Research Integrity and
Peer Review

COMMENT Open Access

Personal experience with AI-generated peer reviews: a case study

Nicholas Lo Vecchio[1]*

Abstract

Background While some recent studies have looked at large language model (LLM) use in peer review at the corpus level, to date there have been few examinations of instances of AI-generated reviews in their social context. The goal of this first-person account is to present my experience of receiving two anonymous peer review reports that I believe were produced using generative AI, as well as lessons learned from that experience.

Methods This is a case report on the timeline of the incident, and my and the journal's actions following it. Supporting evidence includes text patterns in the reports, online AI detection tools and ChatGPT simulations; recommendations are offered for others who may find themselves in a similar situation. The primary research limitation of this article is that it is based on one individual's personal experience.

Results After alleging the use of generative AI in December 2023, two months of back-and-forth ensued between myself and the journal, leading to my withdrawal of the submission. The journal denied any ethical breach, without taking an explicit position on the allegations of LLM use. Based on this experience, I recommend that authors engage in dialogue with journals on AI use in peer review prior to article submission; where undisclosed AI use is suspected, authors should proactively amass evidence, request an investigation protocol, escalate the matter as needed, involve independent bodies where possible, and share their experience with fellow researchers.

Conclusions Journals need to promptly adopt transparent policies on LLM use in peer review, in particular requiring disclosure. Open peer review where identities of all stakeholders are declared might safeguard against LLM misuse, but accountability in the AI era is needed from all parties.

Keywords Peer review, Generative AI, ChatGPT, Large language models, LLMs, Academic misconduct

Introduction

The need for the scholarly community to address the use of generative AI in peer review has become increasingly evident ever since the widespread availability of large language models (LLMs) such as ChatGPT. In the fast-growing body of publications addressing this topic, consensus has emerged that, with certain LLM use in peer review going forward, explicit guidelines and controls are urgently needed to ensure legitimacy of the review process [1–8]. As the quality of LLM output has rapidly improved [1, 3, 7], and as people increasingly rely on these tools [9], it will only become harder to accurately distinguish human from AI text, making it all the more plausible for reviewers to pass off LLM output as human assessment.

A stream of recommendations has been forthcoming from some journals and international publishing organizations, though such guidelines appear to be "scattered" [10, 11] and, notably, disciplinarily uneven, with more from the natural sciences than the humanities or social sciences. The Committee on Publication Ethics (COPE) and the European Association of Science Editors (EASE) state that any AI use in peer review should be declared

*Correspondence:
Nicholas Lo Vecchio
nlovecchio@gmail.com
[1] Independent researcher, Marseille, France

to all relevant stakeholders [3, 12, 13]. The International Association of Scientific, Technical & Medical Publishers (STM) recommends that reviewers should "never" use generative AI in drafting review reports, discouraging even its basic proofreading functions [14]. Due to the intellectual property concerns, guidelines have distinguished between generative AI use by authors themselves and by external reviewers [13–16]. For instance, guidelines from the Journal of the American Medical Association (JAMA) and The Lancet indicate that reviewers are prohibited from inputting author work into language models as this would violate confidentiality; reviewers and authors alike must declare how any LLM was used; in the case of The Lancet, authors themselves must even specify the individual prompts used and for which portions of the manuscript [15, 16].

Some research has addressed the question of LLM use in peer review by looking at population-level text patterns in large corpora [17, 18]. One qualitative study interviewed peer reviewers in various disciplines who have openly, routinely used LLMs to assess human work, though without discussing how or whether they disclose this practice to journals [19]. Missing to date has been the close examination of real peer review reports that could plausibly have been produced using generative AI. Assuming (as we must) that some researchers definitely are using AI to assist with peer review, at various levels of the process [5, 17, 19–23], then discussion is necessary about the actual LLM use in its social context. Due to the convincingly humanlike appearance of much LLM output, and the hybrid human-machine nature of all prompted (and post-edited) AI-generated text, it is very difficult to determine what is human and what is machine for any given text where LLM use is not disclosed. Therefore, clearsighted human interpretation is required to assess the potential use of AI in peer review, as in all scholarship.

The objective of this case report is to describe my own experience of receiving what I believe were two AI-generated peer reviews during the submission of my work to language conference proceedings, outlining some of my actions and considerations following that event.

Background

I submitted a historical linguistics article for consideration to a special issue of *mediAzioni* journal (vol. 43, Oct. 2024; affiliated with the University of Bologna) as part of the proceedings for the Taboo Conference (TaCo) held in Rome in September 2022 [24]. Following its submission in June 2023, the guest editors informed me by email in mid-December 2023 that the paper was "suitable for publication, although significant changes should be made" [25].

Upon reading the two anonymous reports attached, I immediately suspected AI had been used to produce them, based on the following reasons: While the reviews were rather extensive (Reviewer 1 text: 915 words; Reviewer 2 text: 857 words), I felt the points raised were extremely vague, unspecific, formulaic and repetitive [26]. All recommendations involved the formal composition of my article; there was no meaningful engagement with my arguments. The reports were written omnisciently in the third person and were near perfect in terms of English orthographic and syntactic norms. In addition to my own qualitative close readings of the reports, supporting evidence to back up my claims included the use of online AI detectors and comparison to simulations run in ChatGPT; while AI detectors are notoriously unreliable [17, 27–31], when properly controlled they may offer relative indications – not irrefutable proof – that must be contextualized alongside other elements suggestive of AI use, such as the simulations (see Additional file 1: Appendices A and B for further details and discussion). Another way to test the AI basis would have been to compare the anonymous reviewers' reports with sample reviews they had written prior to the public release of LLMs.

As a practical matter, I felt that the review recommendations were so unspecific as to be unusable in the work of revising my research paper. Since some portions of the review reports were, to me, obviously AI generated, the entirety of the reviews had been tainted: there was no way for me to determine which points were human and which were machine. As a matter of principle, I felt that, if an LLM had been used, the review process had been delegitimized, for multiple reasons. Legally, if portions of my paper were ingested into a commercial LLM platform, as I believe, this treatment would raise concerns about the confidentiality of intellectual property, as has been widely acknowledged [1, 3, 5–7]. Far more seriously, in epistemic terms LLM use would raise questions about the source of intellectual authority, assuming a supposed "objectivity" of AI language models and running counter to the fundamental situatedness of human knowledge production. Large language models like ChatGPT are not thinking or reasoning platforms; they are text pattern repeaters [32, 33] which, by extracting their artificial knowledge from nodes of "statistical density" [34], quite literally code normativities into their output. Consequently, they risk reproducing dominant ideologies and biases inherent in their training data [1, 3–6, 9, 32, 33], which poses problems for scholarly inclusivity and innovation. An act perhaps meant in the aim of "saving time" has the epistemic effect of assuming the existence of some disembodied higher intellectual authority, conveniently accessible online via commercial AI platform – the ultimate view from nowhere.

Table 1 Examples of reviewer commentary and my response

Review	Reviewer commentary	My annotations
Review 1	However, while the text is generally well-structured, there are some areas where further elaboration and contextualisation could enhance the clarity of the author's argument.	Which exact specific areas in the text – please cite them line by line?
Review 1	Consequently, the examples provided seem detached from the original historical dimension, making them challenging to interpret.	Please explain point by point where the problems are in my argument.
Review 1	However, the authors [*sic*] should consider breaking down complex concepts into simpler language where possible, especially when introducing them. In particular, this is evident when using specialised terminology.	I have no idea what this refers to, without specific examples. Which terms in particular?
Review 2	The author's line of reasoning occasionally proves challenging to track.	What specific arguments are hard to track?
Review 2	The progression of introduced topics lacks linearity and can be abrupt at times.	Where exactly in the piece?
Review 2	However, the author is encouraged to revise specific sections by providing more comprehensive information on certain concepts.	Which specific sections? What more information is to be provided?

Timeline of actions and journal response

As the TaCo conference took place in September 2022, prior to the public release of ChatGPT (in November 2022), no journal policies regarding AI use were communicated to authors at the time of article commissioning or subsequent submission, in June 2023.

Within twenty-four hours of receiving the reviews, I informed the editors of my suspicions by email, providing a list of specific questions and line-by-line annotations to be presented to the reviewers (which might be seen as exemplifying the call in [35] for authors to be "ready to challenge reviewer comments that are seemingly unrelated"). Table 1 provides some examples of the review text [26], along with my annotations sent to the guest editors on 20 December 2023.

In my initial response, I urged the guest editors to confront the reviewers and to inform the editorial and scientific boards. A day later, on 21 December 2023, I escalated the matter to the editors in chief, to make sure my allegations and requests were known to the journal hierarchy. Thus began a two-month period during which the journal denied any ethical breach and I repeatedly followed up to seek further information on the purported investigation, ultimately withdrawing my paper and appealing by email to all scientific board members. Since both review reports revealed LLM-generated text patterns, I wondered whether higher-level coordination had occurred and felt other submitting authors should be made aware of the situation. Thus I informed all participants of the TaCo conference about the matter by email; no one confided to me any doubts about their own reports.

Table 2 provides a chronological list of my actions taken with respect to the journal and selected outside parties.

The official, final response of 22 February 2024 provided no details about the journal's investigation into my allegations and it denied my request for de-anonymization [36]. It included voluntary responses from each of the two human reviewers. Both reviewers stated that English was not their first language. Composed in highly proficient English (predicted human by online AI detectors; see Additional file 1: Appendix A) and written in the first person, these texts demonstrated to me that they were written by the individuals who presumably used ChatGPT to review my article through a process of individualized AI prompting and post-editing. In their detailed responses, the reviewers stood by the recommendations in the reports without explicitly refuting my claims.

Most notable in the journal response was that the editors likewise took no explicit position on the question of whether generative AI had been used to produce both review reports. They stated: "our journal is not the proper place to conduct a dispute on this issue" [36].

Discussion

The integrity of the peer review process was of primary concern, more decisive in motivating my actions than the individual critiques themselves. I went public with my allegations in the equal aims of establishing accountability and of raising public awareness; rather than being exceptional, surely my case is indicative of what is playing out throughout the academic world. One year on, the situation is all the more serious due to the rapid advances in technology which have made LLM output seem sophisticatedly humanlike, yet still presenting the same fundamental political, epistemic and heuristic problems for knowledge production.

Table 2 Timeline of my communications with the journal and some outside parties

Date	Journal actions	My actions	Interactions with outside parties
19 Dec. 2023	Peer reviews sent with provisional acceptance for publication.		
20 Dec. 2023		Response to guest editors alleging generative AI had been used.	Blog post about AI use in peer review, without naming the journal.
21 Dec. 2023	Indication by guest and *mediAzioni* editors that an investigation would be conducted.	Email to editors in chief requesting official journal investigation and response.	
11 Jan. 2024	Joint indication by guest and *mediAzioni* editors that they found my allegations unsubstantiated.		
15 Jan. 2024		Email to editors to withdraw my paper, with detailed questions about the investigation.	
31 Jan. 2024		Email to all scientific and editorial board members requesting a formal response.	Email informing all conference participants of the matter. Blog post publishing the review reports.
31 Jan. – 5 Apr. 2024			Sporadic contact with journalists in several countries.
2 Feb. 2024		Email reminder to all journal parties, specifying a requested reply date of 5 Feb.	Initial contact with Mario Malički of *Research Integrity and Peer Review*.
3 Feb. 2024			Email to COPE seeking advice.
7 Feb. 2024		Email to all journal parties updating with ChatGPT simulations and reiterating my requests.	Blog update publishing ChatGPT simulations.
8 Feb. 2024			Email to Università di Bologna administrative offices requesting intervention.
20 Feb. 2024			Response from COPE: no assistance possible as the journal was not a member.
22 Feb. 2024	Official journal response by email from editors with replies from the anonymous reviewers.		
26 Feb. 2024			Blog update publishing the journal's email response and new controlled AI detection tests.

In retrospect, there were several things I would have done differently. First of all, I would have requested to receive, within seventy-two hours, a detailed protocol laying out the proposed steps of the investigation. Had one not been provided, or had I found a proposed protocol lacking, that would have been the point at which I escalated the matter to the full scientific board – within days, rather than six weeks. Having an official investigative protocol also would have demonstrated to me the concrete actions I needed to take myself. For instance, I initially relied on the journal to employ digital forensic scrutiny, such as running ChatGPT simulations or comparing the reports to earlier writing samples from the reviewers. This seemed to me the appropriate way to proceed, as I had no access to the reviewers themselves. It was only belatedly that I realized I would have to run such tests myself.

As part of an investigative protocol, I would have asked the journal to specify what, if any, associated external oversight bodies I might have recourse to if needed. As it was, despite relying on its recommendations [36, 37], *mediAzioni* is not a member of COPE and I therefore could not appeal to its jurisdiction [38, 39]. Due to the potentially conflicting political interests between an author (an outside entity) and peer reviewers (who a journal and its scientific board may have an interest in protecting), it is crucial for independent bodies to be able to assess cases such as mine [13]. I remain unaware of any such oversight body, in Italy or elsewhere, that could have intervened.

Lastly, as a linguist who focuses on textual discourse and its underlying ideologies, I believe the text patterns in the peer review reports themselves – and LLMs generally – merit critical examination (see Additional file 1: Appendix B). Amid boosterism and hype about the power of generative AI including in peer review [19, 40–42], candid critique of actual LLM output is needed. Fuller linguistic descriptions of typical LLM output will help to better assess potential use of AI in peer review and scholarship. While several widely cited studies have provided minimal details on known generative AI text patterns, it is not enough to merely analyze them "at scale" or at the "population level" [17, 18], because this removes individual instances from their social context and elides human agency.

Recommendations

Following are several recommendations for authors and editors who may find themselves in a similar situation. By urging sustained, proactive engagement by all parties throughout the review process, these guidelines align with the spirit of those recently issued by EASE [13].

For authors	
Check publication policies and lead the conversation in advance.	As the primary interested parties, authors should take the lead in advance by checking journal or conference policies. Authors concerned about LLM use in peer review should address the topic with editors beforehand, when policies are not clearly stated online or in submissions materials. If allowed by editors, consider adding a disclaimer to submitted work indicating that it may not be ingested into LLMs or analyzed by AI during the process, as a way to put reviewers on notice about the author's position.
In the case of suspicion, amass evidence right away.	Clues to AI review generation may include generic, formulaic, verbose, repetitive text, written omnisciently from the third person and flawlessly in terms of orthographic norms, and focusing on form at the expense of argumentation. With rapidly improving models, this will surely change, though. In the case of suspicion of LLM use, do not rely on the journal to take the lead in substantiating the claim. Run simulations in various LLMs, using the peer review prompts with your paper, to see if responses correspond. Compare the language used to published studies describing AI text patterns. If you run the reviews through online AI detectors, proceed with caution (next point).
Do not rely solely on online AI detectors.	For online AI detectors to have validity, they must be controlled against known human and LLM output; only detectors capable of consistently predicting the latter two should be retained. Various detectors ought to be used, to compare among them (see Additional file 1: Appendix A). However, due to their poor reputation [17, 27–31], these tools easily become a distraction, enabling parties to deny their relevance based on the claim that they "don't work" and foreclosing debate on the real issue of how to substantiate suspected LLM use.
Immediately relay concerns to editors and demand an investigation protocol.	After amassing evidence, inform the editors at once. If the submission is for a guest-edited special issue, also include the editors in chief. Concretely, request a detailed protocol – including a timeframe – laying out all steps so that the terms of a potential investigation are transparent. Journal responses should be timely and forthright. Also insist on the role of the individual human reviewers, supplying a list of questions for them to reply to; in the best-case scenario, their response might provide a control sample against which to compare the disputed reviews. The journal's response at this stage will set the tone for future communications; if the response is contentious or defensive, it may make sense to withdraw your paper so as to avoid further conflicts of interest, still while pursuing accountability for potential misconduct.

If the journal response is lacking, escalate the matter.	If, within a span of days, the investigation protocol is not provided or if it is insufficient, escalate the matter to the full scientific board. Where appropriate, depending on the journal or conference, recourse to outside bodies could serve as an independent backup (such as COPE, ICMJE, EASE, STM, etc.). If higher-level LLM coordination seems plausible in the case of special issues or conference proceedings, inform all other participants to the extent they are known.
Activate institutional resources, if you have access to them.	The role of institutional clout should not be underestimated. For affiliated researchers, access to university lawyers, a press relations department or even just colleagues willing to take a stand can help balance the power differential between an individual author alleging material harm and a publisher or the unnamed reviewers on the other end of it. In contrast to plagiarism, LLM abuse is currently too new and too difficult to prove in a court of law, meaning that softer forms of power may take precedence here.
Go public.	If you are in a position to reasonably do so, share your experience. Due to the unprecedented risks LLMs pose to research integrity, the scholarly community deserves to be considered stakeholders in the matter. Through firsthand accounts we should be able to move beyond mere evocations of the "plausibility" or "likelihood" of AI use in peer review and to assert, as humans, that this is actually happening – and to hold accountable the individuals responsible for it.
For editors and publishers	
Require disclosure of LLM use in peer review and transparently state policies on it.	At a minimum, requiring disclosure of any LLM use seems indispensable to maintaining review integrity, as many commentators have already pointed out [1, 3, 5, 8, 15, 16]. The multifunctionality of LLMs – generating, translating, editing, "enhancing" – means that reviewers who use them to generate portions of reviews can then claim the deniability of having used AI only to edit or translate their original human work. Even minimal LLM use may thus cast doubt over the legitimacy of the entire review process. For this reason, clear guidance is needed from journals on what AI use is acceptable in peer review and what is not – multiply beneficial, in that journals state their policies, reviewers know what is expected from them, and authors may choose to submit only to journals whose values align with their own.
Adopt open peer review.	Disclosure of LLM use, while necessary, is hardly a sufficient solution. Even if some good-faith reviewers will dutifully disclose, it is certain that some actors will not. In light of this, open peer review [43] where author and reviewer identities are declared would ensure transparency and accountability for scholarly assessment in the AI era. This is not to naively suppose that LLM use will disappear once reviewers sign their names to their evaluations, but to require that individual human beings take responsibility for their work, including for any machine output found therein. There are questions to address involving the complexity of academic power dynamics in open review [44, 45], but, on principle, safeguarding against AI misuse is surely by now the strongest argument in favor of de-anonymization.
Where disputes arise, put authors and reviewers into contact without delay.	As a matter of fairness, and in acknowledgment that peer review ought to be a dialogue among equals rather than a top-down imposition of power, if an author disputes an anonymous review, they should promptly be put into contact with the reviewers. This may not change the facts of the matter, but it could potentially defuse a tense situation by balancing the power differential.
Allow flexibility on the language of review reports.	Now that AI has enabled instantaneous and competent translation across many languages, the status of English as the default language of science can, in political terms, no longer credibly go unquestioned. To improve sociolinguistic equity while also potentially reducing recourse to LLM use, in any language, for editing tasks that might later cast doubt about the authenticity of peer reviews, reviewers could be encouraged to respond in whatever language they feel most comfortable in, so long as translation resources are available as needed to editors and authors.
Remember that authors are the primary stakeholders.	Scholarly publishing involves balancing many competing interests, but authors – not reviewers, editors or readers – are always the primary stakeholders where peer review is concerned. Even if some parties are in favor of automation, many authors will not consent to any AI assessment, especially if it occurs under the guise of "blind" review. Acknowledging this fact may help prevent incidents such as the one I have described here.

Conclusion

This incident has convinced me that signed, non-anonymous evaluation is the surest and most responsible way to safeguard against AI misconduct in peer review, especially in the humanities and social sciences: in the absence of practicable controls amid the swiftly evolving technology, asking reviewers to openly stand by their work can counterbalance, though not eliminate, the risks of LLM use in peer review. Yet there is no universal solution across disciplines and research communities. In some fields, innovations involving more agile, interactive but still anonymous review might offer different safeguards against LLM abuse [8, 46].

Much of the discussion to date (as cited here) has focused singularly on peer review in the hard sciences. I hope my case will serve as a reminder that peer review is being impacted by generative AI in all disciplines. The interdisciplinary call to action proposed in [47] aims at restoring trust in peer review precisely by acknowledging its human imperfections. In that spirit, perhaps part of the solution is to candidly acknowledge, too, that the flaws in an inherently social pursuit [1, 4, 44, 47] are not fixed by outsourcing fundamentally human decisions to the machines, which reproduce human error and bias in opaque ways.

Realistically, automation of some reviewing tasks via AI is a certainty going forward – but research integrity need not suffer in the name of quick fixes. To adopt an optimistic view, the AI revolution could be a catalyst for radically rethinking what, in its most entrenched forms, has ostensibly become a broken and inequitable system of knowledge gatekeeping.

Abbreviations

COPE	Committee on Publication Ethics
EASE	European Association of Science Editors
ICMJE	International Committee of Medical Journal Editors
JAMA	Journal of the American Medical Association
LLM	Large language model
STM	International Association of Scientific, Technical & Medical Publishers
TaCo	Taboo Conference

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41073-025-00161-3.

Additional file 1: Appendix A. AI detection tests. Appendix B. ChatGPT simulations

Authors' information

Nicholas Lo Vecchio is an independent linguist based in France and specializing in the historical queer lexicon.

Acknowledgements

I wish to thank Mario Malički for his encouragement on this contribution.

Author contributions

The sole author is responsible for all content.

Funding

Not applicable

Data availability

The materials discussed in this commentary are accessible at the author's website, www.nicospage.eu/publications/ai-peer-review. Private communications not published there may be shared with select parties for transparency purposes.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The author declares to have no competing interests.

Received: 10 February 2025 Accepted: 14 March 2025
Published online: 07 April 2025

References

1. Hosseini M, Horbach SPJM. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. Research Integrity and Peer Review. 2023;8:4.
2. Kousha K, Thelwall M. Artificial intelligence to support publishing and peer review: A summary and review. Learned Publishing. 2024;37:4–12.
3. Levene A. Where next in peer review? Part 2: COPE Commentary. Committee on Publication Ethics. 16 November 2023. https://publicationethics.org/news/where-next-peer-review-ai
4. Schintler LA, McNeely CL, Witte J. A critical examination of the ethics of AI-mediated peer review. arXiv; 2023. https://arxiv.org/abs/2309.12356
5. Mollaki V. Death of a reviewer or death of peer review integrity? The challenges of using AI tools in peer reviewing and the need to go beyond publishing policies. Research Ethics. 2024;20(2):239–50.
6. Resnik DB, Hosseini M. The ethics of using artificial intelligence in scientific research: New guidance needed for a new tool. AI Ethics. 2024. https://doi.org/10.1007/s43681-024-00493-8.
7. Thelwall M. Can ChatGPT evaluate research quality? Journal of Data and Information Science. 2024;9(2):1–21.
8. Zou J. ChatGPT is transforming peer review—how can we use it responsibly? [online headline]. Nature. 2024;635(8037):10.
9. Alvero AJ, Lee J, Regla-Vargas A, Kizilcec RF, Joachims T, Antonio AL. Large language models, social demography, and hegemony: Comparing authorship in human and synthetic text. Journal of Big Data. 2024;11:138.
10. Li Z, Xu H, Cao H, Liu Z, Fei Y, Liu J. Use of artificial intelligence in peer review among top 100 medical journals. JAMA Netw Open. 2024;7(12):e2448609.
11. Cheng K, Sun Z, Liu X, Wu H, Li C. Generative artificial intelligence is infiltrating peer review process. Crit Care. 2024;28:149.
12. Committee on Publication Ethics. Discussion document: Artificial intelligence (AI) in decision making (Version 1). September 2021. https://publicationethics.org/sites/default/files/ai-in-decision-making-discussion-doc.pdf
13. European Association of Science Editors. Recommendations on the use of AI in scholarly communication. 25 September 2024. https://ease.org.uk/2024/09/recommendations-on-the-use-of-ai-in-scholarly-communication/
14. International Association of Scientific, Technical & Medical Publishers. Generative AI in scholarly communications: Ethical and practical guidelines for the use of generative AI in the publication process. December 2023. https://stm-assoc.org/new-white-paper-launch-generative-ai-in-scholarly-communications/
15. Flanagin A, Kendall-Taylor J, Bibbins-Domingo K. Guidance for authors, peer reviewers, and editors on use of AI, language models, and chatbots. JAMA. 2023;330(8):702–3.
16. Bagenal J, Biamis C, Boillot M, Brierley R, Chew M, Dehnel T, Frankish H, Grainger E, Pope J, Prowse J, Samuel D, Slogrove AL, Stacey J, Thapaliya G, Trethewey F, Wang HH, Varley-Reeves J, Kleinert S. Generative AI: Ensuring transparency and emphasising human intelligence and accountability. The Lancet. 2024;404(10468):2142–3.
17. Liang W, Izzo Z, Zhang Y, Lepp H, Cao H, Zhao X, Chen L, Ye H, Liu S, Huang Z, McFarland DA, Zou J. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. Proceedings of the 41st International Conference on Machine Learning. Proc Mach Learn Res. 2024;235:29575–620. https://proceedings.mlr.press/v235/liang24b.html.
18. Liang W, Zhang Y, Wu Z, Lepp H, Ji W, Zhao X, Cao H, Liu S, He S, Huang Z, Yang D, Potts C, Manning CD, Zou JY. Mapping the increasing use of LLMs in scientific papers. arXiv; 2024. https://arxiv.org/abs/2404.01268#
19. Ebadi S, Nejadghanbar H, Salman AR, Khosravi H. Exploring the impact of generative AI on peer review: Insights from journal reviewers. Journal of Academic Ethics. 2025. https://doi.org/10.1007/s10805-025-09604-4.
20. Grove J. "ChatGPT-generated reading list" sparks AI peer review debate. Times Higher Education. 5 April 2023. www.timeshighereducation.com/news/chatgpt-generated-reading-list-sparks-ai-peer-review-debate
21. Maiberg E. ChatGPT looms over the peer-review crisis. 404 Media. 2 April 2024. www.404media.co/chatgpt-looms-over-the-peer-review-crisis/
22. Russo Latona G, Horta Ribeiro M, Davidson TR, Veselovsky V, West R. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. arXiv; 2024. https://arxiv.org/abs/2405.02150
23. Majovsky M. AI-generated responses in peer review pose a growing challenge for reviewers and editors: Call for a reviewer rating system. J Clin Neurosci. 2025;133:111042.
24. Lo Vecchio N. Using translation to chart the early spread of GAY terminology. Lexique 2025;36 (forthcoming).
25. mediAzioni. Email communication from guest editors. 19 December 2023.

26. Anonymous review reports received by author on 19 December 2023, from guest editors for mediAzioni 2024;43. Accessible at www.nicospage.eu
27. Chaka C. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. J Appl Learn Teach. 2023;6(2).
28. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. Patterns. 2023;4(7):100779.
29. Sheinman Orenstrakh M, Karnalim O, Aníbal Suárez C, Liut M. Detecting LLM-generated text in computing education: A comparative study for ChatGPT cases. arXiv; 2023. https://arxiv.org/abs/2307.07411
30. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, Šigut P, Waddington L. Testing of detection tools for AI-generated text. Int J Educ Integr. 2023;19:26.
31. Chakraborty S, Bedi A, Zhu S, An B, Manocha D, Huang F. Position: On the possibilities of AI-generated text detection. Proceedings of the 41st International Conference on Machine Learning. Proc Mach Learn Res. 2024;235:6093–115. https://proceedings.mlr.press/v235/chakraborty24a.html.
32. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery. 2021. p. 610–23.
33. Narayanan A, Kapoor S. AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference. Princeton: Princeton University Press; 2024.
34. Benasayag M with Dowek G, Guillot A. L'IA est-elle une chance? Paris: Philosophie Magazine Éditeur; 2024, p. 55.
35. Donker T. The dangers of using large language models for peer review [letter to the editors with supplementary appendix online]. The Lancet Infectious Diseases 2023;23(7):781. www.thelancet.com/journals/laninf/article/PIIS1473-3099(23)00290-6/fulltext
36. mediAzioni. Journal response. Joint email communication from editors in chief and guest editors. 22 February 2024. Accessible at www.nicospage.eu
37. mediAzioni website. About the Journal. https://mediazioni.unibo.it/about.
38. Committee on Publication Ethics. Email communication with administrator. 20 February 2024.
39. Committee on Publication Ethics website. Members. https://publicationethics.org/members
40. Biswas S, Dobaria D, Cohen HL. ChatGPT and the future of journal reviews: A feasibility study. Yale Journal of Biology and Medicine. 2023;96(3):415–20.
41. Irfanullah H. Ending human-dependent peer review. The Scholarly Kitchen. 29 September 2023. https://scholarlykitchen.sspnet.org/2023/09/29/ending-human-dependent-peer-review/
42. Bauchner H, Rivara FP. Use of artificial intelligence and the future of peer review. Health Aff Sch. 2024;2(5):qxae058.
43. Ross-Hellauer T. What is open peer review? A systematic review [version 2]. F1000Research. 2017;6:588.
44. Lee CJ, Sugimoto CR, Zhang G, Cronin B. Bias in peer review. J Am Soc Inform Sci Technol. 2013;64:2–17.
45. Ross-Hellauer T, Bouter LM, Horbach SPJM. Open peer review urgently requires evidence: A call to action. PLoS Biol. 2023;21(10):e3002255.
46. Levene A. Where next in peer review? Part 1: COPE Commentary. Committee on Publication Ethics. 22 October 2023. https://publicationethics.org/news/where-next-peer-review-part-1
47. Cooke SJ, Young N, Peiman KS, Roche DG, Clements JC, Kadykalo AN, Provencher JF, Raghavan R, DeRosa MC, Lennox RJ, Fayek AR, Cristescu ME, Murray SJ, Quinn J, Cobey KD, Browman HI. A harm reduction approach to improving peer review by acknowledging its imperfections. Facets. 2024;9:1–14.

Publisher's Note

Personal experience with AI-generated peer reviews: a case study
Nicholas Lo Vecchio
Research Integrity and Peer Review 2025
DOI: 10.1186/s41073-025-00161-3

Additional file 1

Table of Contents

Appendix A. AI detection tests

I ran the peer review reports through several online AI detectors, using various controls. The detection tools accurately predicted the pure ChatGPT output and known human output, and predicted that the review reports were machine.

As has been widely reported in papers [17, 27, 28, 29, 30], press [48, 49, 50] and by OpenAI itself [51], early iterations of such commercial detectors have shown notorious unreliability. Their development is ongoing (and improvable [31]) and their use merits caution (including by specialists who have excessively relied on a single detector [22, 52]). When properly controlled, AI detection tools may offer relative, not absolute, indications that help to contextualize a text's origin.

AI detector	**ChatGPT output**	**Reviewer 1**		**Reviewer 2**		**My article**
		Peer review report	**Response to author**	**Peer review report**	**Response to author**	
AI Detector Pro	98% chance AI	98% chance AI	2% chance AI	98% chance AI	3% chance AI	2% chance AI
Content at Scale	"Human probability: **X** Reads like AI!" "Unfortunately, it reads very machine-like"	"Human probability: Hard to tell" "Unfortunately, it reads very machine-like"	"Human probability: Passes as Human!"	"Human probability: **X** Reads like AI!" "Unfortunately, it reads very machine-like"	"Human probability: Passes as Human!"	"Human probability: Passes as Human!"
Copyleaks	100% AI content	73.5% AI content	0% AI content	60% AI content	0% AI content	0% AI content
Crossplag	ChatGPT 3.5: AI Content Index 80% – 100% ChatGPT 4: AI Content Index 0% – 100%	AI Content Index 100% "This text is mainly written by an AI"	AI Content Index 0% "This text is mainly written by a human"	AI Content Index 82% "This text is mainly written by an AI"	AI Content Index 0% "This text is mainly written by a human"	AI Content Index 0% "This text is mainly written by a human"
DetectGPT	Probability Breakdown: Human 0% 84% – 97% probability AI depending on text input	Probability Breakdown: Human 0% Mixed 7% AI 93%	Probability Breakdown: Human 78% Mixed 3% AI 19%	Probability Breakdown: Human 3% Mixed 4% AI 93%	Probability Breakdown: Human 75% Mixed 4% AI 21%	Probability Breakdown: Human 100% Mixed 0% AI 0%
GLTR	higher likelihood	higher likelihood	lower likelihood	higher likelihood	lower likelihood	lower likelihood
GPTZero	95% – 98% probability AI generated depending on text input	Probability Breakdown: Human 24% Mixed 32% AI 44%	Probability Breakdown: Human 93% Mixed 6% AI 0%	Probability Breakdown: Human 83% Mixed 15% AI 2%	Probability Breakdown: Human 90% Mixed 10% AI 0%	Probability Breakdown: Human 91% Mixed 9% AI 0%
Originality AI *(% = confidence level of prediction)*	AI Detection Score 0% Original 100% AI	AI Detection Score 0% Original 100% AI	AI Detection Score 22% Original 78% AI	AI Detection Score 0% Original 100% AI	AI Detection Score 1% Original 99% AI	AI Detection Score 81% Original 19% AI
Sapling	Fake: 98% – 100%	Fake: 100%	Fake: % variable depending on input	Fake: 93.5%	Fake: % variable depending on input	Fake: 0.0% – 2.4%
Winston AI	ChatGPT 3.5: Human Score 0% ChatGPT 4: Human Score 0% – 7%	Human Score 5%	Human Score 100%	Human Score 32%	Human Score 100%	Human Score 100%

Notes:
– Most of the above tests were run on February 23, 2024. DetectGPT results were added on November 24, 2024.
– GPTKit, Plagiarism Check, Scribbr, Undetectable AI, Writer and Writefull were initially tested but were excluded from the analysis due to their inability to consistently detect pure ChatGPT output.
– Such tests provide comparative indications only. Originality AI gives false positives for some human control text, outliers indicating lower reliability of this tool. I did not include the Sapling "% fake" predictions for the reviewer responses due to their inconsistency depending on the text inserted.

Table 3. Results of AI detection tests

Appendix B. ChatGPT simulations

I simulated the use of ChatGPT (3.5 and 4) by feeding versions of my original paper into the model and using the review report questions as prompts [53]. Using the corpus analyzer Sketch Engine, I ran concordances of some of the disproportionately used words and phrases across both the review reports and the ChatGPT simulations [54]. This testing indicated to me that ChatGPT 3.5 was the LLM most likely used to generate the reviews. All related materials are accessible on my website (www.nicospage.eu/publications/ai-peer-review).

The content of the recommendations overlapped between the review reports and the ChatGPT simulations, though not uniformly due to the random nature of each machine response. Globally, the points raised in the review reports are raised in the simulation responses: citing more sources ("existing literature") with less self-citation; providing more "background information"; structure, especially of the introduction and conclusion, and with more signposting and synthesis of points; explanation of the methodological and theoretical basis [26, 53, 54] – all generic aspects, reducible to writing tips, that could be addressed in any paper. One take on this overlap would be that the LLM output aligned with reviewers' independent assessment; a more critical take would be that the reviewers were primed by the LLM responses and signed on to them.

With regard to the text patterns, comparison between the review reports and the ChatGPT simulations revealed many disproportionate uses of the same vocabulary, including multi-token text strings. Table 4 provides some examples (inexhaustive list) [26, 53, 54].

Multi-token text strings	Individual types
articulation of the theoretical and methodological	abrupt
benefit from	broader
deeper understanding	challenging
especially when introducing	coherence
fall short	disjointed
fully explore	elucidate
fully grasp	enhance
further enrich the reader's understanding	enrich
help readers	findings
making it/them challenging	gaps
more robust	progression
persuasiveness of the analysis	rigo(u)r
the author should consider	seamless(ly)
there are instances	solidify
to improve (the) clarity	transition

Table 4. Partial list of matching text patterns found across review reports and ChatGPT simulations

It is striking to note the matching text patterns between my review reports (1772 words) and the tiny corpus (977 words) in the ChatGPT review simulation appended to Donker's letter in *The Lancet Infectious Diseases* [35]: "evaluate the validity of the findings," "reliability of the results/findings," "enhance the credibility," "stronger foundation," "simpler language," "well-structured," along with various other disproportionately repeating lexical types such as "insights," "lack," "contextualize/contextualization," "comprehensive," "thorough," "additionally" [26, 35]. On content, two of the main critiques align as well, as worded in the Donker simulation: "The methodology used in this study is not clearly explained" and "The literature review is not comprehensive" [35; compare to 26]. Whether analyzing work on gay language or infectious disease, ChatGPT says the same things over and over again.

Appendix references

See the main article for other references.

48. Heikkilä M. How to spot AI-generated text. MIT Technology Review. 19 December 2022. www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/

49. Coffey L. Professors cautious of tools to detect AI-generated writing. Inside Higher Ed. 9 February 2024. www.insidehighered.com/news/tech-innovation/artificial-intelligence/2024/02/09/professors-proceed-caution-using-ai#

50. Topinka R. The software says my student cheated using AI. They say they're innocent. Who do I believe? The Guardian. 13 February 2024. www.theguardian.com/commentisfree/2024/feb/13/software-student-cheated-combat-ai

51. OpenAI. New AI classifier for indicating AI-written text. 31 January 2023; updated 20 July 2023 [discontinuation]. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/

52. Gao CA, Howard FM, Markov NS, Dyer E, Ramesh S, Luo Y, Pearson AT. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. npj Digital Medicine. 2023;6:75.

53. OpenAI. ChatGPT 3.5 and ChatGPT 4 responses generated 3 February 2024. Raw output and synthesis document accessible at www.nicospage.eu

54. Sketch Engine concordances of ChatGPT text and anonymous review reports. Accessible at www.nicospage.eu