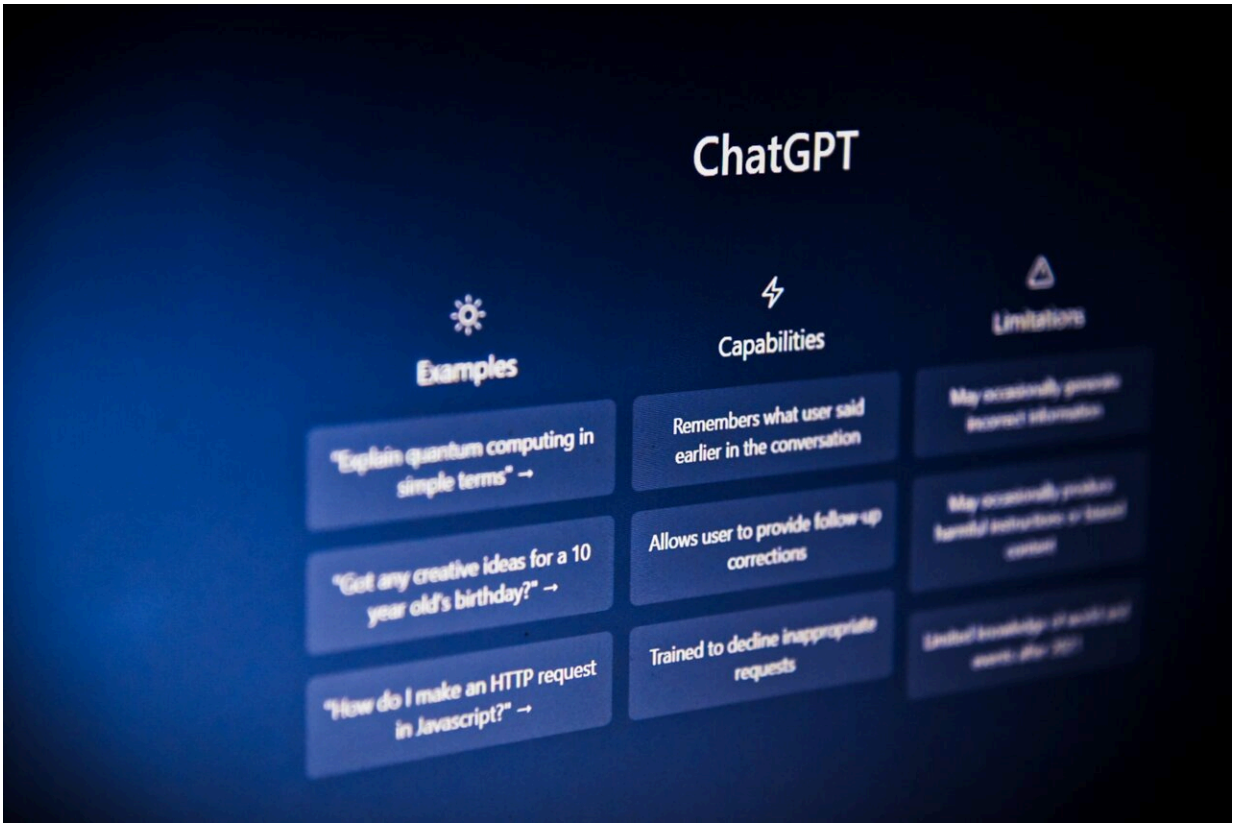
ChatGPT passes radiology board exam

May 16 2023



Credit: Unsplash/CC0 Public Domain

The latest version of ChatGPT passed a radiology board-style exam, highlighting the potential of large language models but also revealing limitations that hinder reliability, according to two new research studies published in *Radiology*.

ChatGPT is an artificial intelligence (AI) chatbot that uses a deep learning model to recognize patterns and relationships between words in its vast training data to generate human-like responses based on a prompt. But since there is no source of truth in its training data, the tool can generate responses that are factually incorrect.

"The use of large language models like ChatGPT is exploding and only going to increase," said lead author Rajesh Bhayana, M.D., FRCPC, an abdominal radiologist and technology lead at University Medical Imaging Toronto, Toronto General Hospital in Toronto, Canada. "Our research provides insight into ChatGPT's performance in a radiology context, highlighting the incredible potential of large language models, along with the current limitations that make it unreliable."

ChatGPT was recently named the fastest growing consumer application in history, and similar chatbots are being incorporated into popular search engines like Google and Bing that physicians and patients use to search for medical information, Dr. Bhayana noted.

To assess its performance on radiology board exam questions and explore strengths and limitations, Dr. Bhayana and colleagues first tested ChatGPT based on GPT-3.5, currently the most commonly used version. The researchers used 150 multiple-choice questions designed to match the style, content and difficulty of the Canadian Royal College and American Board of Radiology exams.

The questions did not include images and were grouped by question type to gain insight into performance: lower-order (knowledge recall, basic understanding) and higher-order (apply, analyze, synthesize) thinking. The higher-order thinking questions were further subclassified by type (description of imaging findings, clinical management, calculation and classification, disease associations).

The performance of ChatGPT was evaluated overall and by question type and topic. Confidence of language in responses was also assessed.

The researchers found that ChatGPT based on GPT-3.5 answered 69% of questions correctly (104 of 150), near the passing grade of 70% used by the Royal College in Canada. The model performed relatively well on questions requiring lower-order thinking (84%, 51 of 61), but struggled with questions involving higher-order thinking (60%, 53 of 89).

More specifically, it struggled with higher-order questions involving description of imaging findings (61%, 28 of 46), calculation and classification (25%, 2 of 8), and application of concepts (30%, 3 of 10). Its poor performance on higher-order thinking questions was not surprising given its lack of radiology-specific pretraining.

GPT-4 was released in March 2023 in limited form to paid users, specifically claiming to have improved advanced reasoning capabilities over GPT-3.5.

In a follow-up study, GPT-4 answered 81% (121 of 150) of the same questions correctly, outperforming GPT-3.5 and exceeding the passing threshold of 70%. GPT-4 performed much better than GPT-3.5 on higher-order thinking questions (81%), more specifically those involving description of imaging findings (85%) and application of concepts (90%).

The findings suggest that GPT-4's claimed improved advanced reasoning capabilities translate to enhanced performance in a radiology context. They also suggest improved contextual understanding of radiology-specific terminology, including imaging descriptions, which is critical to enable future downstream applications.

"Our study demonstrates an impressive improvement in performance of

ChatGPT in radiology over a short time period, highlighting the growing potential of large language models in this context," Dr. Bhayana said.

GPT-4 showed no improvement on lower-order thinking questions (80% vs. 84%) and answered 12 questions incorrectly that GPT-3.5 answered correctly, raising questions related to its reliability for information gathering.

"We were initially surprised by ChatGPT's accurate and confident answers to some challenging radiology questions, but then equally surprised by some very illogical and inaccurate assertions," Dr. Bhayana said. "Of course, given how these models work, the inaccurate responses should not be particularly surprising."

ChatGPT's dangerous tendency to produce inaccurate responses, termed hallucinations, is less frequent in GPT-4 but still limits usability in medical education and practice at present.

Both studies showed that ChatGPT used confident language consistently, even when incorrect. This is particularly dangerous if solely relied on for information, Dr. Bhayana notes, especially for novices who may not recognize confident incorrect responses as inaccurate.

"To me, this is its biggest limitation. At present, ChatGPT is best used to spark ideas, help start the medical writing process and in data summarization. If used for quick information recall, it always needs to be fact-checked," Dr. Bhayana said.

More information: Rajesh Bhayana et al, Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations, *Radiology* (2023). DOI: 10.1148/radiol.230582