AI beats doctors in accurately assessing eye problems

April 17 2024

Thirunavukarasu found that the AI model GPT-4 significantly exceeds the ability of non-specialist doctors to assess eye problems and provide advice. Credit: University of Cambridge

The clinical knowledge and reasoning skills of GPT-4 are approaching the level of specialist eye doctors, a study led by the University of

Cambridge has found.

GPT-4—a [large language model](link)—was tested against doctors at different stages in their careers, including unspecialized junior doctors, and trainee and expert eye doctors. Each was presented with a series of 87 patient scenarios involving a specific eye problem, and asked to give a diagnosis or advise on treatment by selecting from four options.

GPT-4 scored significantly better in the test than unspecialized junior doctors, who are comparable to [general practitioners](link) in their level of specialist eye knowledge.

GPT-4 gained similar scores to trainee and expert eye doctors—although the top performing doctors scored higher.

The researchers say that large language models aren't likely to replace health care professionals, but have the potential to improve health care as part of the clinical workflow.

They say state-of-the-art large language models like GPT-4 could be useful for providing eye-related advice, diagnosis, and management suggestions in well-controlled contexts, like triaging patients, or where access to specialist [health care professionals](link) is limited.

"We could realistically deploy AI in triaging patients with eye issues to decide which cases are emergencies that need to be seen by a specialist immediately, which can be seen by a GP, and which don't need treatment," said Dr. Arun Thirunavukarasu, lead author of the study, which he carried out while a student at the University of Cambridge's School of Clinical Medicine

He added, "The models could follow clear algorithms already in use, and we've found that GPT-4 is as good as expert clinicians at processing eye

symptoms and signs to answer more complicated questions.

"With further development, large language models could also advise GPs who are struggling to get prompt advice from eye doctors. People in the UK are waiting longer than ever for eye care."

Large volumes of clinical text are needed to help fine-tune and develop these models, and work is ongoing around the world to facilitate this.

Thirunavukarasu says AI could help in triaging patients with eye issues, to decide which cases are emergencies that need to be seen by a specialist immediately, which can be seen by a GP, and which don't need treatment. Credit: University of Cambridge

The researchers say that their study is superior to similar, previous studies because they compared the abilities of AI to practicing doctors, rather than to sets of examination results.

"Doctors aren't revising for exams for their whole career. We wanted to see how AI fared when pitted against the on-the-spot knowledge and abilities of practicing doctors, to provide a fair comparison," said Thirunavukarasu, who is now an Academic Foundation Doctor at Oxford University Hospitals NHS Foundation Trust.

He added, "We also need to characterize the capabilities and limitations of commercially available models, as patients may already be using them—rather than the internet—for advice."

The test included questions about a huge range of eye problems, including extreme light sensitivity, decreased vision, lesions, itchy and painful eyes, taken from a textbook used to test trainee eye doctors. This textbook is not freely available on the internet, making it unlikely that its content was included in GPT-4's training datasets.

The results are [published today in the journal *PLOS Digital Health*](link).

"Even taking the future use of AI into account, I think doctors will continue to be in charge of patient care. The most important thing is to empower patients to decide whether they want computer systems to be involved or not. That will be an individual decision for each patient to make," said Thirunavukarasu.

GPT-4 and GPT-3.5—or Generative Pre-trained Transformers—are trained on datasets containing hundreds of billions of words from articles, books, and other internet sources. These are two examples of large language models; others in wide use include Pathways Language Model 2 (PaLM 2) and Large Language Model Meta AI 2 (LLaMA 2).

The study also tested GPT-3.5, PaLM2, and LLaMA with the same set of questions. GPT-4 gave more accurate responses than all of them.

GPT-4 powers the online chatbot ChatGPT to provide bespoke responses to human queries. In recent months, ChatGPT has attracted significant attention in medicine for attaining passing level performance in medical school examinations, and providing more accurate and empathetic messages than human doctors in response to patient queries.

The field of artificially intelligent large language models is moving very rapidly. Since the study was conducted, more advanced models have been released—which may be even closer to the level of expert eye doctors.

More information: Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study, *PLOS Digital Health* (2024). [DOI: 10.1371/journal.pdig.0000341](#)

Provided by University of Cambridge