# Study reveals AI struggles with real-world clinical interactions



A recent study conducted by researchers from Harvard Medical School and Stanford University has explored the efficacy of artificial intelligence (AI) tools, particularly large-language models, in real-world clinical environments. The analysis, published in Nature Medicine on January 2, evaluated various AI models designed to assist in patient interactions, including triaging patients, collecting medical histories, and providing preliminary diagnoses.

These large-language models, such as ChatGPT, are already being employed by patients to interpret their symptoms and understand test results. However, the findings from this study indicate that while these AI technologies excel in controlled settings—such as standardized medical examinations—they struggle significantly in more dynamic, real-world scenarios that require nuanced conversation.

The research team developed the Conversational Reasoning Assessment Framework for Testing in Medicine, commonly referred to as CRAFT-MD. This evaluation framework aimed to simulate authentic patient-doctor interactions, where AI models were tested on their ability to collect comprehensive information regarding symptoms, medications, and family histories before arriving at a diagnosis.

The study revealed a notable discrepancy in performance: all four AI models tested demonstrated proficiency with exam-style questions, typical of what medical students might encounter. However, their efficacy drastically diminished in conversational contexts that better reflect actual clinical encounters. Pranav Rajpurkar, the study’s senior author and an assistant professor of biomedical informatics at Harvard Medical School, noted, "Our work reveals a striking paradox—while these AI models excel at medical board exams, they struggle with the basic back-and-forth of a doctor’s visit."

The research highlighted how these AI systems often faltered in essential conversational aspects, including gathering pertinent patient information, asking critical follow-up questions, and integrating scattered data into coherent diagnoses. These elements are vital in real-world medical practice, where interactions are rarely straightforward. The models’ performance declined sharply when faced with open-ended inquiries or engaged in multiple exchanges, scenarios common in clinical settings.

The authors of the study have proposed several recommendations for both developers and regulators of AI systems in healthcare. They advocate for the design of AI models that utilise open-ended, conversational questioning to better replicate disordered doctor-patient interactions. Additionally, they suggest the need for AI frameworks capable of synthesising both textual and non-textual patient data, such as images and EKG readings, and interpreting non-verbal cues like facial expressions and tone of voice during interactions.

The CRAFT-MD framework not only serves to enhance the assessment of AI models but can also accelerate the development process by processing thousands of conversations in a fraction of the time it would take human evaluators. This efficiency could significantly reduce the costs associated with testing AI tools in healthcare while mitigating the risks associated with deploying unverified systems in actual patient care.

Roxana Daneshjou, co-senior author of the study and an assistant professor at Stanford University, expressed the importance of such frameworks. She stated, “CRAFT-MD creates a framework that more closely mirrors real-world interactions and thus it helps move the field forward when it comes to testing AI model performance in health care.”

The study, involving contributions from various experts from institutions such as Georgetown University and the University of California-Los Angeles, was funded by significant grants intended to innovate in biomedical research and technology integration.

As the integration of AI into healthcare continues to evolve, findings such as these prompt a reevaluation of how AI performance is assessed, ensuring that these tools are equipped to handle the complexities of clinical interactions effectively.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This article supports the claims about the study evaluating AI models' performance in real-world clinical environments, their proficiency in exam-style questions versus conversational contexts, and the development of the CRAFT-MD framework.
2. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This source corroborates the findings that large language models struggle with nuanced conversations and the proposal for AI models to use open-ended, conversational questioning.
3. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This article explains the importance of the CRAFT-MD framework in enhancing the assessment of AI models and accelerating their development process.
4. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This source provides details on the involvement of various experts and the funding for the study, highlighting the collaborative effort and financial support.
5. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This article emphasizes the need for reevaluating AI performance assessment to ensure these tools can handle clinical complexities effectively.
6. <https://med.stanford.edu/news/all-news/2025/01/ai-cancer-prognosis.html> - Although not directly related to the specific study, this article supports the broader context of AI integration in healthcare, particularly in cancer prognosis and treatment, highlighting the importance of multimodal data integration.
7. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This source quotes Pranav Rajpurkar and Roxana Daneshjou, highlighting their insights on the performance paradox of AI models in medical settings.
8. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This article discusses the limitations of AI models in gathering pertinent patient information, asking critical follow-up questions, and integrating scattered data into coherent diagnoses.
9. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This source explains the efficiency and cost reduction potential of the CRAFT-MD framework in testing AI tools in healthcare.
10. <https://hms.harvard.edu/news/how-good-are-ai-clinicians-medical-conversations> - This article mentions the involvement of institutions such as Georgetown University and the University of California-Los Angeles in the study.
11. <https://www.pharmacyupdate.online/how-good-are-ai-doctors-at-medical-conversations/> - Please view link - unable to able to access data