Trust your doctor: Study shows human medical professionals are more reliable than artificial intelligence tools
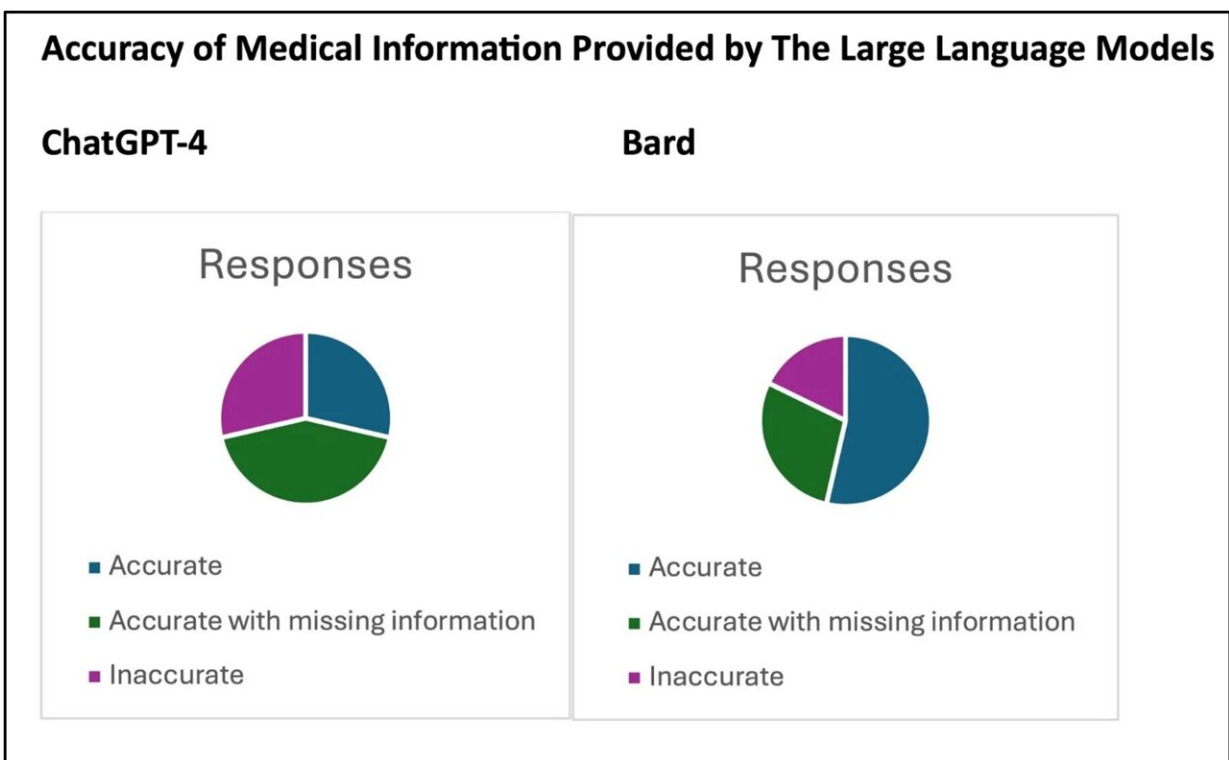
April 2 2024

Accuracy of Medical Information Provided by The Large Language Models

Artificial intelligence (AI) tools are not reliable enough to be substituted for medical professionals in providing accurate medical information, research in the *American Journal of Preventive Medicine* shows. Final assessments found 28.6% of ChatGPT-4's answers accurate, 28.6% inaccurate, and 42.8% partially

accurate but incomplete. Bard performed better, with 53.6% of answers accurate, 17.8% inaccurate, and 28.6% partially accurate. *American Journal of Preventive Medicine* (2024). DOI: 10.1016/j.amepre.2024.02.006

When looking for medical information, people can use web search engines or large language models (LLMs) like ChatGPT-4 or Google Bard. However, these artificial intelligence (AI) tools have their limitations and can sometimes generate incorrect advice or instructions. A new study in the *American Journal of Preventive Medicine* assesses the accuracy and reliability of AI-generated advice against established medical standards and finds that LLMs are not trustworthy enough to replace human medical professionals just yet.

Andrei Brateanu, MD, Department of Internal Medicine, Cleveland Clinic Foundation, says, "Web search engines can provide access to reputable sources of information, offering accurate details on a variety of topics such as preventive measures and general medical questions.

"Similarly, LLMs can offer medical information that may look very accurate and convincing when, in fact, it may be occasionally inaccurate. Therefore, we thought it would be important to compare the answers from LLMs with data obtained from recognized medical organizations. This comparison helps validate the reliability of the medical information by cross-referencing it with trusted health care data."

In the study, 56 questions were posed to ChatGPT-4 and Bard, and their responses were evaluated by two physicians for accuracy, with a third resolving any disagreements. Final assessments found 28.6% of ChatGPT-4's answers accurate, 28.6% inaccurate, and 42.8% partially

accurate but incomplete. Bard performed better, with 53.6% of answers accurate, 17.8% inaccurate, and 28.6% partially accurate.

Dr. Brateanu explains, "All LLMs, including ChatGPT-4 and Bard, operate using complex mathematical algorithms. The fact that both models produced responses with inaccuracies or omitted crucial information highlights the ongoing challenge of developing AI tools that can provide dependable medical advice. This might come as a surprise, considering the advanced technology behind these models and their anticipated role in health care environments."

This research underscores the importance of being cautious and critical of medical information obtained from AI sources, reinforcing the need to consult health care professionals for accurate medical advice. For health care professionals, it points to the potential and limitations of using AI as a supplementary tool in providing patient care and emphasizes the ongoing need for oversight and verification of AI-generated information.

Dr. Brateanu concludes, "AI tools should not be seen as substitutes for medical professionals. Instead, they can be considered as additional resources that, when combined with human expertise, can enhance the overall quality of information provided. As we incorporate AI technology into health care, it's crucial to ensure that the essence of health care continues to be fundamentally human."