI tools can act as reviewers of scientific manuscripts with a level of rigor equal to or greater than that of a human scientist, and 3) to consider how the findings of this study might be translated into practice with a view to reducing publication times and costs, both for publishers and, particularly, for authors.

A growing body of literature has begun to examine whether large language models (LLMs) can perform meaningful peer review functions. Liang et al.[5] conducted the first large-scale empirical analysis using GPT-4 to generate feedback on papers from 15 Nature family journals and the ICLR machine learning conference, finding 31–39% overlap with human reviewer comments and reporting that over half of surveyed authors rated

GPT-4 feedback as “helpful” or “very helpful”. However, differences in study design, particularly the use of already published manuscripts in the present work, limit direct comparability with previous investigations. On the other hand, Bauchner and Rivara[3] argued that LLMs could assist Editors in triaging manuscripts through in-context learning, while acknowledging that questions of effectiveness, fairness, and efficiency remain unresolved. In turn, Kousha and Thelwall[2] reviewed the available technology and concluded that, while AI is demonstrably useful for reviewer identification and initial quality control, evidence is insufficient to support the use of AI as a replacement for human reviewers in full manuscript evaluation. More recently, Zhu et al. [6] employed Claude 2.0 to generate peer review reports for 20 cancer biology manuscripts and found that, while LLM-generated reviews were “somewhat consistent” with human reviews, they lacked depth, particularly in detailed critique. A survey of more than 1,600 academics across 111 countries revealed that over 50% have already used AI tools while conducting peer review, often without explicit journal authorization[7]. Against this backdrop, the present study offers a novel empirical contribution by systematically applying multiple AI tools, across both the desk-decision and full-review stages, to a set of recently published experimental papers in toxicology, a discipline not previously examined in this context. The present work should be interpreted as a proof-of-concept exploratory analysis rather than a formal validation study. Its primary objective is to generate preliminary observations regarding the potential roles of AI tools in editorial workflows, rather than to establish definitive comparative performance metrics.

Methods

Undoubtedly, this is a highly preliminary study, which the author has restricted exclusively to his area of specialization, toxicology, and to a limited number of papers, journals, and AI tools. Given the exploratory nature of the study, no formal sample size calculation or statistical power analysis was performed. Specifically, eight papers were included in the study, and five AI tools were employed as reviewers. These are the eight papers, listed alphabetically:

- Asano T, Ohtani Y, Sugiura S, Taniguchi M, Zaitsu K. Metabolic profiling of the maternal liver in pregnant mice exposed to Di(2-ethylhexyl) phthalate (DEHP). Food Chem Toxicol. 2026 Apr;210:115928. doi: 10.1016/j.fct.2026.115928.
- Carle A, Preizal L, Amyot M, Rosabal M. Acute and Chronic Toxicity of Sediments Containing Platinum and Palladium on Freshwater Benthic Organisms Chironomus riparius and Hyalella azteca. J Appl Toxicol. 2026 Apr;46(4):1151-1163. doi: 10.1002/jat.4933.
- Chen Y, Xu G, Wu Z, Hao C, Yang C, Chen X. Ecotoxicity of Combined Polylactic Acid Microplastics and Thallium Pollution on the Functional Traits of *Folsomia candida. Toxics.* 2026; 14(4):307. doi: 10.3390/toxics14040307.

- Hinojosa M, Johanson G, Norinder U, Forsby A. Classification of industrial chemicals for respiratory chemosensory irritation using the TRPV1-expressing neuronal SH-SY5Y cell model and machine learning. Arch Toxicol. 2026 Apr;100(4):1301-1320. doi: 10.1007/s00204-025-04288-6.
- Liu X, Fan F. A large-scale concordance study of toxicity findings across preclinical species and humans for small molecules and biologics. Front Toxicol. 2026 Apr 2;8:1731947. doi: 10.3389/ftox.2026.1731947.
- Shrestha M, Pappas RS, Gonzalez-Jimenez N, Gray N, Watson CH, Valentín-Blasini L, Otgonsuren M, Taylor KM, Hassink M. Arsenic and Cadmium in Cigar Fillers: A Comparative Study within Cigar Types and with Cigarettes. Chem Res Toxicol. 2026 Apr 20;39(4):510-517. doi: 10.1021/acs.chemrestox.5c00084.
- Takashima H, Makino R, Taguchi H, Ito J, Mishima E, Takenaka Y, Akiyama Y, Sumi D, Conrad M, Tomikoka Y, Toyama T, Saito Y. Arsenite sensitizes to ferroptosis by disrupting selenium metabolism and reducing GPx4 expression. Toxicology. 2026 May;522:154409. doi: 10.1016/j.tox.2026.154409.
- Zeng XX, He WW, Liao W, Xiao X, Tu X, Deng J, Qi XL, Dong YT, Hong W, He Y, Xiao Y, Wei N, Guan ZZ. Alleviating effects of fat mass and obesity-associated protein on fluoride-induced neurotoxicity in rat brain, primary neurons, and SH-SY5Y cells. Toxicol Sci. 2026 Apr 7;209(4):kfag031. doi: 10.1093/toxsci/kfag031.

The selection of the eight journals in which the papers were published was entirely random. Nevertheless, drawing on my extensive experience as Editor-in-Chief, Associate Editor, and Editorial Board member of numerous journals, I considered the specifically selected ones, namely Toxicological Sciences (Oxford), Archives of Toxicology (Springer), Food and Chemical Toxicology (Elsevier), Toxicology (Elsevier), Journal of Applied Toxicology (Wiley), Frontiers in Toxicology (Frontiers), Toxics (MDPI), and Chemical Research in Toxicology (ACS), to enjoy, in general, a good reputation among toxicology researchers, with rather good and/or reasonable bibliometric indicators. The editorial provenance of the selected journals was deliberately diversified, encompassing a total of seven different publishing companies.

Furthermore, given the limited total number of articles included in the study, and to avoid selection bias, the papers evaluated were the first experimental article published in the April 2026 issue of each of the respective journals. All selected articles corresponded to original experimental studies, excluding reviews, commentaries, perspectives, and similar publication types. The sole exception was the paper published in the journal Toxicology, which did not publish a specific April issue, Consequently, the paper selected from that journal was the first article in the May 2026 issue.

Regarding the selection of the AI tools to evaluate the content of the papers, this was based exclusively on the personal experience of the author of the present manuscript. To determine whether the already-published papers, treated as manuscripts submitted to the respective journals, were worthy of being passed to external review, or should instead be rejected at the desk, the AI tools DeepSeek (expert mode), Mistral (Le Chat), and Kimi (K2.6 Instant) were employed. The prompt submitted was identical for all three AI tools: "*I am conducting a*

study on the usefulness of the AIs as potential Reviewers (alternative to human Reviewers) of scientific manuscripts submitted for publication in international journals. My study is aimed at detecting which could be a suitable AI to act as Reviewer of scientific manuscripts. The design of my study is simple. I have selected a few papers recently published in various toxicology journals. You must act as the desk scientific EDITOR of the eight attached papers. For this, you must not consider them as published papers, just as recently submitted manuscripts. Tell me if you feel that each of these papers is worthy of being passed to external review, or it should be rejected "on desk. Therefore, your decisions will be only "Continues to Review" or "Reject without Review". You must act as if the attached paper is a manuscript just submitted to the Journal, but NOT PUBLISHED YET. For making the decisions, you should consider the scientific interest and novelty of the manuscript, potential relevance in the field, appropriate methods including statistics if proceed, appropriate discussion based on the reported results, and updated references. Considering all these issues you should make that initial decision. Proceed indicating for each of them the reasons why you make the respective decisions. I need only a short sentence or paragraph for each one. No more than 4-5 lines. You must proceed avoiding fakes and hallucinations."

Papers that proceeded to peer review were subsequently subjected to evaluation by five AI tools acting as Reviewers. As was the case for the selection of the AI tools that performed the initial screening, the inclusion of the five AI tools chosen to act as Reviewers was based solely on the prior personal experience of the author with their use. The AI tools selected were the free versions of Gemini, Copilot, ChatGPT, Qwen 3.6, and Grok. The prompt submitted was as follows: "*Act as Reviewer of the enclosed paper. You should tell me if you recommendation to the Editor is: Accept in its present form, Minor revisions, Major revisions, or Reject. You must act as if the attached paper is a manuscript just submitted to the Journal, but NOT PUBLISHED YET. You must act as a Reviewer and you must not consider it as a published paper. For you, it is a submitted manuscript, and therefore, you must review it carefully*".

Finally, once the reports from each of the five AI tools were examined, they were submitted to two additional AI tools, which were asked to evaluate the quality and/or scientific value of the comments and suggestions included in the respective reports. Specifically, the AI tools employed in this process were Perplexity Pro and Claude Sonnet 4.6. The prompt submitted to both AI tools was as follows: "*I have considered the attached paper as if was not published yet. You must consider it as a manuscript submitted for publication in a scientific journal. I have asked to five Reviewers to prepare the corresponding reports and decision (recommendations) on the manuscript. You may find them also here attached. Tell me which has been (considering comments and suggestions) for you the best Reviewer and order them starting with the best one. Tell me who have been the most appropriate and useful report for the Editor to help with his/her decision. On the other hand, which would be your own recommendation for the Editor? Indicate: Accept in its present form, Minor revisions, Major revisions, or Reject*".

All AI tools were queried once using identical prompts, without iterative refinement or consensus generation. Therefore, the outputs reflect single-instance responses, consistent with a pragmatic usage scenario. On the

other hand, it is important to remark that all AI evaluations were conducted on already published versions of the manuscripts rather than on the originally submitted versions. Consequently, the comparison between AI-generated recommendations and human editorial decisions is indirect and should not be interpreted as a controlled head-to-head assessment.

Results and Discussion

In the first phase of the study, according to the evaluations performed by DeepSeek, Mistral, and Kimi, all eight papers, treated by the AI tools as manuscripts submitted to the respective journals, were considered worthy of being passed to peer review. Table 1 shows the comments of these three AI tools regarding the manuscript published in Toxicological Sciences, selected as an illustrative example. In relation to this, there was complete and absolute concordance between the initial editorial decisions taken by the Editors of the respective journals and the recommendations made by the three AI tools employed here. However, it should be noted that this concordance reflects decisions made on manuscripts that were ultimately accepted for publication. Therefore, it does not provide information on the capacity of AI tools to identify manuscripts unsuitable for peer review.

Zeng et al. (2026) Alleviating effects of fat mass and obesity-associated protein on fluoride-induced neurotoxicity in rat brain, primary neurons, and SH-SY5Y cells. *Toxicol Sci*. 209(4):kfag031.		
AI Tool	**Recommendation**	**Comments**
DeepSeek	Continues to review	This manuscript presents novel mechanistic insights into FTO-mediated neuroprotection against fluoride toxicity using complementary *in vivo* and *in vitro* approaches. The methodology is robust, statistics are appropriate, and the findings linking m6A demethylation to GluR2 regulation are timely and relevant to neurotoxicology.
Mistral	Continues to review	The study investigates the neuroprotective role of FTO in fluoride-induced neurotoxicity, a topic of significant public health relevance. The methods are robust, including *in vivo* and *in vitro* models, and the discussion is well-supported by the results. The references are current and relevant. The findings could have broad implications for understanding fluoride toxicity and potential therapeutic interventions.
Kimi	Continues to review	This manuscript presents a mechanistically focused study linking fluoride neurotoxicity to epitranscriptomic regulation via FTO/m6A and AMPAR GluR2. The experimental design spans *in vivo* (chronic fluorosis rat model with behavioral assessment), *ex vivo* (primary neurons), and *in vitro* (SH-SY5Y with lentiviral FTO manipulation), providing reasonable cross-validation. The use of actinomycin D chase for mRNA stability and MeRIP-qPCR for site-specific m6A detection on GluR2 adds methodological rigor. However, the sample size is very small (n=8 rats total, n=3-6 for cell work), and the fluoride concentrations used *in vitro* (up to 120 mg/l NaF) raise concerns about non-specific cytotoxicity versus targeted mechanisms. The correlation analyses lack adjustment for multiple comparisons, and the manuscript would benefit from addressing whether FTO's protective effects are specific to fluoride or general antioxidant/anti-apoptotic responses. The discussion appropriately notes the need for *in vivo* FTO conditional knockout models and electrophysiological validation.

Table 1. Editorial Checking: An example for a manuscript published in *Toxicological Sciences*.

The results of the subsequent phase, in which the five selected AI tools generated recommendations for the editors and authors of the respective manuscripts, were surprising. Not a single one of the eight evaluated papers, treated as submitted manuscripts by the AI tools, received a recommendation of “Accept in its present form” from any of them (Table 2). In contrast, the majority of the five AI tools, for most of the eight manuscripts,

recommended decisions of Major revisions. Examining the recommendations of each AI tool individually, the least stringent was Grok, with eight recommendations of Minor revisions and none of Major revisions, while the antithesis was Copilot, with eight recommendations of Major revisions and none of Minor revisions. ChatGPT, Qwen, and Gemini occupied intermediate positions, with 7, 6, and 5 recommendations of Major revisions, respectively. When the reports from each of the five AI tools were submitted to Claude and Perplexity, so that a final recommendation could be established on the basis of the five reports (as if they had been five human reviewers), Claude recommended Major revisions for all manuscripts, with the exception of the papers published in J Appl Toxicol and Chem Res Toxicol, for which it recommended Minor revisions. According to Claude, the best AI acting as a reviewer was ChatGPT, while the least effective was Grok. Copilot ranked second in six of the reviews. With respect to Perplexity, Table 2 shows that only the paper published in Chem Res Toxicol would require Minor revisions, whereas the remaining seven articles should have been subjected to Major revisions prior to publication. Regarding the performance of the five AI tools as reviewers, the most appropriate reports were produced by ChatGPT and Copilot (in four cases each), while according to Perplexity, the least highly rated reports corresponded to Gemini in six cases, and to Grok and Qwen in one case each.

	AI Tools					Claude		Perplexity	
Manuscript published in	**Gemini**	**Copilot**	**ChatGPT**	**Qwen**	**Grok**	**Final recommendation**	**Best report (AI tool) according to Claude**	**Final recommendation**	**Best report (AI tool) according to Perplexity**
Toxicological Sciences	**MINOR**	**MAJOR**	**MAJOR**	**MAJOR**	**MINOR**	**MAJOR**	ChatGPT, Gemini, Qwen, Copilot, Grok	**MAJOR**	ChatGPT, Copilot, Qwen, Grok, Gemini
Archives of Toxicology	**MAJOR**	**MAJOR**	**MAJOR**	**MAJOR**	**MINOR**	**MAJOR**	ChatGPT, Copilot, Qwen, Gemini, Grok	**MAJOR**	ChatGPT, Copilot, Gemini, Qwen, Grok
Food and Chemical Toxicology	**MAJOR**	**MAJOR**	**MAJOR**	**MAJOR**	**MINOR**	**MAJOR**	ChatGPT, Copilot, Qwen, Gemini, Grok	**MAJOR**	ChatGPT, Copilot, Qwen, Grok, Gemini
Toxicology	**MINOR**	**MAJOR**	**MAJOR**	**MAJOR**	**MINOR**	**MAJOR**	ChatGPT, Copilot, Gemini, Grok, Qwen	**MAJOR**	ChatGPT, Copilot, Qwen, Grok, Gemini
Journal of Applied Toxicology	**MINOR**	**MAJOR**	**MAJOR**	**MAJOR**	**MINOR**	**MINOR**	ChatGPT, Copilot, Gemini, Grok, Qwen	**MAJOR**	Copilot, ChatGPT, Grok, Gemini, Qwen
Frontiers in Toxicology	**MINOR**	**MAJOR**	**MAJOR**	**MINOR**	**MINOR**	**MAJOR**	ChatGPT, Copilot,	**MAJOR**	Copilot, ChatGPT,

doi.org/10.32388/LRQPSD

	AI Tools					Claude		Perplexity	
Manuscript published in	**Gemini**	**Copilot**	**ChatGPT**	**Qwen**	**Grok**	**Final recommendation**	**Best report (AI tool) according to Claude**	**Final recommendation**	**Best report (AI tool) according to Perplexity**
							Qwen, Gemini, Grok		Qwen, Grok, Gemini
Toxics	**MAJOR**	**MAJOR**	**MAJOR**	**MAJOR**	**MINOR**	**MAJOR**	ChatGPT, Copilot, Qwen, Gemini, Grok	**MAJOR**	Copilot, ChatGPT, Qwen, Grok, Gemini
Chemical Research in Toxicology	**MINOR**	**MAJOR**	**MINOR**	**MINOR**	**MINOR**	**MINOR**	ChatGPT, Qwen, Copilot, Gemini, Grok	**MINOR**	Copilot, ChatGPT, Qwen, Grok, Gemini

Table 2. AI tools acting as reviewers of various manuscripts published in international toxicology journals.

Major: Major revision; Minor: Minor revision.Overall, recommendations of "Major revisions" clearly predominated across AI tools, indicating a consistent tendency toward conservative evaluative outcomes.

The results of this exploratory analysis suggest a tendency for AI tools to generate more conservative evaluative recommendations compared to the outcomes implied by final editorial decisions. However, there was a clear correspondence between the initial decisions of the Editors of the respective journals (who chose to proceed to peer review with the original manuscripts) and the recommendations of the AI tools. Notwithstanding, given that the manuscripts as originally submitted were not available, it cannot be established whether the AI tools would have recommended desk rejection for any of the eight papers examined here. Nevertheless, given the nature of the revisions suggested (Major in many cases), this remains plausible.

The finding that AI tools were more demanding than the original human reviewers is consistent with observations already reported. Thus, Liang et al.[5] noted that LLM-generated feedback tended to be more comprehensive and systematic than human reviews, particularly for weaker manuscripts. The higher stringency observed in the present study may reflect the absence of social and cognitive biases that can affect human reviewers, such as familiarity with the research group, prior exposure to the topic, or time constraints. On the other hand, it is important to acknowledge that the AI tools evaluated here lacked access to the original submitted versions of the manuscripts and therefore reviewed already-revised, accepted, and published papers. This is a factor that must be considered when interpreting the discrepancy between AI recommendations (Major revisions in most cases) and the decisions of the original human reviewers (acceptance for publication). The variation in performance across AI tools is also noteworthy. Consistent with recent reports[2][6], no single tool proved uniformly superior, and performance differences across manuscripts suggest that an ensemble approach (combining outputs from multiple AI tools and adjudicated by a human Editor) may represent the most prudent model for AI-assisted peer review. Concerns regarding AI-specific limitations in this context have also been raised in the recent literature, including susceptibility to prompt injection attacks embedded within manuscripts[8], potential hallucinations of references or experimental details[6], as well as a tendency toward superficial rather than deep methodological critique[6]. These risks reinforce the conclusion that AI tools should be regarded as decision-support instruments rather than autonomous reviewers. It should be emphasized that a higher frequency of “Major revision” recommendations does not necessarily indicate superior review quality but may instead reflect a systematic tendency of LLMs toward conservative or risk-averse evaluative outputs.

Another limitation of the current work is that the evaluation of reviewer report quality was conducted using AI tools rather than human experts, which may introduce additional layers of bias and reduces the availability of an external reference standard.

Conclusions

Despite the very significant limitations of this preliminary study, given that none of the evaluated manuscripts received a recommendation of “Accept in its present form”, the findings suggest that AI tools may tend to produce more conservative recommendations than those reflected in final editorial decisions. Among the AI tools used as reviewers, ChatGPT and Copilot were most frequently identified by the meta-reviewer tools as producing comparatively useful reports. In contrast, Grok, according to the results obtained, would be the least reliable for acting as a scientific reviewer. These observations should not be interpreted as evidence that AI tools outperform human reviewers, but rather as indicative of their potential role as complementary decision-support systems

In any case, the results obtained highlight the need to conduct a large-scale study evaluating a substantial number of already-published articles (access to the original submitted versions would be ideal), with multiple AI tools acting as Editors for desk decisions, and as reviewers to formulate comments and suggestions, and to recommend corresponding final decisions. If it is confirmed that AI can replace human activity in this domain, the implications would involve changes of a tremendous and undetermined magnitude. However, it must be emphasized that the use of AI tools in peer review raises important ethical questions regarding transparency, accountability, confidentiality, and potential bias that will need to be addressed through clear policy frameworks before any large-scale implementation is contemplated[3][9]. Therefore, the present study should be regarded as a proof-of-concept preliminary investigation, and its findings interpreted accordingly.

Statements and Declarations

Funding

This article was conducted without specific funding support.

Potential Competing Interests

The author declares no conflicts of interest related to the preparation of this manuscript.

Ethics

This study did not involve human participants, human-derived data, or animals. It used exclusively previously published, publicly available scientific articles as test material. Accordingly, no ethics approval or informed consent was required.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. ^Lee J, Lee J, Yoo JJ (2025). "The Role of Large Language Models in the Peer-Review Process: Opportunities and Challenges for Medical Journal Reviewers and Editors." J Educ Eval Health Prof. 22:4. doi:10.3352/jeehp.2025.22.4.
2. a, b, c, dKousha K, Thelwall M (2024). "Artificial Intelligence to Support Publishing and Peer Review: A Summary and Review." Learned Publishing. 37(1):4–12. doi:10.1002/leap.1570.

3. a, b, c Bauchner H, Rivara FP (2024). "Use of Artificial Intelligence and the Future of Peer Review." Health Aff Scholar. **2**(5):qxae058. doi:10.1093/haschl/qxae058.
4. ^ Domingo JL (2024). "To Publish Scientific Journals: For Some, the Big Business of the Century." Qeios. **6**(2). doi:10.32388/YTFEF2.2.
5. a, b Liang W, Zhang Y, Cao H, Wang B, Ding DY, Yang X, Vodrahalli K, He S, Smith D, Yin Y, McFarland D, Zou J (2024). "Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis." NEJM AI. **1**(8). doi:10.1056/AIoa2400196.
6. a, b, c, d Zhu L, Lai Y, Xie J, Mou W, Huang L, Qi C, Tang Y, Jiang A, Gan W, Zeng D, Tang B, Xiao M, Chu G, Liu Z, Cheng Q, Lin A, Luo P (2025). "Evaluating the Potential Risks of Employing Large Language Models in Peer Review." Clin Transl Discov. **5**:e70067. doi:10.1002/ctd2.70067.
7. ^ Naddaf M (2026). "More Than Half of Researchers Now Use AI for Peer Review—Often Against Guidance." Nature. **649**(8096):273–274. doi:10.1038/d41586-025-04066-5.
8. ^ Bauchner H, Rivara F (2026). "Using AI to Improve Peer Review and Research Integrity in Scientific Journals." Hea lth Aff Scholar. **4**(2):qxag028. doi:10.1093/haschl/qxag028.
9. ^ Li ZQ, Xu HL, Cao HJ, Liu ZL, Fei YT, Liu JP (2024). "Use of Artificial Intelligence in Peer Review Among Top 100 Medical Journals." JAMA Netw Open. **7**(12):e2448609. doi:10.1001/jamanetworkopen.2024.48609.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.