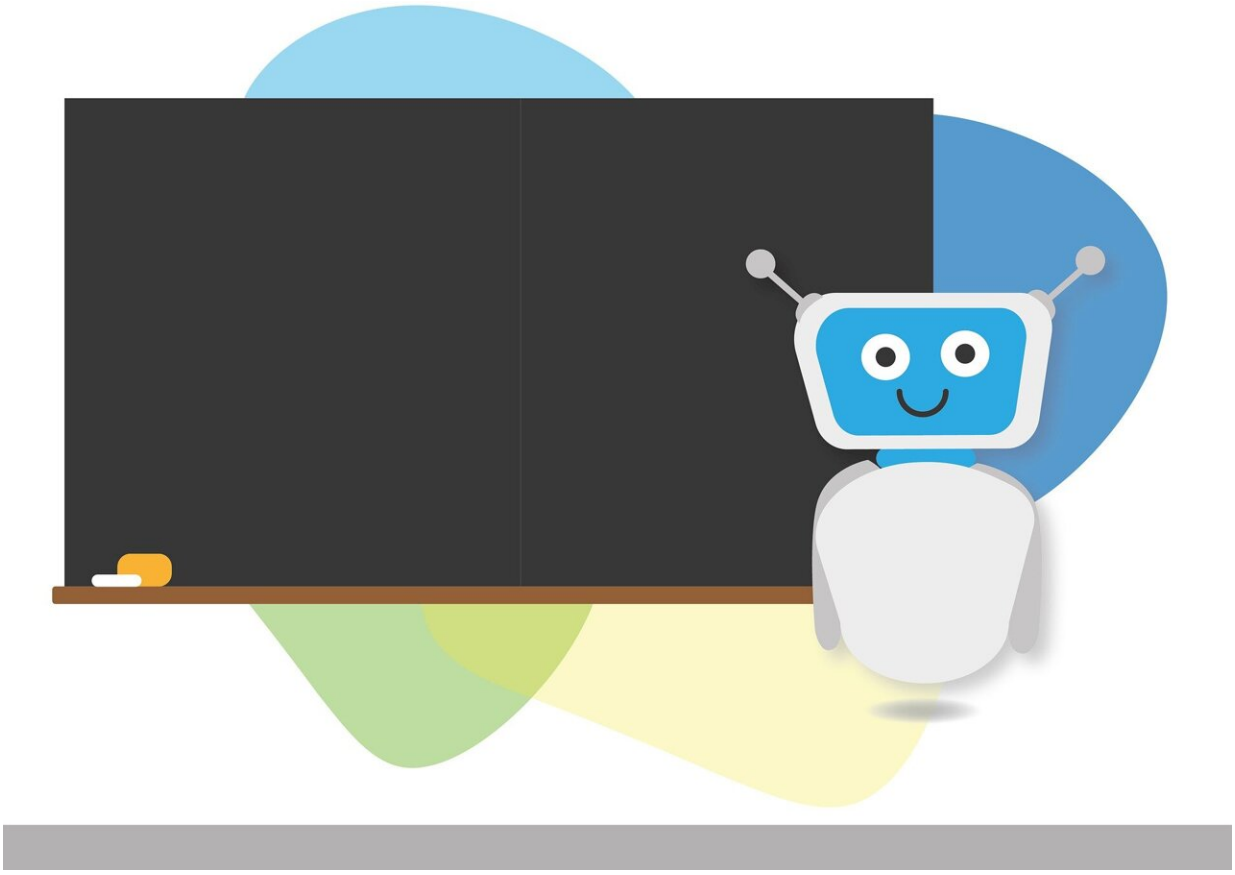
Medical Xpress

ChatGPT outperforms trainee doctors in assessing complex respiratory illness in children

September 8 2024



Credit: Pixabay/CC0 Public Domain

The chatbot ChatGPT performed better than trainee doctors in assessing

complex cases of respiratory disease in areas such as cystic fibrosis, asthma and chest infections in a study presented at the European Respiratory Society (ERS) Congress in Vienna, Austria.

The study also showed that Google's chatbot Bard performed better than trainees in some aspects and Microsoft's Bing chatbot performed as well as trainees.

The research suggests that these large language models (LLMs) could be used to support trainee doctors, nurses and general practitioners to triage patients more quickly and ease pressure on health services.

The study was presented by Dr. Manjith Narayanan, a consultant in pediatric pulmonology at the Royal Hospital for Children and Young People, Edinburgh and honorary senior clinical lecturer at the University of Edinburgh, UK. He said, "Large language models like ChatGPT have come into prominence in the last year and a half with their ability to seemingly understand natural language and provide responses that can adequately simulate a human-like conversation. These tools have several potential applications in medicine. My motivation to carry out this research was to assess how well LLMs are able to assist clinicians in real life."

To investigate this, Dr. Narayanan used clinical scenarios that occur frequently in pediatric respiratory medicine. The scenarios were provided by six other experts in pediatric respiratory medicine and covered topics like cystic fibrosis, asthma, sleep disordered breathing, breathlessness and chest infections. They were all scenarios where there is no obvious diagnosis, and where there is no published evidence, guidelines or expert consensus that points to a specific diagnosis or plan.

Ten trainee doctors with less than four months of clinical experience in pediatrics were given an hour where they could use the internet, but not

any chatbots, to solve each scenario with a descriptive answer of 200 to 400 words. Each scenario was also presented to the three chatbots.

All the responses were scored by six pediatric respiratory experts for correctness, comprehensiveness, usefulness, plausibility, and coherence. They were also asked to say whether they thought each response was human- or chatbot-generated and to give each response an overall score out of nine.

Solutions provided by ChatGPT version 3.5 scored an average of seven out of nine overall and were believed to be more human-like than responses from the other chatbots. Bard scored an average of six out of nine and was scored as more "coherent" than trainee doctors, but in other respects was no better or worse than trainee doctors. Bing scored an average of four out of nine—the same as trainee doctors overall. Experts reliably identified Bing and Bard responses as non-human.

Dr. Narayanan said, "Our study is the first, to our knowledge, to test LLMs against trainee doctors in situations that reflect real-life clinical practice. We did this by allowing the trainee doctors to have full access to resources available on the internet, as they would in real life. This moves the focus away from testing memory, where there is a clear advantage for LLMs. Therefore, this study shows us another way we could be using LLMs and how close we are to regular day-to-day clinical application.

"We have not directly tested how LLMs would work in patient-facing roles. However, it could be used by triage nurses, trainee doctors and primary care physicians, who are often the first to review a patient."

The researchers did not find any obvious instances of "hallucinations" (seemingly made-up information) with any of the three LLMs.

"Even though in our study we did not see any instance of hallucination by LLMs, we need to be aware of this possibility and build mitigations against this," Dr. Narayanan added. Answers that were judged to be irrelevant to the context were occasionally given by Bing, Bard and the trainee doctors.

Dr. Narayanan and his colleagues are now planning to test chatbots against more senior doctors and to look at newer and more advanced LLMs.

Hilary Pinnock is ERS Education Council Chair and Professor of Primary Care Respiratory Medicine at The University of Edinburgh, UK, and was not involved in the research. She says, "This is a fascinating study. It is encouraging, but maybe also a bit scary, to see how a widely available AI tool like ChatGPT can provide solutions to complex cases of respiratory illness in children. It certainly points the way to a brave new world of AI-supported care.

"However, as the researchers point out, before we start to use AI in routine clinical practice, we need to be confident that it will not create errors either through 'hallucinating' fake information or because it has been trained on data that does not equitably represent the population we serve. As the researchers have demonstrated, AI holds out the promise of a new way of working, but we need extensive testing of clinical accuracy and safety, pragmatic assessment of organizational efficiency, and exploration of the societal implications before we embed this technology in routine care."

More information: Abstract no: OA2762 "Clinical scenarios in paediatric pulmonology: Can large language models fare better than trainee doctors?", by Manjith Narayanan et al; Presented in session, "Respiratory care in the digital age: innovative applications and their evidence" at 09:30-10:45 CEST on Monday 9 September 2024.

[k4.ersnet.org/prod/v2/Front/Pr … ?e=549&session=17916]

Provided by European Respiratory Society