ChatGPT flunks self-assessment test for urologists

June 6 2023

Credit: Pixabay/CC0 Public Domain

At a time of growing interest in the potential role of artificial intelligence (AI) technology in medicine and healthcare, a new study reported in *Urology Practice* finds that the groundbreaking ChatGPT

chatbot performs poorly on a major specialty self-assessment tool.

ChatGPT achieved less than a 30% rate of correct answers on the AUA's widely used Self-Assessment Study Program for Urology (SASP). "ChatGPT not only has a low rate of correct answers regarding clinical questions in urologic practice, but also makes certain types of errors that pose a risk of spreading medical misinformation," comment Christopher M. Deibert, MD, MPH, and colleagues of University of Nebraska Medical Center.

Can AI-trained chatbot pass a test of clinical urology knowledge?

Recent advances in large language models (LLMs) provide opportunities for adapting AI technology as a tool for mediating human interaction. "With adequate training and application, these AI systems can process complex information, analyze relationships between ideas, and generate coherent responses to an inquiry," note the authors.

ChatGPT (Chat Generative Pre-Trained Transformer) is an innovative LLM chatbot that has spurred interest in use in a wide range of settings—including health and medicine. In one recent study, ChatGPT scored at or near passing levels on all three steps of the United States Medical Licensing Examination (USMLE), without any special training or feedback on medical topics. Could this innovative AI-trained tool perform similarly well on a more advanced test of clinical knowledge in a surgical specialty?

To find out, Dr. Deibert and colleagues evaluated ChatGPT's performance on the AUA's Self-Assessment Study Program (SASP)—a 150-question practice examination addressing the core curriculum of medical knowledge in urology. The SASP is a valuable test of clinical

knowledge for urologists in training and practicing specialists preparing for Board certification. The study excluded 15 questions containing visual information such as pictures or graphs.

ChatGPT scores low on SASP, with 'redundant and cyclical' explanations

Overall, ChatGPT gave correct answers to less than 30% of SASP questions: 28.2% of multiple-choice questions and 26.7% of open-ended questions. The [chatbot](#) provided "indeterminate" responses to several questions. On these questions, accuracy was decreased when the LLM model was asked to regenerate its answers.

For most open-ended questions, ChatGPT provided an explanation for the selected answer. The explanations provided by ChatGPT were longer than those provided by SASP, but "frequently redundant and cyclical in nature," according to the authors.

"Overall, ChatGPT often gave vague justifications with broad statements and rarely commented on specifics," Dr. Deibert and colleagues write. Even when given feedback, "ChatGPT continuously reiterated the original explanation despite it being inaccurate."

ChatGPT's poor accuracy on the SASP contrasts with its performance on the USMLE and other graduate-level exams. The authors suggest that while ChatGPT may do well on tests requiring recall of facts, it seems to fall short on questions pertaining to clinical medicine, which require "simultaneous weighing of multiple overlapping facts, situations and outcomes."

"Given that LLMs are limited by their human training, further research is needed to understand their limitations and capabilities across multiple

disciplines before it is made available for general use," Dr. Deibert and colleagues conclude. "As is, utilization of ChatGPT in urology has a high likelihood of facilitating medical misinformation for the untrained user."

  More information: Linda My Huynh et al, New Artificial Intelligence ChatGPT Performs Poorly on the 2022 Self-assessment Study Program for Urology, *Urology Practice* (2023). DOI: 10.1097/UPJ.0000000000000406

Provided by Wolters Kluwer Health