Perspectives

AI-Assisted Peer Review

A framework for more efficient and reliable scientific evaluation

Spiros Kitsinelis

Association of Greek Chemists

eex@eex.gr

DOI: 10.62579/JAGC0021

Abstract

Peer review is central to scientific quality assurance, yet it faces persistent challenges, including reviewer workload, delays, variability in assessment and potential bias. Artificial intelligence (AI), particularly large language models, may support selected aspects of manuscript evaluation, but its reliability, limitations and appropriate role in editorial decision-making require careful consideration. This Perspectives article examines the potential of AI-assisted peer review, drawing on preliminary observations from an exploratory exercise in which two independent AI systems evaluated a number of scientific manuscripts. Both systems identified potential weaknesses in methodological reporting, data presentation and statistical analysis, while also offering differing emphases on clarity, interpretation and context. Because the exercise was not benchmarked against expert human reviewers and did not use a formal quantitative evaluation protocol, these observations are illustrative rather than evidence of validated reviewing performance. Building on these observations, we propose a twelve-criterion framework for structured manuscript assessment and a human-supervised workflow in which AI supports screening and the identification of potential problems, while expert reviewers and editors retain responsibility for evaluating scientific significance, interpreting disputed findings and making publication decisions. The framework emphasizes independent verification, confidentiality, transparency and accountability. Blinded, controlled studies are needed to establish the accuracy, consistency, fairness and practical value of AI-assisted review. Responsible integration should therefore be guided by empirical validation rather than assumptions that AI is inherently objective or that faster generation of a review necessarily improves the publication process.

Keywords: artificial intelligence; peer review; scientific publishing; research integrity; reproducibility; editorial governance

1. Introduction

Peer review remains a central mechanism through which scientific manuscripts are evaluated before publication. It provides an opportunity to examine the suitability of methods, the interpretation of results, the relationship between evidence and conclusions, and the contribution a study may make to its field. However, the conventional process depends on the availability of qualified reviewers, whose

time is also committed to research, teaching, clinical or professional duties, and other editorial responsibilities. Review times can be prolonged and the depth and focus of reviews may vary with expertise, workload and individual judgment.

Artificial intelligence (AI), including large language models (LLMs), has created new possibilities for supporting parts of manuscript assessment. Such systems can rapidly process text, organize feedback around predefined criteria, flag missing methodological information, identify inconsistencies between sections and suggest questions that a reviewer may wish to investigate. These functions could be useful in a workflow where reviewers and editors must decide how to allocate limited attention. However, fluent and detailed output should not be mistaken for accuracy. AI systems can overlook substantive flaws, produce incorrect explanations, misinterpret technical details or express unwarranted confidence.

Guidance from organizations concerned with scholarly publishing emphasizes human responsibility, transparency and the need to protect confidential manuscript material when AI tools are considered for editorial work (WAME, 2023; COPE Council, 2024). Current publisher policies also differ, making it necessary for reviewers and editors to follow the specific rules of the journal concerned. The relevant question is therefore not whether AI can simply replace human reviewers but which tasks it may support, under what safeguards and how its contribution should be evaluated.

This article presents preliminary observations from an exploratory exercise involving two independent AI systems, proposes a twelve-criterion framework for AI-assisted manuscript assessment and outlines a human-supervised workflow and research agenda. The pilot is presented as a proof-of-concept observation, not as a formal validation study.

2. Preliminary observations from independent AI assessments

In an exploratory exercise, two AI systems independently evaluated a number of randomly selected scientific manuscripts. The systems identified overlapping potential issues, including gaps in methodological detail, insufficient explanation of experimental parameters, weaknesses in data presentation, and unclear or insufficiently justified statistical analysis. Their reports also differed in emphasis as one tended to focus more on technical rigor and reproducibility while the other gave greater attention to clarity, presentation and contextual interpretation. Both produced structured comments that could, in principle, help authors or reviewers identify points requiring further examination.

These observations motivated the framework proposed below, but their evidential status must be stated carefully. The exercise was not designed as a controlled comparison with expert human reviewers. The number and characteristics of the manuscripts, the prompts, model versions, assessment conditions and evaluation procedures were not documented in a way that would support independent replication or quantitative estimates of performance. There was also no independent adjudication establishing whether the AI-identified problems were correct, whether important problems were missed or whether agreement between the systems reflected accuracy rather than shared error.

Accordingly, the pilot cannot establish that AI reviews are equivalent or superior to human reviews, that agreement between AI systems indicates correctness, or that AI use improves publication quality or turnaround time. Agreement is a useful signal for prioritizing checks, but it is not a substitute for verification. Two systems may share limitations, rely on similar patterns in their training data, or

overlook the same domain-specific issue. Conversely, disagreement may identify ambiguity that merits expert attention rather than indicate that one system is necessarily wrong. The exercise suggests that AI systems can generate structured, potentially useful questions about manuscript quality and that comparing independent outputs may reveal different areas for scrutiny.

3. A twelve-criterion framework for AI-assisted manuscript assessment

The following framework is intended as a practical organizing structure for AI-assisted screening and reviewer support. It is a proposed framework rather than a validated instrument. The criteria should be adapted to the discipline, study design and journal requirements; not every criterion will be equally relevant to every manuscript. AI output should be treated as a set of claims or prompts for verification, not as an authoritative score or decision.

Criterion	Potential AI contribution	Human oversight required
1. Originality and novelty	Identify the manuscript's stated contribution and compare it with relevant literature where reliable, traceable sources are available.	Human experts assess whether the contribution is genuinely new and meaningful within the field; AI-generated literature claims and citations must be verified.
2. Significance and impact	Summarize the claimed scientific, practical or societal relevance and identify whether the manuscript supports those claims.	Judgments about importance, likely influence and field priorities require disciplinary and editorial judgment.
3. Methodology and experimental design	Flag missing parameters, unclear procedures, absent controls, inconsistent descriptions or gaps in reporting.	A qualified expert must determine whether the design is appropriate and whether limitations invalidate the conclusions.
4. Data presentation and visualization	Check labels, units, legends, stated sample sizes, consistency between figures and text, and clarity of reported uncertainty.	Visual or statistical interpretation must be checked against the original data and the norms of the discipline.
5. Statistical analysis	Flag unclear methods, inconsistencies in reported values, missing information about assumptions, or possible mismatches between analyses and claims.	Statistical correctness requires verification by a suitably qualified reviewer, especially for complex analyses or when raw data are unavailable.
6. Interpretation of results	Identify claims that appear stronger than the results presented or conclusions that do not clearly follow from the evidence.	Domain experts must evaluate alternative explanations, mechanisms and the strength of the inference.
7. Reproducibility and transparency	Check whether materials, procedures, data, code and relevant analytical details are reported or made available as required.	Experts determine which omissions materially impede replication and whether discipline-specific standards have been met.
8. Literature context and references	Flag incomplete context, potentially relevant missing literature, inconsistent citations and references that require checking.	Every reference must be verified; AI can generate nonexistent or inaccurate citations and may overlook important competing evidence.
9. Clarity and readability	Suggest improvements to structure, terminology, logical flow and consistency between the abstract, results and conclusions.	Stylistic preferences should not be confused with scientific flaws, and language differences should not unfairly disadvantage authors.
10. Ethical considerations	Flag missing statements about approvals, consent, safety, conflicts of interest, data availability or other required declarations.	AI cannot certify ethical compliance. Editors and qualified specialists must

		assess the relevant requirements and any concerns.
11. Energy and resource efficiency, where relevant	Check whether claims about energy, cost, materials or sustainability are supported by quantitative evidence and clearly stated assumptions.	The importance of these measures and the adequacy of comparisons depend on the research question and field.
12. Overall recommendation	Organize the identified strengths, limitations and unresolved questions into a concise summary for the human reviewer or editor.	AI should not independently accept or reject a manuscript. Publication recommendations and decisions remain the responsibility of human reviewers and editors.

Across these criteria, AI may be particularly useful for structured checks of reporting completeness, internal consistency and the identification of issues that can be explicitly located in the manuscript. Its role is more uncertain when assessment depends on deep field knowledge, tacit disciplinary standards, interpretation of ambiguous evidence or judgments about novelty and significance. Even apparently objective checks can fail if a model misreads a figure, misunderstands a method or invents a requirement. The framework should therefore encourage traceable, manuscript-specific comments and require the reviewer to verify each substantive criticism.

4. A proposed human-supervised workflow

A practical model could use AI as a support layer within the journal's established editorial process rather than as an autonomous reviewer. First, the journal would determine whether AI assistance is permitted and ensure that the chosen system meets its confidentiality, security and data-protection requirements. Second, an authorized tool could screen the manuscript against a predefined checklist, producing comments linked to specific passages, tables or figures where possible. Third, a second independent assessment or a separate verification step could be used to identify agreement, disagreement and potentially unsupported claims. Multiple AI outputs should not automatically be treated as independent expert opinions, because the systems may share training data, design assumptions or failure modes. Fourth, a human reviewer would examine the AI-generated points, reject incorrect or irrelevant comments, and independently assess the scientific argument. Reviewers should be able to distinguish confirmed problems from questions that need investigation. Fifth, the editor would consider the human reviews, the authors' evidence and any verified technical concerns before reaching a decision. The system should preserve a clear record of how AI was used, what was verified, and who is accountable for the final review and editorial outcome.

This workflow could help reviewers focus attention on issues that require specialist reasoning while reducing the effort needed to organize routine checks. Whether it actually reduces total workload or review time, however, remains an empirical question as verification of inaccurate AI comments may create additional work and a rapid AI-generated report does not by itself shorten the full editorial process.

5. Challenges and ethical considerations

5.1 Confidentiality and intellectual property

Unpublished manuscripts may contain original data, methods, figures, commercially sensitive information or intellectual property. Uploading such material to a general-purpose, third-party AI service may violate journal policy, confidentiality obligations or the authors' reasonable expectations about how their work is handled. WAME guidance warns editors and reviewers about the confidentiality implications of submitting manuscript content to chatbots, and the American Psychological Association's current policy instructs reviewers not to upload submission materials into AI tools (WAME, 2023; APA, 2026). Requirements vary by publisher, so the journal's own policy must take precedence.

Where a journal authorizes AI-assisted review, it should use a secure, institutionally approved environment with clearly defined data-retention, access, training-use and deletion arrangements. A closed or enterprise environment may reduce some risks, but the label alone does not guarantee security and technical and contractual protections must be assessed. If those protections cannot be confirmed, confidential manuscript content should not be entered into the system. Authors and reviewers should also be informed, in accordance with journal policy, about whether and how AI is used.

5.2 Bias and uneven performance

AI systems can generate fabricated references, misstate methods, overlook important flaws or offer plausible but incorrect criticisms. Their performance may vary across disciplines, languages, writing styles and methodological traditions. If AI feedback is accepted uncritically, these limitations could introduce new forms of bias rather than reduce existing ones. A system might, for example, mistake unfamiliar terminology for poor writing or favor familiar research approaches over valid but less common methods. Agreement between systems should therefore be interpreted as a reason to check a point, not as proof that it is correct.

Safeguards should include verification of factual claims and citations, disclosure of AI use where required, human accountability for the review report, and monitoring for differential error rates across fields and author groups. AI-generated scores should not be used as a sole basis for rejecting a manuscript, and reviewers should retain the ability to disregard recommendations that are unsupported or inconsistent with disciplinary expertise.

5.3 Accountability and editorial responsibility

AI systems cannot assume professional or ethical responsibility for a review. Reviewers remain responsible for the accuracy and fairness of their reports, and editors remain responsible for editorial decisions. Journals should define permitted uses, prohibited uses, disclosure expectations, record-keeping procedures and routes for handling disputes. Any use of AI to generate substantive assessment should be subject to meaningful human review, not merely a cursory approval of the model's output. This approach is consistent with broader publication-ethics guidance that emphasizes human responsibility and transparency (COPE Council, 2024; APA, 2026).

6. A research agenda for validating AI-assisted peer review

The next step should be controlled evaluation rather than broader claims about effectiveness. Journal editors, publishers, research-integrity organizations and independent investigators could collaborate on blinded comparative studies. A useful study design would include a defined and appropriately diverse sample of manuscripts, multiple human reviewers, and one or more AI systems operating under documented and reproducible conditions. Manuscripts and reviews should be anonymized where feasible, and assessors should not know which reports were generated by AI when rating review quality.

Evaluation should distinguish several outcomes. These include (1) the validity of criticisms, including the proportion that are accurate, inaccurate or unverifiable, (2) the sensitivity of each approach to important flaws and the number of serious issues missed, (3) the relevance, specificity and usefulness of recommendations, (4) agreement with expert adjudication and editorial decisions, (5) consistency across repeated runs, disciplines and manuscript types, (6) potential bias and differences in error rates and (7) total time and effort required, including human verification. Inter-reviewer agreement alone is insufficient because reviewers can agree and still be wrong, while a final editorial decision is itself an imperfect reference standard.

Such studies should predefine their criteria, document model names and versions, prompts, dates, manuscript selection and scoring procedures, and report limitations transparently. Confidentiality and consent requirements must be addressed before any manuscript is submitted to an external system. Where possible, studies should compare conventional review with AI-assisted review, rather than comparing AI output with an idealized assumption that human reviews are always correct. The objective should be to establish which tasks, if any, benefit from AI assistance and under what conditions.

7. Conclusions and call to action

AI-assisted peer review may offer a useful way to structure manuscript screening, flag reporting gaps and help reviewers identify issues that warrant closer examination. The exploratory exercise described here suggests that independent AI systems can produce overlapping as well as complementary observations. However, because the exercise was not benchmarked against human reviewers and did not use a formal accuracy assessment, it should be regarded as a preliminary proof of concept rather than evidence that AI reviews are reliable, objective or superior to conventional peer review.

The twelve-criterion framework proposed in this article is intended to support systematic discussion and future testing. Its value will depend on whether it can be operationalized consistently, adapted to different disciplines, and shown to improve the quality or efficiency of editorial assessment without compromising fairness, confidentiality or scientific judgment. AI should be treated as a fallible assistant whose claims require verification, not as an independent authority on scientific merit.

Journal editors and publishers are encouraged to pilot transparent, secure and human-supervised workflows and to evaluate them through blinded, controlled comparisons with conventional review. Such evaluations should measure both benefits and harms, including the accuracy of feedback, important errors missed, reviewer workload, total turnaround time and possible bias. Until evidence establishes where AI adds value, final judgments about scientific validity, significance and publication

must remain with accountable human reviewers and editors. The goal should not be to automate peer review for its own sake but to determine whether carefully governed AI assistance can make scientific evaluation more rigorous, fair and effective.

References

American Psychological Association (APA). (2026). AI policy for peer reviewers. https://www.apa.org/pubs/journals/resources/ai-peer-reviewers

Committee on Publication Ethics (COPE) Council. (2024). COPE position: Authorship and AI. https://doi.org/10.24318/cCVRZBms

Zielinski, C., Winker, M. A., Aggarwal, R., Ferris, L. E., Heinemann, M., Lapeña, J. F., Jr., Pai, S. A., Ing, E., Citrome, L., Alam, M., Voight, M., & Habibzadeh, F. (2023). Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. Colombian Medical Journal, 54(3), e1015868. https://doi.org/10.25100/cm.v54i3.5868