Editorial Commentary

Editorial on large language models

Christopher M. Deibert

Division of Urologic Surgery, University of Nebraska Medical Center, Omaha, NE, USA

Correspondence to: Christopher M. Deibert, MD, MPH. Division of Urologic Surgery, University of Nebraska Medical Center, 984110 Nebraska Medical Center, Emile Street, Omaha, NE 68198, USA. Email: Christopher.deibert@unmc.edu.

Comment on: Davis R, Eppler M, Ayo-Ajibola O, et al. Evaluating the Effectiveness of Artificial Intelligence-powered Large Language Models Application in Disseminating Appropriate and Readable Health Information in Urology. J Urol 2023;210:688-94.

Keywords: Artificial intelligence (AI); urology; trust

Submitted Nov 17, 2023. Accepted for publication Mar 08, 2024. Published online Apr 23, 2024.
doi: 10.21037/tau-23-591
View this article at: https://dx.doi.org/10.21037/tau-23-591

Artificial intelligence (AI) has long captivated the imagination. Most recently, a number of large language models (LLM) have been launched, including ChatGPT (OpenAI/Bing), Bard (Google/Alphabet), and LLaMA (Meta). There has been much scientific excitement about these models. In particular, their potential value in the healthcare environment. The authors from the University of Southern California have quickly positioned themselves as leaders in this area of exploration. They recently discussed the role of AI with direct patient medical queries. Since at least 60% of Americans use the Internet to find health information, it's natural to assume that many will turn to these artificial intelligent chat bots for questions about their health (1).

The authors used Google Trends to identify 18 common online urologic searches (2). They then formatted these searches into questions and posed them to ChatGPT version 3.5. These were assessed in several domains including: accuracy, comprehensiveness, clarity, and readability. In general, the authors used well established quality metrics to assess the chatbot outcomes, which strengthens the quality of this study.

Overall, the outputs were of high quality and clarity. All answers were graded to be clear. However, the accuracy and comprehensiveness were lower. Fourteen of 18 responses were deemed appropriate by the reviewing authors. Although there was no statistical significance, the chat bot seemed to perform better on treatment questions and benign urologic conditions compared to oncologic and emergency type topics.

On their assessment of readability, most outputs were at a grade 13, whereas the American Medical Association recommends patient education materials be geared towards a 6th grade reading level. This means that the chatbot gave information and potentially advice at a reading level higher than patients might understand. This is one of several areas of concern. When patients are seeking medical care online, only if a patient understands what they read can they hope to implement the recommendation. The AI companies could of course improve these outputs through training. For instance, when a medical question is asked of their LLM, it could automatically define the output to be at a 6th grade reading level. When we as clinicians input the LLM prompt, you can also ask the chatbot to "Produce an output at 6th grade reading level". Of note, 17 of the 18 queries did respond with the appropriate recommendation to seek medical care for the prompted health query.

The natural evolution of this type of work is to present it directly to our patients. How do our patients perceive these LLM systems? Do they yet trust these programs to answer medical information accurately? Should they? How would patients rate the readability, clarity, and appropriateness of the response to their question? In a similar project, 100 urology case studies were presented to ChatGPT 3.5 and found only 52% of responses appropriate. This suggests that the models have a long way to go for trustworthy direct patient facing clinical responses (3).

Much of the next stage of AI LLMs will be fine tuning and using pre-existing datasets to build stronger more accurate output. This often includes reinforcement learning from human feedback (RHF). Here, humans essentially review the outputs and then provide feedback that pushes

the LLM in a direction of improving the response quality. Clinicians will likely play a role here in fine tuning health care related material. Additionally, a process termed retrieval-augmented generation has LLMs use specific datasets or trusted technical or policy documents as external sources that can be cited by the LLM output. For instance, a urology specific chat bot was developed based on the European Association of Urology Oncology guidelines. When built directly on guidelines, it is no surprise that accuracy of response to patient queries improved (4). Such training can focus on using only certain prereviewed medical websites or datasets in any healthcare training model. While the current models aren't quite ready for prime time in healthcare, the growth and improvement has been staggering in just the past year. It's easy to imagine both our patients and frankly us, using these systems regularly, even within the next year. If you have yet to personally interact with one of these LLM systems, I strongly encourage you to do so, to learn for yourself.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Translational Andrology and Urology*. The article has undergone external peer review.

Peer Review File: Available at https://tau.amegroups.com/article/view/10.21037/tau-23-591/prf

Conflicts of Interest: The author has completed the ICMJE uniform disclosure form (available at https://tau.amegroups.com/article/view/10.21037/tau-23-591/coif). The author has no conflicts of interest to declare.

Ethical Statement: The author is 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.

References

1. Pew Research Center. Survey. The Internet and Health 2009. 2013. Available online: https://www.pewresearch.org/internet/2013/02/12/the-internet-and-health/
2. Davis R, Eppler M, Ayo-Ajibola O, et al. Evaluating the Effectiveness of Artificial Intelligence-powered Large Language Models Application in Disseminating Appropriate and Readable Health Information in Urology. J Urol 2023;210:688-94.
3. Cocci A, Pezzoli M, Lo Re M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer Prostatic Dis 2024;27:103-8.
4. Khene ZE, Bigot P, Mathieu R, et al. Development of a Personalized Chat Model Based on the European Association of Urology Oncology Guidelines: Harnessing the Power of Generative Artificial Intelligence in Clinical Practice. Eur Urol Oncol 2024;7:160-2.

Cite this article as: Deibert CM. Editorial on large language models. Transl Androl Urol 2024. doi: 10.21037/tau-23-591