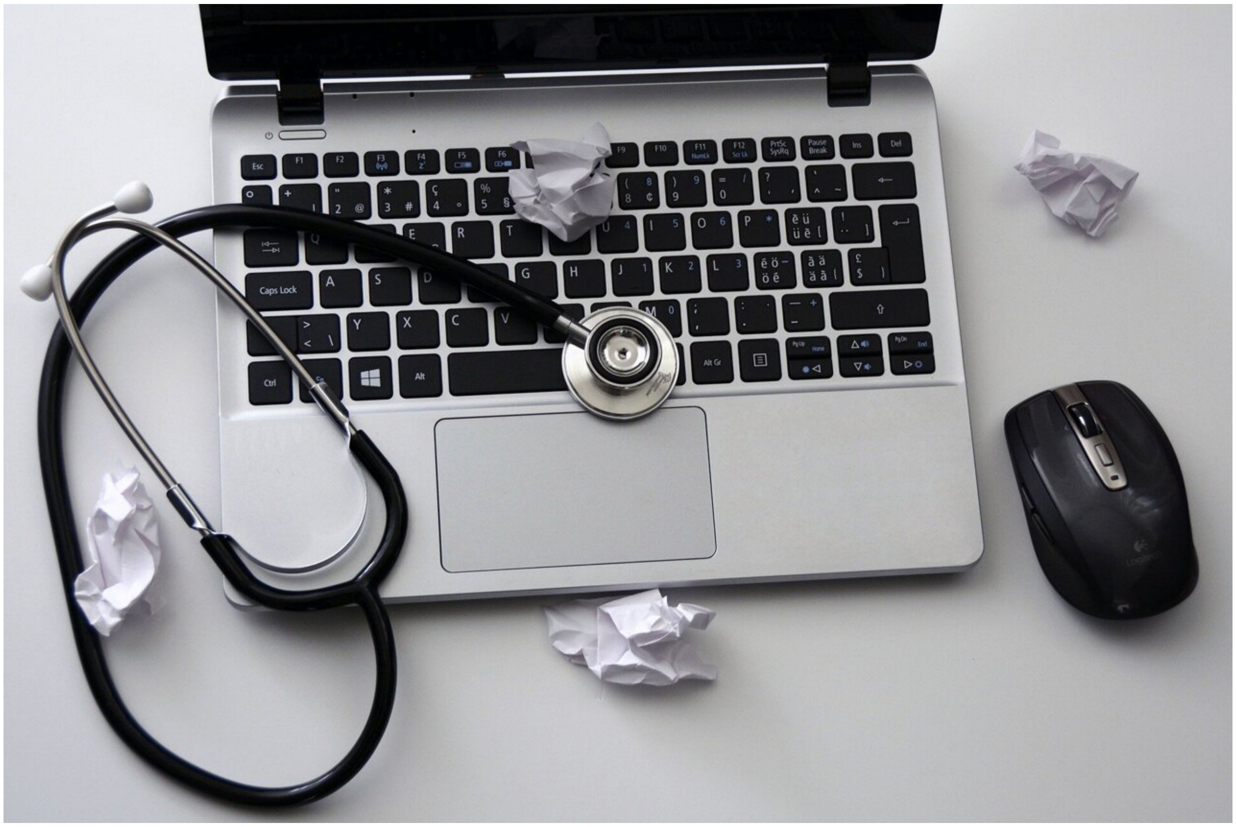
Medical Xpress

ChatGPT outscores med students on complex clinical exam questions

July 19 2023, by Adam Hadhazy



Credit: Pixabay/CC0 Public Domain

ChatGPT can outperform first- and second-year medical students in answering challenging clinical care exam questions, a new study by Stanford researchers has revealed. The findings highlight the accelerating impact of artificial intelligence (AI) on medical education and clinical practice and suggest the need for a new approach to teaching tomorrow's doctors.

ChatGPT is the best known of the large language model AI systems that have captivated the world over the past several months. The systems are trained on the entire corpus of internet content and function as online chatbots, allowing users to input text and then quickly receive automatically generated, human-like text in response.

Recent studies have shown that ChatGPT can successfully handle multiple-choice questions on the United States Medical License Examination (USMLE), which doctors must pass in order to practice medicine. The Stanford authors wanted to explore how the AI system could handle harder, open-ended questions used to assess the clinical reasoning skills of first- and second-year students at Stanford. These questions reveal the details of a patient case in discrete passages separated by questions that ask students to perform clinical reasoning skills, such as coming up with possible diagnoses.

In their newly published article in *JAMA Internal Medicine*, the researchers find that the model on average scored more than four points higher than students on this case-report portion of the exam.

"We were very surprised at how well ChatGPT did on these kinds of free-response medical reasoning questions by exceeding the scores of the human test-takers," says Eric Strong, a hospitalist and clinical associate professor at Stanford School of Medicine and an author of the study.

"With these kinds of results, we're seeing the nature of teaching and testing medical reasoning through written text being upended by new tools," says co-author Alicia DiGiammarino, the Practice of Medicine Year 2 Education manager at the School of Medicine. "ChatGPT and other programs like it are changing how we teach and ultimately practice medicine."

AI is a successful student

The new study used the latest version of ChatGPT, called GPT-4, which was released in March 2023. The study follows up on a prior study that Strong and DiGiammarino led involving the predecessor version, GPT-3.5, which was released by its San Francisco-based maker, OpenAI, in November 2022.

For both studies, the Stanford researchers compiled 14 clinical reasoning cases. The cases, with text descriptions ranging in length from several hundred words to a thousand words, contain myriad extraneous details, such as unrelated chronic medical conditions and medications, just like real-life patient medical charts. During the exam, test-takers must write out paragraph-long answers to a set of questions posed after each case report.

Analyzing the text and composing original answers in this manner contrasts with the comparative simplicity of USMLE multiple-choice test questions. Those questions consist of a short passage, a query, and five possible answers. Nearly all the information provided is relevant to the right answer.

"It's not hugely surprising that ChatGPT and programs like it would do well on multiple-choice questions," says Strong. "Everything test-takers are being told is a central part of the question, so it's mostly information recall. A far harder hill to climb is an open-ended, free-response question."

One small assist that ChatGPT needed, though, before fielding the case-based questions was prompt engineering. Because ChatGPT draws upon the entire internet, it may not correctly construe health care-centric terms used in the test. An example is "problem list," which refers to patients' past and present medical issues but can appear in other non-medical contexts.

After tweaking some questions accordingly, the Stanford researchers input the information into ChatGPT, recorded the chatbot's responses, and passed them on to experienced faculty graders. The AI program's grades were then compared with first- and second-year medical students who had tackled the same cases.

In the prior study, GPT-3.5 was "borderline passing" in its responses, Strong says. In the new study with GPT-4, however, the chatbot scored an average of 4.2 points higher than the students and posted passing grade rates 93 percent of the time versus the students' 85 percent.

For as well as ChatGPT performed, however, it was not flawless. A particularly concerning issue that did significantly lessen with GPT-4 versus 3.5 was confabulation—the adding-in of false details, like a patient having a fever when in fact the patient did not in a particular case study. The confabulatory "false memories" may stem from conflation, where ChatGPT is pulling in information from similar cases.

Rethinking medical education

With regard to test-taking integrity and curricula design, ChatGPT's influence is already being felt at Stanford School of Medicine. This past semester, school administrators decided to switch exams from open book—meaning with internet access to ChatGPT—to closed book. Students must now reason through questions based entirely on memory. While this approach has its merits, the major con, DiGiammarino says, is that the exams no longer assess students' abilities to gather information from sources—a crucial skill in clinical care.

Keenly aware of this issue, School of Medicine faculty and staff have started convening as an AI working group. The group is considering curricula updates that will incorporate AI tools to supplement student learning, all with the goal of pedagogically preparing future clinicians.

"We don't want doctors who were so reliant on AI at school that they failed to learn how to reason through cases on their own," says DiGiammarino. "But I'm more scared of a world where doctors aren't trained to effectively use AI and find it prevalent in modern practice."

"We may be decades away from anything like the wholesale replacing of doctors," adds Strong. "But we're only a few years away from having to incorporate AI into everyday medicine."

More information: Eric Strong et al, Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations, *JAMA Internal Medicine* (2023). DOI: 10.1001/jamainternmed.2023.2909

Provided by Stanford University

Citation: ChatGPT outscores med students on complex clinical exam questions (2023, July 19) retrieved 10 May 2024 from https://medicalxpress.com/news/2023-07-chatgpt-outscores-med-students-complex.html