

From Hallucination to Precision: A Longitudinal Analysis of Reference Accuracy and Plagiarism in AI-Generated Medical Literature (2024–2026)

Mevlüt Okan AYDIN¹, Alper VATANSEVER², Sezer Erer KAFA³

¹ Department of Medical Education, Faculty of Medicine, Bursa Uludağ University, Bursa, Türkiye.

² Department of Anatomy, Faculty of Medicine, Bursa Uludağ University, Bursa, Türkiye.

³ Department of Medical Ethics and History of Medicine, Faculty of Medicine, Bursa Uludağ University, Bursa, Türkiye.

ABSTRACT

This study aims to evaluate the longitudinal development of general-purpose and specialized artificial intelligence tools in terms of reliability in academic writing and citation accuracy. Eight platforms (ChatGPT, Gemini, QuillBot, Claude, Microsoft Copilot, Elicit, Consensus, and SciSpace) were analyzed using five standardized medical prompts in November 2024 and January 2026. The generated introductions were assessed for reference authenticity using PubMed, Google Scholar, and Web of Science, and for plagiarism using iThenticate. Findings revealed that in November 2024, general-purpose Large Language Models exhibited high hallucination rates, with ChatGPT and Claude providing zero authentic references for certain prompts. Conversely, specialized academic tools like Elicit and SciSpace maintained near-perfect accuracy from the outset. By January 2026, a dramatic improvement was observed, with general-purpose tools like ChatGPT achieving 100% reference accuracy across all categories. Although plagiarism rates were typically below 15%, Gemini recorded a peak of 45% in 2024 before stabilizing. Specialized tools demonstrated superior capacity to manage larger citation volumes, such as SciSpace, which provided 31 verified references in a single output in 2026. While both general and specialized tools have matured significantly, researchers should still exercise caution and use verification protocols. The results indicate that artificial intelligence tools have rapidly transitioned from being prone to academic hallucinations to becoming highly reliable instruments for scholarly literature synthesis.

Keywords: Artificial intelligence. Scientific writing. Authorship. Plagiarism. Large language model.

Halüsinasyondan Hassasiyete: Yapay Zeka Tarafından Oluşturulan Tıbbi Literatürde Referans Doğruluğu ve İntihal Üzerine Boylamsal Bir Değerlendirme (2024–2026)

ÖZET

Bu çalışma, genel amaçlı ve uzmanlaşmış yapay zeka araçlarının akademik yazım ve atıf doğruluğu konusundaki güvenilirliğinin zamana bağlı gelişimini değerlendirmeyi amaçlamaktadır. Kasım 2024 ve Ocak 2026 tarihlerinde sekiz farklı platform (ChatGPT, Gemini, QuillBot, Claude, Microsoft Copilot, Elicit, Consensus ve SciSpace), tıp odaklı beş standart istem kullanılarak analiz edilmiştir. Üretilen giriş bölümleri, PubMed, Google Scholar ve Web of Science üzerinden referans doğruluğu ve iThenticate aracılığıyla intihal oranları açısından incelenmiştir. Bulgular, Kasım 2024'te ChatGPT ve Claude gibi genel amaçlı modellerin bazı istemlerde sıfır doğru referans vererek yüksek halüsinasyon oranları sergilediğini, Elicit ve SciSpace gibi akademik odaklı araçların ise başlangıçtan itibaren tam doğruluk sağladığını göstermiştir. Ocak 2026 itibarıyla, ChatGPT'nin tüm istemlerde tam doğruluğa ulaşmasıyla genel amaçlı araçlarda büyük bir iyileşme gözlemlenmiştir. İntihal oranları genellikle %15'in altında kalsa da, Gemini 2024 yılında bir istemde %45 ile en yüksek seviyeye ulaşmıştır. SciSpace gibi akademik araçlar, 2026'da tek bir metinde 31 doğrulanmış referans sunarak yüksek hacimli atıf yönetiminde üstünlük göstermiştir. Hem genel hem de uzmanlaşmış yapay zeka araçları önemli ölçüde olgunlaşmış olsa da, araştırmacıların doğrulama protokollerini kullanmaya devam etmeleri önerilmektedir. Sonuç olarak, yapay zeka araçları akademik halüsinasyonlara eğilimli bir yapıdan, bilimsel literatür sentezi için oldukça güvenilir enstrümanlara hızla dönüşmüştür.

Keywords: Yapay zeka. Akademik yazım. Yazarlık. İntihal. Büyük dil modelleri.

Date Received: 23 January 2026

Date Accepted: 9 March 2026

Dr. Mevlüt Okan AYDIN
Bursa Uludağ University, Faculty of Medicine, Department
of Medical Education, Bursa, Türkiye.
E-mail: okanaydin@uludag.edu.tr

AUTHORS' ORCID INFORMATION

Mevlüt Okan AYDIN: 0000-0002-8060-8803
Alper VATANSEVER: 0000-0002-3632-1020
Sezer ERER KAFA: 0000-0001-8926-996X

Academic writing transcends mere communication; it is the essential currency of knowledge exchange and the primary engine of career advancement within the competitive "publish or perish" paradigm of modern academia^{1,2}. Proficiency in its complex conventions, which demand precision, analytical rigor, and adherence to disciplinary norms, is a non-negotiable determinant of a researcher's credibility, influence, and professional trajectory^{3,4}. However, this very sophistication creates a formidable barrier to entry.

The acquisition of academic writing competency represents a significant, often stressful challenge, particularly for early-career researchers and non-native English speakers who must master an opaque code while under intense pressure to produce and publish original contributions⁵.

The recent advent of sophisticated generative artificial intelligence (AI) and Natural Language Processing (NLP) models offers a transformative potential to assist in this process⁶. These tools promise to accelerate research, overcome language barriers, and foster creativity^{7,8}. Consequently, their adoption in academia is rapidly increasing. However, this integration is fraught with significant risks, including the generation of inaccurate information ("hallucinations") and fabricated research, the propagation of bias, and ethical concerns over plagiarism^{9–11}. This fundamental tension between promise and peril has sparked intense ethical debate and prompted scientific journals to begin implementing AI-detection tools to safeguard integrity^{12,13}.

Consequently, the academic community faces a pressing need for clear, evidence-based guidelines¹⁴. Current discourse is thus polarized between leveraging AI for efficiency and ensuring the integrity of scholarly work^{7,15}. While studies often discuss this balance in theory, there is a critical lack of empirical, longitudinal data comparing the performance of different AI models on specific, measurable parameters of academic integrity. Key questions remain unanswered: How reliable are the citations generated by these tools? To what extent do they

produce plagiarized content? Moreover, as these models evolve at an extraordinary pace, it is unknown whether their performance on these critical metrics improves or worsens over time. This knowledge gap makes it difficult for researchers and institutions to formulate evidence-based guidelines for responsible use.

To address this gap, this study presents a longitudinal evaluation of leading AI writing tools. The primary aim was to empirically evaluate and compare the performance of these models at two distinct time points (November 2024 and January 2026) based on two fundamental metrics of scholarly integrity: citation hallucination (reference accuracy) and plagiarism. By systematically testing multiple models with standardized academic prompts, this research seeks to identify temporal trends, model-specific strengths and weaknesses, and provide actionable insights for the responsible integration of AI into academic writing practices.

Materials and Methods

Study Design and Data Collection Timeline

This study employed a descriptive longitudinal design to assess changes in the performance of AI tools over a 14-month period. Data were collected at two distinct time points: November 20–24, 2024 (Phase 1) and January 5–9, 2026 (Phase 2). A total of eight AI platforms were evaluated using five standardized academic prompts (ChatGPT, Gemini, Claude,

Table 1. Longitudinal assessment of reference accuracy and plagiarism rates between general-purpose LLMs and academic AI tools (Nov 2024 vs. Jan 2026).

AI Tool	1st prompt (Nov 2024)	1st prompt (Jan 2026)	2nd prompt (Nov 2024)	2nd prompt (Jan 2026)	3rd prompt (Nov 2024)	3rd prompt (Jan 2026)	4th prompt (Nov 2024)	4th prompt (Jan 2026)	5th prompt (Nov 2024)	5th prompt (Jan 2026)
	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)	References (real/ total) Plagiarism (%)
ChatGPT	0 / 0 0%	4 / 4 0%	2 / 4 0%	6 / 6 5%	0 / 4 0%	9 / 9 3%	3 / 3 0%	7 / 7 0%	0 / 3 0%	5 / 5 8%
Gemini	2 / 3 0%	3 / 5 0%	1 / 4 45%	3 / 3 9%	6 / 8 0%	6 / 6 12%	0 / 1 15%	4 / 6 13%	0 / 0 4%	5 / 6 0%
QuillBot	0 / 2 0%	5 / 5 0%	0 / 3 20%	8 / 8 2%	0 / 0 0%	6 / 6 4%	0 / 2 0%	4 / 4 0%	0 / 0 0%	5 / 5 4%
Claude	0 / 7 3%	6 / 8 0%	0 / 3 19%	N / A* 11%	3 / 5 3%	N / A* 11%	1 / 4 0%	N / A* 13%	0 / 4 4%	N / A* 5%
Microsoft Copilot	2 / 3 0%	3 / 3 0%	3 / 3 7%	2 / 2 0%	2 / 2 8%	3 / 3 0%	2 / 2 26%	3 / 3 8%	2 / 3 15%	3 / 3 0%
Elicit	4 / 4 0%	10 / 10 0%	4 / 4 0%	10 / 10 5%	4 / 4 11%	10 / 10 9%	4 / 4 0%	10 / 10 14%	4 / 4 0%	10 / 10 7%
Consensus	5 / 10 10%	18 / 18 0%	10 / 10 11%	20 / 20 3%	10 / 10 4%	20 / 20 2%	10 / 10 0%	19 / 19 2%	10 / 10 11%	10 / 16 0%
SciSpace	5 / 5 0%	7 / 7 0%	5 / 5 0%	31 / 31 8%	5 / 5 0%	7 / 7 0%	5 / 5 0%	6 / 6 17%	5 / 5 0%	8 / 8 4%

* Indicates that the data was not available or could not be generated by the tool at the time of testing

Artificial Intelligence in Academic Writing

Microsoft Copilot, Elicit, Consensus, QuillBot, and SciSpace). For each prompt, the AI-generated introduction section was recorded. To ensure the independence of each response and prevent contextual bias, all previous chat histories and cache data were cleared before entering each prompt. No follow-up or clarifying prompts were used; only the initial, unmodified outputs were included in the analysis.

AI Tools and Prompting Strategy

Eight AI platforms were selected for analysis, representing a combination of the most widely utilized general-purpose Large Language Models (LLMs) and specialized academic-oriented AI tools. The general-purpose LLMs included ChatGPT, Gemini, Claude, and Microsoft Copilot, while Elicit, Consensus, QuillBot, and SciSpace were selected as specialized academic platforms. Each tool was tasked with writing the introduction section of a scientific article using five standardized prompts:

- Write an introduction for a scientific paper by citing published articles about medical students' knowledge of ethical principles of using cadavers in anatomy laboratories.
- Write an introduction for a scientific paper by citing published articles on the use of artificial intelligence in medical education.
- Write an introduction for a scientific paper by citing published articles about problem-based learning in medical education.
- Write an introduction for a scientific paper by citing published articles about simulation-based learning in medical education.
- Write an introduction for a scientific paper by citing published articles about case-based learning in medical education.

The direct outputs from these tools were recorded immediately without the use of follow-up or clarifying prompts to maintain the objectivity of the initial response.

Evaluation Metrics and Verification

The generated introduction sections were evaluated based on two primary parameters: reference accuracy and originality.

- **Reference Verification:** The total number of citations provided by each AI tool was recorded. Each reference was manually cross-checked using Google Scholar, PubMed, and Web of Science databases to verify its authenticity. References were categorized as either "real" or "generated" (hallucinated).
- **Plagiarism Assessment:** Each AI-generated section (excluding the reference list) underwent a formal similarity analysis using the iThenticate Plagiarism Checker.

All parameters, including the real/total reference ratio and plagiarism percentages, were evaluated objectively based on the generated reports to compare the performance across different tools and time periods

Results

Reference Accuracy and Reliability Trends

The evaluation of AI tools from November 2024 to January 2026 reveals a significant shift in the reliability of generated academic references. In November 2024, general-purpose Large Language Models (LLMs) exhibited substantial "hallucination" rates. For instance, ChatGPT and QuillBot frequently failed to produce any authentic references for several prompts (e.g., 0/4 and 0/3 for ChatGPT in prompts 3 and 5, respectively). Similarly, Claude performed poorly in the initial period, yielding 0/7 real references for the first prompt.

However, by January 2026, a dramatic improvement in accuracy was observed across all general-purpose tools. ChatGPT achieved a 100% accuracy rate across all tested prompts in 2026, successfully identifying 4/4, 6/6, 9/9, 7/7, and 5/5 real references. Gemini also showed progress, evolving from 0/1 real references in prompt 4 (Nov 2024) to 4/6 in January 2026 (Table I).

Performance of Specialized Academic AI Tools

Unlike general LLMs, specialized academic tools such as Elicit, Consensus, and SciSpace maintained high precision from the outset. Elicit provided 100% authentic references (4/4) for all prompts in November 2024 and scaled this performance to 10/10 references per prompt in January 2026. SciSpace demonstrated exceptional consistency, providing entirely real references in both periods, with a notable volume increase in the second prompt from 5/5 to 31/31. Consensus also exhibited a significant increase in reference volume and reliability; for the first prompt, its success rate grew from 50% (5/10) in 2024 to 100% (18/18) in 2026.

Plagiarism Analysis

Plagiarism rates remained generally low across the majority of the tools, typically fluctuating between 0% and 15%. However, specific outliers were identified in the November 2024 data. Gemini recorded the highest plagiarism rate of 45% in the second prompt, while Microsoft Copilot reached 26% in the fourth prompt during the same period. By January 2026, these rates generally stabilized, though minor fluctuations persisted, such as SciSpace reaching 17% in the fourth prompt and Elicit reaching 14% in the same category.

Impact of Prompt Complexity

The data suggests that the nature of the prompt influenced the tools' performance, particularly in the earlier assessment phase. In November 2024, tools like Consensus attempted to provide a high volume of references (e.g., 10/10) but occasionally struggled with authenticity compared to their January 2026 performance. Specialized tools (Elicit, SciSpace) were less affected by prompt variations, maintaining near-perfect reliability across queries.

Discussion and Conclusion

The integration of artificial intelligence into academic writing is advancing at an unprecedented pace, presenting both transformative opportunities and significant challenges for scholarly integrity^{6,10}. This longitudinal study, conducted at two distinct time points, provides empirical evidence that, while AI tools are rapidly evolving, critical issues of citation reliability and text originality persist and may change over time. Our findings underscore that the outputs of these tools cannot be taken at face value and require meticulous verification, a caution that applies equally to researchers, educators, and publishers^{16,17}.

The mostly used AI tools offer daily limited free and pro/premium subscription options. Our study aimed to investigate the performance of these commonly used AI tools in academic writing, using only their daily free usage options. A primary concern identified was citation hallucination, where AI-generated references are either non-existent or riddled with inaccuracies. This issue is well-documented. Controlled studies have reported alarmingly high error rates in AI-only writing, with one finding up to 70% incorrect references¹⁸ and another noting a 44.2% error rate¹⁹. Comparative research further refines our understanding. For instance, a 2024 study directly compared two major models, finding that Gemini significantly outperformed GPT-4 in reference precision (77.2% vs. 54.0% correctness), though both models were still prone to generating unreferenced facts²⁰. These findings underscore that citation reliability varies markedly between different AI systems, a variance that our study expands upon by including a broader array of tools and a longitudinal dimension.

Our longitudinal, multi-model comparison adds a critical temporal perspective to this landscape and corroborates the significant variability in model performance observed in prior comparative studies. This variability manifests in two key dimensions: improvement across generations of the same tool and stark differences between distinct models. It has been shown that a dramatic reduction in fabricated citations from ChatGPT-3.5 (55%) to GPT-4 (18%)²¹. This

intra-tool progress is further detailed by another study, which systematically compared GPT-3.5, GPT-4, and Bard (an earlier version of Gemini), finding that GPT-4's precision (13.4%) and hallucination rate (28.6%) were superior to those of GPT-3.5 (9.4% precision, 39.6% hallucination), while Bard performed worst with 0% precision and a 91.4% hallucination rate²². These studies establish a clear, snapshot hierarchy of model accuracy. Our findings at two distinct time points reveal a dynamic and non-uniform evolution in citation reliability across the AI ecosystem. We observed that the prevalence and nature of hallucinations varied significantly both between models and over the 14-month period. Notably, while some tools showed a reduction in blatantly fabricated citations by 2026, others exhibited a relative increase in more subtle metadata errors, such as incorrect volume numbers or DOIs. This shift in error profile suggests that the challenge of citation hallucination is evolving in form, rather than being uniformly resolved.

The findings on plagiarism reveal a more nuanced and persistent challenge than those for citation accuracy. Unlike the marked linear improvement in reference reliability, plagiarism rates exhibited a volatile and less predictable pattern over time. Alarmingly high initial rates, such as Gemini at 45%, expose a core vulnerability where models often function as sophisticated "stochastic parrots," recomposing training data with insufficient transformation²³. This represents a critical threat to academic integrity, given the severe consequences of plagiarism for scholarly careers²⁴. Although a general stabilization at lower rates was observed by 2026, this trend is neither uniform nor complete. Controlled studies complicate the picture; for instance, a study found that an AI-assisted draft had a higher similarity index (36%) than either a human-only (10%) or an AI-only draft (9%), suggesting that human-AI collaboration can inadvertently amplify text reuse¹⁹. The persistence of non-zero plagiarism, including occasional spikes, underscores that originality constitutes a distinct and complex problem. This challenge is compounded by the fact that AI-generated "pseudo-original" paraphrasing may evade standard detection software, implying that true plagiarism rates are likely under-reported^{9,15}.

The use of AI tools raises questions about authorship and accountability in scientific writing. As AI increasingly integrates into the writing process, a debate is emerging regarding whether AI should be recognized as an author or simply acknowledged as a tool utilized in research^{25,26}. Scholarly consensus indicates that although AI can aid in content generation, it cannot assume the responsibilities of authorship, including accountability for the work produced^{16,27,28}. During the academic writing process,

Artificial Intelligence in Academic Writing

posing pertinent inquiries, obtaining credible information, and delivering precise data are ethical responsibilities of any researcher. The successful execution of these steps transforms "written material" into "validated academic work." Exclusively depending on AI for the composition of articles or studies may impede critical thinking and analytical abilities, reduce creativity, and result in an excessive volume of publications lacking authentic effort. It may challenge the conventional definition of "authorship," similar to an "artificial author"^{16,28}. Many journals now require explicit disclosure of AI use, but policies are heterogeneous and sometimes conflicting, creating confusion for authors^{29,30}. Consequently, attention must be directed not only to the advantages AI offers but also to ensuring its ethical and responsible application, accompanied by ongoing oversight.

Artificial intelligence tools can enhance data processing and literature reviews, enabling researchers to concentrate on the analytical dimensions of their work^{15,27}. Nonetheless, reliance on AI for composition and citation may lead to a cursory understanding of the subject matter, as researchers may become overly reliant on these tools for content generation and reference management³¹. This dependence may jeopardize the critical thinking abilities vital for proficient scientific investigation. Composing an essay is not merely an "ability" but a "process," an extensive procedure that encompasses problem definition to result reporting, underpinned by a meticulous literature review.

While this study provides valuable longitudinal insights, several limitations must be acknowledged. First, the analysis was based on a single output per tool for each prompt at each time point. As large language models are inherently non-deterministic, repeated generation of the same prompt could yield variations in both text and citations. Therefore, our findings represent a snapshot of performance rather than a probabilistic range of possible outputs. The evaluation was confined exclusively to the introduction section. AI performance might vary when drafting other sections of a paper (e.g., Methods or Discussion) or when responding to prompts from different academic disciplines. The evaluation was conducted at only two time points. Although they captured a period of rapid evolution, a more granular timeline could reveal non-linear progression or temporary regressions in model performance. The plagiarism assessment relied on a specific text-matching algorithm. Such tools may not detect all forms of paraphrasing, plagiarism, or conceptual similarity, potentially underestimating the issue. Conversely, they may flag common academic phrases as plagiarized, slightly inflating scores. Finally, while manual verification of references is thorough, it inherently leaves a minimal margin for human error.

Despite these limitations, the study's controlled descriptive design and focus on core integrity metrics yield robust, actionable findings.

This research opens several avenues for future inquiry. Subsequent studies should expand the scope to include AI-generated abstracts, literature reviews, and discussion sections, thereby creating a comprehensive map of strengths and weaknesses across manuscript types. Investigating a wider array of discipline-specific prompts would clarify whether AI tools are more reliable in certain fields. Furthermore, extending the longitudinal framework to include quarterly or biannual checkpoints would allow for a finer-grained analysis of development trajectories and help predict future trends. From a technical perspective, research is needed to develop and benchmark AI models that integrate robust, real-time academic database queries (akin to specialized tools) with advanced drafting capabilities (akin to general LLMs), effectively merging the strengths observed in this study. Finally, educational research should focus on developing and evaluating training programs that teach researchers effective prompt engineering and critical verification skills specific to AI-assisted writing.

This longitudinal assessment demonstrates that the landscape of AI-assisted academic writing is undergoing rapid and significant transformation. The most encouraging finding is the substantial improvement in citation accuracy among general-purpose LLMs from 2024 to 2026, suggesting that the technical challenge of reducing "hallucinations" is being actively and successfully addressed. However, the persistent and more erratic nature of plagiarism risk underscores a deeper, more complex problem related to originality and synthesis. The study conclusively highlights a critical dichotomy: specialized academic AI tools are inherently more reliable for literature discovery and citation generation, while advanced general-purpose LLMs require vigilant, expert oversight when used for text production. As these technologies continue to evolve, the academic community must prioritize fostering digital literacy that emphasizes verification and ethical use over uncritical reliance. The ultimate goal is not to avoid AI, but to integrate it as a rigorously scrutinized tool that augments, rather than compromises, scholarly rigor and integrity.

Researcher Contribution Statement:

Idea and design: M.O.A., A.V., S.E.K.; Data collection and processing: M.O.A., A.V.; Analysis and interpretation of data: M.O.A., A.V., S.E.K.; Writing of significant parts of the article: M.O.A., A.V., S.E.K.

Support and Acknowledgement Statement:

No funding recieved

Conflict of Interest Statement:

The authors of the article have no conflict of interest declarations.

Ethics Committee Approval Information:

Not applicable

References

- Busse CE, Anderson EW, Endale T, et al. Strengthening research capacity: a systematic review of manuscript writing and publishing interventions for researchers in low-income and middle-income countries. *BMJ Glob Heal*. 2022;7(2):8059. doi:10.1136/BMJGH-2021-008059
- Aronson JK. When I use a word . . . “Publish or perish”: adverse effects. *BMJ*. 2025;390. doi:10.1136/BMJ.R1577
- Jefferies D, McNally S, Roberts K, et al. The importance of academic literacy for undergraduate nursing students and its relationship to future professional clinical practice: A systematic review. *Nurse Educ Today*. 2018;60:84-91. doi:10.1016/J.NEDT.2017.09.020
- Sweet L, Borrelli S, Matvienko-Sikar K, Kuipers Y, Fox D. Writing for publication: The basics. *Women and Birth*. 2025;38(2):101878. doi:10.1016/J.WOMBI.2025.101878
- Zhang Y, Eto H, Cui J. Linguistic challenges of writing papers in English for scholarly publication: Perceptions of Chinese academics in science and engineering. *PLoS One*. 2025;20(5):e0324760. doi:10.1371/JOURNAL.PONE.0324760
- Huang J, Tan M. The role of ChatGPT in scientific communication: writing better scientific review articles. *Am J Cancer Res*. 2023;13(4):1148. Accessed January 21, 2026. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10164801/>
- Liebrezn M, Schleifer R, Buadze A, Bhugra D, Smith A. Generating scholarly content with ChatGPT: ethical challenges for medical publishing. *Lancet Digit Heal*. 2023;5(3):e105-e106. doi:10.1016/S2589-7500(23)00019-5
- Giglio A Del, da Costa MUP. The use of artificial intelligence to improve the scientific writing of non-native english speakers. *Rev Assoc Med Bras*. 2023;69(9):e20230560. doi:10.1590/1806-9282.20230560
- Elali FR, Rachid LN. AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns*. 2023;4(3):100706. doi:10.1016/j.patter.2023.100706
- Chen Z, Chen C, Yang G, et al. Research integrity in the era of artificial intelligence: Challenges and responses. *Med (United States)*. 2024;103(27):e38811. doi:10.1097/MD.00000000000038811
- Agrawal A. Fairness in AI-Driven Oncology: Investigating Racial and Gender Biases in Large Language Models. *Cureus*. 2024;16(9):e69541. doi:10.7759/CUREUS.69541
- 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(3):275-278. doi:10.1177/02537176241247934
- Sajid M, Sanaullah M, Fuzail M, Malik TS, Shuhidan SM. Comparative analysis of text-based plagiarism detection techniques. *PLoS One*. 2025;20(4):e0319551. doi:10.1371/JOURNAL.PONE.0319551
- Ganjavi C, Eppler MB, Pekcan A, et al. Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. *BMJ*. 2024;384. doi:10.1136/BMJ-2023-077192
- Salvagno M, Taccone FS, Gerli AG. Can artificial intelligence help for scientific writing? *Crit Care* 2023 271. 2023;27(1):75-. doi:10.1186/S13054-023-04380-2
- Leung TI, de Azevedo Cardoso T, Mavragani A, Eysenbach G. Best Practices for Using AI Tools as an Author, Peer Reviewer, or Editor. *J Med Internet Res*. 2023;25(1):e51584. doi:10.2196/51584
- Xu X, Chen Y, Miao J. Opportunities, challenges, and future directions of large language models, including ChatGPT in medical education: a systematic scoping review. *J Educ Eval Health Prof*. 2024;21. doi:10.3352/JEEHP.2024.21.6
- Kacena MA, Plotkin LI, Fehrenbacher JC. The Use of Artificial Intelligence in Writing Scientific Review Articles. *Curr Osteoporos Reports* 2024 221. 2024;22(1):115-121. doi:10.1007/S11914-023-00852-0
- Margetts TJ, Karnik SJ, Wang HS, et al. Use of AI Language Engine ChatGPT 4.0 to Write a Scientific Review Article Examining the Intersection of Alzheimer's Disease and Bone. *Curr Osteoporos Reports* 2024 221. 2024;22(1):177-181. doi:10.1007/S11914-023-00853-Z
- Omar M, Nassar S, Hijazi K, Glicksberg BS, Nadkarni GN, Klang E. Generating credible referenced medical research: A comparative study of openAI's GPT-4 and Google's gemini. *Comput Biol Med*. 2025;185:109545. doi:10.1016/J.COMPBIOMED.2024.109545
- Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Reports* 2023 131. 2023;13(1):14045-. doi:10.1038/s41598-023-41032-5
- Chelli M, Descamps J, Lavoué V, et al. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis. *J Med Internet Res*. 2024;26(1):e53164. doi:10.2196/53164
- Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? *FAccT 2021 - Proc 2021 ACM Conf Fairness, Accountability, Transpar*. 2021;21:610-623. doi:10.1145/3442188.3445922
- Kampa RK, Padhan DK, Karna N, Gouda J. Identifying the factors influencing plagiarism in higher education: An evidence-based review of the literature. *Account Res*. 2025;32(2):83-98. doi:10.1080/08989621.2024.2311212
- Jenkins R, Lin P. AI-Assisted Authorship: How to Assign Credit in Synthetic Scholarship. *SSRN Electron J*. Published online 30 January 2023. doi:10.2139/SSRN.4342909
- Flanagin A, Bibbins-Domingo K, Berkswits M, Christiansen SL. Nonhuman “Authors” and Implications for the Integrity of Scientific Publication and Medical Knowledge. *JAMA*. 2023;329(8):637-639. doi:10.1001/JAMA.2023.1344
- Oermann MH, Owens JK, Carter-Templeton H, Peterson G, Bailey HE. Using Artificial Intelligence for Scholarly Writing. *Am J Nurs*. 2025;125(11):52-55. doi:10.1097/AJN.0000000000000179
- Lee JY. Can an artificial intelligence chatbot be the author of a scholarly article? *J Educ Eval Health Prof*. 2023;20. doi:10.3352/JEEHP.2023.20.6
- Bobier C, Rodger D, Hurst D. Artificial intelligence policies in bioethics and health humanities: a comparative analysis of publishers and journals. *BMC Med Ethics* 2025 261. 2025;26(1):79-. doi:10.1186/S12910-025-01239-9
- Tang A, Li KK, Kwok KO, Cao L, Luong S, Tam W. The importance of transparency: Declaring the use of generative artificial intelligence (AI) in academic writing. *J Nurs Scholarsh*. 2024;56(2):314-318. doi:10.1111/JNU.12938
- Dergaa I, Chamari K, Zmijewski P, Saad H Ben. From human writing to artificial intelligence generated text: examining the prospects and potential threats of ChatGPT in academic writing. *Biol Sport*. 2023;40(2):615-622. doi:10.5114/BIOLOSPORT.2023.125623