frontiers | Frontiers in Artificial Intelligence

TYPE Perspective
PUBLISHED 02 September 2025
DOI 10.3389/frai.2025.1644098

OPEN ACCESS

EDITED BY
Song Yanhui,
Hangzhou Dianzi University, China

REVIEWED BY
Timothy Ros,
McKendree University, United States

*CORRESPONDENCE
Mohamed Helmy
✉ mohamed.helmy@usask.ca

RECEIVED 09 June 2025
ACCEPTED 15 August 2025
PUBLISHED 02 September 2025

CITATION
Pellegrina D and Helmy M (2025) AI for scientific integrity: detecting ethical breaches, errors, and misconduct in manuscripts. *Front. Artif. Intell.* 8:1644098. doi: 10.3389/frai.2025.1644098

AI for scientific integrity: detecting ethical breaches, errors, and misconduct in manuscripts

Diogo Pellegrina[1] and Mohamed Helmy[1,2,3,4,5]*

[1]Vaccine and Infectious Diseases Organization (VIDO), University of Saskatchewan, Saskatoon, SK, Canada, [2]Vaccinology and Immunotherapeutics Program, School of Public Health, University of Saskatchewan, Saskatoon, SK, Canada, [3]Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada, [4]Department of Computer Science, Idaho State University, Pocatello, ID, United States, [5]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore

The use of Generative AI (GenAI) in scientific writing has grown rapidly, offering tools for manuscript drafting, literature summarization, and data analysis. However, these benefits are accompanied by risks, including undisclosed AI authorship, manipulated content, and the emergence of papermills. This perspective examines two key strategies for maintaining research integrity in the GenAI era: (1) detecting unethical or inappropriate use of GenAI in scientific manuscripts and (2) using AI tools to identify mistakes in scientific literature, such as statistical errors, image manipulation, and incorrect citations. We reviewed the capabilities and limitations of existing AI detectors designed to differentiate human-written (HWT) from machine-generated text (MGT), highlighting performance gaps, genre sensitivity, and vulnerability to adversarial attacks. We also investigate emerging AI-powered systems aimed at identifying errors in published research, including tools for statistical verification, citation validation, and image manipulation detection. Additionally, we discuss recent publishing industry initiatives to AI-driven papermills. Our investigation shows that these developments are not yet sufficiently accurate or reliable yet for use in academic assessment, they mark an early but promising steps toward scalable, AI-assisted quality control in scholarly publishing.

KEYWORDS

artificial intelligence, generative AI, research integrity, research ethics, responsible research, AI detection

Introduction

Within the scientific domain, GenAI offers opportunities to streamline research and writing processes (Wells, 2024). Since the introduction of ChatGPT 3.5 (Marr, 2023), late 2022, the applications of GenAI in the scientific research process have proliferated, and it has now become challenging to survey all of them.

One significant application of Generative AI (GenAI) is in drafting and editing scientific manuscripts. These tools assist in generating introductions, summarizing findings, aligning content with journal guidelines, and automating literature reviews (Gauckler and Werner, 2024; Kim, 2023). GenAI is also used in grant writing, helping structure proposals and improving clarity. Beyond writing, large language models (LLMs) are increasingly used in data processing and mining of unstructured text, enabling hypothesis generation, experimental design, and data visualization. Tools like ChatGPT and DeepSeek (DeepSeek-AI, 2024) can generate code and workflows from complex datasets, while platforms such as GitHub Copilot suggest context-aware programming solutions. However, such tools can pose risks when used by novice developers who may not recognize flawed or suboptimal outputs.

While GenAI offers transformative benefits to the scientific research process, its widespread adoption also raises critical concerns that warrant careful scrutiny. Among the most pressing are the unethical use of GenAI in writing, such as undisclosed AI usage, manipulation of scientific content or the proliferation of papermills (Pérez-Neri et al., 2022), which risk eroding trust in scholarly communication. At the same time, AI presents promising opportunities for strengthening research integrity by identifying mistakes in manuscripts, including factual inaccuracies, statistical errors, and subtle inconsistencies that peer review may overlook. This perspective focuses on both aspects: the detection of unethical or inappropriate use of GenAI in scientific writing, and the application of AI tools to identify and correct errors in scientific literature.

GenAI detection tools

As LLM-generated texts became increasingly better, a need has emerged to create tools that could detect whether a text is Human Written Text (HWT), or Machine Generated Text (MGT). Thus, several tools were developed to perform this task to help identify MGT.

An early attempt of evaluating the performance of AI detectors was in a 2023 study that compared ChatGPT and university students answering questions from tests in 32 university courses (Ibrahim et al., 2023), and tested how well they can be classified by two tools [GPTZero (Tian, 2023) and OpenAI's Text Classifier (OpenAI, n.d.)], with a False Negative Rate (FNR) (AI texts classified as human) of, respectively, 32 and 49% on average. To test the robustness of these detectors they used QuillBot's Paraphraser (n.d.), a popular tool that automatically paraphrases texts, and showed that they can be exploited, increasing the average FNR of both Algorithms to 95 and 98%. Since July 2023, Open AI removed its texts classifier tool, citing low accuracy concerns (OpenAI, n.d.) and has not released a new version.

In order to compare how different detectors were able to differentiate HWT and MGT from different LLMs (ChatGLM, Dolly, ChatGPTturbo, GPT4All, StableLM, and Claude) and across different types of corpora (academic essays, short stories, and news articles), BenchGPT (He et al., 2024) created datasets of each genus containing 1,000 HWT and 1,000 MGT (from those LLMs). The study showed that all detectors are sensitive to changes in the selection of their training dataset. There is a trade-off where detectors that are robust against genre changes like ConDA (Bhattacharjee et al., 2023) (F1-score when trained with news dropped from 0.99 to 0.67 when testing essays) are poor at detecting MGT created with a model different than the one it was trained on, when trained with StableLM (Hugging Face, n.d.) the F1-score testing Claude drops to 0.00. On the other hand, detectors like DEMASQ (Kumari et al., 2023) that are robust against changes in LLM (F0-score drop from 0.92 to 0.71 when trained in ChatGPT-Turbo testing MGT from StableLM) fail when there is a change in genre (F0-score of 0.23 when trained on news and testing essays).

Another factor that contributes to the low accuracy in MGT detection is bias in the training datasets used by AI detectors, which can impact their ability to classify some types of HWT. An earlier study claimed that several detectors exhibited higher false positive rates (FPR) on texts written by non-native English speakers (Liang et al., 2023). This was attributed to non-native writing showing lower text perplexity, making it appear, paradoxically, more machine-like. Perplexity, a metric used in several detectors, measures how difficult each word is to predict given the preceding text. Texts with lower perplexity are easier for a language model to predict and are thus assumed to be more "machine-like," especially if the detector is based on models like GPT.

This result, however, is counterintuitive. Non-native speakers are expected to use more loan words, construct sentences with non-standard syntax, and make more grammatical errors, traits that would typically increase perplexity, not decrease it. These features make their writing appear less natural and less similar to the LLMs' training corpus, and therefore harder to predict. A more recent and rigorous study used a larger dataset and perplexity estimations using unpublished detectors based on GPT-2 to revisit this issue" looks better, but then again it might be due to my foreigner non-standard syntax (Jiang et al., 2024). It analyzed a mixed dataset of native and non-native English GRE writing assessments containing both HWT and MGT. Contrary to the earlier claims, this analysis showed that non-native texts had the highest perplexity, while MGTs consistently had much lower perplexity. Using this feature alone, the authors reported 99.9% accuracy in detecting MGTs. These conflicting findings may be explained by differences in dataset composition, detector models, or evaluation design. The earlier study might have used small or biased datasets, or misinterpreted correlations between writing style and perplexity. The later study's use of real educational writing and unpublished detectors with stricter evaluation may offer a more accurate reflection of cross-linguistic variation. This contrast highlights the need for careful consideration of language background in AI detector evaluation, and it raises important concerns about cross-linguistic generalizability and fairness in MGT detection.

Another factor that contributes to the low accuracy in MGT detection is biases in the training datasets used by AI detectors, which can impact their ability to classify some types of HWT. A study claimed that several detectors had larger FPR on texts written by non-native English speakers since they show a smaller text perplexity (Liang et al., 2023). Perplexity, a metric used in several detectors, is a measurement of how hard each word is to predict when LLMs are given the text that precedes it, so a text in which an LLM can more often predict words was likely written on a similar method. Therefore, it is quite counterintuitive that non-natives would write with lower perplexity, as they should use more loan words and often structure phrases in non-standard ways. A more recent study used a larger dataset, and non-published detectors based on perplexity derived from GPT-2 to classify MGTs mixed in a dataset containing English GRE writing assessments from natives and non-natives. This analysis showed that non-natives had the highest perplexity, and that MGTs had much smaller perplexity, being able to detect MGTs with 99.9% accuracy using perplexity alone (Jiang et al., 2024).

A recent study (Fishchuk and Braun, 2024) compared the capabilities of the latest generation of commercial detectors against several types of attacks, like prompt engineering, hyperparameter-tweaking, character mutations, translation, and paraphrasing. Although no detector is completely invulnerable to adversarial attacks, the authors show that a newer version Copyleaks (n.d.) resisted most types of attacks more often than not but lacked proper statistics.

Like GenAIs, their detectors evolve over time, making them more accurate in their classification, but simultaneously attacks against such detectors also evolves (Figure 1). In order to test the current state of

Abbreviations: GenAI, Generative Artificial Intelligence; HWT, Human-written; MGT, Machine-generated text; LLMs, Large language models.

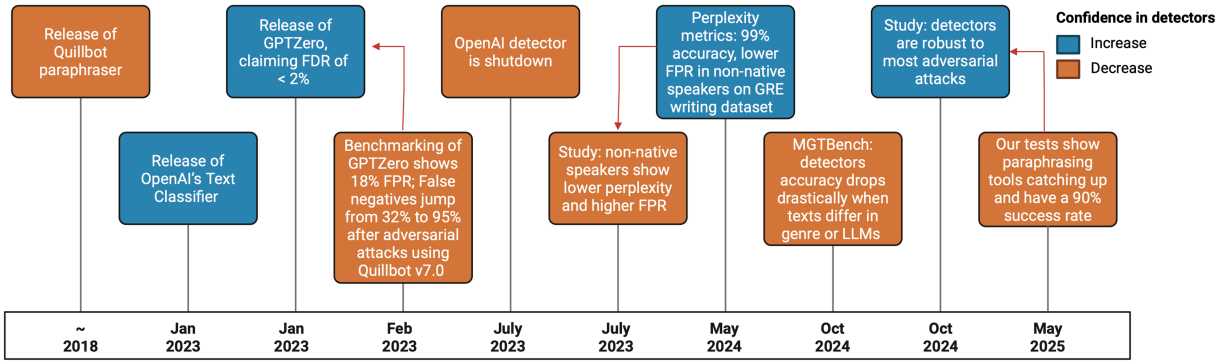

FIGURE 1
A timeline showing major breakthroughs and setbacks related to the detection of AI generated text and the tools that humanize text to avoid detection. Red arrows indicate results that counter previous ones. Colors indicate if the event increased or decreased how confident the public is in MGT detectors. Before MGT detectors were released, Quillbot was already used to obfuscate plagiarism. GPTZero was released with plenty of media coverage, but without any evidence to back its efficacy, one month later a benchmarking study found it to be very inaccurate. Open AI's detector came with similar criticism, but the developers decided it was better to shelf the detector than to develop it further to be more accurate. In 2023 a study found that AI detectors were less accurate against certain non-native speakers, but a follow-up showed that algorithms trained on GRE questions was able to classify GRE questions with 99% accuracy, giving great confidence in detectors. But another study showed that such accuracy could not be maintained in broader scopes of text. Finally, the last red arrow points to a study that evaluated how detectors performed against several attacks, concluding that they ahead of the most common obfuscation techniques, but at present time tools like AIUndetect have an almost perfect success rate.

MGTs and adversary attacks, we prompted DeepSeek to "generate 10 abstracts for made-up computer science papers with at least 100 words and no more than 250 words each" in May 2025 (Supplementary File 1). The resulting texts were evaluated by the two easily accessible AI detectors, GPTZero and Copyleaks. Both detected confidently that 9 out of 10 abstracts were MGT, GPTZero was uncertain about one abstract, and Copyleaks evaluated another one as 0% AI. After each abstract was obfuscated by AI Undetect (n.d.) (using the settings Manual/Balance/Academic), and the tests were repeated and GPTZero was moderately confident one abstract was AI while 9 were considered highly confident to be human. Copyleaks evaluated all of them as 0% AI (Figure 2; Supplementary File 1). We also tried humanizing the texts from within ChatGPT, by giving it samples of our previous abstracts and by asking it explicitly to create texts that look human, both detectors were not disrupted by this approach.

AI detection of mistakes in scientific literature

Mistakes in scientific literature are an enduring challenge in research. Since science is a human endeavor, errors are an inevitable part of the process. Some mistakes stem from honest miscalculations, misinterpretations, or technical oversights, while others arise from fraudulent practices, including data fabrication, image manipulation, and methodological misreporting (Conroy, 2025). Regardless of their origin, such errors can have a lasting impact, misleading subsequent research, influencing policy decisions, and eroding public trust in science.

Recognizing the importance of quality control, the scientific community has long sought ways to identify mistakes more systematically. Notably, efforts to automate error detection began well before the current surge in GenAI. One early initiative is Statcheck, a tool developed for automatically scanning published articles to verify the consistency of reported statistical values (e.g., *p*-values and test statistics). Statcheck compares reported values to recalculated ones and flags potential inconsistencies, helping journals and readers spot statistical errors that might otherwise go unnoticed (Nuijten and Polanin, 2020). A second early example comes from Retraction Watch (n.d.) and the accompanying Retraction Database, which tracks retracted papers and the reasons behind them. Although not an automation tool per se, it has been instrumental in documenting patterns of misconduct and unintentional errors, laying the groundwork for data-driven approaches to understanding mistakes in the literature.

The recent developments in AI enable new levels of automated error detection with higher accuracy and scale. One promising approach is the using LLMs for the detection of reference errors. A recent study demonstrated that LLMs can detect incorrect or misattributed citations with limited context, offering a valuable layer of quality control for reference accuracy, an area often overlooked during peer review (Zhang and Abernethy, 2024).

Image manipulation detection is another critical area where AI tools have proven effective. Platforms such as Imagetwin (n.d.) and Proofig AI (n.d.), and community-driven services like PubPeer (n.d.) use computer vision and machine learning to identify duplicated, rotated, or altered images across publications. These tools have contributed to the exposure of widespread image duplication, leading to retractions and corrections in prominent journals.

In addition, newer initiatives such as The Black Spatula Project (n.d.) and YesNoError (n.d.) aim to identify a broader range of mistakes in published literature, including mathematical inconsistencies, incorrect units, and flawed experimental logic (Gibney, 2025). These platforms use a combination of natural language processing and rule-based systems to scan large corpora of scientific texts and flag anomalies for expert review. While not yet foolproof, they represent a shift toward scalable, AI-assisted post-publication review.

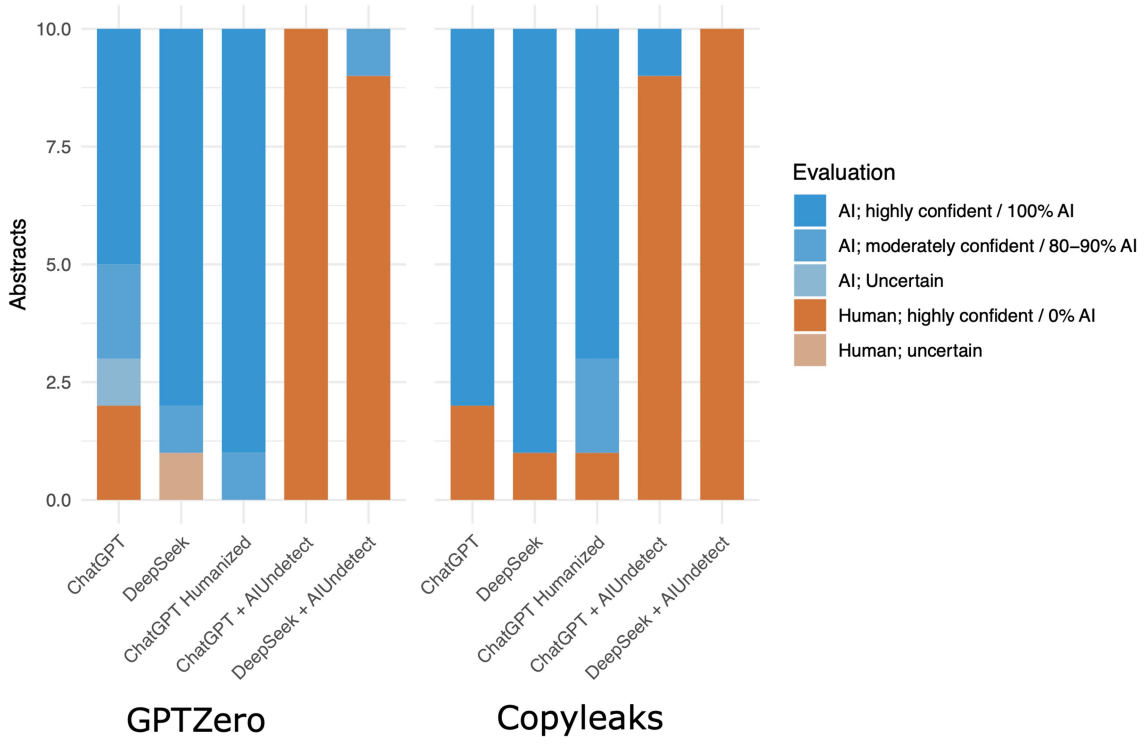

FIGURE 2
GenAI text detectors and adversary attacks. Brief evaluation on the ability of GPTZero and Copyleaks to correctly classify MGT. We tested texts from ChatGPT and from DeepSeek. To test how easy it is to evade those detectors we used AIUndetect to paraphrase the text. We also gave ChatGPT a sample of real abstracts from previous papers published before 2022, we then asked it to create texts that looked human and that were based on that style. Copyleaks are given in AI percentages. GPTZero results were classified as Human or AI with different levels of confidence. Although the detectors had acceptable accuracy on the unedited MGTs, AIUndetect was able to fool them almost all times.

Together, these tools highlight the growing potential of AI not only to improve writing and analysis but also to serve as a critical safeguard against errors in scientific communication. As these systems continue to mature, their integration into editorial workflows and post-publication monitoring could substantially enhance the integrity and reliability of the scientific record.

AI detection of papermills

As discussed earlier, the academic publishing industry is highly impacted by the proliferated use of GenAI tools, which have significantly contributed to the rise of papermills, fabricated data, and manipulated figures. These unethical practices undermine scientific credibility and erode trust in peer-reviewed literature. In response, publishers are taking active countermeasures to mitigate the damage, including the adoption of AI detectors to help identify suspicious content. For instance, Wiley recently announced a pilot of a new AI-powered Papermill Detection service, although the specific tools or technologies behind this effort have not been publicly disclosed (Wiley, n.d.). Such tools are anticipated to assist journal editors and peer reviewers in detecting AI-generated or AI-manipulated submissions before they reach publication.

Several other major publishers and research integrity organizations have launched similar initiatives. Springer Nature, in collaboration with Slimmer AI's Science division, has developed two AI tools focused on detecting fraudulent submissions. These tools are designed to flag fabricated or low-quality content and help distinguish legitimate scientific work from papermill outputs. Another industry-wide initiative, the STM Integrity Hub, offers a centralized cloud-based platform for publishers to share intelligence and detect papermill-generated manuscripts through automated screening applications. These systems can identify key indicators of papermill involvement, such as reused templates, duplicated phrases, or unnatural statistical patterns. Complementing these efforts are specialized tools focused on detecting image and data manipulation. Services like Proofig and Imagetwin (discussed above) apply machine learning to identify duplicated or altered figures, common hallmarks of papermill submissions.

While promising, these technologies are still in the early stages of integration into journal workflows, and their accuracy and reliability are being actively evaluated. The broader deployment and refinement of these tools will take time, and their effectiveness in real-world editorial settings remains to be fully assessed. Nonetheless, these developments represent important first steps in leveraging AI to protect the integrity of scientific publishing, and their impact will become clearer as the technologies mature and are tested at scale.

Conclusion and future perspectives

In this perspective, we explored how artificial intelligence, particularly GenAI and LLM-based methods, is beginning to play a

dual role in scientific publishing: both as a source of new risks to research integrity and as a powerful set of tools for enhancing transparency and error detection. We reviewed emerging AI-based systems for detecting MGT, identifying reference errors, uncovering image manipulation, and flagging methodological or statistical flaws in the scientific literature. These technologies highlight the transformative potential of AI to support editorial workflows, peer review, and post-publication auditing.

At the same time, our analysis reveals that many of these tools are still in early stages of development. Current AI-generated text detectors vary in accuracy and remain vulnerable to paraphrasing and genre-shifting adversarial attacks. Similarly, tools designed to detect scientific errors or misconduct require further validation before they can be reliably applied in high-stakes settings such as manuscript screening or academic evaluations. Biases in training data and variation across disciplines also pose challenges to their generalizability.

In conclusion, the development of tools and technologies for detecting the unethical use of generative AI and identifying errors in scientific literature represents a promising step toward safeguarding research integrity in the AI era. These systems offer valuable support for editors, reviewers, and institutions by flagging potential issues and streamlining quality control. However, they continue to face significant limitations in terms of accuracy, consistency, and contextual understanding. As such, they should not yet be relied upon to automate the evaluation or judgment of researchers' work. Human oversight remains essential, and these technologies should serve as complementary aids rather than standalone solutions in research assessment and editorial decision-making.

In the short term, publishers, editors, and peer reviewers should consider integrating AI-assisted tools into existing workflows as optional aids rather than mandatory screening mechanisms. Training programs should be offered to help editorial staff and reviewers interpret AI-generated flags appropriately, using them as prompts for further human investigation rather than definitive judgments. Peer reviewers could also benefit from voluntary access to specialized tools, such as image analysis or citation-checking software, during the review process to enhance the detection of overlooked issues.

In the longer term, scholarly publishing stakeholders should collaborate to develop shared benchmarks, open datasets, and validation protocols for AI-based integrity tools. Such efforts would improve reliability, reduce duplication, and promote transparency in design and limitations. Publishers should also work toward seamless integration of validated tools into manuscript management systems, enabling consistent quality control from submission through post-publication monitoring. These steps will help ensure that AI technologies evolve into trusted partners in safeguarding research integrity while maintaining the central role of expert human judgment.

Looking ahead, we anticipate that AI tools will become an indispensable part of the scientific publishing ecosystem. To be seamlessly integrated into submission and review platforms, future tools must be more robust, context-aware, and transparent in their design. We expect increased collaboration between publishers, AI developers, and research integrity bodies to ensure these systems are used ethically and effectively. As the field evolves, rigorous benchmarking, open evaluation, and interdisciplinary oversight will be crucial to fully harness the potential of AI in promoting scientific integrity.

Data availability statement

The datasets presented in this study can be found in the article/Supplementary material.

Author contributions

DP: Writing – original draft, Methodology, Visualization, Data curation, Writing – review & editing, Formal analysis. MH: Writing – review & editing, Investigation, Supervision, Visualization, Writing – original draft, Conceptualization, Methodology.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by the core research fund of Vaccine and Infectious Disease Organization (VIDO), University of Saskatchewan. VIDO receives operational funding from the Government of Saskatchewan through Innovation Saskatchewan and the Ministry of Agriculture and from the Canada Foundation for Innovation through the Major Science Initiatives Fund. Published as VIDO manuscript series no. 1116.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. Generative AI tools, specifically OpenAI's ChatGPT and DeepSeek, were used in the preparation of this manuscript to support idea organization, language refinement, and editorial assistance. These tools assisted in drafting, revising, and polishing sections of the manuscript under the direct supervision of the authors. Both ChatGPT and DeepSeek were used to generate text that was used for the analysis reported in Figure 2. All factual claims, references, and interpretations were independently verified by the authors to ensure accuracy and integrity. At no stage were AI tools used to generate false data, invent citations, or replace the critical analysis and intellectual contributions of the authors. The use of AI followed ethical guidelines for responsible authorship and transparency, and the final content reflects original work that was conceived, developed, and approved by the authors.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2025.1644098/full#supplementary-material

References

AI Undetect. Undetectable AI, AI Rewriter, Rewording tool. (n.d.). Available online at: https://www.aiundetect.com/ (accessed May 21, 2025).

Bhattacharjee, A., Kumarage, T., Moraffah, R., and Liu, H. (2023). ConDA: contrastive domain adaptation for AI-generated text detection, vol. *1*. Nusa Dua, Bali: The Association for Computational Linguistics, 598–610.

Conroy, G. (2025). Retractions caused by honest mistakes are extremely stressful, say researchers. *Nature*. doi: 10.1038/D41586-025-00026-1

Copyleaks. AI Detector—Free AI Checker for ChatGPT, GPT-4, Gemini & More. (n.d.). Available online at: https://copyleaks.com/ai-content-detector (accessed May 21, 2025).

DeepSeek-AI. DeepSeek-V3 Technical Report (2024). arXiv. Available at: https://arxiv.org/abs/2412.19437

Fishchuk, V., and Braun, D. (2024). Robustness of generative AI detection: adversarial attacks on black-box neural text detectors. *Int. J. Speech Technol.* 27, 861–874. doi: 10.1007/S10772-024-10144-2/TABLES/4

Gauckler, C., and Werner, M. H. (2024). Artificial intelligence: a challenge to scientific communication. *Klin. Monatsbl. Augenheilkd.* 241, 1309–1321. doi: 10.1055/A-2418-5238

Gibney, E. (2025). AI tools are spotting errors in research papers: inside a growing movement. *Nature*. doi: 10.1038/D41586-025-00648-5

He, X., Shen, X., Chen, Z., Backes, M., and Zhang, Y. (2024). MGTBench: Benchmarking Machine-Generated Text Detection. CCS 2024—Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security:2251–65.

Hugging Face. StabilityAI/stablelm-tuned-alpha-7b. (n.d.). Available online at: https://huggingface.co/stabilityai/stablelm-tuned-alpha-7b (accessed May 21, 2025).

Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., et al. (2023). Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Sci. Rep.* 13:1. doi: 10.1038/s41598-023-38964-3

Imagetwin. Beta Version for Detecting AI-Generated Images (n.d.). Available online at: https://imagetwin.ai/posts/ai-image-detection-beta (accessed May 21, 2025).

Jiang, Y., Hao, J., Fauss, M., and Li, C. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: is there a bias against non-native English speakers? *Comput. Educ.* 217:105070. doi: 10.1016/J.COMPEDU.2024.105070

Kim, S. G. (2023). Using chatgpt for language editing in scientific articles. *Maxillofac. Plast. Reconstr. Surg.* 45, 1–2. doi: 10.1186/S40902-023-00381-X /METRICS

Kumari, K., Pegoraro, A., Fereidooni, H., and Sadeghi, A.-R. (2023). DEMASQ: unmasking the ChatGPT wordsmith. *arXiv*. doi: 10.14722/ndss.2024.231190

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., and Zou, J. (2023). GPT detectors are biased against non-native English writers. 4, 100779. doi: 10.1016/j.patter.2023.100779

Marr, B.. (2023). A Short History Of ChatGPT: How We Got To Where We Are Today. Available online at: https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/ (accessed May 21, 2025).

Nuijten, M. B., and Polanin, J. R. (2020). "Statcheck": automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Res. Synth. Methods* 11, 574–579. doi: 10.1002/JRSM.1408

OpenAI (n.d.). New AI classifier for indicating AI-written text. https://openai.com/index/new-ai-classifier-for-indicating-ai-written-text/ (accessed January 31, 2025).

Pérez-Neri, I., Pineda, C., and Sandoval, H. (2022). Threats to scholarly research integrity arising from paper mills: a rapid scoping review. *Clin. Rheumatol.* 41, 2241–2248. doi: 10.1007/S10067-022-06198-9

Proofig AI. Image Integrity Risks in Life Science Publications. (n.d.). Available online at: https://www.proofig.com/post/image-integrity-risks-in-life-science-publications (accessed May 21, 2025).

PubPeer. PubPeer 2.0. (n.d.). Available online at: https://blog.pubpeer.com/publications/pubpeer2#0 (accessed May 21, 2025).

QuillBot's Paraphraser: The best AI paraphrasing tool (n.d.). Available online at: https://quillbot.com/blog/quillbot-tools/quillbots-paraphraser-best-ai-paraphrasing-tool/ (accessed May 21, 2025).

Retraction Watch. Tracking retractions as a window into the scientific process (n.d.). Available online at: https://retractionwatch.com/ (accessed May 21, 2025).

The Black Spatula Project. Website for The Black Spatula Project. (n.d.). Available online at: https://the-black-spatula-project.github.io/ (accessed May 21, 2025).

Tian, E.. (2023). gptzero update v1—by Edward Tian—GPTZero 2023. Available online at: https://gptzero.substack.com/p/gptzero-update-v1 (accessed May 21, 2025).

Wells, S. (2024). Can AI shake-up translational research? *Nature*. 16. doi: 10.1038/D41586-024-03318-0

Wiley. Wiley announces pilot of new AI-powered Papermill Detection service | John Wiley & Sons, Inc. (n.d.) Available online at: https://newsroom.wiley.com/press-releases/press-release-details/2024/Wiley-announces-pilot-of-new-AI-powered-Papermill-Detection-service/default.aspx (accessed January 31, 2025).

YesNoError. (n.d.). Available online at: https://yesnoerror.com/ (accessed May 21, 2025).

Zhang, T. M., and Abernethy, N. F.. Detecting reference errors in scientific literature with large language models (2024).