






PERSPECTIVE

Threats and opportunities of using ChatGPT in scientific writing—The risk of getting spineless

Luca Ambrosio^{1,2,3}  | Jordy Schol³  | Vincenzo Amedeo La Pietra⁴ | Fabrizio Russo^{1,2}  | Gianluca Vadalà^{1,2}  | Daisuke Sakai³ 

¹Operative Research Unit of Orthopaedic and Trauma Surgery, Fondazione Policlinico Universitario Campus Bio-Medico, Rome, Italy

²Research Unit of Orthopaedic and Trauma Surgery, Department of Medicine and Surgery, Università Campus Bio-Medico di Roma, Rome, Italy

³Department of Orthopaedic Surgery, Tokai University School of Medicine, Isehara, Japan

⁴Faculty of Law, La Sapienza University of Rome, Rome, Italy

Correspondence

Daisuke Sakai, Department of Orthopaedic Surgery, Tokai University School of Medicine, 143 Shimokasuya, Isehara, Kanagawa 259-1143, Japan.
Email: daisakai@is.icc.u-tokai.ac.jp

Since the end of 2022, ChatGPT (OpenAI, San Francisco, CA, USA) and other chatbots have become increasingly popular worldwide for their incredible capacity to synthesize a vast body of information and express it in an outstandingly comprehensible way.¹ Considering these features, the use of ChatGPT has been quickly implemented in several fields of education and research, including data analysis and interpretation, production of digital content, and scientific writing.²

Briefly, ChatGPT is an artificial intelligence (AI)-based large language model (LLM) platform developed with a technique called Reinforcement Learning from Human Feedback.³ Basically, the software has been (and is continuously) trained with several text datasets to learn how to generate conversational, human-like responses, which fall under the Natural Language Processing (NLP) domain of AI.⁴ As a result, the user can literally ask whatever question—ranging from simple queries to detailed investigations—and receive an answer in a matter of seconds. What is more intriguing is the possibility to further tune and adjust the response, for example, in terms of length, linguistic register, text composition, and so forth. Needless to say, such characteristics have been increasingly used by researchers in several different ways which have been shown to potentially revolutionize clinical practice and scientific research. For example, ChatGPT has been able to effectively generate a patient discharge summary, simplify and implement a radiology report, identify potential targets for drug design, and even pass a medical licensing exam.²

Intuitively, the capacity of ChatGPT to search and summarize a large amount of data in a few seconds makes it a very efficient author. A rising number of editorials and letters to editors from eminent journals, such as *Nature*⁵ and *Science*,³ have pointed out the inherent risks concerning the applications of ChatGPT in medical writing. These include not only the act of misappropriating original content from an external source other than the author (which can be assimilated to plagiarism) but also relying on the correctness of reported data, often without verifying its truthfulness. Indeed, previous reports have already demonstrated that ChatGPT may reference inaccurate or even inexistent citations.² For example, when the authors asked ChatGPT to provide references on the cell density within the nucleus pulposus, the chatbot provided five different citations of which the first two were correct, while the rest included DOIs redirecting to other articles, fictitious titles, and/or wrong authors (Figure 1; the full conversation can be accessed here: <https://chat.openai.com/share/6fb8952a-31fb-4432-8d6a-159e0cbdc88b>).

Interestingly, despite the fictional nature of the fake references, all the authors mentioned are renowned experts in the intervertebral disc field with a strong publication track on similar topics. This demonstrates how ChatGPT may easily generate incorrect information which can be inadvertently reported, resulting in the propagation of inaccuracies as well as trivial circumstances.⁶ In a recent study by Walters et al.,⁷ the authors asked ChatGPT to generate 42 short essays on several different topics and provide

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *JOR Spine* published by Wiley Periodicals LLC on behalf of Orthopaedic Research Society.



Sure, here are a few references related to cell density of the nucleus pulposus and how it changes with aging:

1. Urban, J. P. (2002). The role of the physicochemical environment in determining disc cell behaviour. *Biochemical Society Transactions*, 30(6), 858-864. doi: 10.1042/bst0300858
2. Risbud, M. V., & Shapiro, I. M. (2014). Role of cytokines in intervertebral disc degeneration: pain and disc content. *Nature Reviews Rheumatology*, 10(1), 44-56. doi: 10.1038/nrrheum.2013.160
3. Li, Z., Lang, G., Karfeld-Sulzer, L. S., Ibrahim, Z. A., Fuchs, T., Gantenbein-Ritter, B., & Grad, S. (2018). Age-related changes in the cellular, mechanical, and biochemical properties of human cervical spine intervertebral discs. *PLoS One*, 13(5), e0197248. doi: 10.1371/journal.pone.0197248
4. Roberts, S., Menage, J., Urban, J. P., & Boyd, A. (1989). Matrix-metalloproteinases and aggrecanase: their role in disorders of the human intervertebral disc. *Spine*, 14(6), 601-610. doi: 10.1097/00007632-198906000-00014
5. Wuertz, K., & Haglund, L. (2013). Inflammatory mediators in intervertebral disk degeneration and discogenic pain. *Global Spine Journal*, 3(3), 175-184. doi: 10.1055/s-0033-1347365

FIGURE 1 Response of ChatGPT after asking for references related to cell density in the nucleus pulposus. While citations number 1 and 2 were correct, the remaining three contained fictitious authors' lists, titles, and/or DOI. For the full conservation access: <https://chat.openai.com/share/6fb8952a-31fb-4432-8d6a-159e0cbdc8b8>.

citations, using both the free access GPT-3.5 version and the premium subscription GPT-4 version. Among the 636 references generated, 55% (GPT-3.5) and 18% (GPT-4) were completely fabricated, meaning they had never been published, presented, or otherwise disseminated. When considering the real citations, 43% (GPT-3.5) and <7% (GPT-4) contained substantial errors in author names, titles, dates, journal titles, volume/issue/page numbers, publishers, or hyperlinks. Nonetheless, >40% of the references generated by both versions of the chatbot included minor formatting errors. Despite the higher performance of GPT-4, it is likely that most researchers utilize the free GPT-3.5 version, thus incurring an unacceptably high risk of generating misinformation. However, when balancing the advantages of producing such a large output in a matter of seconds against the risk of inaccurate citation reporting, the additional time needed for data validation and correction may still be worth the use of these tools.⁸

Due to serious threats to the integrity of scientific literature, the World Association of Medical Editors has recently released specific recommendations on the use of chatbots in scholarly publications.⁹ More specifically, it is stated that chatbots cannot be listed as authors, their use should be clearly stated in the acknowledgments of the manuscript, and that content generated or altered by AI should be actively inspected by editors and peer reviewers after submission. Some eminent journals have even prohibited the use of any AI-generated text or content in their published manuscripts altogether.¹⁰ Indeed, the concerns around the use of AI-based chatbots in research are multifaceted and can be imputed to several factors.

ChatGPT, and arguably most of the intelligent systems based on machine learning algorithms, suffer from what is referred to as the "black box problem." With this expression, experts usually refer to the intrinsic phenomenon of the inexplicability of the outputs of any given intelligent system. Developers and users are basically

unable to explain and justify how a system came to a given result. The problematic aspect of this peculiarity is that it can cause numerous issues. One of the major problems regarding scientific literature is plagiarism. As discussed above, ChatGPT can attribute contents to the wrong authors or even fabricate both the content and the author, through a collage of data and information which is practically impossible to predict or manage. This issue is aggravated by the bias in which the system can incur, due to the incorrectness or trustworthiness of the training data.¹¹ By definition, ChatGPT is intended for the average user and employs a natural, colloquial language to answer common questions. Therefore, the risk of inaccuracies is substantially higher in specialized and technical contexts, considering that the main task of the chatbot is not to provide a correct answer in absolute terms, but the one the user expects to be plausible based on the composition and tone of the query.¹² Nonetheless, while the volume of scientific publications increases steadily on a daily basis, creating new knowledge and/or revising preexisting concepts, ChatGPT has a default knowledge cut-off set in September 2021, meaning that the chatbot does not have access to real-time information beyond that date.¹³ Consequently, injudiciously relying on ChatGPT's outputs may generate misinformation and encourage confirmation bias.

In other conditions, inaccurate answers may not depend on the chatbot's limited knowledge or fabricated outputs, but rather on inherent limitations of the AI model itself. These may lead to comprehension errors (failure to understand the query context and intention), factualness errors (the model lacks the necessary supporting facts to generate a correct answer), specificity errors (failure to address the question at the appropriate level of specificity), and inference errors (the chatbot has the necessary knowledge to answer the query, but fails to produce the correct answer).¹⁴ Interestingly, these errors often occur when the user formulates laconic, unspecific, or poorly

contextualized questions. According to a recent study, the truthfulness of the model can be improved by providing exhaustive background information, specific external knowledge, and decomposing complex problems into subproblems.¹⁴

Therefore, the authors strongly urge members of the spine research community to exercise utmost vigilance when utilizing ChatGPT and similar AI chatbots for research and writing purposes.

1 | SUGGESTION FOR THE JOR SPINE

As members of the JOR Spine editorial and scientific advisory team, we acknowledge the potential advantages of utilizing ChatGPT and similar AI chatbots as valuable tools to support the research process. However, it is crucial to exercise caution and prudence when employing these technologies. We would strongly suggest that papers written (solely) by ChatGPT should not be accepted for publication in our journal nor should AI chatbots be included as an author. Furthermore, any data or references derived from such chatbots should only be included in manuscripts if they have been duly validated and properly referenced. Our journal might want to consider requesting authors to confirm that their manuscript did not employ AI chatbots for data or text generation, as part of the manuscript submission process. Nevertheless, we do encourage the use of ChatGPT as an assistant to aid in manuscript writing, particularly for tasks such as spelling, style, and grammar checks. We firmly believe that ChatGPT can serve as a valuable tool to facilitate and promote accessibility and equality in publishing, particularly for individuals from non-native English-speaking countries. Nevertheless, it is imperative to ensure that the meaning and statements generated by ChatGPT are carefully validated and checked prior to any manuscript submission. With this perspective, we hope to start a discussion in our community on how to deal with these upcoming technologies. We are hopeful that the ORS Spine community can design clear guidelines and expectations with regard to these next-generation technologies to safeguard the transparency, accuracy, and value of science in our field.

ACKNOWLEDGMENTS

Not applicable.

CONFLICT OF INTEREST STATEMENT

Daisuke Sakai is one of the Editors-in-Chief of JOR Spine, while Gianluca Vadalà and Jordy Schol are members of the Editorial Board. They were all excluded from editorial decision-making related to the acceptance of this article for publication in the journal.

ORCID

Luca Ambrosio  <https://orcid.org/0000-0003-2424-1274>

Jordy Schol  <https://orcid.org/0000-0001-5489-2591>

Fabrizio Russo  <https://orcid.org/0000-0002-8566-8952>

Gianluca Vadalà  <https://orcid.org/0000-0001-7142-1660>

Daisuke Sakai  <https://orcid.org/0000-0003-4189-9270>

REFERENCES

1. OpenAI. ChatGPT. Accessed July 9, 2023. <https://chat.openai.com/>
2. Sallam M. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare*. 2023;11(6):887.
3. Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379:313.
4. Bacco L, Russo F, Ambrosio L, et al. Natural language processing in low back pain and spine diseases: a systematic review. *Front Surg*. 2022;9:957085.
5. Stokel-Walker C. ChatGPT listed as author on research papers: many scientists disapprove. *Nature*. 2023;613:620-621.
6. Twitter. #ChatGPT wrote the 'Discussion' section of an @MDPIOpenAccess paper in the Toxins journal?! 2023. Accessed July 9, 2023. https://twitter.com/gcabanac/status/1655129160245821440?s=61&t=CHg-VZHLal2AMdiEqTRQGQ&fbclid=IwAR3_9xGrkjAQgiogphul7vvSOFLvLleEaakaBxEIpYGKGtQtFKfRL95XkOY
7. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023;13:14045.
8. Chen T-J. ChatGPT and other artificial intelligence applications speed up scientific writing. *J Chin Med Assoc*. 2023;86:351-353.
9. WAME. Chatbots, Generative AI, and Scholarly Manuscripts WAME Recommendations on Chatbots and Generative Artificial Intelligence in Relation to Scholarly Publications. 2023. Accessed July 10, 2023. <https://wame.org/page3.php?id=106>
10. Science. Science Journals Editorial Policies. 2023. Accessed July 11, 2023. https://www.science.org/content/page/science-journals-editorial-policies?adobe_mc=MC MID%3D06834777952383523013640686692414937187%7CMCORGID%3D242B6472541199F70A4C98A6%2540AdobeOrg%7CTS%3D1688923565#authorship
11. Quinn TP, Jacobs S, Senadeera M, le V, Coghlan S. The three ghosts of medical AI: can the black-box present deliver? *Artif Intell Med*. 2022;124:102158.
12. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595.
13. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023;307:e230163.
14. Zheng S, Huang J, Chang KC-C. Why Does ChatGPT Fall Short in Providing Truthful Answers? 2023. arXiv 2304.10513.

How to cite this article: Ambrosio, L., Schol, J., La Pietra, V. A., Russo, F., Vadalà, G., & Sakai, D. (2024). Threats and opportunities of using ChatGPT in scientific writing—The risk of getting spineless. *JOR Spine*, 7(1), e1296. <https://doi.org/10.1002/jsp2.1296>