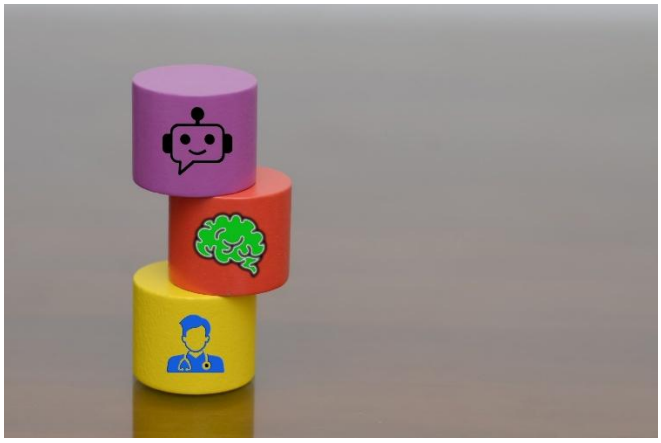
To Guide Real Patient Decisions Large Language Models Excel in Tests yet Struggle

Researchers investigated whether large language models (LLMs) can effectively support the general public in navigating common health scenarios.

Although LLMs performed very well when evaluated on their own, participants who used LLMs did not perform better than those using traditional resources and often performed worse. This highlights significant failures in human-LLM interaction that are not captured by standard benchmarks or simulations, and suggests that strong standalone medical knowledge does not necessarily translate into effective real-world user support or improved clinical decision-making by lay users.



Study

Researchers systematically evaluated how members of the general public interact with LLMs when making medical decisions and assessed whether these interactions improve understanding and decision-making compared with conventional approaches.

They conducted a randomized controlled study involving 1,298 adult participants living in the United Kingdom. Participants were presented with realistic medical scenarios designed by physicians, each requiring the user to identify possible underlying medical conditions and select an appropriate disposition or course of action on a five-point scale, with options ranging from calling an ambulance (the most serious response) to self-care (the least serious response).

Three doctors developed and validated the scenarios and agreed on the correct disposition, while four additional doctors generated gold standard lists of relevant diagnoses. Each participant was randomly allocated to one of four groups. Three treatment groups each used a different LLM, GPT 4o, Llama 3, or Command R plus, and a control group could use any resources they normally would, such as internet searches.

Demographic stratification ensured that each group resembled the national adult population. Each participant completed up to two scenarios, resulting in 600 responses per experimental condition. Performance was evaluated by comparing participants' selected dispositions with physician-agreed answers and by assessing whether listed conditions matched the gold-standard diagnostic lists.

To contextualize the findings, the researchers also evaluated LLMs operating alone on the same scenarios, tested performance on medical knowledge benchmarks, and conducted simulated user experiments in which LLMs replaced human participants.

Results

When evaluated independently, all three LLMs performed strongly. Each successfully identified one or more relevant conditions in over 90 percent of cases, although correct disposition recommendations were more moderate, averaging roughly half of cases rather than approaching expert-level clinical reliability.

However, when paired with human users, performance declined substantially. Participants using LLMs identified relevant medical conditions in fewer than 35 percent of cases, significantly worse than the control group, which relied on traditional resources such as search engines and established health websites.

Disposition accuracy did not differ significantly between LLM-assisted participants and controls, with both groups frequently underestimating the seriousness of conditions. Analyses of interaction transcripts revealed multiple points of failure. Users often provided incomplete or poorly structured information; LLMs sometimes misinterpreted inputs or generated misleading suggestions; and participants frequently failed to adopt correct recommendations, even when they were presented during the conversation.

Benchmark evaluations using medical licensing style questions showed that high LLM scores did not predict success in interactive settings. Similarly, simulations using LLM-generated patients overestimated real-world performance and failed to capture human variability, with simulated users generally performing better than real participants. Overall, the results demonstrate a consistent gap between LLMs' medical knowledge and their effectiveness when used interactively by non-expert users.

Conclusion

This study demonstrates that current LLMs tested in this research setting are not ready for safe deployment as public-facing medical advisors, despite strong standalone performance on medical tasks. A key strength of the work is its focus on real human-LLM interactions, which reveal failures that are invisible to traditional benchmarks and simulations.

The findings suggest that the main challenge lies not only in medical knowledge but also in communication, consistency, and user interpretation. Users struggled to know what information to provide, how to evaluate multiple suggestions, and when to trust or act on LLM output, even when correct information appeared during the interaction. This highlights limitations in both AI communication design and human interpretation skills.

Limitations include the use of hypothetical scenarios rather than real clinical encounters and a focus on common conditions, which may underestimate the challenges posed by more complex cases. Additionally, newer or specialized models may perform differently, and future improvements in conversational design or clinical fine-tuning could alter performance, although whether such gains translate into real-world benefit remains uncertain.

Nonetheless, the study provides a critical lower bound on real-world performance and highlights the risks of premature deployment. The authors concluded that systematic testing with diverse

human users should be a foundational requirement before LLMs are used in healthcare. Medical expertise alone is insufficient to ensure safe and effective patient support, and real user evaluation should complement traditional benchmark testing before clinical integration.

Source:

https://www.news-medical.net/news/20260210/Large-language-models-excel-in-tests-yet-struggle-to-guide-real-patient-decisions.aspx