Performance of artificial intelligence on a simulated Canadian urology board exam: Is CHATGPT ready for primetime?

Naji J. Touma, Jessica Caterini, Kiera Liblk
Queen's University, Kingston, ON, Canada

Acknowledgement: *QUEST is supported by an unrestricted educational grant from the CUA*

Cite as: 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 June 10; Epub ahead of print. http://dx.doi.org/10.5489/cuaj.8800

Published online June 10, 2024

Corresponding author: Dr. Naji J. Touma, Queen's University, Kingston, ON, Canada; Naji.Touma@kingstonhsc.ca

***

ABSTRACT

Introduction: Generative artificial intelligence (AI) has proven to be a powerful tool with increasing applications in clinical care and medical education. CHATGPT has performed adequately on many specialty certification and knowledge assessment exams. The objective of this study was to assess the performance of CHATGPT 4 on a multiple-choice exam meant to simulate the Canadian urology board exam.
Methods: Graduating urology residents representing all Canadian training programs gather yearly for a mock exam that simulates their upcoming board-certifying exam. The exam consists of written multiple-choice questions (MCQs) and an oral objective structured clinical examination (OSCE). The 2022 exam was taken by 29 graduating residents and was administered to CHATGPT 4.
Results: CHATGPT 4 scored 46% on the MCQ exam, whereas the mean and median scores of graduating urology residents were 62.6%, and 62.7%, respectively. This would place CHATGPT's score 1.8 standard deviations from the median. The percentile rank of CHATGPT would be in the sixth percentile. CHATGPT scores on different topics of the exam were as follows: oncology 35%, andrology/benign prostatic hyperplasia 62%, physiology/anatomy 67%, incontinence/female urology 23%, infections 71%, urolithiasis 57%, and trauma/reconstruction 17%, with ChatGPT 4's oncology performance being significantly below that of postgraduate year 5 residents.

Conclusions: CHATGPT 4 underperforms on an MCQ exam meant to simulate the Canadian board exam. Ongoing assessments of the capability of generative AI is needed as these models evolve and are trained on additional urology content.

INTRODUCTION

The public release of OpenAI's ChatGPT (Chat Generative Pretrained Transformer) in November 2022 has facilitated layperson access to artificial intelligence (AI) and medical professionals alike. Its adoption surpassed the current record holder, TikTok, reaching 100 million users just two months after its launch. Its popularity stems largely from its user-friendly interface with a question-and-answer format, allowing users with limited technical backgrounds to interact with the transformer.

While a thorough understanding of the technicalities behind ChatGPT is not necessary for its use, some basics of terminology and mechanics may help to clarify the workings of this novel and powerful tool. Natural Language Processing (NLP) is an interdisciplinary research field that aims to develop algorithms for the computational understanding of written and spoken languages. Applications of NLP have become familiar tools in everyday smartphones and other devices, and they include text classification, question answering, speech recognition, language translation, chatbots, and the generation or summarization of texts. (1)

Large Language Models (LLMs) refer to massive Transformer models trained on extensive datasets. The Bidirectional Encoder Representations from Transformers (BERT) [2], a state-of-the-art model at its introduction in 2019, possessed 110 million parameters. Aided by massive computational resources, GPT-3 [3] had already reached 175 billion parameters by 2021. The training of LLMs unfolds in two distinct phases. In the pre-training phase, models are exposed to a massive set of unlabeled text data, learning by predicting subsequent tokens in a given text. Subsequently, the model undergoes refinement by broadening its understanding and application of various facts and concepts acquired during pre-training.

Naturally, there has been a spike in interest in leveraging the capabilities of ChatGPT in healthcare and medicine. A recent review has categorized these applications in various topics, including: medical education, consultation, and research, as well as in various scenarios in the clinical workflow, such as diagnosis, decision-making, and clinical documentation. (4) Medical education, in particular, has proven to be fertile ground for evaluating the capabilities of various LLMs, such as ChatGPT. The fund of knowledge required of medical trainees is ever expanding, and the time needed to study is a limited resource. (5,6) Leveraging technology to assist in training and testing has the potential to enhance knowledge acquisition and skill mastery. (7) An early form of LLMs use has been to explore the capacity of these models to pass standardized exams. The hypothesis is that if LLMs are capable of performing well on such exams, they will prove to be useful tools for learners.

In a urology context, a few reports have explored LLMs' capacity to take urology exams with mixed results. However, no previous report has explored the capacity of ChatGPT on a Canadian exam by comparing it to Canadian test takers, nor has there been any exploration of ChatGPT's performance on various urology subtopics. This study aims to test ChatGPT 4 on a Canadian multiple choice exam (MCQ) and compare it to the performance of graduating residents a few months before their Royal College Exam. We will also explore the performance of ChatGPT 4 on various subtopics of the exam.

METHODS

Graduating urology residents (PGY-5) representing all Canadian training programs gather yearly for a mock exam that simulates their upcoming Royal College certifying exam. The Queen's Urology Exam Skill Test (QUEST) occurs over one weekend about three months before the Royal College Exam. It is meant to simulate the certifying exam, providing practice and feedback to candidates. (8) The exam consists of a written multiple-choice questions (MCQs) exam and an oral Objective Structured Clinical Exam (OSCE). For the 2022 cohort, 29 candidates attended QUEST and sat for both exam components. The MCQ portion of the exam consisted of 100 questions with three distractors and one key and correct answer each, arranged randomly. The exam was administered over the Elentra® software (Kingston, Canada), in person, with two hours allocated for completion. The software also compiled the candidates' results. The same exam was then administered to ChatGPT 4 in August 2023. The prompts provided to ChatGPT were similar to what was provided to the candidates. The model was asked to provide the best answer to the following multiple choice questions. The exam was then copied and pasted in the query box. The software completed the entire exam in less than 1 minute, with the main time-limiting factor being the ability to input the questions into the query box. Descriptive statistics were then used to describe the relative performance of ChatGPT to the cohort of graduating residents. In addition, an analysis of ChatGPT 4's performance on different subtopics of the exam was completed. The chi Square and the fisher's exact tests were used to calculate differences on subtopics of the exam.

RESULTS

The mean and median scores of graduating urology residents on the MCQ exam were 62.6%, and 62.7%, respectively. ChatGPT 4 scored 46% on the same exam, placing it 1.8 standard deviations below the median score. Figure 1 is a histogram illustrating the frequency of each score range for the exam. ChatGPT 4's score is in the lowest range (42%-52%), placing it with three other candidates. The Percentile rank of ChatGPT 4 would be in the 6th percentile.Table 1 compares the scores of ChatGPT 4 to the cohort of candidates on different subtopics of the exam. The scores of ChatGPT 4 were markedly inferior on questions related to oncology.

DISCUSSION

This study demonstrates that ChatGPT 4 performs poorly on a MCQ exam meant to simulate the Royal College exam in urology. This study goes beyond evaluating the rate of correct answers of

ChatGPT 4 on a urology exam. We compare the performance of the latest version of the LLM of OpenAI to a cohort of Canadian urology PGY-5 test takers. When put in this context, the performance of ChatGPT 4 is indeed underwhelming. Its overall score is 1.8 standard deviation below the median score for the cohort, and would place it in the sixth percentile amongst test takers. For the first time, the performance of a LLM is evaluated on a Canadian urological exam, as is an exploration of a LLM's score on different subtopics of a urology exam.

Previous reports have shown that earlier versions of ChatGPT perform poorly on urology exams. ChatGPT 3 was evaluated by submitting the American Urological Association Self-assessment Study Program (AUA-SASP) exams in 2021 and 2022. It performed slightly better on the 2021 exam with a 42.1% rate of correct answers versus 30% on the 2022 exam. (9) Interestingly, a now-retracted report examined the performance of ChatGPT 3.5 on the 2022 AUA-SASP exam and showed a slightly inferior performance with a 26.7% score on the open-ended questions and 28.2% on the multiple-choice questions. (10) Beyond the observation that there was a slight decline in the performance with a newer version of the software on the 2022 AUA-SASP, different answers can be given to the same questions with different queries. It is also important to point out that the report was retracted because the authors did not receive permission to use the copyrighted questions of the AUA-SASP in their study. Conversely, the questions used in this study were designed by the author (NJT), and copyright issues are not germane. However, care must always be exercised when utilizing copyrighted material to query any LLM. The legal limits of this new area of law are currently under consideration in courts, with publishers such as the New York Times trying to shield their copyrighted work from LLMs hoovering up training material. (11)

Another report specifically explored the difference in the performance of various versions of ChatGPT (3.5 Vs 4), and Bing AI. They found that ChatGPT 4 and Bing AI significantly outperformed ChatGPT 3.5 in the rate of correct answers on multiple choice questions of the European Board of Urology In-Service Exam. (12) In addition to the comparison across models, each LLM was queried the questions three times, 48 hours apart. The correct answer rate was again different for each model depending on the iteration, yielding less-than-perfect agreement scores for each LLM. Finally, this report also found that the rate of correct answers declined with increasing question complexity. In our report, we found that the performance of ChatGPT 4 was significantly lower on urological oncology. This is not so much related to complexity as to the possibility that certain topics were more likely to be present in the data on which the LLM was trained. This, however, is speculative as we can't be sure what data the model has been trained on. One limitation here is that the number of questions on different sub topics is limited and more sampling will be necessary in the future to ascertain differences, if any.

Another comparison of various LLM's performance on the European Board of Urology (EBU) In-Service Exam evaluated the ranking of ChaGPT 3.5, ChatGPT 4, and Google Bard to 736 test takers. (13) Final year residents and urologists in an EBU country are eligible to be fellows of the EBU, and can therefore participate in the in-service exam on their website. It

found that ChatGPT 3.5 ranked 570th in this pool of examinees, whereas ChatGPT 4 was 80th, and Google Bard 340th. This placed the performance of ChatGPT 4 in almost the top 10 percent of test takers, contrasting dramatically with our finding that ChatGPT 4 is in the bottom 6th percentile. The variability of results in LLM performance among different exams must be kept in mind and highlights the fact that there remains significant ambiguity on what these models are trained on. LLMs are incapable of reasoning, and their ability to select the correct choice is based on predicting the next words in sequence according to probability. It is, therefore, perfectly possible to arrive at the correct answer based on nonsensical justifications. ChatGPT cannot distinguish between real and fake information fed into it. Consequently, its answers can be misleading, fabricated, or biased, while simultaneously conveyed in an assertive tone. Therefore, its answers should always be verified by human experts before adoption. Artificial hallucination might emerge as a serious concern, especially when the model is trained using large amounts of unsupervised data. This can be resolved by training the system using a diverse and representative data set.

Beyond urology, the performance of ChatGPT has been explored on exams such as the US Medical Licencing Exam (USMLE) (14,15) and an undergraduate exam in parasitology for a Doctor of Veterinary Medicine degree. (16) In those exams, ChatGPT achieved a passing grade or scores that are comparable to test takers. Such exams that usually feature first-order questions that rely on recall are particularly suited to an LLM, whereas exams such as QUEST that mostly feature complex clinical scenarios appear to prove more difficult to ChatGPT's current capabilities.Finally, the wide variability in the performance of ChatGPT and other LLMs on exams highlights the fact that these models are not open-sourced, and there is no transparency around what data sets they are trained on and therefore, what they can handle in terms of topics of conversation. Some have suggested investing in truly open LLMs that perform on par with proprietary products like ChatGPT to address these limitations fully. (17) For healthcare applications, specialized AI models trained on biomedical datasets, such as BioGPT, are always more desirable than ChatGPT. (18) Open AI seems to have recognized this limitation by recently offering the possibility of a customizable LLM that can be trained on a specific dataset. (19) In a urology context, this would likely place entities with propriety over a vast amount of copyrighted urological material such as large established journals in a privileged position to create such models.

Some limitations of this study is that it is limited to 100 MCQs in total. The questions related to some subtopics are even more limited to as few as 6 MCQs, and therefore those results are to be considered a signal. Confirmation with a greater number of questions will be needed to draw meaningful conclusions.

CONCLUSIONS

CHATGPT 4 underperforms on an MCQ exam meant to simulate the Canadian Royal College exam. Its performance compares unfavourably to a cohort of PGY-5 Canadian urology residents about to undergo their licensing exam. It seems to have more trouble with questions related to

oncology. Ongoing assessments of the capability of generative AI is needed as these models evolve and are trained on additional urology content.

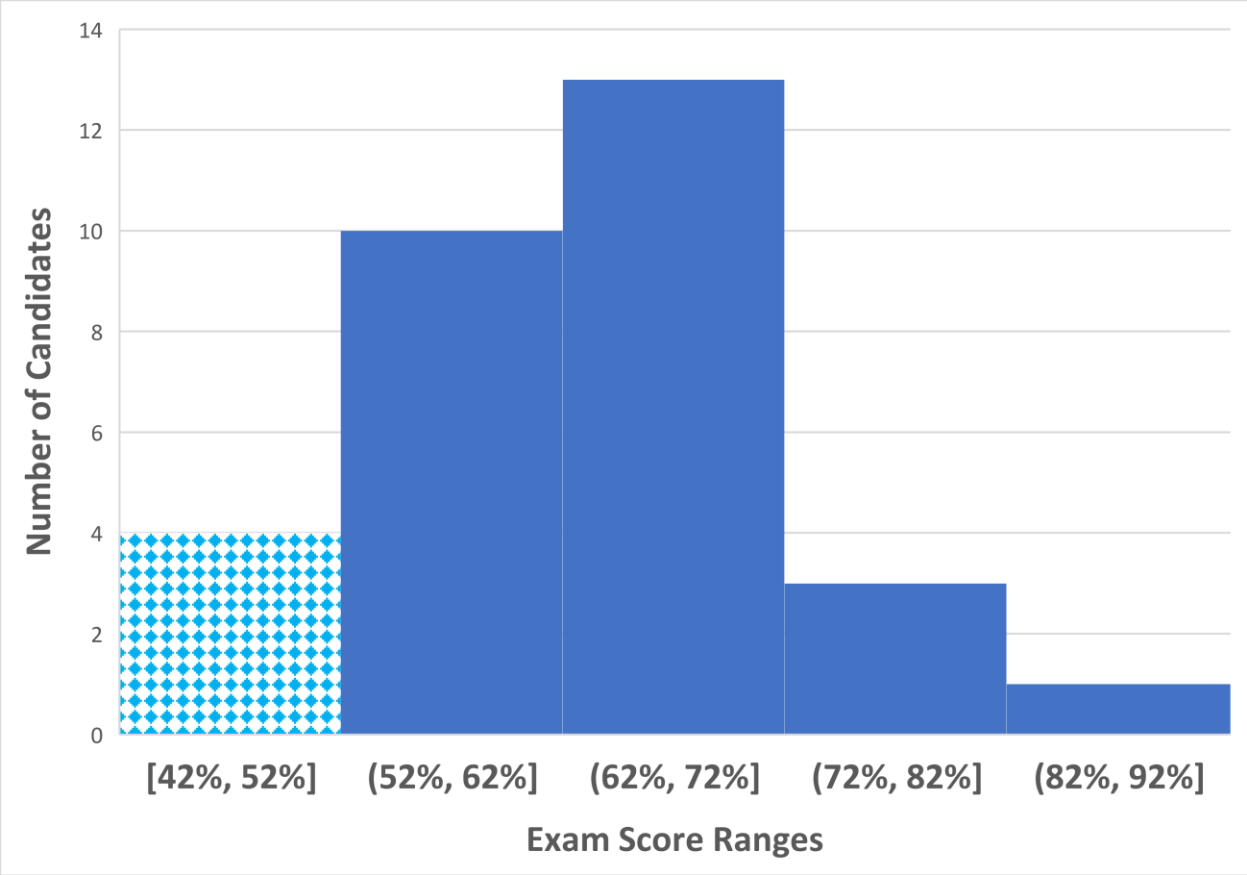
REFERENCES

1. KR Chowdhary. Natural language processing, in: Fundamentals of Artificial Intelligence, 2020, pp. 603-649 https://doi.org/10.1007/978-81-322-3972-7_19
2. Devlin J, et al. Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint, arXiv:1810.04805, 2018.
3. Brown T, et al. Language models are few-shot learners, in: H. Larochelle, et al. (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 1877-1901, https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.
4. Li J, Dada A, Puladi B, et al. ChatGPT in healthcare: A taxonomy and systematic review. *Comput Methods Programs Biomed* 2024;245:108013. https://doi.org/10.1016/j.cmpb.2024.108013
5. Touma NJ, Beiko DT, Macneily AE, et al. Impact of a training program on the performance of graduating Canadian residents on a national urology exam: Results of the last 20 years. *Can Urol Assoc J* 2019;13:39-42. https://doi.org/10.5489/cuaj.5386
6. 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
7. Jenkins D, Nashed JY, Touma NJ. Virtual OSCE examinations during COVID-19 A 360 satisfaction assessment from examiners and candidates. *Can Urol Assoc J* 2023;17:E315-8. https://doi.org/10.5489/cuaj.8332
8. Touma NJ, Leveridge MJ, Beiko D, et al. QUEST at 25: An enduring innovation in Canadian urology. *Can Urol Assoc J* 2022;16:79-80. https://doi.org/10.5489/cuaj.7855
9. 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
10. Huynh LM, Bonebrake BT, Schultis K, et al. New artificial intelligence ChatGPT performs poorly on the 2022 self-assessment study program for urology. *Urol Pract* 2023;10:409-15. https://doi.org/10.1097/UPJ.0000000000000406
11. Grynbaum MM, Mac R. The times sues OpenAI and Microsoft over AI use of copyrighted work. nytimes.com; Dec 27, 2023
12. 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/s00345-023-04749-6

13. Mesnard B, Schirmann A, Branchereau J, et al Artificial intelligence: Ready to pass the European board examinations in urology? *Eur Urol Open Sci* 2024;60:44-6. https://doi.org/10.1016/j.euros.2024.01.002
14. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination (USMLE)? The implications of large language models for medical education and knowledge assessment *JMIR Med Educ* 2023;9:e45312 https://doi.org/10.2196/45312
15. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health* 2023;2:e0000198. https://doi.org/10.1371/journal.pdig.0000198
16. Šlapeta J. Are ChatGPT and other pretrained language models good parasitologists? *Trends Parasitol* 2023;39:314-6. https://doi.org/10.1016/j.pt.2023.02.006
17. Van Dis EA, Bollen J, Zuidema W, et al. ChatGPT: Five priorities for research. *Nature* 2023;614:224-6. https://doi.org/10.1038/d41586-023-00288-7
18. Luo R, Sun L, Xia Y, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Brief Bioinform* 2022;23:bbac409. https://doi.org/10.1093/bib/bbac409
19. Cascella M, Semeraro F, Montomoli J, et al. The breakthrough of large language models release for medical applications: 1-year timeline and perspectives. *J Med Syst* 2024;48:22. https://doi.org/10.1007/s10916-024-02045-3

FIGURES AND TABLES

Figure 1. Distribution of multiple-choice question exam scores of the cohort of candidates and ChatGPT 4.



Gradient fill indicates the range where ChatGPT 4's performance falls.

Table 1. Performance of candidates and ChatGPT 4 on different topics of the exam				
Topic	**Number of questions**	**Mean PGY5 cohort ccores (%)**	**ChatGPT 4 score (%)**	**p**
Oncology	20	71	35	<0.05
Andrology/benign prostatic hyperplasia	13	57	62	NS
Physiology/anatomy	18	61	57	NS
Female urology/incontinence	13	75	23	NS
Pediatrics	16	52	56	NS
Infections	7	52	71	NS
Stone disease	7	63	57	NS
Trauma/reconstruction	6	63	17	NS