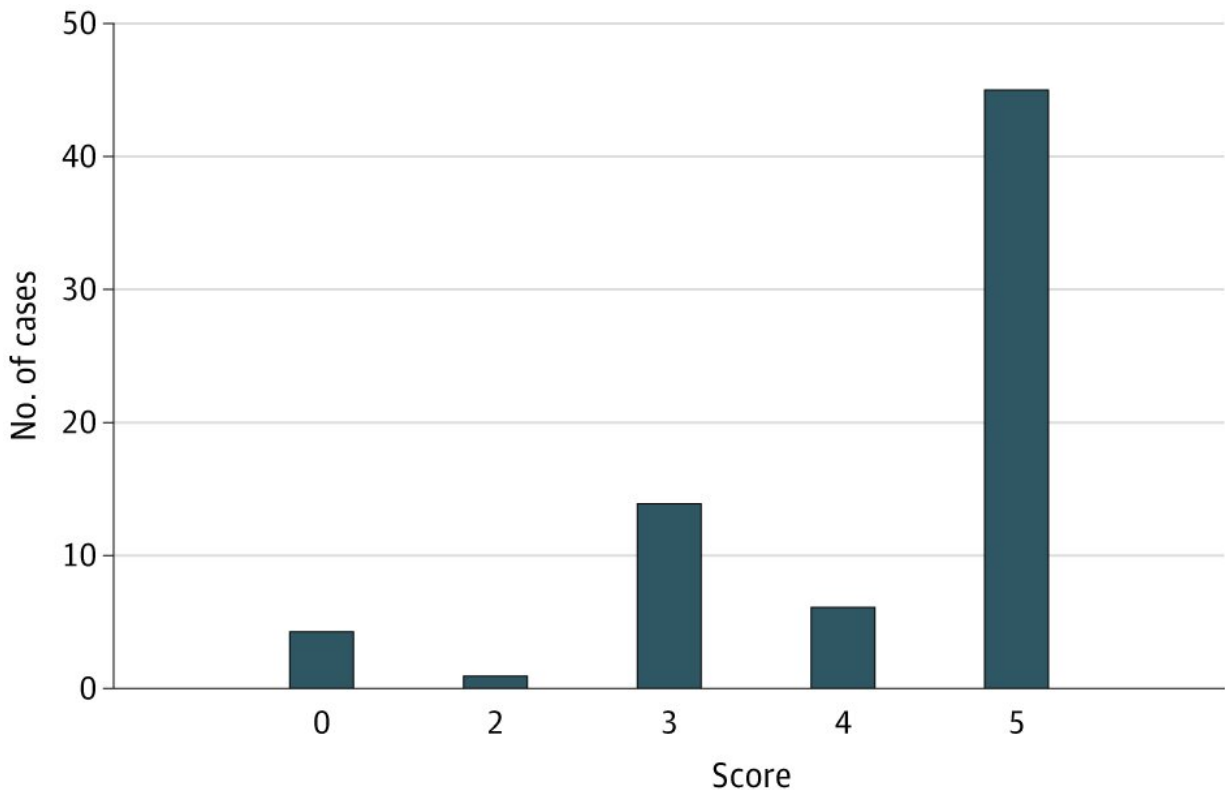# Researchers test AI-powered chatbot's medical diagnostic ability

June 15 2023, by Jacqueline Mitchell

Performance of Generative Pre-trained Transformer 4 (GPT-4) Histogram of GPT-4's performance. Performance scale scores (Bond et al[2]): 5 = the actual diagnosis was suggested in the differential; 4 = the suggestions included something very close, but not exact; 3 = the suggestions included something closely related that might have been helpful; 2 = the suggestions included something related, but unlikely to be helpful; 0 = no suggestions close to the target diagnosis. (The scale does not contain a score of 1.). Credit: *JAMA* (2023). DOI: 10.1001/jama.2023.8288

In a recent experiment published in *JAMA*, physician-researchers at Beth Israel Deaconess Medical Center (BIDMC) tested one well-known publicly available chatbot's ability to make accurate diagnoses in challenging medical cases. The team found that the generative AI, Chat-GPT 4, selected the correct diagnosis as its top diagnosis nearly 40 percent of the time and provided the correct diagnosis in its list of potential diagnoses in two-thirds of challenging cases.

Generative AI refers to a type of artificial intelligence that uses patterns and information it has been trained on to create new content, rather than simply processing and analyzing existing data. Some of the most well-known examples of generative AI are so-called chatbots, which use a branch of artificial intelligence called natural language processing (NLP) that allows computers to understand, interpret and generate human-like language.

Generative AI chatbots are powerful tools poised to revolutionize creative industries, education, customer service and more. However, little is known about their potential performance in the clinical setting, such as complex diagnostic reasoning.

"Recent advances in artificial intelligence have led to generative AI models that are capable of detailed text-based responses that score highly in standardized medical examinations," said Adam Rodman, MD, MPH, co-director of the Innovations in Media and Education Delivery (iMED) Initiative at BIDMC and an instructor in medicine at Harvard Medical School.

"We wanted to know if such a generative model could 'think' like a doctor, so we asked one to solve standardized complex diagnostic cases used for educational purposes. It did really, really well."

To assess the chatbot's diagnostic skills, Rodman and colleagues used

clinicopathological case conferences (CPCs), a series of complex and challenging patient cases including relevant clinical and laboratory data, imaging studies, and histopathological findings published in the *New England Journal of Medicine* for educational purposes.

Evaluating 70 CPC cases, the artificial intelligence exactly matched the final CPC diagnosis in 27 (39 percent) of cases. In 64 percent of the cases, the final CPC diagnosis was included in the AI's differential—a list of possible conditions that could account for a patient's symptoms, medical history, clinical findings and laboratory or imaging results.

"While Chatbots cannot replace the expertise and knowledge of a trained medical professional, generative AI is a promising potential adjunct to human cognition in diagnosis," said first author Zahir Kanjee, MD, MPH, a hospitalist at BIDMC and assistant professor of medicine at Harvard Medical School.

"It has the potential to help physicians make sense of complex medical data and broaden or refine our diagnostic thinking. We need more research on the optimal uses, benefits and limits of this technology, and a lot of privacy issues need sorting out, but these are exciting findings for the future of diagnosis and patient care."

"Our study adds to a growing body of literature demonstrating the promising capabilities of AI technology," said co-author Byron Crowe, MD, an internal medicine physician at BIDMC and an instructor in medicine at Harvard Medical School.

"Further investigation will help us better understand how these new AI models might transform health care delivery."

  More information: Zahir Kanjee et al, Accuracy of a Generative Artificial Intelligence Model in a Complex Diagnostic Challenge, *JAMA*

Provided by Beth Israel Deaconess Medical Center