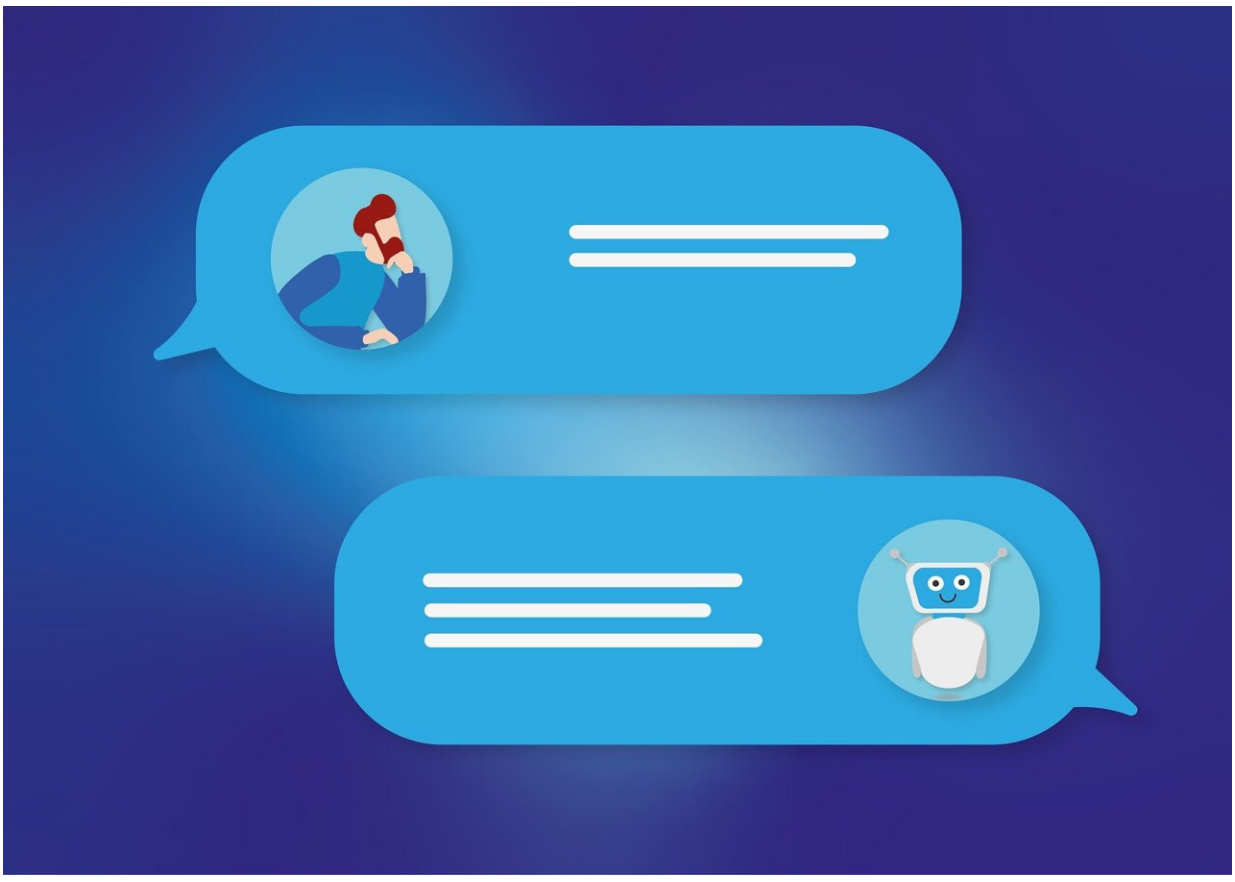
Chatbot outperforms physicians in clinical reasoning, but also underperforms against residents on many occasions

April 1 2024



Credit: Pixabay/CC0 Public Domain

A recent review shows that ChatGPT-4, an artificial intelligence

program designed to understand and generate human-like text, has outperformed internal medicine residents and attending physicians at two academic medical centers at processing medical data and demonstrating clinical reasoning.

In a research letter published in *JAMA Internal Medicine*, physician-scientists at Beth Israel Deaconess Medical Center (BIDMC) compared a large language model's (LLM) reasoning abilities directly against human performance using standards developed to assess physicians.

"It became clear very early on that LLMs can make diagnoses, but anybody who practices medicine knows there's a lot more to medicine than that," said Adam Rodman MD, an internal medicine physician and investigator in the department of medicine at BIDMC.

"There are multiple steps behind a diagnosis, so we wanted to evaluate whether LLMs are as good as physicians at doing that kind of clinical reasoning. It's a surprising finding that these things are capable of showing the equivalent or better reasoning than people throughout the evolution of clinical case."

Rodman and colleagues used a previously validated tool developed to assess physicians' clinical reasoning called the revised-IDEA (r-IDEA) score. The investigators recruited 21 attending physicians and 18 residents who each worked through one of 20 selected clinical cases comprised of four sequential stages of diagnostic reasoning.

The authors instructed physicians to write out and justify their differential diagnoses at each stage. The chatbot GPT-4 was given a prompt with identical instructions and ran all 20 clinical cases. Their answers were then scored for clinical reasoning (r-IDEA score) and several other measures of reasoning.

"The first stage is the triage data, when the patient tells you what's bothering them and you obtain vital signs," said lead author Stephanie Cabral, MD, a third-year internal medicine resident at BIDMC. "The second stage is the system review, when you obtain additional information from the patient. The third stage is the physical exam, and the fourth is diagnostic testing and imaging."

Rodman, Cabral and their colleagues found that the chatbot earned the highest r-IDEA scores, with a median score of 10 out of 10 for the LLM, 9 for attending physicians and 8 for residents. It was more of a draw between the humans and the bot when it came to diagnostic accuracy—how high up the correct diagnosis was on the list of diagnosis they provided—and correct clinical reasoning.

But the bots were also "just plain wrong"—had more instances of incorrect reasoning in their answers—significantly more often than residents, the researchers found. The finding underscores the notion that AI will likely be most useful as a tool to augment but not replace the human reasoning process.

"Further studies are needed to determine how LLMs can best be integrated into clinical practice, but even now, they could be useful as a checkpoint, helping us make sure we don't miss something," Cabral said. "My ultimate hope is that AI will improve the patient-physician interaction by reducing some of the inefficiencies we currently have and allow us to focus more on the conversation we're having with our patients.

"Early studies suggested AI could make diagnoses, if all the information was handed to it," Rodman said. "What our study shows is that AI demonstrates real reasoning—maybe better reasoning than people through multiple steps of the process. We have a unique chance to improve the quality and experience of health care for patients."

Co-authors included Zahir Kanjee, MD, Philip Wilson, MD, and Byron Crowe, MD, of BIDMC; Daniel Restrepo, MD, of Massachusetts General Hospital; and Raja-Elie Abdulnour, MD, of Brigham and Women's Hospital.

More information: Stephanie Cabral et al, Clinical Reasoning of a Generative Artificial Intelligence Model Compared With Physicians, *JAMA Internal Medicine* (2024). DOI: 10.1001/jamainternmed.2024.0295

Provided by Beth Israel Deaconess Medical Center

Citation: Chatbot outperforms physicians in clinical reasoning, but also underperforms against residents on many occasions (2024, April 1) retrieved 20 May 2024 from https://medicalxpress.com/news/2024-04-chatbot-outperforms-physicians-clinical-underperforms.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.