PRINCIPLES AND PRACTICE OF CLINICAL RESEARCH
A Global Journal of Clinical Research

Editorial - The Use of Large Language Models in Science: Opportunities and Challenges

Bassel Almarie[1], Paulo E. P. Teixeira[1], Kevin Pacheco-Barrios[1], Carlos Augusto Rossetti[1], Felipe Fregni[1*]

[1]Neuromodulation Center and Center for Clinical Research Learning, Spaulding Rehabilitation Hospital, Massachusetts General Hospital, Harvard Medical School, Boston, MA, United States.

Introduction

A large language model (LLM) is a narrow artificial intelligence (AI) system that has been trained on a massive amount of text data to interpret natural language and generate human-like responses to text-based prompts or questions (Kasneci et al., 2023). These models are typically based on deep neural networks that have been trained on vast amounts of data, such as entire books, articles, or web pages, to learn patterns in language and extract meaning from text (Kasneci et al., 2023). The most advanced large language models can generate high-quality text that is difficult to distinguish from human-generated content, and they are used for a wide range of natural language processing tasks, such as language translation, text summarization, question-answering, and content generation (Kasneci et al., 2023).

One of the most popular large language model today is ChatGPT, also known as the Chat Generative Pre-trained Transformer. It is a deep learning model that uses generative artificial intelligence algorithms to provide personalized responses to each end-user. It was launched in November 2022 and gained wide global attention. ChatGPT leverages natural language processing techniques and retains information from previous exchanges to provide human-like interactions that can be used for several different purposes. Despite its potential benefits, concerns about its misuse have been raised. Recently, ChatGPT was employed in scientific manuscript publications, which has sparked debates (Stokel-Walker, 2023) about the ethical implications of utilizing this new technology. In this editorial, we will discuss the application of LLMs, using ChatGPT as an example, in scientific endeavors, including scientific writing, evidence synthesis, and clinical research tasks. In addition, we will define authorship in science in the age of AI and highlight the appropriate use of these tools in manuscript submission and peer-reviewing processes.

The science and art of medical writing: understanding authorship criteria and limitations of AI tools

Medical writing is science and art (Sharma, 2010). Science involves understanding the complexity of the field, identifying scientific gaps, and communicating clinical data and research findings. The art lies in delivering these complex findings in a clear and coherent fashion that guides the reader to a smooth and logical understanding. Scientists have been recognized and rewarded for such skills. To present research findings through well-written, peer-reviewed scientific manuscripts is among the ultimate goals of any investigational research study. Therefore, taking credit for impactful research papers is a common trophy in academia, whether achieved by a single author or a team of co-authors. Consequently, authorship has always been an important aspect of medical writing, as it reflects contributions made by individuals involved in the research project. The Principles and Practice of Clinical Research adheres to the guidelines regarding authorship set by the International Committee of Medical Journal Editors (ICMJE) ("ICMJE | Recommendations | Defining the Role of Authors and Contributors," n.d.), which states that authors must meet a complete set of criteria to be entitled to all rights and privileges that comes with authorship. These criteria involve (1) contribution to the conception and design or data analysis and interpretation; (2) article drafting or revision, which

*Corresponding author: fregni.felipe@mgh.harvard.edu
Received: April 05, 2023 **Accepted:** May 04, 2023
Published: July 10, 2023
Editor: Aurore Thibaut
Keywords: artificial intelligence, plagiarism, clinical research
DOI: http://dx.doi.org/10.21801/ppcrj.2023.91.1

is critically important for intellectual content; (3) final version approval before publication; and (4) agreement to be accountable for all aspects of the work including accuracy or integrity of any part of the work. Consequently, some may argue that LLMs such as ChatGPT may not qualify for authorship as the ICMJE guidelines for authorship cannot be met for these tools. Generative AI tools operate on statistical patterns using data they were trained on. Thus, it can be averred that these tools lack critical judgment and the human ability to intellectually revise contents. They are also unable to approve a manuscript or agree to be accountable to its content.

Likewise, AI tools do not qualify to peer review a scientific manuscript. The peer-review process is the heart (Smith, 2006) and gold standard of the scientific publication process (Mayden, 2012). Peer review is a common process that can grant allocation, academic promotion, textbook writing, and Nobel prize determination (Smith, 2006). Peer reviewers must be experts in the field with the ability to identify the research gap and evaluate the consistency and novelty of a particular piece of work. A peer reviewer must have complex analytical skills to evaluate data, advance science, and improve outcomes (Bearinger, 2006). Therefore, using AI tools in the peer-reviewing process should be limited to grammatical editing and proofreading. Materials retrieved through AI tools should be transparent and verified by the reviewer to ensure the integrity and standards of the peer-reviewing process. Furthermore, if LLMs and related tools are used, they must be fully acknowledged in the peer-review summary or in the published manuscript.

Advantages and appropriate use LLMs tools in science

LLMs can be powerful tools to reduce the amount of time spent on language editing and proofreading in academia (Pividori and Greene, 2023). They can improve the readability of drafts, especially for non-native speakers. This can accelerate the submission and publishing process. However, LLMs are associated with automated biases (Skitka et al., 2000), and may spread misinformation, and errors based on the data the model was trained on (van Dis et al., 2023). The use of these models must be transparent, where authors clearly state how the tools were utilized, acknowledging the AI tool in the text, and conducting a thorough review and approval of the final manuscript before submission. Additionally, the literal use of text produced by LLMs must be avoided. To prevent plagiarism, the text must be edited, paraphrased, or reported using quotation marks. Ideally, we recommend the authors save the chat history and report it as supplementary material if the text is part of a scientific manuscript seeking peer-review publication.

Furthermore, LLMs can be a valuable tool to brainstorm, organize ideas, and generate titles. It can also be used to search literature and generate literature reviews (Aydın and Karaarslan, 2022). It can potentially be integrated in the systematic reviews pipelines to facilitate narrative synthesis of included studies and to support data extraction from clinical trials. These usages can be implemented in living synthesis (automatic updated systems) of evidence to serve as a communication interface for the development of evidence-based summaries and clinical practice guidelines (Oliveira et al., 2014).

However, it is unable to provide reliable sources of the data it is generating ("The AI writing on the wall," 2023). LLMs can also assist with various aspects of data analysis. It can generate codes to create graphs and visualize results. It can also generate code for various programming languages such as Stata, R, and Python. Nonetheless, the generated codes must be verified and modified properly as the language model may produce incorrect output. Similarly, there are several clinical research tasks that requires constant interaction between potential study participants and the research team, including study recruitment and adherence. The LLMs systems can be implemented as initial screening contact that provides accurate information to participants to confirm inclusion and exclusion criteria. Furthermore, LLMs can be used as personalized reminder systems to increase adherence and study participation. The advantages and appropriate use of LLMs in science are summarized in Table 1.

The writing process has evolved throughout history, from the invention of typewriters to the use of word processors and grammar checkers. Through proper education, we maximized benefits to outweigh risks of these tools. AI-generative models can also be utilized in the education system by prioritizing the education on their advantages and limitations. Instead of focusing solely on outlining texts, which can easily lead to academic misconduct, tasks can be shifted to evaluate discrepancies between AI output and human original work through identifying errors and inaccuracies. This process can enhance critical thinking and innovative abilities.

Disadvantages and risks of using LLMs

LLMs are trained on data that contains encoded biases, some of which have been identified as harmful and carry potential risks. Moreover, the coherent responses generated by these models may create a false perception that the content is meaningful and

Advantage	Appropriate Use
Improve text quality	▪ Final text must be reviewed and approved by the author ▪ AI-tool used for language editing or proofreading must be acknowledged in the text e.g. (ChatGPT, OpenAI, California, USA)
Brainstorm and organize ideas	▪ Provide prompt as starting point ▪ Guide the human-machine interactions ▪ Improve brainstormed material by adding elements AI tools unable to generate such as context and innovation
Coding	▪ Verify generated codes ▪ Modify as needed
Evidence synthesis	▪ Identification of included studies in systematic reviews ▪ Automatization of narrative synthesis and data extraction
Clinical research tasks	▪ Initial eligibility screening for potential study participants ▪ Reminder systems to increase adherence and study participation
Improve human critical thinking	▪ Generate AI text ▪ Detect errors and misinformation ▪ Provide feedback on how to elevate the generated text into high-quality content

Table 1: *Utilizing large language models in scientific contexts*

corresponds to an accountable entity, leading the end-user to trust the information produced without question (Bender et al., 2021). LLMs can accelerate the spread of misinformation and create new misinformation. Conversely, human writing has been shown to have numerous benefits, including promoting learning and enhancing higher-order thinking (Kim et al., 2021), as well as having therapeutic benefits for wellbeing ("Writing well: health and the power to make images | Medical Humanities," n.d.). Relying on AI-generated texts may compromise these benefits and have potential negative impacts on individuals and society.

Complex and multi-faceted fields such as medicine and law require a thorough understanding of specific knowledge and a high level of emotional intelligence to make tailored ethical decisions in specific circumstances. Language models such as ChatGPT lack the expertise to be considered reliable or provide meaningful prompts in these fields. They are also limited by outdated data that may not reflect present-day information. Nonetheless, the false sense of interaction may lead end-users to seek medical or legal advice that can result in potential harm.

Modern plagiarism in the age of AI

The National Library of Medicine introduced the term "plagiarism" to its controlled vocabulary thesaurus, MeSH (Medical Subject Headings) in 1990. The term was defined as passing off as one's own the work of another without credit ("Plagiarism - MeSH - NCBI," n.d.). Plagiarism is a form of academic misconduct ("Federal Research Misconduct Policy | ORI - The Office of Research Integrity," n.d.). It violates research integrity and undermines academic standards. Tools such as ChatGPT can result in academic misconduct. While such AI tools can be considered mediums assisting in the creation process ("Burrow-Giles Lithographic Company v. Sarony, 111 U.S. 53 (1884)," n.d.; Hristov, 2016), failing to cite these tools properly, indicates that the authors wrote a text they did not write. This modern form of misconduct can be defined as AI-assisted plagiarism. Therefore, it is essential to credit all tools used in the manuscript publishing process to ensure academic integrity as well as meeting the criteria for authorship. To combat AI plagiarism, tools have been developed to detect AI output ("GPT-2 Output Detector," n.d.; "GPT-2 Output Detector," n.d.; "GPTZero," n.d.). Such tools can be helpful to account for the reliability of the resulting content and maintain academic honesty and standards. However, some AI-output detectors may not be reliable as newer tools are now available to make AI-generated text undetectable by AI-output detectors.

Conclusions

We agree and envision that several tasks during the editorial workflow can be optimized and implemented using LLMs, such as the verification of article completion ("initial quality check"), a communication bridge between the editorial team and authors, references verification, and proofreading. Moreover, numerous scientific tasks can also be improved by using LLMs (brainstorming, manuscript writing, evidence synthesis, etc.). However, we believe that LLMs cannot be considered authors in the strict sense in the context of scientific endeavors and must be considered tools that must be carefully used and disclosed under international AI governance protocols.

Funding

FF is funded by NIH RO1 grant (1R01HD082302-01A1). KPB is supported by a Spaulding Research Institute Grant (Jay Stroke Award).

References

- Aydın, Ö., Karaarslan, E., 2022. OpenAI ChatGPT Generated Literature Review: Digital Twin in Healthcare. https://doi.org/10.2139/ssrn.4308687
- Bearinger, L.H., 2006. Beyond objective and balanced: Writing constructive manuscript reviews. Res. Nurs. Health 29, 71–73.

https://doi.org/10.1002/nur.20119
- Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21. Association for Computing Machinery, New York, NY, USA, pp. 610–623. https://doi.org/10.1145/3442188.3445922
- Burrow-Giles Lithographic Company v. Sarony, 111 U.S. 53 (1884) [WWW Document], n.d. . Justia Law. URL https://supreme.justia.com/cases/federal/us/111/53/ (accessed 3.3.23).
- Federal Research Misconduct Policy | ORI - The Office of Research Integrity [WWW Document], n.d. URL https://ori.hhs.gov/federal-research-misconduct-policy (accessed 3.3.23).
- GPT-2 Output Detector [WWW Document], n.d. URL https://openai-openai-detector.hf.space/ (accessed 3.4.23).
- GPTZero [WWW Document], n.d. URL https://gptzero.me/ (accessed 3.4.23).
- Hristov, K., 2016. Artificial Intelligence and the Copyright Dilemma.
- ICMJE | Recommendations | Defining the Role of Authors and Contributors [WWW Document], n.d. URL
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.
- Kim, S., Yang, J.W., Lim, J., Lee, S., Ihm, J., Park, J., 2021. The impact of writing on academic performance for medical students. BMC Med. Educ. 21, 61. https://doi.org/10.1186/s12909-021-02485-2
- Mayden, K.D., 2012. Peer Review: Publication's Gold Standard. J. Adv. Pract. Oncol. 3, 117–122.
- Oliveira, T., Novais, P., and Neves, J. (2014). Development and implementation of clinical guidelines: an artificial intelligence perspective. Artificial Intelligence Review, 42, 999-1027.
- Pividori, M., Greene, C.S., 2023. A publishing infrastructure for AI-assisted academic authoring. https://doi.org/10.1101/2023.01.21.525030
- Plagiarism - MeSH - NCBI [WWW Document], n.d. URL https://www.ncbi.nlm.nih.gov/mesh (accessed 3.3.23).
- Sharma, S., 2010. How to Become a Competent Medical Writer? Perspect. Clin. Res. 1, 33–37.
- Skitka, L.J., Mosier, K., Burdick, M.D., 2000. Accountability and automation bias. Int. J. Hum.-Comput. Stud. 52, 701–717. https://doi.org/10.1006/ijhc.1999.0349
- Smith, R., 2006. Peer review: a flawed process at the heart of science and journals. J. R. Soc. Med. 99, 178–182. https://doi.org/10.1177/014107680609900414
- Stokel-Walker, C., 2023. ChatGPT listed as author on research papers: many scientists disapprove. Nature 613, 620–621. https://doi.org/10.1038/d41586-023-00107-z
- Kim, S., Yang, J.W., Lim, J., Lee, S., Ihm, J., Park, J., 2021. The impact of writing on academic performance for medical students. BMC Med. Educ. 21, 61. https://doi.org/10.1186/s12909-021-02485-2
- The AI writing on the wall, 2023. . Nat. Mach. Intell. 5, 1–1. https://doi.org/10.1038/s42256-023-00613-9
- van Dis, E.A.M., Bollen, J., Zuidema, W., van Rooij, R., Bockting, C.L., 2023. ChatGPT: five priorities for research. Nature 614, 224–226. https://doi.org/10.1038/d41586-023-00288-7
- Writing well: health and the power to make images | Medical Humanities [WWW Document], n.d. URL https://mh.bmj.com/content/26/2/79 (accessed 3.3.23).