

Research Article

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

Jose L. Domingo¹¹. School of Medicine, Laboratory of Toxicology and Environmental Health, Universitat Rovira i Virgili, Spain

Scientific peer review is under increasing pressure because manuscript submissions continue to rise while the number of available expert reviewers declines. Artificial intelligence (AI) tools may help support editorial decisions, but their ability to act as independent reviewers has not been adequately evaluated. This exploratory study assessed how current AI systems evaluate scientific manuscripts across three disciplines: toxicology, environmental sciences, and food science and technology. Twenty-four recently published papers were evaluated using a structured review protocol. Three AI tools (DeepSeek, Mistral, and Kimi) performed initial desk screening; five tools (Gemini, Copilot, ChatGPT, Qwen, and Grok) generated peer-review reports and recommendations; and two tools (Claude and Perplexity) assessed the quality of those reports and provided final recommendations. Across all disciplines, AI reviewers most frequently recommended major revision, although every manuscript had already undergone conventional peer review and had been accepted for publication. Recommendations to accept manuscripts without revision were uncommon and were made only by Gemini. Review patterns were generally consistent within each AI system across disciplines, suggesting that the type of recommendation depended more on the tool than on the research field. Importantly, stricter recommendations were not necessarily linked to more useful reviews. ChatGPT and Qwen produced the most detailed and constructive comments, whereas some more permissive tools generated less informative reports. These findings suggest that current AI systems tend to evaluate published scientific papers more critically than human reviewers. AI may therefore have an important role in improving the rigor, consistency, and efficiency of peer review, while helping reduce reviewer workload and limiting some interpersonal biases. However, this study is exploratory and based on a limited number of articles. Larger prospective studies should directly compare AI and human reviews of newly submitted manuscripts. If their recommendations are sufficiently consistent, AI-assisted peer review could help address reviewer shortages and reduce publication delays.

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

Introduction

Throughout the present century, the scientific publication system has undergone profound transformations affecting all stakeholders in scholarly communication, including authors, reviewers, editors, publishers, and readers. One of the most visible changes has been the progressive transition from traditional subscription-based publishing toward Open Access (OA) models. While many journals continue to operate under hybrid systems, a substantial proportion is now published exclusively under OA arrangements, often involving considerable article-processing charges (APCs). Alongside these developments, the rapid expansion of new journals in recent years has raised concerns regarding their scientific rigor, indexing status, and overall contribution to the advancement of knowledge.

Parallel to these structural and economic changes, the volume of scientific output has dramatically increased. Major publishers now receive unprecedented numbers of manuscript submissions, placing substantial pressure on editorial systems. For example, Elsevier alone receives more than one million submissions annually^[1]. This expansion has created a critical imbalance between submission rates and reviewer availability, giving rise to one of the most pressing challenges in contemporary scientific publishing: the difficulty of identifying and recruiting qualified reviewers willing to dedicate the time required for rigorous manuscript evaluation.

Extensive editorial experience, supported by the available literature, indicates that securing appropriate reviewers has become increasingly problematic. It is no longer unusual for editors to invite 20, 30, or even more potential reviewers before obtaining the acceptance of only a small (1-3) number, often resulting in delays of several months before a first decision is reached. Although editorial processes are now fully digital, review times have not decreased accordingly. Even in some fields, they have lengthened. Survey-based evidence confirms that peer review is widely perceived as time-consuming and difficult to reconcile with growing academic workloads^[2].

These challenges extend beyond reviewer recruitment to the broader editorial decision-making process. Editors are frequently required to handle large volumes of submissions under significant time constraints, which may introduce elements of arbitrariness or bias, particularly during initial desk evaluations. Additional complications arise from ethical concerns associated with reviewer suggestions and from evolving attitudes toward peer review, including reluctance to review for certain OA journals without compensation^[3]. Collectively, these factors highlight the need for innovative approaches capable of supporting editorial workflows and improving the efficiency and consistency of manuscript evaluation. Beyond workload and logistical constraints, human peer review is also exposed to relational and interpersonal biases, which artificial intelligence (AI)-based reviewers might help to mitigate, provided their own algorithmic biases are carefully managed.

Recent advances in AI, particularly large language models (LLMs), have generated significant interest regarding their potential role in scientific publishing. The integration of AI tools into academic practice has accelerated rapidly, and many journals now require authors to disclose their use during manuscript preparation. However, a

more fundamental question remains unresolved: to what extent can AI systems meaningfully assist editors and reviewers in the evaluation of scientific manuscripts? Regarding this, a growing body of evidence suggests that AI tools may provide useful support in the peer-review process. Liang et al.^[4] conducted a large-scale analysis using GPT-4 and reported a 31–39% overlap between AI-generated and human reviewer comments, with many authors rating the feedback as helpful. In turn, Bauchner and Rivara^[5] proposed that AI could assist in editorial triage and manuscript screening, while Kousha and Thelwall^[1] concluded that AI is already useful for reviewer identification and quality control, although insufficient evidence supports its use as a full replacement for human reviewers. In a more recent contribution, Bauchner and Rivara^[6] further argued that, with the biomedical literature now exceeding three million published manuscripts per year, human editorial and peer review alone can no longer safeguard research integrity, making AI assistance increasingly necessary. Similarly, Zhu et al.^[7] found that AI-generated reviews can resemble human assessments but often lack depth, especially in detailed methodological critique. Additionally, a recent large-scale survey indicates that more than 50% of researchers have already used AI tools in peer review, frequently without explicit authorization from journals^[8]. At the journal-policy level, Li et al.^[9] surveyed the top 100 medical journals and found that, although most now provide explicit guidance on AI-assisted peer review, a substantial fraction still prohibits it, primarily owing to confidentiality concerns.

Based on this background, the present study has analyzed a total of twenty-four papers, published in 24 journals belonging to three disciplines (toxicology, environmental and food sciences), using consistent procedures and AI tools. The specific objectives of this study are: (a) to evaluate the capacity of contemporary AI systems to assist editors during initial manuscript screening, (b) to assess the recommendations generated by AI systems acting as manuscript reviewers, (c) to compare the quality and usefulness of reviewer reports produced by different AI tools, and (d) to determine whether the reviewer profiles exhibited by individual AI systems remain stable across multiple scientific disciplines. The results should be interpreted as a proof-of-concept exploratory investigation aimed at generating hypotheses and preliminary observations regarding the potential role of AI in peer review, rather than as a definitive assessment of performance.

Methods

Study design and AI tools

This is a preliminary study, restricted to a limited number of papers, journals, and AI tools, which has been conducted in three stages. The first stage was restricted to toxicology, while the second stage used the same design to Earth and environmental sciences. Finally, the third stage extends the study to food science and technology. In each stage, eight papers were included and the same AI tools were used, giving a combined total of twenty-four papers. Given the exploratory nature of the work, no formal sample size calculation or statistical power analysis was performed.

The same three AI tools were used to emulate the desk Editor making the initial decision (Continue to Review, or Reject without Review): DeepSeek (expert mode), Mistral (Le Chat), and Kimi (K2.6 Instant). The same five AI tools were used to act as Reviewers of the papers that proceeded to peer review: the free versions of Gemini, Copilot, ChatGPT, Qwen 3.6, and Grok. Finally, the same two AI tools were used as meta-reviewers to evaluate the quality and scientific value of the five reviewer reports and to issue an overall recommendation: Claude Opus 4.8 and Perplexity Pro. The selection of all AI tools was based exclusively on the prior personal experience of the author. 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.

The prompt submitted to the three desk-Editor tools was identical and asked each tool to act as the desk scientific Editor of the attached papers, treating them “not as published papers but as recently submitted manuscripts”, and to decide only “Continue to Review” or “Reject without Review” based on scientific interest and novelty, potential relevance, appropriateness of the methods (including statistics where applicable), adequacy of the discussion in relation to the reported results, and currency of the references, providing a short justification of 4–5 lines for each decision, for which to avoid fakes and hallucinations was especially required. The prompt submitted to the five reviewer tools asked each to act as Reviewer of the enclosed paper, treating it as a just-submitted (not yet published) manuscript, and to recommend one of: Accept in its present form, Minor revisions, Major revisions, or Rejection. The prompt submitted to the two meta-reviewer tools asked them to consider the manuscript together with the five reviewer reports, to rank the five reviewers from best to worst (considering the usefulness of their comments and suggestions for the Editor), and to issue their own overall recommendation (Accept, Minor, Major, or Reject). It is important to remark once again, that in all stages, the AI evaluations were conducted on the already-published versions of the papers 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.

To avoid an overly long Methods section and the repeated use of full bibliographic citations within the running text, the twenty-four assessed papers (eight per discipline) are not cited individually in the text by their complete bibliographic details. Instead, each is referred to in the abbreviated author-year form (Author et al., year). The full bibliographic details of all 24 assessed papers are provided, together with the rest of citations, in the list of References.

Toxicology papers

Eight papers published in eight different toxicology journals were evaluated, encompassing seven different publishing companies. To avoid selection bias, the paper chosen from each journal was the first experimental article published in its April 2026 issue (the sole exception being Toxicology, which did not publish a specific April issue, so the first article of the May 2026 issue was used). All were original experimental studies, whereas reviews,

commentaries, perspectives, and similar publication types were excluded. The eight toxicology papers, with their journals, were: Zeng et al.^[10] in *Toxicological Sciences* (Oxford); Hinojosa et al.^[11] in *Archives of Toxicology* (Springer); Asano et al.^[12] in *Food and Chemical Toxicology* (Elsevier); Takashima et al.^[13] in *Toxicology* (Elsevier); Carle et al.^[14] in *Journal of Applied Toxicology* (Wiley); Liu and Fan^[15] in *Frontiers in Toxicology* (Frontiers); Chen et al.^[16] in *Toxics* (MDPI); and Shrestha et al.^[17] in *Chemical Research in Toxicology* (ACS).

Earth/environmental-science papers

Eight papers were also evaluated, each published in a different and reputable Earth/environmental-science journal. The selected papers were the first articles published in the June issues of the respective journals. With the single exception of the systematic review and meta-analysis published in *The Lancet Planetary Health*^[18], all selected articles were original research papers, spanning epidemiological, field, analytical, materials, and methodological topics, which provides a markedly different thematic and methodological profile from the toxicology set. The same desk-screening, reviewing, and meta-reviewing procedure described above was applied without modification, and the eight environmental-science papers were carried forward to the peer-review phase. The eight environmental-science papers, with their journals, were: Laue et al.^[19] in *Environmental Health Perspectives* (ACS); Megraso et al.^[20] in *Environmental Science and Pollution Research* (Springer); Liang et al.^[21] in *Environmental Science & Technology* (ACS); Menone et al.^[22] in *Archives of Environmental Contamination and Toxicology* (Springer); Peng et al.^[23] in *Environment International* (Elsevier); Chatzitoulousis et al.^[24] in *Environmental Pollution* (Elsevier); Gocer and Ramezannia^[25] in *Science of the Total Environment* (Elsevier); and Bonell et al.^[18] in *The Lancet Planetary Health* (Elsevier/The Lancet).

Food-science papers

The same protocol was applied to a third field, food science and technology. Eight further papers were evaluated, each published in a different and reputable food-science journal and selected following the same criterion used with the previous fields. All eight were recently published (mostly, June 1-3, 2026). All were original research articles spanning food chemistry, nutrition, food microbiology and safety, analytical and detection methods, and food-processing topics, again providing a thematic and methodological profile distinct from the two earlier sets. The same desk-screening, reviewing, and meta-reviewing procedure was applied without modification. As in the previous stages, all eight papers were carried forward to the reviewing phase so that the complete set could be evaluated by the five reviewer tools. The eight food-science papers, with their journals, were: Guo et al.^[26] in *Food Chemistry* (Elsevier); Cha et al.^[27] in *Food & Function* (Royal Society of Chemistry); Xing et al.^[28] in *Food Research International* (Elsevier); Luo et al.^[29] in *Food Science & Nutrition* (Wiley); Nunes et al.^[30] in *Foods* (MDPI); Li et al.

[31] in Journal of Agricultural and Food Chemistry (ACS); Luedke et al.[32] in Journal of Food Protection (Elsevier/IAFP); and Cai et al.[33] in npj Science of Food (Springer Nature).

Results and Discussion

The present study is focused on examining the behavior of multiple contemporary AI tools acting as editors, reviewers, and meta-reviewers across three distinct scientific domains: toxicology, environmental sciences, and food science and technology. By applying a consistent protocol to twenty-four recently published papers, the study provides a structured framework to explore whether AI-assisted manuscript evaluation exhibits tool-dependent or discipline-dependent patterns, and whether such patterns remain stable across methodologically heterogeneous fields.

Across all three disciplines, a consistent and robust observation was the predominance of recommendations for “Major revisions” at the reviewer level (Tables 1–3). This tendency was particularly evident in toxicology (Table 1) and food science (Table 3), while environmental sciences (Table 2) showed a somewhat less stringent profile, with a higher proportion of “Minor revisions” and the only cases in which “Accept in its present form” recommendations were issued. Although this difference may suggest variations in disciplinary conventions, manuscript characteristics, or methodological expectations, it should be interpreted cautiously given the limited and non-equivalent datasets. More importantly, it must be emphasized that all AI evaluations were conducted on “final, published versions of the manuscripts” rather than on their originally submitted forms. Consequently, the relatively high frequency of critical recommendations cannot be interpreted as evidence that AI systems outperform human reviewers but may instead reflect a combination of conservative evaluative tendencies and the intrinsic characteristics of revised manuscripts.

Manuscript published in	AI Tools					Claude		Perplexity	
	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, Qwen,	MAJOR	Copilot, ChatGPT, Qwen,

Manuscript published in	AI Tools					Claude		Perplexity	
	Gemini	Copilot	ChatGPT	Qwen	Grok	Final recommendation	Best report (AI tool) according to Claude	Final recommendation	Best report (AI tool) according to Perplexity
							Gemini, Grok		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 1. AI tools acting as reviewers of eight manuscripts published in international toxicology journals.

Manuscript published in	AI Tools					Claude		Perplexity	
	Gemini	Copilot	ChatGPT	Qwen	Grok	Final recommendation	Best report (AI tool) according to Claude	Final recommendation	Best report (AI tool) according to Perplexity
Environmental Health Perspective	ACCEPT	MAJOR	MINOR	MINOR	MINOR	MINOR	Qwen, Grok, ChatGPT, Copilot, Gemini	MINOR	Qwen, Grok, ChatGPT, Copilot, Gemini
Environmental Science and Pollution Research	MINOR	MAJOR	MAJOR	MAJOR	MINOR	MAJOR	Qwen, ChatGPT, Copilot, Grok, Gemini	MAJOR	Qwen, ChatGPT, Copilot, Grok, Gemini
Environmental Science and Technology	MINOR	MAJOR	MINOR	MINOR	MAJOR	MINOR (border Major)	Qwen, Grok, Copilot, ChatGPT, Gemini	MAJOR	Copilot, Grok, ChatGPT, Qwen, Gemini
Archives of Environmental Contamination and Toxicology	MINOR	MAJOR	MAJOR	MAJOR	MAJOR	MAJOR	Qwen, ChatGPT, Grok, Copilot, Gemini	MAJOR	ChatGPT, Copilot, Grok, Qwen, Gemini
Environment International	MAJOR	MAJOR	MAJOR	MAJOR	MINOR	MAJOR	ChatGPT, Qwen, Copilot, Gemini, Grok	MAJOR	ChatGPT, Gemini, Grok, Qwen, Copilot
Environmental Pollution	ACCEPT	MINOR	MINOR	MINOR	MINOR	MINOR	Qwen, ChatGPT, Grok,	MINOR	Grok, ChatGPT, Copilot,

Manuscript published in	AI Tools					Claude		Perplexity	
	Gemini	Copilot	ChatGPT	Qwen	Grok	Final recommendation	Best report (AI tool) according to Claude	Final recommendation	Best report (AI tool) according to Perplexity
							Copilot, Gemini		Qwen, Gemini
Science of the Total Environment	MINOR	MAJOR	MAJOR	MAJOR	MAJOR	MAJOR	Copilot, ChatGPT, Qwen, Grok, Gemini	MAJOR	ChatGPT, Qwen, Grok, Copilot, Gemini
Lancet Planetary Health	MINOR	MAJOR	MAJOR	MAJOR	MINOR	MAJOR	Qwen, ChatGPT, Grok, Copilot, Gemini	MAJOR	ChatGPT, Qwen, Copilot, Grok, Gemini

Table 2. AI tools acting as reviewers of eight manuscripts published in international Earth/environmental sciences journals.

Manuscript published in	AI Tools					Claude		Perplexity	
	Gemini	Copilot	ChatGPT	Qwen	Grok	Final recommendation	Best report (AI tool) according to Claude	Final recommendation	Best report (AI tool) according to Perplexity
Food Chemistry	MAJOR	MAJOR	MAJOR	MINOR	MAJOR	MAJOR	ChatGPT, Copilot, Qwen, Grok, Gemini	MAJOR	Copilot, Grok, ChatGPT, Qwen, Gemini
Food & Function	MINOR	MAJOR	MINOR	MINOR	MINOR	MAJOR (border Minor/Major)	Copilot, ChatGPT, Qwen, Grok, Gemini	MAJOR	Copilot, Grok, ChatGPT, Qwen, Gemini
Food Research International	MINOR	MAJOR	MAJOR	MINOR	MAJOR	MAJOR	Qwen, ChatGPT, Copilot, Grok, Gemini	MAJOR	Copilot, ChatGPT, Grok, Qwen, Gemini
Food Science & Nutrition	MAJOR	MAJOR	MAJOR	MAJOR	MAJOR	MAJOR (with a caveat to move to reject)	Qwen, ChatGPT, Grok, Copilot, Gemini	MAJOR	Qwen, Copilot, Grok, ChatGPT, Gemini
Foods	MINOR	MINOR	MAJOR	MAJOR	MAJOR	MAJOR	Qwen, ChatGPT, Grok, Copilot, Gemini	MAJOR	Qwen, ChatGPT, Grok, Copilot, Gemini
Journal of Agricultural	ACCEPT	MAJOR	MINOR	MINOR	MINOR	MINOR	ChatGPT, Qwen, Grok,	MINOR	Copilot, ChatGPT, Grok,

Manuscript published in	AI Tools					Claude		Perplexity	
	Gemini	Copilot	ChatGPT	Qwen	Grok	Final recommendation	Best report (AI tool) according to Claude	Final recommendation	Best report (AI tool) according to Perplexity
and Food Chemistry							Copilot, Gemini		Qwen, Gemini
Journal of Food Protection	MINOR	MAJOR	MAJOR	MAJOR	MINOR	MAJOR	Qwen, ChatGPT, Copilot, Grok, Gemini	MAJOR	Qwen, Copilot, ChatGPT, Grok, Gemini
npj Science of Food	MINOR	MAJOR	MAJOR	MAJOR	MINOR	MAJOR	Qwen, ChatGPT, Copilot, Grok, Gemini	MAJOR	ChatGPT, Copilot, Qwen, Grok, Gemini

Table 3. AI tools acting as reviewers of eight manuscripts published in international food sciences journals.

A second notable finding is the consistency of reviewer profiles across disciplines, which emerges as one of the most stable patterns in the study. Copilot consistently exhibited the most stringent behavior, issuing “Major revisions” recommendations in most cases across all three fields. In contrast, Gemini displayed the most permissive profile and was the sole source of all “Accept in its present form” recommendations observed in the study. ChatGPT and Qwen showed relatively balanced and stable performance across disciplines, while Grok exhibited a transition from higher permissiveness in toxicology to more moderate behavior in the other fields (Table 4). The persistence of these relative profiles across methodologically diverse domains suggests that the evaluative tendencies of individual AI tools may be largely intrinsic to the systems themselves, rather than primarily determined by disciplinary content. While this interpretation is necessarily preliminary, it may have practical implications for the design of AI-assisted editorial workflows, particularly in relation to tool selection and combination.

AI tool (reviewer)	Toxicology (n = 8)			Environmental sciences (n = 8)			Food sciences (n = 8)		
	Accept	Minor	Major	Accept	Minor	Major	Accept	Minor	Major
Gemini	0	5	3	2	5	1	1	5	2
Copilot	0	0	8	0	1	7	0	1	7
ChatGPT	0	1	7	0	3	5	0	2	6
Qwen	0	2	6	0	3	5	0	4	4
Grok	0	8	0	0	5	3	0	4	4
Claude (meta-reviewer)	0	2	6	0	3	5	0	1	7
Perplexity (meta-reviewer)	0	1	7	0	2	6	0	1	7

Table 4. Distribution of recommendations issued by each AI tool acting as reviewer, by discipline (number of papers out of eight in each field).

In addition to recommendation patterns, the study provides insight into the perceived quality and usefulness of reviewer reports, as assessed by the two meta-reviewer systems. Across the three disciplines, reports generated by ChatGPT and Qwen were most frequently ranked as the most useful, whereas Gemini's reports were consistently ranked lowest despite its more favorable recommendations. It highlights a notable dissociation between recommendation severity and report quality, indicating that the informativeness and utility of reviewer feedback cannot be inferred solely from the final recommendation. The consistency of this pattern across disciplines reinforces the need to evaluate AI-generated reviews along multiple dimensions, including clarity, specificity, and constructiveness.

When considered jointly, the results suggest that current AI tools can reproduce certain structural features of the peer-review process, including differentiation in stringency, variability in recommendations, and generation of structured critiques. However, no single system demonstrated uniformly superior performance across all dimensions. This observation aligns with previous studies indicating that AI-generated reviews can be informative but may lack depth or consistency in methodological critique. In this context, the results support the notion that an ensemble approach, combining outputs from multiple AI tools and interpreted by human editors, may offer a more balanced and reliable strategy than reliance on individual systems.

The comparison across disciplines also provides some indication that contextual factors may influence overall stringency, although to a lesser extent than tool-specific characteristics. Environmental-science manuscripts

appeared, on average, to elicit slightly less stringent evaluations than toxicology and food-science papers. Whether this reflects inherent differences in reporting standards, study designs, or simply the specific selection of manuscripts cannot be determined within the present study. Nonetheless, the fact that tool-specific profiles were preserved across all three domains suggests that inter-tool variability is a more dominant factor than disciplinary variability, at least within the scope of the current dataset.

Several important limitations must be considered when interpreting the current findings. First, the study is based on a relatively small number of manuscripts, journals, and AI tools, and does not include formal statistical analysis. Therefore, the results are merely descriptive and should be regarded as exploratory. Second, the use of published manuscripts introduces an unavoidable constraint, as these texts do not correspond to the versions originally evaluated during peer review. Third, each AI tool was queried only once using fixed prompts, so the analysis does not account for potential variability arising from stochastic outputs or prompt sensitivity. Fourth, the evaluation of reviewer report quality was conducted by AI-based meta-reviewers rather than by independent human experts, which may introduce additional layers of bias. Finally, although the three disciplinary groups were analyzed using an identical protocol, they are not strictly equivalent in composition, as they encompass different study designs and levels of heterogeneity.

Taken together, these limitations indicate that the present findings should be interpreted as hypothesis-generating observations rather than definitive conclusions regarding the capabilities of AI tools in peer review. Future studies involving larger datasets, repeated queries, direct comparison with human reviewer reports, and access to originally submitted manuscripts are necessary to confirm and extend these observations. Such research would also help clarify whether the patterns identified here persist under more controlled and reproducible conditions.

Building on this, the single most relevant study that could be undertaken at present would be a prospective, head-to-head comparison between conventional peer review (conducted by two, three, or four human reviewers) and an equivalent number of AI tools evaluating the same submitted manuscripts. Such a study could realistically be carried out only by Publishers willing to extend it across several of their journals, using manuscripts at the point of initial submission and with the indispensable collaboration of the corresponding Editors. These Editors would invite human reviewers in the usual manner, while simultaneously submitting the same manuscripts to a predefined set of AI tools. If this design were applied to a sufficiently large and diverse set of manuscripts and journals, and the recommendations issued by humans and AI systems proved to be broadly concordant, the currently acute problem of identifying suitable reviewers willing to complete high-quality evaluations within a reasonable timeframe could be effectively resolved. If AI tools were to perform adequately as reviewers, it would seem paradoxical that, with the technologies currently available, the time required for a manuscript to be reviewed and eventually published should not be comparable to that which was customary in the 1980s, when the entire editorial workflow relied on ordinary postal mail.

Conclusions

The present study provides an exploratory cross-disciplinary assessment of the performance of contemporary AI tools acting as manuscript reviewers within a structured and consistent evaluation framework. By analyzing twenty-four papers across three distinct scientific domains, the results reveal several consistent and informative patterns regarding how these systems evaluate scientific manuscripts. A central and robust finding is that AI tools systematically produced more stringent evaluations than might be inferred from the actual editorial outcomes of the manuscripts. Across all three disciplines, recommendations of “Major revisions” clearly predominated, despite the fact that all evaluated manuscripts had already undergone conventional peer review and had been accepted for publication. Importantly, this pattern emerged even though the AI systems assessed “final, published versions”, which would reasonably be expected to elicit more favorable evaluations. This observation strongly suggests that current AI tools tend to apply more demanding and less permissive evaluative criteria than those typically observed in human peer review.

In this context, the results support the view that AI systems, at least in their current state, can represent not only complementary tools, but also potential alternatives to human reviewers, particularly in terms of maintaining or increasing the level of critical scrutiny applied to scientific manuscripts. Their apparent tendency toward stricter evaluation may contribute to reducing the acceptance of manuscripts that do not meet sufficiently rigorous methodological or interpretative standards, thereby potentially enhancing the overall robustness of the peer-review process. An additional advantage of the potential use of AI tools as reviewers, compared with traditional peer review conducted by scientists, is their neutrality. Editors are well aware that the selection of invited reviewers is one of the critical steps in the process, particularly in highly specialized fields where the pool of true experts is very limited. Consequently, it is not uncommon for invited reviewers to be either closely aligned with the authors or, conversely, individuals with whom scientific or even personal relationships may not be sufficiently neutral to ensure an unbiased evaluation. These circumstances, which often remain beyond the editors’ awareness, are largely unavoidable, as in any human-driven activity. However, in the context of scientific manuscript review, they can introduce substantial bias into the recommendations, an issue that would, in principle, be minimized if reviews were conducted by AI tools.

A second key finding is the existence of stable reviewer profiles across disciplines, indicating that the behavior of AI tools is largely determined by their intrinsic characteristics rather than by the subject matter being evaluated. This consistency reinforces the relevance of AI systems as reproducible evaluative agents, in contrast with the well-documented variability often observed among human reviewers. Furthermore, the study shows that the quality and usefulness of reviewer reports are not directly linked to recommendation severity, highlighting that some AI tools can generate detailed and constructive critiques while maintaining a relatively stringent evaluative stance.

This combination of structured feedback and critical assessment may represent a valuable asset in editorial workflows.

However, these findings should be interpreted considering the exploratory nature of the study and its methodological constraints. It includes the limited sample size, the use of published manuscripts, and the reliance on single-instance AI outputs. Nevertheless, rather than weakening the main observation, the fact that AI tools remained highly critical even when evaluating already accepted papers can reinforce the conclusion that their evaluative approach is inherently conservative.

In summary, the present results suggest that contemporary AI systems have reached a level of performance at which they can realistically be considered viable participants in the peer-review process, not only as supportive instruments but also as potentially more stringent and consistent evaluators than human reviewers in certain contexts. While further large-scale and controlled studies are required to confirm these findings, the evidence provided here indicates that AI-assisted (or even AI-driven) peer review may play an increasingly important role in ensuring the rigor and reliability of scientific publishing. Regarding this, a prospective head-to-head comparison between human reviewers and AI tools, conducted by Publishers across several of their journals on newly submitted manuscripts, would constitute the decisive next step. Should human and AI recommendations prove broadly concordant, the persistent difficulty of securing timely, high-quality reviewers could be largely overcome, and review and publication times could realistically avoid those of the pre-digital era.

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. 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. ^a Kousha 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](https://doi.org/10.1002/leap.1570).
2. ^a 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](https://doi.org/10.3352/jeehp.2025.22.4).
3. ^a Domingo JL (2024). "To Publish Scientific Journals: For Some, the Big Business of the Century." *Qeios*. doi:[10.32388/YTFEF2.2](https://doi.org/10.32388/YTFEF2.2).
4. ^a 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](https://doi.org/10.1056/AIoa2400196).
5. ^a Bauchner H, Rivara FP (2024). "Use of Artificial Intelligence and the Future of Peer Review." *Health Affairs Scholar*. 2(5):qxae058. doi:[10.1093/haschl/qxae058](https://doi.org/10.1093/haschl/qxae058).
6. ^a Bauchner H, Rivara F (2026). "Using AI to Improve Peer Review and Research Integrity in Scientific Journals." *Health Affairs Scholar*. 4(2):qxag028. doi:[10.1093/haschl/qxag028](https://doi.org/10.1093/haschl/qxag028).
7. ^a 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." *Clinical and Translational Discovery*. 5:e70067. doi:[10.1002/ctd2.70067](https://doi.org/10.1002/ctd2.70067).
8. ^a 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](https://doi.org/10.1038/d41586-025-04066-5).
9. ^a 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 Network Open*. 7(12):e2448609. doi:[10.1001/jamanetworkopen.2024.48609](https://doi.org/10.1001/jamanetworkopen.2024.48609).
10. ^a 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 (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. doi:[10.1093/toxsci/kfag031](https://doi.org/10.1093/toxsci/kfag031).
11. ^a Hinojosa M, Johanson G, Norinder U, Forsby A (2026). "Classification of Industrial Chemicals for Respiratory Chemosensory Irritation Using the TRPV1-Expressing Neuronal SH-SY5Y Cell Model and Machine Learning." *Arch Toxicol*. 100(4):1301–1320. doi:[10.1007/s00204-025-04288-6](https://doi.org/10.1007/s00204-025-04288-6).
12. ^a Asano T, Ohtani Y, Sugiura S, Taniguchi M, Zaitzu K (2026). "Metabolic Profiling of the Maternal Liver in Pregnant Mice Exposed to Di(2-ethylhexyl) Phthalate (DEHP)." *Food Chem Toxicol*. 210:115928. doi:[10.1016/j.fct.2026.115928](https://doi.org/10.1016/j.fct.2026.115928).
13. ^a Takashima H, Makino R, Taguchi H, Ito J, Mishima E, Takenaka Y, Akiyama Y, Sumi D, Conrad M, Tomioka Y, Toyama T, Saito Y (2026). "Arsenite Sensitizes to Ferroptosis by Disrupting Selenium Metabolism and Reducing GPx4 Expression." *Toxicology*. 522:154409. doi:[10.1016/j.tox.2026.154409](https://doi.org/10.1016/j.tox.2026.154409).

14. ^ΔCarle A, Preizal L, Amyot M, Rosabal M (2026). "Acute and Chronic Toxicity of Sediments Containing Platinum and Palladium on Freshwater Benthic Organisms *Chironomus riparius* and *Hyaella azteca*." *J Appl Toxicol*. **46**(4):1151–1163. doi:[10.1002/jat.4933](https://doi.org/10.1002/jat.4933).
15. ^ΔLiu X, Fan F (2026). "A Large-Scale Concordance Study of Toxicity Findings Across Preclinical Species and Humans for Small Molecules and Biologics." *Front Toxicol*. **8**:1731947. doi:[10.3389/ftox.2026.1731947](https://doi.org/10.3389/ftox.2026.1731947).
16. ^ΔChen Y, Xu G, Wu Z, Hao C, Yang C, Chen X (2026). "Ecotoxicity of Combined Polylactic Acid Microplastics and Thallium Pollution on the Functional Traits of *Folsomia candida*." *Toxics*. **14**(4):307. doi:[10.3390/toxics14040307](https://doi.org/10.3390/toxics14040307).
17. ^ΔShrestha M, Pappas RS, Gonzalez-Jimenez N, Gray N, Watson CH, Valentin-Blasini L, Otgonsuren M, Taylor KM, Hassink M (2026). "Arsenic and Cadmium in Cigar Fillers: A Comparative Study Within Cigar Types and with Cigarettes." *Chem Res Toxicol*. **39**(4):510–517. doi:[10.1021/acs.chemrestox.5c00084](https://doi.org/10.1021/acs.chemrestox.5c00084).
18. ^ΔBonell A, Qureshi AAM, Flouris AD, Tsoutsoubi L, Sferruzzi-Perri AN, Das JK, Hirst JE, Ioannou LG (2026). "Physiological and Biological Response to Heat Exposure in Pregnancy: A Systematic Review and Meta-Analysis." *Lancet Planet Health*. **10**(4):101431. doi:[10.1016/j.lanplh.2026.101431](https://doi.org/10.1016/j.lanplh.2026.101431).
19. ^ΔLaue HE, Fleury ES, Jackson-Browne MS, Calafat AM, Chen A, Yolton K, Cecil KM, Newman NC, Buckley J, Lanphear BP, Braun JM (2025). "Associations of Gestational and Childhood Urinary Triclosan Concentrations with Atopic and Allergic Symptoms in Health Outcomes and Measures of the Environment (HOME) Study Participants Ages 1–12 Years." *Environ Health Perspect*. doi:[10.1289/EHP16710](https://doi.org/10.1289/EHP16710).
20. ^ΔMegraso CCF, Batulan KRT, Lauron JC, Suico JDC, Sumagang AJKP, Quitayen JR, Lim PJY (2026). "Utilizing Tropical Tree Rings to Trace the Spatial and Temporal Distribution of Mercury Levels in Cebu, Philippines." *Environ Sci Pollut Res Int*. doi:[10.1007/s11356-026-37874-x](https://doi.org/10.1007/s11356-026-37874-x).
21. ^ΔLiang R, Song G, Zhang X, Sun J, Fu C, Ren X, Chai Y, Liu J, Cheng Z, Sires I, Zhou M (2026). "Unlocking Electro-Enhanced Peroxymonosulfate Activation Mechanism with a Single-Atom Cobalt Monolithic Electrode Toward Efficient and Sustainable Water Decontamination." *Environ Sci Technol*. **60**(23):16998–17012. doi:[10.1021/acs.est.6c02643](https://doi.org/10.1021/acs.est.6c02643).
22. ^ΔMenone ML, Crupkin AC, Villagran DM, Taglioretti V, Medici SK, Aparicio VC, Gonzalo Mayoral E, Echegoyen MA, Lombardero LR, Dopchiz MC, Landivar L, Albani CM, Samarco E, Fabregas Lengard S (2026). "Is the Presence of Pesticides in the Environment a Problem for Everyone? A Comparison of People and Their Environment in the City and Near Crop Fields in the Austral Pampas Region of Argentina." *Arch Environ Contam Toxicol*. **91**(1):1. doi:[10.1007/s00244-026-01199-0](https://doi.org/10.1007/s00244-026-01199-0).
23. ^ΔPeng C, Liggio J, Moussa SG, Tang Y, Johannessen C, Zhang X, Zhang Z, Li L, Harner T, Saini A, Hung H, Wentzell J, Wheeler MJ, Dabek-Zlotorzynska E, Yassine M, Liu Q, Tang M, Friel-Bartlett Z, Abdul-Sater AA (2026). "Atmospheric Aging of Tire Rubber Antioxidants Forms Complex Mixtures that Trigger Inflammation in Human Macrophages." *Environ Int*. **212**:110307. doi:[10.1016/j.envint.2026.110307](https://doi.org/10.1016/j.envint.2026.110307).
24. ^ΔChatzitoulousis P, Oswald S, Ploskas N, Tinnevelt GH, Jansen JJ, Niemann H, Collas F, Ghaffari M (2026). "A Pipeline for Meso- and Microplastic Identification in Aquatic Systems Using Machine Learning and Hyperspectral Imaging." *Env*

iron Pollut. 398:127995. doi:[10.1016/j.envpol.2026.127995](https://doi.org/10.1016/j.envpol.2026.127995).

25. [△]Gocer O, Ramezannia A (2026). "Circular Economy Approaches for Improving Environmental and Economic Performance of Felt-Based Living Walls." *Sci Total Environ.* 1033:181831. doi:[10.1016/j.scitotenv.2026.181831](https://doi.org/10.1016/j.scitotenv.2026.181831).
26. [△]Guo X, Jiang Q, Zhang J, Jin Y, Du S, Shao B (2026). "Soybean Protein Complexes as Diglyceride Nanostructured Lipid Carriers for Lycopene Delivery: Preparation via Metal-Phenol Network Formation." *Food Chem.* 513:148968. doi:[10.1016/j.foodchem.2026.148968](https://doi.org/10.1016/j.foodchem.2026.148968).
27. [△]Cha S, Bell L, Eastwood J, Fisher DR, Shukitt-Hale B, Zhang Z, Rodriguez-Mateos A, Williams CM (2026). "A Randomised Controlled Study to Investigate the Cognitive, Mood, Metabolic and Anti-Inflammatory Effects of Chronic Oyster Mushroom Intervention in Healthy Older Adults." *Food Funct.* doi:[10.1039/d5fo04951b](https://doi.org/10.1039/d5fo04951b).
28. [△]Xing Y, Zhang H, Bo Y, Chen Y, Li X, Liu C, Shi X, Liu C, Zhou Y (2026). "Nitrogen Regulates Starch Biosynthesis and Multiscale Structural Properties of Sorghum Under Low-Light Stress." *Food Res Int.* 233(Pt 1):118939. doi:[10.1016/j.foodres.2026.118939](https://doi.org/10.1016/j.foodres.2026.118939).
29. [△]Luo T, Zhu Y, Xu F, Xie J, Chen H, Qin L, Ren T (2026). "Hydroxy- α -sanshools from *Zanthoxylum bungeanum* Maxim. Alleviate Obesity in Mice via the Regulation of Appetite and Gut Microbiota." *Food Sci Nutr.* 14(6):e71952. doi:[10.1002/fsn3.71952](https://doi.org/10.1002/fsn3.71952).
30. [△]Nunes JA, da Silva Gonçalves AH, Mclaine Duarte de Freitas J, Santos do Nascimento J, Aparecido Meireles Grillo L, Fontes Landell M, Juarez Vieira Pereira H (2026). "A Novel Chymotrypsin-Like Protease from *Trichoderma koningii* FFT13 with Efficient Milk-Clotting Activity." *Foods.* 15(11):1940. doi:[10.3390/foods15111940](https://doi.org/10.3390/foods15111940).
31. [△]Li J, Xu X, Wu X, Xu L, Kuang H, Xu C, Guo L (2026). "Computer-Aided Rational Hapten Design of Mepiquat Chloride: High-Affinity Antibody and Sensitive Immunoassay for Food and Environmental Samples." *J Agric Food Chem.* 74(21):16312–16323. doi:[10.1021/acs.jafc.6c04402](https://doi.org/10.1021/acs.jafc.6c04402).
32. [△]Luedke SG, Lawrence TE, Lucher LW, Perkins TL, Posey D, MacDonald D, Schwartz TE (2026). "Evaluation of Foreign Metal Material Detection Methods for Hide-On and Hide-Off Beef Cattle." *J Food Prot.* 89(6):100771. doi:[10.1016/j.jfp.2026.100771](https://doi.org/10.1016/j.jfp.2026.100771).
33. [△]Cai L, Huang P, Cui C (2026). "Acid Precipitation Salting Out as an Efficient Strategy for High Purity Potato Trypsin Inhibitor Recovery." *NPJ Sci Food.* doi:[10.1038/s41538-026-00912-6](https://doi.org/10.1038/s41538-026-00912-6).

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.