Jethro C.C. Kwong[1,2], David-Dan Nguyen[1,3], Adree Khondker[1], Tiange Li[4]

[1]Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada; [2]Temerty Centre for AI Research and Education in Medicine, University of Toronto, Toronto, ON, Canada; [3]Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada; [4]Institute of Medical Science, University of Toronto, Toronto, ON, Canada

COMMENTARY

Cite as: Kwong JCC, Nguyen DD, Khondker A, et al. *Can Urol Assoc J* 2024;18(10):333-4. http://dx.doi.org/10.5489/cuaj.8987

See related article on page 329

Beyond the hype: Unveiling the challenges of large language models in urology

Large language models (LLMs), like ChatGPT, have gained much attention in urology in recent years, with applications ranging from patient-facing chatbots to summarizing clinical notes.[1] In medical education, several groups have explored whether these LLMs can successfully pass urology certification exams, such as the American Urological association self-assessment study program and the European Board of Urology in-service assessment.

Performance of these models has varied widely, with scores ranging from 27–81%.[2-4] In this month's CUAJ, Touma et al evaluated the performance of ChatGPT 4 on the 2022 Queen's Urology Exam Skill Test (QUEST) multiple-choice questions (MCQ), which simulates the Canadian Royal College urology licensing exam.[5] The authors found that, disappointingly, ChatGPT 4 achieved only a 46% accuracy rate, placing it in the sixth percentile, compared to a 63% average score among final-year Canadian residents, with the disparity in performance being most pronounced for oncology-related questions.

The wide range of performance across these studies is likely multifactorial. First, LLMs are heavily dependent on their training data, which grows with each model update; however, there is limited transparency regarding these information sources, and LLMs often lack access to content behind paywalls or restricted memberships. Moreover, outdated training data may contribute to lower performance on rapidly evolving topics like oncology. Similarly, there is no information regarding how LLMs handle questions with conflicting evidence or situations when recommendations differ across guidelines.

Second, although LLMs excel in questions that involve factual recall or basic clinical scenarios, they may struggle with more complex cases that require consideration of patient comorbidities, contraindications, or interpretation of clinical data. Kollitsch and colleagues found that ChatGPT 4 performed worse with increasing question difficulty, with the LLM achieving 100% accuracy on the easiest questions but only 32% on the most challenging cases.[2]

Third, LLMs are prone to generating hallucinations, which are fabricated and often incorrect responses produced without any supporting evidence. Unlike humans, who are capable of reasoning based on provided information, LLMs may generate hallucinations because they focus on predicting the next most likely sequence of words based on user prompts. In healthcare and medical education, these hallucinations are particularly dangerous because they can be sporadic yet delivered with the same level of confidence as factually correct responses. Therefore, careful design of the LLM prompts is vital to minimizing hallucinations and enhancing response accuracy.

This iterative process, known as prompt engineering, can range from simply submitting the question as-is to ChatGPT to employing more advanced techniques, such as chain-of-verification and degree of uncertainty (Figure 1).[6] Chain-of-verification provides a structured approach for the LLM to cross-check its output against trusted data sources, such as clinical practice guidelines. The degree of uncertainty enables the LLM to indicate its confidence level in its response, thereby building greater trust in its outputs.

Last, transparent reporting remains a widespread cncern among artificial intelligence (AI) studies in urology.[7,8] Given the rapid advancements of LLMs, clear documentation of LLM versions, date of analysis, prompts used, and hallucination rates should be encouraged to enhance reproducibility. Recent guidelines, such as the CANGARU checklist and TRIPOD-LLM, offer valuable guidance on reporting key elements specific to LLMs in clinical applications.[9,10]

Much work remains before LLMs are ready for prime time in urology. Their inconsistent performance emphasizes the need for regular updates to training data, careful prompt engineering, and transparent reporting. As the integration of AI into urology continues to grow, adhering to best practices in LLM reporting will be essential for ensuring their effective and responsible use in both clinical practice and medical education.

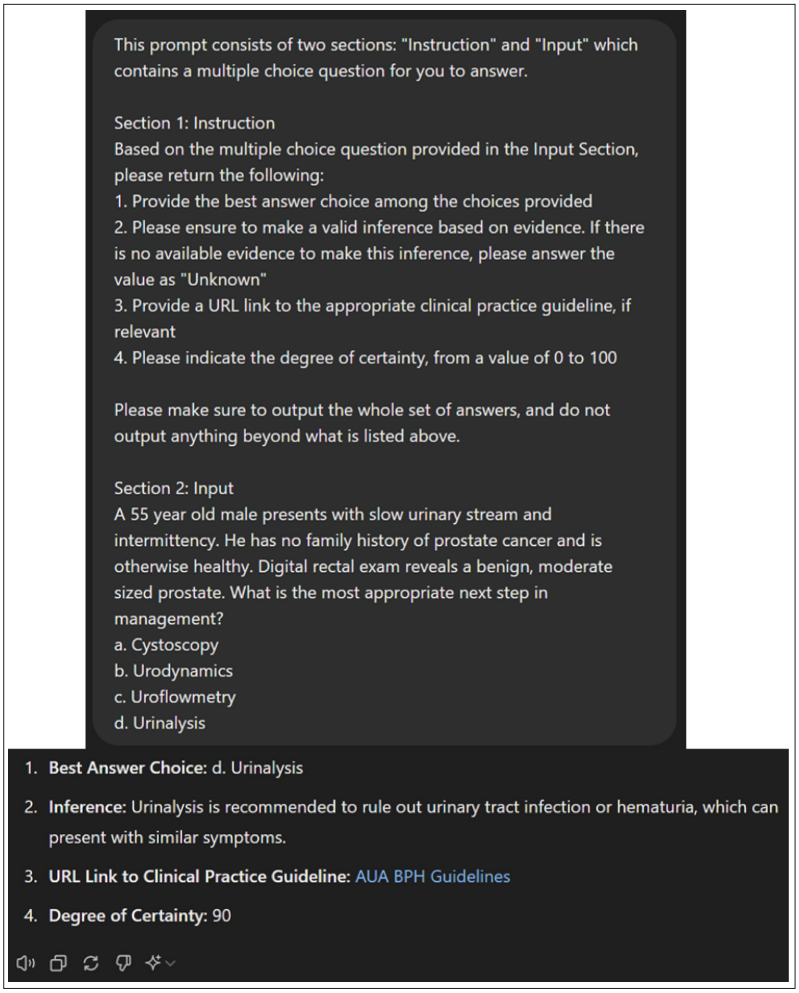

Figure 1. Example of prompt engineering strategies to improve accuracy of LLM responses. Additional strategies to mitigate hallucinations are described by Kwong et al.[6]

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

REFERENCES

1. Eppler M, Ganjavi C, Ramacciotti LS, et al. Awareness and use of ChatGPT and large language models: A prospective cross-sectional global survey in urology. *Eur Urol* 2024;85:146-53 https://doi.org/10.1016/j.eururo.2023.10.014
2. Kollitsch L, Eredics K, Marszalek M, et al. How does artificial intelligence master urological board examinations? A comparative analysis of different large language models' accuracy and reliability in the 2022 in-service assessment of the European Board of Urology. *World J Urol* 2024;42:20. https://doi.org/10.1007/s00 345-023-04749-6
3. May M, Körner-Riffard K, Marszalek M, et al. Would Uro_Chat, a newly developed generative artificial intelligence large language model, have successfully passed the in-service assessment questions of the European Board of Urology in 2022? *Euro Urol Oncol* 2024;7:155-6. https://doi.org/10.1016/j.euo.2023.08.013
4. Deebel NA, Terlecki R. ChatGPT Performance on the American Urological Association self-assessment study program and the potential influence of artificial intelligence in urologic training. *Urology* 2023;177:29-33. https://doi.org/10.1016/j.urology.2023.05.010
5. Touma NJ, Caterini J, Liblk K. Performance of artificial intelligence on a simulated Canadian urology board exam: Is CHATGPT ready for primetime? *Can Urol Assoc J* 2024;18:329-32. https://doi.org/10.5489/cuaj.8800
6. Kwong JCC, Wang SCY, Nickel GC, et al. The long but necessary road to responsible use of large language models in healthcare research. *NPJ Digit Med* 2024;7:177. https://doi.org/10.1038/s41746-024-01180-y
7. Kwong JCC, Wu J, Malik S, et al. Predicting non-muscle invasive bladder cancer outcomes using artificial intelligence: A systematic review using APPRAISE-AI. *NPJ Digit Med* 2024;7:98. https://doi.org/10.1038/s41746-024-01088-7
8. Khondker A, Kwong JCC, Rickard M, et al. Application of STREAM-URO and APPRAISE-AI reporting standards for artificial intelligence studies in pediatric urology: A case example with pediatric hydronephrosis. *J Pediatr Urol* 2024;20:455-67. https://doi.org/10.1016/j.jpurol.2024.01.020
9. Cacciamani GE, Eppler MB, Ganjavi C, et al. Development of the ChatGPT, generative artificial intelligence and natural large language models for accountable reporting and use (CANGARU) guidelines [Internet]. 2023. Available at: http://arxiv.org/abs/2307.08974. Accessed Feb. 4, 2024.
10. Gallifant J, Afshar M, Ameen S, et al. The TRIPOD-LLM statement: A targeted guideline for reporting large language models use [Internet]. medRxiv; 2024 Available at: https://www.medrxiv.org/content/10.1101/2024.07.24.24310930v1. Accessed Aug 26, 2024.

CORRESPONDENCE: Dr. Jethro C.C. Kwong, Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada; jethro.kwong@mail.utoronto.cae