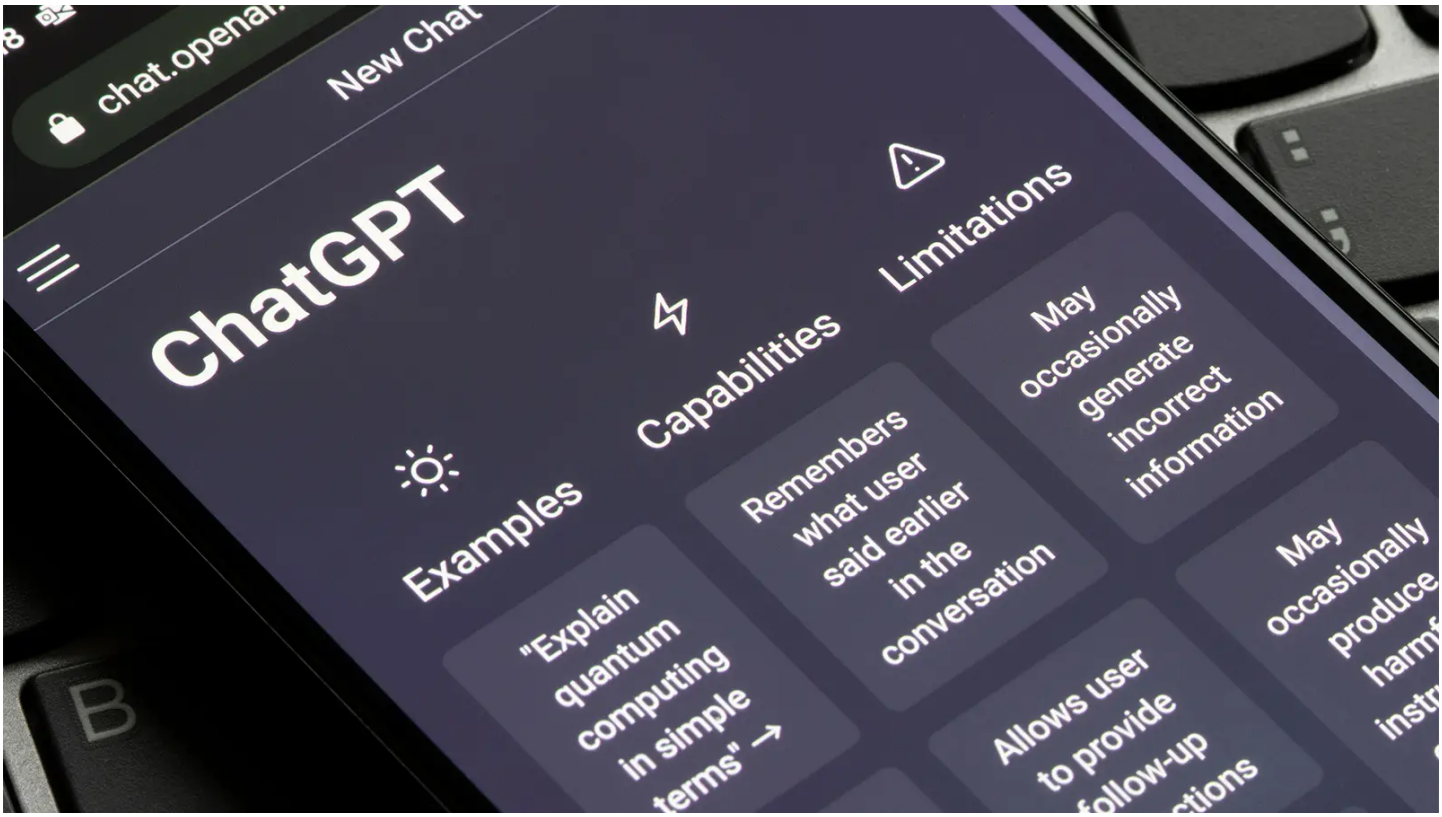
AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by Michael DePeau-Wilson, Enterprise & Investigative Writer, MedPage Today
January 19, 2023



Two artificial intelligence (AI) programs -- including ChatGPT -- have passed the U.S. Medical Licensing Examination (USMLE), according to two recent papers.

The papers highlighted different approaches to using large language models to take the USMLE, which is comprised of three exams: Step 1, Step 2 CK, and Step 3.

ChatGPT is an artificial intelligence (AI) search tool that mimics long-form writing based on prompts from human users. It was developed by OpenAI, and became popular after several social media posts showed potential uses for the tool in clinical practice, often with mixed results.

The first paper, published on *medRxiv* in December, investigated ChatGPT's performance on the USMLE without any special training or reinforcement prior to the exams. According to Victor Tseng, MD, of Ansible Health in Mountain View, California, and colleagues, the results showed "new and surprising evidence" that this AI tool was up to the challenge.

Tseng and team noted that ChatGPT was able to perform at >50% accuracy across all of the exams, and even achieved 60% in most of their analyses. While the USMLE passing threshold does vary between years, the authors said that passing is approximately 60% most years.

"ChatGPT performed at or near the passing threshold for all three exams without any specialized training or reinforcement," they wrote, noting that the tool was able to demonstrate "a high level of concordance and insight in its explanations."

"These results suggest that large language models may have the potential to assist with medical education, and potentially, clinical decision-making," they concluded.

The second paper, published on *arXiv*, also in December, evaluated the performance of another large language model, Flan-PaLM, on the USMLE. The key difference between the two models was that this model was heavily modified to prepare for the exams, using a collection of medical question-answering databases called the MultiMedQA, explained Vivek Natarajan, an AI researcher, and colleagues.

Flan-PaLM achieved 67.6% accuracy in answering the USMLE questions, which was about 17 percentage points higher than the previous best performance conducted using PubMed GPT.

Natarajan and team concluded that large language models "present a significant opportunity to rethink the development of medical AI and make it easier, safer and more equitable to use."

ChatGPT, along with other AI programs, have been showing up as the subject -- and sometimes as the co-author -- of new research papers focused on testing the technology's usefulness in medicine.

DNA identifies remains of Florida teen who disappeared in 1972. Detectives now say she may have been killed by a serial killer cop.

Of course, healthcare professionals have also expressed concerns over these developments, especially when ChatGPT is being listed as an author on research papers. A recent article from *Nature* highlighted the uneasiness from would-be colleagues and co-authors of the emerging technology.

One objection to the use of AI programs in research was based on whether they can be truly capable of making meaningful scholarly contributions to a paper, while another objection emphasized that AI tools can't consent to be a co-author in the first place.

The editor of one of the papers that listed ChatGPT as an author said it was an error that would be corrected, according to the *Nature* article. Still, researchers have published several papers now touting these AI programs as useful tools in medical education, research, and even clinical decision making.

Natarajan and colleagues concluded in their paper that large language models could become a beneficial tool in medicine, but their first hope was that their findings would "spark further conversations and collaborations between patients, consumers, AI researchers, clinicians, social scientists, ethicists, policymakers and other interested people in order to responsibly translate these early research findings to improve healthcare."

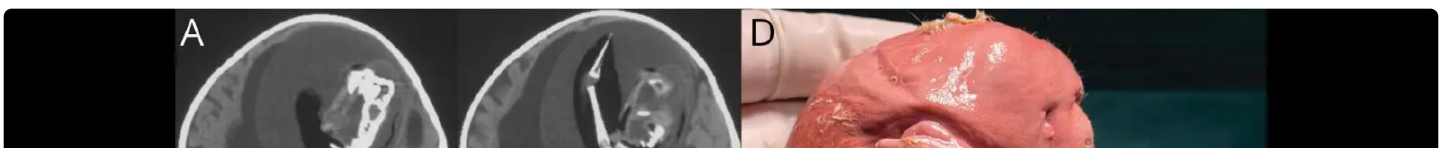
Michael DePeau-Wilson is a reporter on MedPage Today's enterprise & investigative team. He covers psychiatry, long covid, and infectious diseases, among other relevant U.S. clinical news. Follow

Primary Source
medRxiv
Source Reference: *Kung TH, et al "Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models" medRxiv 2022; DOI: 10.1101/2022.12.19.22283643.*

Secondary Source
arXiv
Source Reference: *Singhal K, et al "Large language models encode clinical knowledge" arXiv 2022; DOI: 10.48550/arXiv.2212.13138.*

53 Comments

Recommended For You



Oncology/Hematology

Jimmy Carter's Melanoma; Post-Op Osimertinib Boosts OS; $83K for Unproven Therapy

---

Public Health & Policy

Celeb Warns on Butt Shots; 27 Drugs Face Rebate Penalty; 'Shocking Amount of Misery'

Second Opinions

How Overturning Roe v. Wade Changed Match Day 2023

Oncology/Hematology

Shortage of Prostate Cancer Drug; Win for DBT Mammo; Double Lung Transplant

Infectious Disease

Evidence COVID Began in Raccoon Dog; Migraine Drug Recall; FDA Mum on Spring Booster

Medical News From Around the Web

KAISER HEALTH NEWS

Black Women Have Much at Stake in States Where Abortion Access May Vanish

BBC

Brain cancer research plea made by family of Ollie Gardiner

CBS NEWS

DNA identifies remains of Florida teen who disappeared in 1972. Detectives now say she may have been killed by a serial killer cop.

THE WASHINGTON POST

How my dryer door gave me a concussion

JOURNAL OF ADDICTIVE DISEASES

Pain, smoking, and moderating effect of gender in a large, representative sample of Danish adults.

THE JOURNAL OF ORTHOPAEDIC AND SPORTS PHYSICAL THERAPY

All MCIDs Are Wrong, But Some May be Useful.