Naji J. Touma
Department of Urology, Queen's University, Kingston, ON, Canada

Cite as: Touma NJ. Resident assessment in urology: ChatGPT, take the wheel? *Can Urol Assoc J* 2025;19(6):188. http://dx.doi.org/10.5489/cuaj.9256

See related article on page 182

Resident assessment in urology: ChatGPT, take the wheel?

Since the release of Open AI's ChatGPT in 2022, large language models (LLMs) have found an increasing role in streamlining tasks and improving efficiency in many fields, including medical education. Initially, gauging the ability of a LLM to pass a board exam acted as a barometer of the readiness of such models to reflect the state of knowledge in that field.[1] More recently, ChatGPT's capability to generate questions, specifically multiple-choice questions (MCQs), for such exams has been evaluated. An earlier report suggested that ChatGPT-4 generated questions with a lower discrimination index (DI) compared to those generated by faculty, with the possibility of hallucinations being ever present.[2]

In the study published in this month's CUAJ, the authors' innovation is that they were able to generate MCQs with a better DI, albeit this will need to be confirmed with larger numbers;[3] however, their technique of creating a customizable LLM trained on specific urology content is promising and is likely to yield improved testing material. Prompt engineering is also likely to enhance any LLM-generated questions.

With the accelerated evolution of LLM capability, it appears that a threat is manifesting on the horizon, jeopardizing the role of the educator; however, a few limits remain in the short-to-intermediate-term, preserving the role of human content creators.

One concept in clinical assessments is Miller's pyramid, where broad knowledge is at the bottom and higher-up skills, such as competence and performance, are layered on top.[4] The top of the pyramid is best evaluated by objective structured clinical examinations (OSCEs) and simulation, and although not yet available, artificial intelligence tools could be developed in the future for these assessments. MCQs can evaluate disparate and diverse topics and can, therefore, address skills at the bottom of the pyramid.

One limit of a customizable LLM is that it is based on inputs from publicly available and not copyright-restricted content, in this case, Canadian Urological Association and European Association of Urology guidelines; however, testable resources that are expected of Canadian graduates include other material, such as the Campbell's Walsh textbook and American Urological Association guidelines.[5] By restricting the input that a LLM is trained on, we may be compromising the base of the pyramid by selectively evaluating concepts that are only in the public domain. It may also be that organizations such as the CUA and the EAU will develop more restrictive policies about the use of their content for training a LLM in the future. One solution seems to be that publishers could demand a fee for the use of their content in LLM training, but this area of the law remains in flux.[6]

No matter how improved a LLM is, supervision of outputs by an expert will remain needed in the near-to-intermediate term. It remains to be seen whether taking the role of editor of content will save human examiners significant time, but this can be easily tested.

With the introduction of LLM tools in clinical assessments, the future is open to a brave new world (hopefully, not in a dystopian sense) of residency training and evaluation. We are at the dawn of understanding how to rigorously and ethically integrate such tools; however, we are not quite likely at the stage of relinquishing the wheel.

COMPETING INTERESTS: The author does not report any competing personal or financial interests related to this work.

REFERENCES

1. Touma NJ, Caterini J, Liblk K. Is ChatGPT ready for primetime? Performance of artificial intelligence on a simulated Canadian urology board exam. *Can Urol Assoc J* 2024;18:329-32. https://doi.org/10.5489/cuaj.8800
2. Touma NJ, Patel R, Skinner T, et al. Artificial intelligence as a discriminator of competence in urological training: Are we there? *J Urol* 2025;213:504-11. https://doi.org/10.1097/JU.0000000000004357
3. Kim JK, Chua M, Lorenzo A, et al. Use of AI (GPT-4)-generated multiple-choice questions for the examination of surgical subspecialty residents: Report of feasibility and psychometric analysis. *Can Urol Assoc J* 2025;19:182-7. https://doi.org/10.5489/cuaj.9020
4. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63-7. https://doi.org/10.1097/00001888-199009000-00045
5. Skinner TAA, Ho L, Touma NJ. Study habits of Canadian urology residents: Implications for development of a competence by design curriculum. *Can Urol Assoc J* 2017;11:83-7. https://doi.org/10.5489/cuaj.4132
6. Grynbaum MM, Mac R. The Times sues OpenAI and Microsoft over AI use of copyrighted work. December 27. 2023. Available at nytimes.com. Accessed April 17. 2025

CORRESPONDENCE: Dr. Naji J. Touma, Department of Urology, Queen's University, Kingston, ON, Canada; naji.touma@kingstonhsc.ca