LETTER TO THE EDITOR

Misclassification of text in AI detection: A serious limitation of AI detectors and Its threats to human-based scholarly writing

Noureen Durrani[1], Yusra Nasir[2], Sobia Ali[3]

Dear Editor, The rapid integration of AI into academic writing has necessitated tools for detecting AI-generated content such as iThenticate ZeroGPT, Turnitin, Phrasly AI, Open AI text Classifier, Writer, Copy leaks to differentiate between human and machine-generated text.[1] However, current AI detection tools suffer from significant misclassification rates, generating false positives that wrongly accuse human authors and false negatives that allow AI-generated text to evade detection.[2] This unreliability unduly impacts non-native English speakers, who often utilise AI tools for paraphrasing and grammar correction to ensure their work meets academic standards; however, these legitimate linguistic refinements are frequently misconstrued by detection software as evidence of AI-generated content.[3] This problem is heightened by the fact that many AI detection tools, primarily trained on English corpora, struggle to accurately assess texts with diverse linguistic structures and stylistic conventions, thereby increasing the risk of false positives for non-English speaking scholars.[2] Such false positives, where human-written manuscripts are incorrectly labelled as AI-generated , severely threaten scholarly psychological safety by fostering distrust and creating an environment of anxiety and unfair allegations. This systemic issue fundamentally challenges academic integrity, undermining the credibility of authors and the foundation of human-based scholarly writing.[4,5] This letter highlights the urgent need for human centric approach to AI detection, advocating for strategies that prioritise human-AI collaboration over sole reliance on fallible automated systems.

A fundamental flaw of current AI detectors lies in their documented inconsistency and unreliability. Studies consistently demonstrate high rates of both false positives, where human-written text is erroneously identified as AI-generated, and false negatives, where AI-generated content evades detection.[6,7] For instance, literature indicated that both free and commercially available AI detection tools can incorrectly classify human-written content as AI-generated with rates ranging from 43.3% to 83.3%.[7,8]

The ethical concerns arise directly because of incorrectly labelling human-written manuscripts as AI-generated and vice versa. Such errors lead to unfair allegations, rejection of genuine work, unwarranted accusations of academic misconduct, and reputational damage for authors.[9,10] When authors must modify their writing or use "humaniser" tools to avoid false detection, it paradoxically increases AI involvement and further obscures human-machine authorship boundaries.[8] Furthermore, the ease with which AI-generated text can be altered allows it to bypass current detection methods, turning the process into a counterproductive 'cat-and-mouse' game.[1] This eventually weakens the very goal of identifying AI misuse while simultaneously unjustly burdening diligent human scholars.[4]

In light of these critical limitations, academic institutions and journals must reconsider their reliance on AI detection tools as the sole arbiters of academic integrity. We strongly recommend adopting a human-centric approach. This involves integrating expert human judgment and contextual understanding into the review process, acknowledging the inherent unreliability of current detection technologies.[1,4] Crucially, the developers and computer software companies responsible for designing free and commercially available AI detection tools must prioritise significant improvements in their accuracy and reliability. This includes rigorously minimising misclassification rates and enhancing their ability to precisely differentiate a human-generated content from AI-generated text without bias, thereby addressing the current limitations and ethical concerns. Following the precedent set by major publishers, such as Elsevier, which mandate the disclosure of AI tool usage at the end of submitted papers, we recommend uniformity across the board among local publishers and journals to establish comprehensive policies.[5] These policies should include robust human

[1]Department of Biostatistics, Liaquat National Hospital, Karachi, Pakistan. [2,3]Department of Health Professions Education, Liaquat National Hospital, Karachi, Pakistan.

Correspondence: Yusra Nasir. **Email:** yusra.nasir@live.com

ORCID ID: 0009-0009-8940-6659

Submission complete: 07-03-2026 **First Revision received:** 17-03-2026

Acceptance: 08-04-2026 **Last Revision received:** 07-04-2026

review processes, clear appeal mechanisms for authors, and opportunities for scholars to defend their work. It is therefore essential not to rely solely on AI tools; instead, a balanced approach combining human assessment and improved AI detection efforts, with appropriate weight given to both, should be adopted.

DOI: https://doi.org/10.47391/JPMA.40793

Acknowledgment: ChatGPT was used solely for grammar refinement and language polishing. The content and idea presented in this manuscript are entirely the work and cognition of the authors.

Disclaimer: None.

Conflict of Interest: None.

Source of Funding: None.

References

1. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, et al. Testing of detection tools for AI-generated text. Int J Educ Integr. 2023;19:26. doi:10.1007/s40979-023-00146-z.
2. Shamsi A, Wang T, Amraei M, Raju NV. Evaluating AI text detection tools for distinguishing human-written from AI-generated abstracts in Persian-language journals of library and information science. Acad Inf Process. 2026;15:126–34. doi:10.18267/j.aip.293.
3. Scarfe P, Watcham K, Clarke A, Roesch E. A real-world test of artificial intelligence infiltration of a university examinations system: a "Turing test" case study. PLoS One. 2024;19:e0305354. doi:10.1371/journal.pone.0305354.
4. Kar SK, Bansal T, Modi S, Singh A. How sensitive are the free AI-detector tools in detecting AI-generated texts? A comparison of popular AI-detector tools. Indian J Psychol Med. 2025;47:275–8. doi:10.1177/02537176241247934.
5. Chemaya N, Martin D. Perceptions and detection of AI use in manuscript preparation for academic journals. PLoS One. 2024;19:e0304807. doi:10.1371/journal.pone.0304807.
6. Ladha N, Yadav K, Rathore P. AI-generated content detectors: boon or bane for scientific writing. Indian J Sci Technol. 2023;16:3435–9. doi:10.17485/IJST/v16i39.1632.
7. Cheng A, Lin Y, Reedy G, Joseph C, Wirkowski S, Mallette V, et al. Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content. Adv Simul. 2025;10:66. doi:10.1186/s41077-025-00396-6.
8. Giray L. The problem with false positives: AI detection unfairly accuses scholars of AI plagiarism. Ser Libr. 2024;85:181–9. doi:10.1080/0361526X.2024.2433256.
9. Erol G, Ergen A, Gülşen Erol B, Kaya Ergen Ş, Bora TS, Çölgeçen AD, et al. Can we trust academic AI detectives? Accuracy and limitations of AI-output detectors. Acta Neurochir (Wien). 2025;167:214. doi:10.1007/s00701-025-06622-4.
10. Blau W, Cerf VG, Enriquez J, Francisco JS, Gasser U, Gray ML, et al. Protecting scientific integrity in an age of generative AI. Proc Natl Acad Sci U S A. 2024;121:e2407886121. doi:10.1073/pnas.2407886121.

AUTHORS' CONTRIBUTIONS:
ND: Concept, design, literature review and drafting.
YN: Concept, design, literature review, drafting, revision, final approval and agreement to be accountable for all aspects of the work.
SA: Drafting, revision and final approval.