Benchmarking large language model-based agent systems for clinical decision tasks

Info

PUBLIZIERT AM
2026-02-18

JOURNAL
npj Digital Medicine

LINK
https://doi.org/10.1038/s41746-026-02443-6

Autoren

Yunsong Liu Zunamys I. Carrero Xiaofeng Jiang Dyke Ferber Georg Wölflein Li Zhang Sanddhya Jayabalan Tim Lenz Zhouguang Hui Jakob Nikolas Kather

Abstract

Agentic artificial intelligence (AI) systems, designed to autonomously reason, plan, and invoke tools, have shown promise in healthcare, yet systematic benchmarking of their real-world performance remains limited. In this study, we evaluate two such systems: the open-source OpenManus, built on Meta's Llama-4 and extended with medically customized agents; and Manus, a proprietary agent system employing a multistep planner-executor-verifier architecture. Both systems were assessed across three benchmark families: AgentClinic, a stepwise dialog-based diagnostic simulation; MedAgentsBench, a knowledge-intensive medical QA dataset; and Humanity's Last Exam (HLE), a suite of challenging text-only and multimodal questions. Despite access to advanced tools (e.g., web browsing, code development and execution, and text file editing) agent systems yielded only modest accuracy gains over baseline LLMs, reaching 60.3% and 28.0% in AgentClinic MedQA and MIMIC, 30.3% on MedAgentsBench, and 8.6% on HLE text. Multimodal accuracy remained low (15.5% on multimodal HLE, 29.2% on AgentClinic NEJM), while resource demands increased substantially, with >10× token usage and >2× latency. Although 89.9% of hallucinations were filtered by in-agent safeguards, hallucinations remained prevalent. These findings reveal that current agentic designs offer modest performance benefits at significant computational and workflow cost, underscoring the need for more accurate, efficient, and clinically viable agent systems.