Ethics of large language models in medicine and medical research

Published Online
April 27, 2023
https://doi.org/10.1016/
S2589-7500(23)00083-3

Large language models (LLMs) are a type of deep learning model that are trained on vast amounts of text data with the goal of generating new text that closely resembles human responses. The release of ChatGPT (OpenAI, San Francisco, CA, USA), an LLM-based chatbot, on Nov 30, 2022, propelled LLMs to the forefront of public attention and made them accessible to millions of people to experiment with. Since then, medical practitioners and researchers have been exploring potential applications of LLMs, as much of medical practice and research revolve around large text-based tasks, such as presentations, publications, documentation, and reporting. The use of LLMs to aid or simplify these tasks can lead to significant time saving, and enable clinicians and researchers toward other efforts. There are multiple additional LLMs in various stages of development, including BioGPT (Massachusetts Institute of Technology, Boston, MA, USA), LaMDA (Google, Mountainview, CA, USA), Sparrow (Deepmind AI, London, UK), Pangu Alpha (Huawei, Shenzen, China), OPT-IML (Meta, Menlo Park, CA, USA), and Megataron Turing MLG (Nvidia, Santa Clara, CA, USA). Some of the newer adaptations, such as BioGPT, which focuses on biomedical text generation and mining after training on PubMed articles, could have major implications for the future of medicine and medical research. As with any emerging, disruptive technology, it is important to consider the ethics of use and prioritise responsible and beneficial applications that serve the best interests of society.

Use of LLMs in medicine and medical research can be used for text generation, text summarisation, and text correction. LLMs can be used to generate large amounts of text de novo; for example, templates for clinical documentation or standardised reporting, presentation outlines, or cover letter examples for manuscript submissions. LLMs can be used to condense complex academic papers into succinct summaries, enabling authors to comprehend difficult concepts in manuscripts, or automatically generate abstracts. Additionally, LLMs have the potential to summarise large amounts of clinical data; for example, for the purposes of turning clinical notes into concise summary statements for patient workups. LLMs can be used to enhance the grammar, readability, and conciseness of written content while maintaining the overall message and context. This has applications in academic writing to improve the clarity and coherence of manuscripts and grant proposals. Furthermore, they can help convert written works into various formats suitable for different journal publications. Clinically, LLMs can help rephrase text to be more patient friendly. For example, they can help generate letters to patients or insurance companies with appropriate verbiage instead of medical jargon.

Ethical considerations for the use of LLMs in academic medicine and medical research should be addressed. First, in safety-critical domains, such as medicine and medical research, hidden bias in LLMs could have serious consequences for patient outcomes. LLMs produce text that are reflections of their training data and thus could perpetuate biases pertaining to race, sex, language, and culture.[1–3] Moreover, the body of knowledge used to train the models typically arises from well funded institutions in high-income, English-speaking countries. Thus, there is a significant under-representation of perspectives from other regions of the world, leading to mechanistic models of health and disease biased towards understandings of these processes of high-income countries. For example, a clinician in Africa using LLMs to generate an outline of a presentation for treatment options in diabetes could lead towards focusing on treatment paradigms applicable only in high-income countries. This could limit the scope of discussion of different treatments popular in other regions of the world or those that might be more relevant to that country's patient population.

Second, the use of LLMs in medical writing disrupts the traditional notion of trust, which encompasses the dependability of information and credibility of sources or authors. LLM outputs are untraceable, difficult to discern from the voices of actual authors, and at times might be completely inaccurate.[4,5] Furthermore, the ground truth in medicine is constantly evolving, making it difficult to determine whether LLMs reflect the most current data, and current models do not evaluate the quality or provide a measure of uncertainty for their

Panel: **International Committee of Medical Journal Editors Authorship Guidelines**

- Substantial contributions to the conception or design of the work; or the acquisition, analysis, or interpretation of data for the work; AND
- Drafting the work or revising it critically for important intellectual content; AND
- Final approval of the version to be published; AND
- Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved

outputs. Although this could be a future feature in LLMs, current hesitation regarding LLMs in medical research is justified and evidenced by journals now requiring authors to disclose the use of LLMs or prohibiting their use entirely. Interestingly, this raises the question of whether the use of LLMs will be stigmatised or discriminated against, particularly in areas of great consequence to patient outcomes, such as medical research. Will society forfeit the immense potential of LLMs due to our lack of trust in its use? Or should we embrace this technology but demand a higher level of scrutiny when assessing content generated by LLMs?

Third, authorship is critically important in medicine and medical research as there are legal implications for the authors of medical documentation and there is immense trust attached to the credibility of the research authors. If LLMs are used in clinical and research tasks, where do we draw the line between an assistive tool and the author? According to the International Committee of Medical Journal Editors (ICMJE), authorship must meet all criteria (panel). However, LLMs are a type of technology that the framers of the ICMJE never anticipated. For example, authorship is predicated on contributing to aspects such as the conception, design, analysis, and drafting of a work. While LLMs seem capable of these aspects, they do not exactly know what they are doing and do not have intentionality. What is assured, however, is that they can neither approve a final version or be held accountable for the integrity of the work.[6] Therefore, a case can be made that LLMs are in the same category as Microsoft Word (Microsoft, Redmond, WA, USA), Endnote (Clarivate, London, UK), or a thesaurus—as assistance tools. But to the author, LLMs appear to offer much more than the other assistive technology tools. And thus, these guidelines are insufficient to truly characterise contribution in terms of authorship, and further discussions and new guidelines might be required.

Fourth, there are varying payment models with emerging LLMs. For example, to train a model using ChatGPT's base LLM, there is a tiered payment system based on performance and usage volume. Other LLMs have followed suit with variations, such as upfront subscription fees for access, use, or training. We raise the concern that cost of access widen existing digital divides for those with less academic support or funding. Thus, it is important to ensure that LLMs are affordable and accessible with provisions in subsidies or grants to support users at less-funded institutions. Conversely, LLMs can enhance productivity and potentially introduce equal opportunity in academia for individuals with limited proficiency in the English language, and for those who might not have strong editorial skills. The use case of text rephrasing allows researchers to create a more polished and professional product, allowing individuals to clearly communicate their ideas on par with that of experienced academics.

Fifth, the use of LLMs raises the ethical concern relating to the collection, use, and potential dissemination of the data inputted into LLMs. Text input into LLM application programming interfaces might contain sensitive protected health information or unpublished data that might be at risk by being available to their company employees or potential hackers. Given the absence of transparency in how commercial companies use or store their input information, a user must consider whether it is ethical to risk putting sensitive data into this black box. Thus, it will be important for individuals and institutions to implement strict controls for the de-identification of data and obtaining informed consent for the use of protected health information submitted to LLM application programming interfaces.

In conclusion, the integration of LLMs, such as ChatGPT, in medical practice and research raises crucial ethical issues concerning bias, trust, authorship, equitability, and privacy. Although it is undeniable that this technology has the power to revolutionise medicine and medical research, being mindful of its potential consequences is essential. An outright ban on the use of this technology would be shortsighted. Instead, establishing guidelines that aim to responsibly and effectively use LLMS is crucial.

We declare no competing interests. We would like to acknowledge the use of ChatGPT version 3.5 February, 2023 (OpenAI, San Francisco, CA, USA) in the editing of this Comment, specifically in the discussion of equitability and in the conclusion section (appendix). The use of a large language model has greatly assisted us in rephrasing and ensuring the clarity and effectiveness of our language, and highlights the potential of language models in aiding researchers and health-care professionals save valuable time by effectively and efficiently communicate in written words the concepts that they wish to express.

*Hanzhou Li, John T Moon, Saptarshi Purkayastha, Leo Anthony Celi, Hari Trivedi, Judy W Gichoya
hli277@emory.edu

Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta, GA 30322, USA (HL, JTM, HT, JWG); School of Informatics and Computing, Indiana University Purdue University, Indianapolis, IN, USA (SP); Massachusetts Institute of Technology, Cambridge, MA, USA (LAC)

1 Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories: proceedings of the third workshop on narrative understanding. Stroudsburg, PA: Association for Computational Linguistics, 2021.
2 Guo W, Caliskan A. Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY: Association for Computing Machinery, 2021.
3 Abid A, Farooqi M, Zou J. Large language models associate Muslims with violence. *Nat Mach Intell* 2021; **3:** 461–63.
4 Holzinger A. The next frontier: AI we can really trust. In: Kamp M, Koprinska I, Bibal A, et al, eds. Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Cham: Springer International Publishing, 2021: 427–40.
5 International Association of Scientific, Technical, and Medical Publishers. AI ethics in scholarly communication: STM best practice principles for ethical, trustworthy and human-centric AI. International Association of Scientific, Technical, and Medical Publishers, 2021.
6 The Lancet Digital Health. ChatGPT: friend or foe? *Lancet Digit Health* 2023; **5:** e102.

See **Online** for appendix