Letter to the Editor

Authors' Reply: Citation Accuracy Challenges Posed by Large Language Models

Mohamad-Hani Temsah[1], MD; Ayman Al-Eyadhy[1], MD; Amr Jamal[2], MBBS; Khalid Alhasan[1], MBBS; Khalid H Malki[3], PhD

[1]Pediatric Department, College of Medicine, King Saud University, Riyadh, Saudi Arabia
[2]Department of Family and Community Medicine, King Saud University Medical City, Riyadh, Saudi Arabia
[3]Research Chair of Voice, Swallowing, and Communication Disorders, Department of Otolaryngology-Head and Neck Surgery, College of Medicine, King Saud University, Riyadh, Saudi Arabia

Corresponding Author:

Mohamad-Hani Temsah, MD
Pediatric Department
College of Medicine, King Saud University
King Abdullah Road
Riyadh, 11424
Saudi Arabia
Phone: 966 114692002
Email: mtemsah@ksu.edu.sa

Related Articles:
Comment on: https://mededu.jmir.org/2025/1/e72998
Comment on: https://mededu.jmir.org/2025/1/e63400

JMIR Med Educ 2025;11:e73698; doi: 10.2196/73698

Keywords: ChatGPT; Gemini; DeepSeek; medical education; AI; artificial intelligence; Saudi Arabia; perceptions; medical students; faculty; LLM; chatbot; qualitative study; thematic analysis; satisfaction; RAG retrieval-augmented generation

We appreciate the thoughtful critique of our manuscript "Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study" [1] by Zhao and Zhang [2]. Concerns over the generation of hallucinated citations by large language models (LLMs), such as OpenAI's ChatGPT, Google's Gemini, and Hangzhou's DeepSeek, warrant exploring advanced and novel methodologies to ensure citation accuracy and overall output integrity [3].

The LLMs have demonstrated a propensity to generate well-formatted yet fictitious references—a limitation largely attributed to restricted access to subscription-based databases and their reliance on probabilistic text generation [4]. As LLMs evolve, future iterations may integrate more reliable retrieval-based architectures, enhancing their capacity to cite legitimate sources while reducing fabricated references [4,5]. However, until such improvements are systematically validated, scholars must remain cautious.

One suggested enhancement is using retrieval-augmented generation (RAG) [6]. This approach integrates up-to-date external information, substantially improving real-world applicability. However, even RAG-based systems can misinterpret or distort source content under high-trust conditions. To address this, the authors developed Hallucination-Aware Tuning (HAT) [6]. HAT trains dedicated detection models to generate labels and detailed descriptions of identified hallucinations. These descriptions are then used by GPT-4 to correct discrepancies. The combination of corrected and original outputs forms a preference dataset that, when used for Direct Preference Optimization training, yields LLMs with reduced hallucination rates and improved answer quality [6].

We also propose another solution aimed at fundamentally reducing citation errors: the development of "Reference-Accurate" academic LLM by major global publishers. Leading journals could develop their own specialized LLM, trained exclusively on rigorously verified academic literature from robust databases. This targeted training would ensure that every generated reference is accurate and directly traceable to published work. Ideally, these publisher-backed LLMs would be made freely available to promote open science.

Therefore, we recommend a dual approach that combines advanced RAG methodologies with publisher-developed academic LLMs. Comparative studies should be conducted to evaluate the citation accuracy, factual consistency, and

overall performance of RAG-HAT-tuned models against these publisher-specific models. Collaborative efforts among academic institutions, publishers, and AI developers are essential to establish standardized protocols and reliable training datasets. Such partnerships would not only enhance the reliability of LLM-generated outputs but also foster greater trust in AI-assisted scholarly communication.

Moreover, the broader academic community bears responsibility for critically appraising AI-generated content. While LLMs can streamline information retrieval and synthesis, human oversight remains indispensable for safeguarding academic integrity. Rather than dismissing AI-driven tools due to their current flaws, we advocate for further research to ensure greater alignment with evidence-based scholarship and authentic publications. Future LLM iterations may rapidly overcome these limitations, but until then, transparency, responsible usage, and ongoing improvements in AI training remain imperative.

In conclusion, while RAG augmented by HAT represents a potential advancement in reducing hallucinations, the development of specialized, reference-accurate academic LLMs by publishers may offer a promising pathway. By integrating both strategies and ensuring human oversight, the academic community can ensure that AI-driven tools reliably support the rigor and transparency essential to scholarly research.

Conflicts of Interest

None declared.

References

1. Abouammoh N, Alhasan K, Aljamaan F, et al. Perceptions and earliest experiences of medical students and faculty with ChatGPT in medical education: qualitative study. JMIR Med Educ. Feb 20, 2025;11:e63400. [doi: 10.2196/63400] [Medline: 39977012]
2. Zhang M, Zhao T. Citation accuracy challenges posed by large language models. JMIR Med Educ. 2025. URL: https://mededu.jmir.org/2025/1/e72998 [doi: 10.2196/72998]
3. Temsah A, Alhasan K, Altamimi I, et al. DeepSeek in healthcare: revealing opportunities and steering challenges of a new open-source artificial intelligence frontier. Cureus. Feb 2025;17(2):e79221. [doi: 10.7759/cureus.79221] [Medline: 39974299]
4. Aljamaan F, Temsah MH, Altamimi I, et al. Reference hallucination score for medical artificial intelligence chatbots: development and usability study. JMIR Med Inform. Jul 31, 2024;12:e54345. [doi: 10.2196/54345] [Medline: 39083799]
5. Howard A, Hope W, Gerada A. ChatGPT and antimicrobial advice: the end of the consulting infection doctor? Lancet Infect Dis. Apr 2023;23(4):405-406. [doi: 10.1016/S1473-3099(23)00113-5] [Medline: 36822213]
6. Song J, Wang X, Zhu J, et al. RAG-HAT: a hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. Presented at: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Miami, Florida, US. 2024.[doi: 10.18653/v1/2024.emnlp-industry.113]

Abbreviations

HAT: Hallucination-Aware Tuning
LLM: large language model
RAG: retrieval-augmented generation

Edited by Surya Nedunchezhiyan; This is a non–peer-reviewed article; submitted 10.03.2025; accepted 12.03.2025; published 02.04.2025

Please cite as:
Temsah MH, Al-Eyadhy A, Jamal A, Alhasan K, Malki KH
Authors' Reply: Citation Accuracy Challenges Posed by Large Language Models
JMIR Med Educ 2025;11:e73698
URL: https://mededu.jmir.org/2025/1/e73698
doi: 10.2196/73698