Reviews & Letters in Social Sciences

A Narrative Review on Use of AI Tools in Research Writing: Benefits, Risks and Ethics

Dr. M. A. Pasha, Ph. D.*
Professor (Retired), Department of Chemistry, Bangalore University, Bengaluru-560056, INDIA

*Present address: Chief-Editor, ESRA Publications Pvt. Ltd., Bengaluru-560094, INDIA
e-mail: chief.editoresra@gmail.com

ARTICLE INFO

Prepared 12th October 2026, www.esrapublications.com

Edited & Published 18th October 2026

ABSTRACT

Generative artificial intelligence (AI) tools based on large language models (LLMs) are now widely used to draft, edit and summarise scientific text. This narrative review examines the benefits, risks and ethical rules governing their use in research writing. An analysis of more than 15 million PubMed abstracts concluded that at least 13.5% of abstracts published in 2024 had been processed with LLMs, rising to about 40% in some subcorpora. The main benefit is for authors writing in a second language: non-native English speakers report spending up to 51% more time writing papers and up to 91% more time reading them, and LLM feedback on manuscripts overlapped with human reviewers' comments about as much as two human reviewers overlapped with each other. The risks are substantial. In one study, 55% of citations generated by GPT-3.5 and 18% by GPT-4 were fabricated, and 43% and 24% of the real citations contained substantive errors; LLM summaries of scientific papers were nearly five times more likely than human summaries to overgeneralise (odds ratio 4.85); and AI-generated figures have led to a retraction. Detection is not a solution: blinded reviewers identified only 68% of ChatGPT-generated abstracts, and detectors misclassified 61% of non-native English essays as AI-generated. Editorial bodies (ICMJE, COPE, WAME) agree that AI cannot be an author, that use must be disclosed, and that authors remain fully responsible, while funders such as the US National Institutes of Health prohibit generative AI in peer review for confidentiality reasons. We propose a five-step responsible-use pathway: check policy, choose low-risk tasks, protect data, verify everything and disclose. AI can support clearer writing, but it cannot carry accountability.

Key words: Artificial intelligence; large language models; research writing; scientific publishing; research ethics; hallucination; peer review; authorship.

.

1. Introduction

Large language models can generate and revise text at a quality that is often hard to distinguish from human writing. Within a short time, they have become part of how many researchers prepare manuscripts. The change is visible in the literature itself: a study of more than 15 million biomedical abstracts from 2010 to 2024 found an abrupt rise in certain style words after LLMs became available, and estimated that at least 13.5% of abstracts published in 2024 had been processed with LLMs, a lower bound that reached about 40% in some sub corpora [1].

The questions raised are practical as much as philosophical. Can an AI tool help a researcher write more clearly without compromising integrity? What can go wrong? And what do journals, funders and ethics bodies expect? This review addresses these questions for an audience of authors, reviewers and editors, with attention to authors in India and other settings where English is not the first language. It is a narrative review: evidence was selected by the author from major empirical studies, editorial policies and official notices, without a systematic search, and the field is changing quickly, so specific numbers should be read as snapshots.

2. How AI Writing Tools Work, and Why They Err

LLMs generate text by predicting likely continuations of a sequence of words, learned from very large text corpora. They are optimised for fluency and plausibility rather than for truth, so they can produce fluent statements that are wrong. This behaviour is called hallucination. In the natural language generation literature, hallucinated text is classified as intrinsic, when the output contradicts the source material provided, or extrinsic, when it cannot be verified from the source at all [2]. For research writing the second category is the more dangerous, because a statement or reference may look entirely credible while having no basis in the literature.

Copyright: The author © 2026. This is an open-access article distributed under Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Figure 1 maps the stages of the writing workflow at which AI tools are used and the main risk at each stage. The risk is lowest in language editing, where the author supplies the content and can check that the meaning is unchanged, and highest where the tool is asked to supply facts, references, images or judgements.

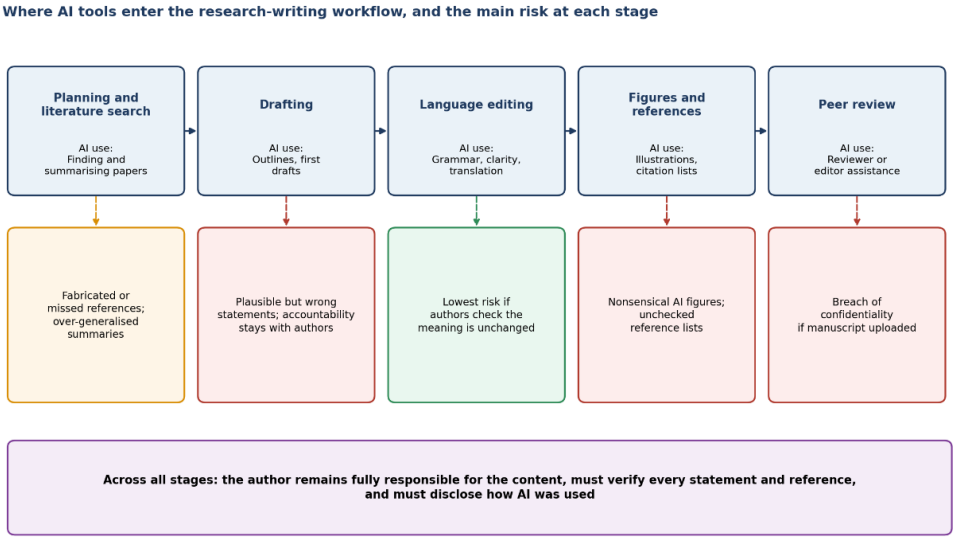


Figure 1. Stages of the research-writing workflow at which AI tools are used, with the main risk at each stage [2–6].

3. Benefits

3.1 Language support for non-native English speakers

English is the dominant language of science, and the burden on non-native speakers is large. In a survey of more than 900 environmental scientists, non-native English speakers reported needing up to 51% more time to write a paper than native speakers, spent 47% (moderate proficiency) to 91% (low proficiency) more time reading papers in English, and sought proofreading help about 50% more often [7]. AI tools that improve grammar and clarity can reduce this burden, and the rapid uptake seen in the literature suggests that many authors already find them useful [1]. In India, where most researchers write in English as a second or later language, this is a practical benefit.

3.2 Feedback and productivity

LLMs can also give feedback on manuscripts. In a large analysis, GPT-4 comments on 3,096 papers from 15 Nature-family journals and 1,709 papers from a machine-learning conference overlapped with human reviewers' points (average overlap 30.85% and 39.23%) at a level comparable to the overlap between two human reviewers (28.58% and 35.25%); in a prospective user study, 57.4% of researchers found the feedback helpful or very helpful, though the authors identified several limitations [8]. These findings suggest that AI feedback can complement, but not replace, human review (Figure 2).

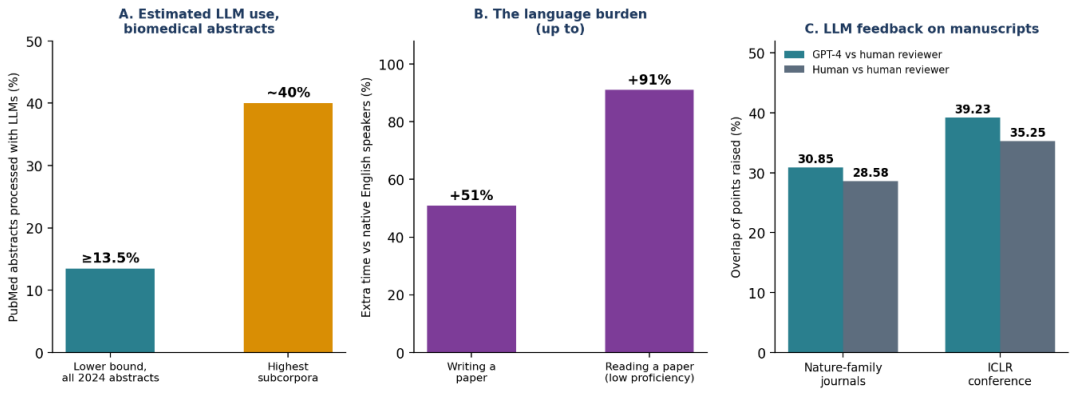


Figure 2. Use and benefits: (A). Estimated LLM use in PubMed abstracts [1]; (B). Extra time reported by non-native English speakers in science [7]; (C). Overlap between GPT-4 and human reviewer comments compared with overlap between two human reviewers [8]. Graphs prepared from published values.

4. Risks

4.1 Fabricated and erroneous references

Reference fabrication is the best-documented risk. Early reports showed that references generated by ChatGPT for a medical topic had titles that did not exist [9]. A systematic study asked GPT-3.5 and GPT-4 to write short literature reviews on 42 topics and checked all 636 citations: 55% of GPT-3.5 citations and 18% of GPT-4 citations were fabricated, and among real citations 43% (GPT-3.5) and 24% (GPT-4) contained substantive errors [3]. Fabricated citations tend to look legitimate at first glance, so they are easily missed by authors, reviewers and readers [3]. Newer models and retrieval-based tools may perform differently, and error rates should be re-measured, but the lesson is that no AI-generated reference should be used without checking it in an independent database.

4.2 Over-generalised summaries

AI tools are often used to summarise papers. A study of ten widely used LLMs found that, even when prompted for accuracy, most produced broader generalisations than the original texts: DeepSeek, ChatGPT-4o and LLaMA 3.3 70B over-generalised in 26–73% of cases, LLM summaries were nearly five times more likely than human-written summaries to contain broad generalisations (OR 4.85, 95% CI 3.06–7.70), and newer models tended to do worse than older ones [4]. This matters for literature reviews, where nuance and limitations are essential.

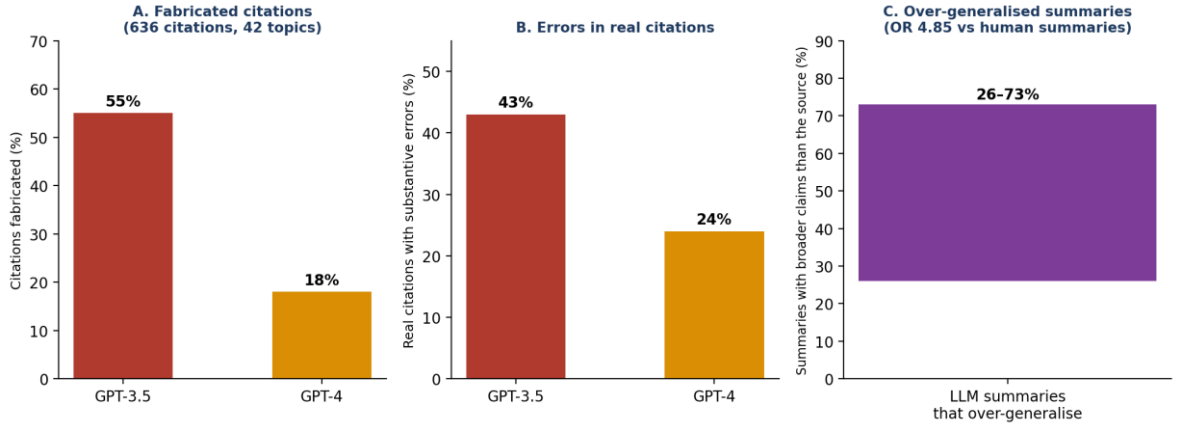


Figure 3. Accuracy risks: (A). Citations fabricated by ChatGPT; (B). Substantive errors in real citations; (C). Share of LLM summaries that over-generalise [3,4]. Graphs prepared from published values.

4.3 Figures, images and integrity

AI image generators can produce illustrations that look professional but are scientifically meaningless. A paper in Frontiers in Cell and Developmental Biology that used Midjourney figures with garbled labels was retracted on 16 February 2024, with the publisher stating that it did not meet standards of editorial and scientific rigour [5]. The case shows that disclosure alone does not make AI-generated content acceptable, and that peer review can fail to catch it.

4.4 Confidentiality in peer review

Uploading an unpublished manuscript or grant application to a public AI tool discloses confidential material to a third party. The US National Institutes of Health prohibits its peer reviewers from using LLMs or other generative AI to analyse applications or formulate critiques, because uploading content to such tools violates peer review confidentiality [6]. Journal reviewers face the same issue.

4.5 Detection is unreliable

Because AI text is hard to spot, detection is tempting, but it is unreliable and can be unfair. When blinded human reviewers were given a mix of real and ChatGPT-generated abstracts, they correctly identified 68% of generated abstracts as generated but flagged 14% of real abstracts as generated; they found the generated abstracts vaguer and more formulaic, and a free AI-output detector separated real from generated abstracts better than the reviewers did [10]. More worryingly, in a test of seven GPT detectors, 61.22% of essays by non-native English speakers were misclassified as AI-generated, compared with 5.19% of essays by US students, and rewriting the non-native essays with richer

vocabulary reduced the false-positive rate to 11.77% [11]. Detectors should therefore never be used as the sole basis for an accusation, especially against non-native writers (Figure 4).

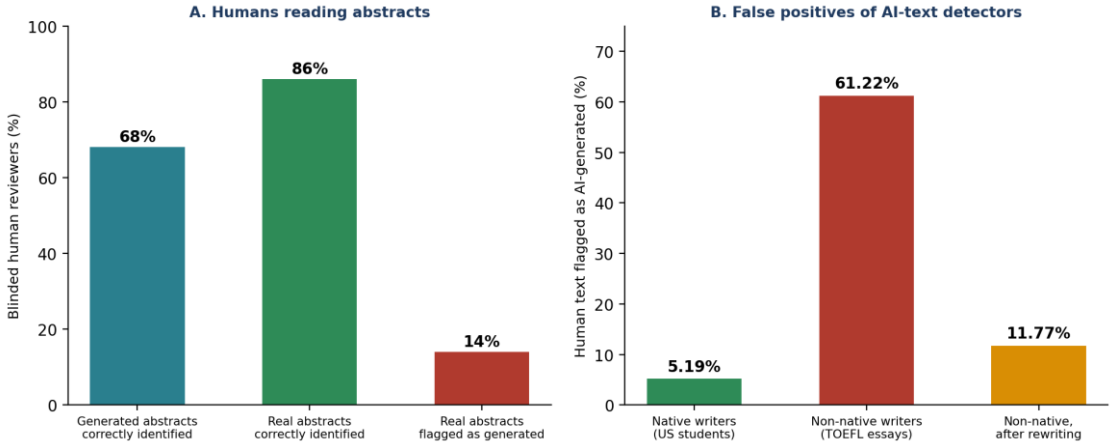


Figure 4. Detection of AI-generated text: (A). Blinded human reviewers reading real and ChatGPT-generated abstracts [10]; (B). False-positive rates of GPT detectors for native and non-native English writing [11]. Graphs prepared from published values.

Table 1. Benefits and risks of AI tools by writing task.

Task	Benefit	Risk	Ref.
Language editing	Reduces extra time and cost borne by non-native English speakers	Meaning may shift; authors must check the result	[1,7]
Feedback on drafts	GPT-4 comments overlapped with human reviewers about as much as two humans did; 57.4% of users found them helpful	Limitations identified; not a substitute for expert review	[8]
Literature summaries	Fast overview of many papers	26–73% of summaries over-generalised in one study; OR 4.85 vs human	[4]
Reference lists	Formatting help	55% (GPT-3.5) and 18% (GPT-4) of citations fabricated; 24%–43% of real ones had errors	[3,9]
Figures	Quick illustration	Nonsensical AI figures led to retraction	[5]
Peer review	Possible help in structuring feedback	Uploading manuscripts breaches confidentiality	[6]

5. Ethics and Policy

Editorial bodies have converged on a small set of rules (Table 2, Figure 5). The International Committee of Medical Journal Editors (ICMJE) states that AI tools should not be listed as authors because they cannot be responsible for the accuracy, integrity and originality of the work. Journals should require authors to disclose AI use at submission and to describe in both the cover letter and the manuscript how the tools were used. Humans are responsible for any material produced with AI, must check that it is correct and free of plagiarism, and must ensure proper attribution of sources [12]. The Committee on Publication Ethics (COPE) takes the same position: AI tools cannot be authors because they cannot take responsibility, manage copyright or declare conflicts of interest, authors must disclose which tool was used and how, and they are fully responsible for the content [13]. The World Association of Medical Editors (WAME) has published recommendations along similar lines [14].

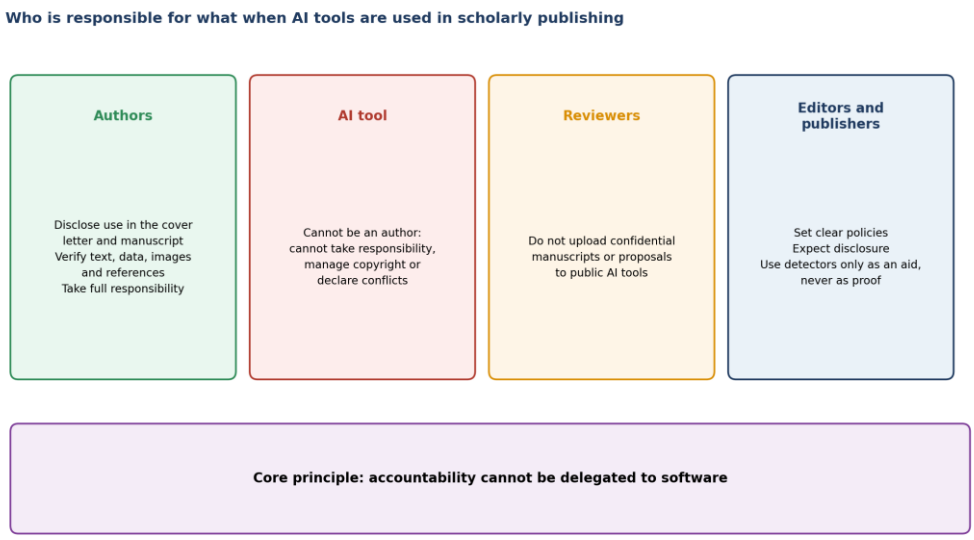


Figure 5. Allocation of responsibility when AI tools are used in scholarly publishing, based on refs [6,12,13]. Scheme prepared by the author.

Table 2. Key policy positions on AI in scholarly publishing.

Body	Position	Concerns	Ref.
ICMJE (updated January 2025)	AI cannot be an author; disclose use in cover letter and manuscript; humans responsible; verify output and attribution	Authors and reviewers	[12]
COPE (February 2023)	AI cannot be an author; disclose tool and use; authors liable for breaches of publication ethics	Authors	[13]
WAME (2023)	Recommendations on chatbots and generative AI in scholarly publications	Authors and journals	[14]
NIH (June 2023)	Generative AI prohibited in NIH peer review because uploading content breaches confidentiality	Reviewers	[6]
ICMR, India (2023)	Ethical guidelines for AI in biomedical research and healthcare	Researchers and institutions	[15]

India's ICMR guidelines address the ethical application of AI in biomedical research and healthcare and are relevant to Indian authors and institutions using AI in research, though they are not specific to manuscript writing [15]. Institutions and journals should translate them into clear, practical rules for writing, review and editing.

6. Evidence Summary

Table 3 summarises the key empirical studies. Three points stand out. First, usage is high and growing. Second, accuracy problems are real and measured, though they depend on the model and task. Third, detection tools are not reliable enough to enforce policy and may be biased.

Table 3. Key empirical studies on AI tools in research writing.

Study	Design	Key finding	Ref.
Kobak et al., 2025	Excess-vocabulary analysis of >15 million PubMed abstracts	≥13.5% of 2024 abstracts processed with LLMs; ~40% in some subcorpora	[1]
Amano et al., 2023	Survey of >900 environmental scientists	Up to 51% more writing time; 47–91% more reading time for non-native speakers	[7]
Liang et al., 2024	GPT-4 feedback on 4,805 papers; user study	Overlap with human reviewers comparable to human–human; 57.4% found it helpful	[8]
Walters and Wilder, 2023	636 citations from GPT-3.5 and GPT-4 reviews	55% and 18% fabricated; 43% and 24% of real citations erroneous	[3]
Peters and Chin-Yee, 2025	Ten LLMs summarising papers	Over-generalised in 26–73%; OR 4.85 vs human summaries	[4]
Gao et al., 2023	Blinded reviewers and detectors on abstracts	Humans identified 68% of generated abstracts; 14% of real ones flagged	[10]
Liang et al., 2023	Seven GPT detectors on native and non-native essays	False positives 61.22% (non-native) vs 5.19% (native)	[11]

7. Practical Guidance for Authors

Taken together, the evidence supports a cautious but practical approach (Figure 6, Table 4). Authors should first check the policy of the journal and their institution. They should prefer low-risk tasks such as language editing and brainstorming, where they supply the content and can judge the output, over tasks that generate facts, references or images. They should not upload confidential, unpublished or personal data to public tools [6]. They should verify every statement and every reference against the original source, since fabricated and erroneous citations are common [3]. And they should disclose the tool, the purpose and the extent of its use, keeping accountability with the named authors [12,13].

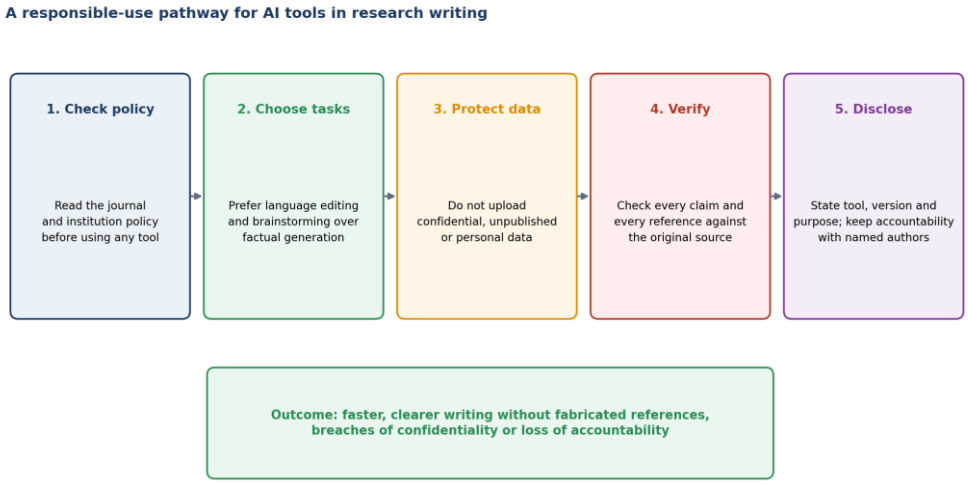


Figure 6. A responsible-use pathway for AI tools in research writing, refs [3,4,6,12,13].

Table 4. Checklist for responsible use of AI tools in research writing.

Step	Action	Why	Ref.
Check policy	Read the author instructions of the target journal and institutional rules	Policies differ between journals and institutions	[12]
Choose tasks	Prefer language editing and feedback; avoid asking AI to supply facts, references or images	Fabrication and over-generalisation	[3,4]
Protect data	Do not upload unpublished data, confidential manuscripts or personal information	Confidentiality breach	[6]
Verify	Check every claim and every reference in an independent database; inspect all figures	Fabricated citations; unusable figures	[3,5]
Disclose	State tool, purpose and extent in cover letter and manuscript; AI is never an author	Journal and ethics requirements	[12,13]
Be fair	Never judge authors by detector scores alone	Detectors biased against non-native writers	[10,11]

8. Challenges

Several challenges remain. Policies are not harmonised across journals, publishers and institutions, so authors face uncertainty about what to disclose. Studies of accuracy age quickly as models change; the fabrication rates reported for GPT-3.5 and GPT-4 [3] may not apply to newer systems or to tools that retrieve real documents, and need to be re-measured. Detection tools are unreliable and biased [10,11], so enforcement cannot depend on them. Equity is a concern: AI may help authors from non-English-speaking countries [7], yet the same authors are most at risk from false accusations [11]. Finally, overreliance on AI summaries may spread over-generalised claims through the literature [4].

9. Future Directions

Future work should include regularly updated audits of reference accuracy for current tools, including retrieval-based systems; trials that test whether AI-assisted drafting changes the quality of manuscripts and peer review outcomes; and research on disclosure practices and compliance. Journals should adopt clear, standard disclosure statements, publishers should develop verification tools for references and images rather than relying on AI-text detectors, and institutions should train researchers, including in India, in the responsible use of AI [12,15]. Concerted work on fair and transparent policies will matter more than any single technology.

10. Conclusions

AI tools can help researchers write more clearly, especially in a second language, and can offer useful feedback. But they also fabricate references, over-generalise findings, produce unusable figures and threaten confidentiality when misused in peer review, and the tools that claim to detect AI text are unreliable and biased. There is already consensus among editorial bodies that AI cannot be an author, that use must be disclosed, and that authors remain fully responsible. The practical message for authors is simple: use AI for low-risk tasks, protect confidential data, verify everything and disclose use. Accountability for scientific text cannot be delegated to software.

Declaration, Conflict of Interest, Funding, and other Statements

Use of Artificial Intelligence (AI): Artificial intelligence (AI) tools were used to assist in the organization, development, and presentation of this narrative review on the benefits, risks, and ethical considerations associated with the use of AI tools in research writing. The author is responsible for critically reviewing and verifying the content, assessing the accuracy and reliability of the information presented, and ensuring the scientific integrity and appropriate attribution of sources throughout the manuscript.

Funding: No specific funding was received from any public, commercial, or non-profit organization for the preparation of this review.
Conflict of Interest: The author declares no conflict of interest related to the preparation or publication of this manuscript.
Ethics Approval: Not applicable. This article is a narrative literature review based on previously published literature and does not involve original research on human participants or animals.
Data Availability: No new datasets were generated or analyzed during the preparation of this review. The discussion and conclusions are based on previously published literature and other sources cited in the manuscript.

Author Note

This manuscript is a narrative review intended to provide researchers, academicians, students, editors, reviewers, and other scholarly professionals with an overview of the potential benefits, limitations, risks, and ethical considerations associated with the use of artificial intelligence tools in research writing and scholarly publishing.

The review aims to promote the responsible, transparent, and informed use of AI tools while emphasizing the importance of human judgment, critical evaluation, factual accuracy, originality, proper attribution, confidentiality, and accountability in the research and publication process.

AI tools should be used to support, rather than replace, the intellectual contribution, critical thinking, and scholarly responsibility of authors. Their use does not diminish the author's responsibility for the accuracy, integrity, originality, and ethical acceptability of the submitted manuscript. Authors should disclose AI assistance in accordance with the applicable policies of the target journal and relevant publication ethics guidelines.

References

1. Kobak D, González-Márquez R, Horvát EÁ, Lause J. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Sci Adv*., **2025**;11:eadt3813. doi:10.1126/sciadv.adt3813
2. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*., **2023**;55(12):248.
3. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*., **2023**;13:14045. doi:10.1038/s41598-023-41032-5
4. Peters U, Chin-Yee B. Generalization bias in large language model summarization of scientific research. *R Soc Open Sci*., **2025**;12(4):241776. doi:10.1098/rsos.241776
5. Frontiers in Cell and Developmental Biology. Retraction: Cellular functions of spermatogonial stem cells in relation to JAK/STAT signaling pathway. Retraction notice, 16 February **2024**.
6. National Institutes of Health. The use of generative artificial intelligence technologies is prohibited for the NIH peer review process. Notice NOT-OD-23-149. 23 June **2023**.
7. Amano T, Ramírez-Castañeda V, Berdejo-Espinola V, Borokini I, Chowdhury S, Golivets M, et al. The manifold costs of being a non-native English speaker in science. *PLoS Biol*., **2023**;21(7):e3002184.

8. Liang W, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*., **2024**;1(8).
9. Alkaissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*, **2023**;15(2):e35179.
10. Gao CA, Howard FM, Markov NS, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *npj Digit Med*., **2023**;6:75.
11. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. *Patterns*, **2023**;4(7):100779.
12. International Committee of Medical Journal Editors. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals: Use of AI by authors (updated January **2025**). Available at: icmje.org.
13. Committee on Publication Ethics. Authorship and AI tools: COPE position statement. 13 February **2023**.
14. Zielinski C, Winker MA, Aggarwal R, et al. Chatbots, generative AI, and scholarly manuscripts: WAME recommendations on chatbots and generative artificial intelligence in relation to scholarly publications. World Association of Medical Editors; **2023**.
15. Indian Council of Medical Research. Ethical guidelines for application of artificial intelligence in biomedical research and healthcare. New Delhi: ICMR; **2023**. ISBN 978-93-5811-343-3.