News and Perspectives

Can Humanlike Reasoning Be Replicated in Large Language Models for Clinical Decision-Making?

Shalini Kathuria Narang, JMIR Correspondent

Abstract

A recent study comparing physicians and large language models on clinical reasoning tasks has received widespread attention. In this *News and Perspectives* article, JMIR Correspondent Shalini Kathuria Narang speaks to one of the study's researchers and reports on the real-world implications that can and cannot be drawn from its findings.

Key Takeaways:

- A recent study suggests that OpenAI's o1 preview model outperforms physicians across some common clinical reasoning tasks.
- While it's clear that large language models have improved from prior generations, the results do not suggest that AI systems are ready to practice medicine independently or that physicians can be removed from diagnostic processes.
- Further prospective trials are needed to examine how these models might be safely integrated into clinical workflows.

Large language models (LLMs) are rapidly being integrated into health care to support clinical decision-making. Although applying artificial intelligence (AI) to assist with clinical decision support is sometimes viewed as a high-risk endeavor, greater use of these tools might serve to mitigate the human and financial costs of diagnostic errors, delays, and lack of access.

Reasoning LLMs that can sequentially work through problems—mirroring structured thinking—have done well in assessing clinical cases. Whether these tools could match or perhaps exceed physicians' clinical reasoning for real diagnoses has been an ongoing question.

In a study led by researchers at Harvard Medical School and Stanford University, the OpenAI o1 series LLM, which was designed with more advanced reasoning capabilities, was compared to physicians with different levels of experience on clinical reasoning tasks. Clinical cases ranged from published patient studies to diagnosis and clinical care tasks with new patients in internal medicine.

A follow-up interview with Adam Rodman, MD, MPH, FACP, helps contextualize the findings of this study after the flurry of media coverage on its publication. Rodman is a general internist and medical educator at Beth Israel Deaconess Medical Center, assistant professor at Harvard Medical School, and Director of AI Programs for the Carl J. Shapiro Center for Education and Research.

The Study and Its Implications

In the study, the LLM was pitted against the physicians in experiments of clinical reasoning, including cases utilizing real and unstructured clinical data taken from the health records in an emergency room (ER). These diagnostic touchpoints mirror the high-stakes, time-sensitive decisions taken by health care professionals with limited information.

Human experts and LLM second opinions were compared for randomly selected patients in the ER of a Boston hospital at 3 stages: initial triage at arrival, first contact with a physician, and admission to the medical floor or intensive care unit. The model was evaluated on 76 cases by 2 doctors who did not know if the case assessments had been made by the AI model or by attending physicians.

The LLM outperformed physicians across tasks like rapid assessment, recommendations for medications, caring goals, and overall case management. It matched or exceeded human performance across each stage, with the widest gap at initial ER triage, where there is the least information available. Humans, GPT-4, and the OpenAI o1 series all improved as more information became available.

While the study results have face validity, the tasks that were studied, namely providing second opinions at predefined touchpoints, are best thought of as a proof of concept. Decisions in the ER are often centered around triage, disposition, and immediate management rather than diagnostic accuracy.

"The study is a validation of the diagnostic performance of these [LLM] models. The basic claim is that the diagnostic performance of the models is not just an artifact of the evaluation mechanisms like vignettes, but holds with real clinical data. That does not mean that just deploying it makes a difference in patient care. It's more, like, watching over like a second set of eyes," says Rodman.

Caveats and Cautions

While the study established a foundation for evaluation across text-based tasks, a substantial caveat is that clinical practice involves visual and auditory cues, such as findings from physical examinations.

Existing studies suggest that current foundational models are more limited in reasoning over nontext inputs. Future research is needed to build on preliminary work assessing how humans and machines may collaborate by using nontext signals, which requires new benchmarks, trials, and technology to more faithfully measure clinical encounters. Such studies may help further investigate an important limitation of existing benchmarks of medical AI: their reliance on the careful work of clinicians to curate and "clean up" cases and their reliance on questions originally developed for educational purposes, which can lead to the performance of AI models being overstated compared to when using "messy" data in more realistic clinical workflows.

"There's a lot of desire to say that we could use technology to replace doctors, but that's not at all what I think it's capable of," says Rodman. "The models are very capable in assisting with diagnosis, but they are not good at integrating information from many different sources. Doing a physical exam, I talk to a patient, look at them, hear the hesitation in their voices. I'm situated in the room with them. Not just getting information from them directly…looking at the medical record, calling other people, and doing an investigation. The LLMs are really good at integrating information that a physician curates from other sources, or they're really good at collecting that information verbally from a patient. They are not at all good at all those other parts of the diagnostic process, which are just as important."

> *There is always the human touch in healing, which AI would not be able to provide.*
>
> Jasmeen Pombra

Jasmeen Pombra—assistant chief of Hospital Operations and chief of Neuro Hospital Medicine at Kaiser Permanente, Redwood City, California—agrees.

"AI [medical tools] can help make diagnosis and be an adjunct to help provide faster care but cannot replace clinical gestalt which comes from years of clinical practice and more from treating patients by observing subtle signs," she says. "AI is very useful in tedious tasks and may help decrease workload for physicians by generating notes and being able to sift through patients' health summary to make relevant data available for clinicians for making diagnosis and treatment plans. There is always the human touch in healing, which AI would not be able to provide."

Further, although large numbers and varieties of cases were included in the study, all were focused on internal medicine and emergency care. It is not representative of broader medical practice, which includes multiple specialties that require varying skill sets. Performance may vary according to diagnoses, patient characteristics, or practice locations that were not interrogated in this study. Dozens of other tasks that have greater impact on actual clinical care can be studied.

Moving Forward

The study makes the case that medical AI is ready to be studied the same way as all new medical interventions—through carefully controlled clinical trials in multiple real care settings.

The researchers are clear that their results do not suggest that AI systems are ready to practice medicine autonomously or that physicians can be removed from the diagnostic process.

Instead, they emphasize the importance of conducting prospective trials in real-world settings to evaluate the latest models of the LLM used in the study, which has since been superseded by newer models (like OpenAI's o3 model) with improved reasoning capabilities, deliberation time, and ability to process multimodal inputs.

They also emphasize the need for health care systems to invest in computing infrastructure and design for clinician-AI interaction, to facilitate safe integration of AI tools into patient-care workflows. This includes the development of monitoring frameworks to oversee the broader implementation of AI clinical decision support systems, monitoring not just final diagnostic accuracy but also other metrics for successful deployment, including safety, efficiency, and cost.

"We've largely moved on with our research. The most obvious use case for this specific technology is trying to identify errors before they happen. Looking for when doctors may be going in the wrong direction, and using that as an early signal," says Rodman. "This could be something called an E trigger or in the form of a second opinion."

Keywords: clinical reasoning; large language models; artificial intelligence; diagnostic decision support; internal medicine; OpenAI model

Please cite as:
Narang SK
Can Humanlike Reasoning Be Replicated in Large Language Models for Clinical Decision-Making?
J Med Internet Res 2026;28:e103526
URL: https://www.jmir.org/2026/1/e103526
doi: 10.2196/103526