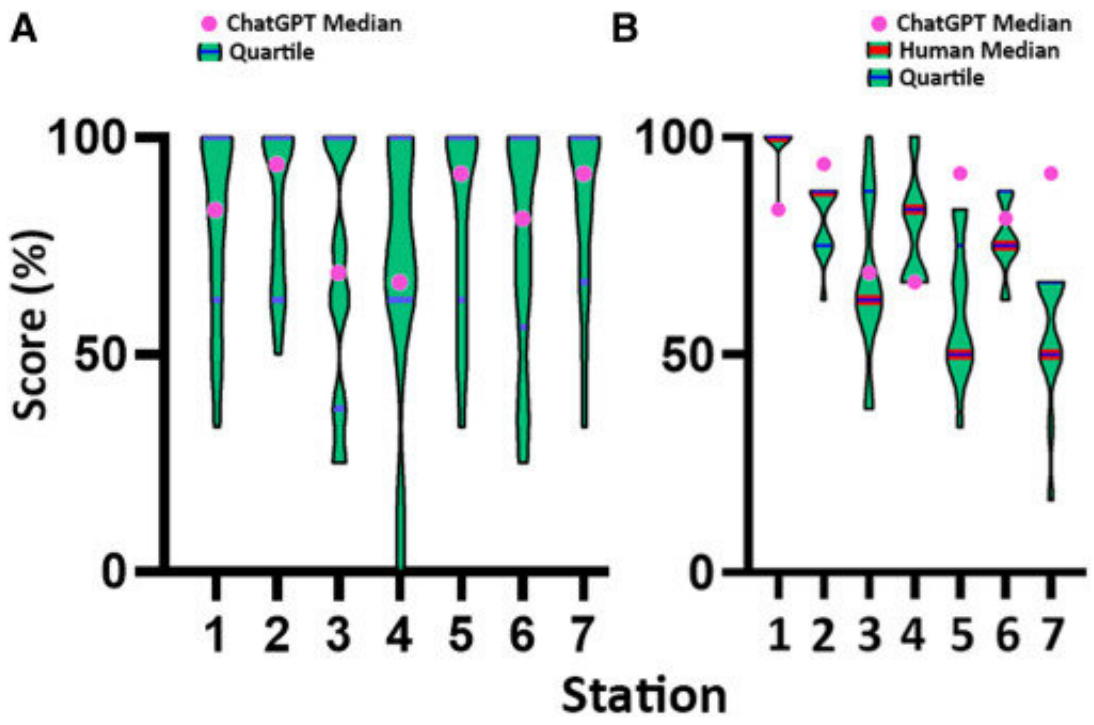
ChatGPT candidate performs well in obstetrics and gynecology clinical examination, compared to human candidates

June 12 2023



Violin plots of ChatGPT and human scores from 7 stations. A, ChatGPT median and quartile values from seven stations. B, The ChatGPT median values (fuchsia dots) overly human scores. Station codes: (1) early pregnancy, (2) postpartum management, (3) urogynecology and pelvic floor problems, (4) core surgical skills, (5) labor management, (6) gynecologic oncology, and (7) postoperative care. Credit: *American Journal of Obstetrics and Gynecology* (2023). DOI: 10.1016/j.ajog.2023.04.020

In a study to determine how the Chat Generative Pre-Trained Transformer or ChatGPT would fare in medical specialist examinations compared to human candidates without additional training, the Artificial Intelligence chatbot performed better than human candidates in a mock Obstetrics and Gynecology (O&G) specialist clinical examination, used to assess the eligibility of individuals to become O&G specialists.

The results from the mock clinical examination detailed that ChatGPT also achieved high scores in empathetic communication, information-gathering and clinical reasoning. The study is published in the *American Journal of Obstetrics and Gynecology*.

The tabulated results showed that ChatGPT attained a higher average score of 77.2%, compared to the human candidates who scored an average of 73.7%. It was also recorded that ChatGPT took an average of two minutes and 54 seconds to complete each station, markedly ahead of the stipulated 10 minutes given. Even though ChatGPT completed the stations in record time, ChatGPT did not outperform all the individuals in each cohort. To minimize bias, the responses of all three candidates were submitted to the examination panel, while concealing the true identity of ChatGPT.

In the study, the team selected seven stations that were in the objective structured clinical examinations (OSCEs) that had been run in the actual mock examinations in the two previous years, all similar in scope and difficulty, with no inclusion of visual interpretations to cater to the current limitations of ChatGPT present at the time of study. Each station has multiple layers of evolving questions based on initial data presented and subsequent responses from the candidate. The OSCE is a criterion-based assessment, where each candidate is assessed on their clinical competencies by completing a series of circuit stations in a simulated environment.

Given 10 minutes to complete each station, the candidate is introduced to an unfamiliar clinical scenario, coupled with the necessary information which would aid them to make an informed clinical decision. The candidate is expected to articulate a care plan, while demonstrating expertise such as communication, information gathering, application of clinical knowledge and patient safety within the time limit. The stations were introduced in an identical format and in the same order to two human candidates—Candidates A and B, and ChatGPT, known as Candidate C.

The study team from the Department of O&G at the Yong Loo Lin School of Medicine (NUS Medicine), led by Associate Professor Mahesh Choolani, Head of the Department of O&G, also conducted an analysis of the answers and found that ChatGPT scored very well in the empathetic communication domain. It was able to skillfully and rapidly generate factually accurate and contextually relevant answers to evolving clinical questions, based on unfamiliar data in the shortest time possible, a feat that would take an average intelligent person more than 10 years of clinical training to be able to understand the questions in this type of highly-complex examinations and answer them appropriately.

It is laudable that Generative AI, which is at present only in its infancy stage, has the prowess to consolidate and interpret huge chunks of general content quickly, and organize it into coherent and concise conversational-type responses, something that would not come naturally to non-native English speakers or candidates facing examination stress. Despite best efforts to blind the examination panel, examiners were generally able to identify the responses from ChatGPT, but not in all cases.

From the mix of answers from human candidates and ChatGPT that were transcribed verbatim and assessed by 14 trained clinician examiners, it was also observed that even though English was used

throughout, there was also an infusion of Singlish or words loaned from Malay, Tamil and Chinese dialects, that was included extensively by human candidates. The intonation and unique vocabulary are very familiar and endearing to Singaporeans or long-term residents in Singapore. This method of communication would very well serve as a bridge to initiate closeness and build trust, while helping to ease nervousness in patients, compared to the more articulately-scripted answers of ChatGPT. The lack of local ethnic knowledge is one of the major limitations of ChatGPT, on top of the lack of up-to-date medical references and data, which in turn causes hallucinations in ChatGPT, compelling it to churn out irrelevant or incorrect answers and conclusions at times.

Crucially, the study results also revealed that ChatGPT is less able to handle subjects that have multiple changes of scenarios, within the question itself, that require open interpretation. The stations with multiple-changing scenarios would require additional training in context-specific medical knowledge in highly-specialized topics. This would be manageable for a highly-trained human candidate who has cultivated higher-level discernment and flexible reasoning needed to tackle ambiguities within these questions. ChatGPT was found to outperform human candidates in several knowledge areas, including labor management, gynecologic oncology and postoperative care, topics or stations that largely focused on standard protocol-driven decision-making, but not in highly contextual situations.

"The arrival and increased use of ChatGPT has proven that it can be a viable resource in guiding medical education, possibly provide adjunct support for clinical care in real time, and even support the monitoring of medical treatment in patients. In an era where accurate knowledge and information is instantly accessible, and these capabilities could be embedded within appropriate context by Generative AI in the foreseeable future, the need for future generations of medical doctors to

clearly demonstrate the value and importance of the human touch is now saliently obvious. As doctors and medical educators, we need to strongly emphasize and exemplify the use of soft skills, compassionate communication and knowledge application in medical training and clinical care," said Associate Professor Mahesh Choolani.

More information: Sarah W. Li et al, ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology, *American Journal of Obstetrics and Gynecology* (2023). DOI: 10.1016/j.ajog.2023.04.020

Provided by National University of Singapore