Qeios

Peer Review

# Review of: "MedAgentBench: A Realistic Virtual EHR Environment to Benchmark Medical LLM Agents"

**Sebastian Spethmann**[1]

1. Charité Universitätsmedizin Berlin, Berlin, Germany

MedAgentBench fills a critical gap in AI evaluation for healthcare by providing a comprehensive benchmark that assesses AI agents in realistic medical scenarios. The benchmark includes 300 clinically derived tasks across 10 categories, 100 patient profiles with over 700,000 data elements, and a FHIR-compliant interactive environment. This robust framework enables a more accurate evaluation of AI capabilities beyond traditional question-answering approaches.

By assessing AI agents' ability to perform complex medical tasks, MedAgentBench supports the development of AI systems that can enhance clinical workflows, reduce administrative burdens, and improve patient care. The evaluation of 12 state-of-the-art large language models (LLMs) using MedAgentBench demonstrates promising capabilities but also highlights that these models are not yet reliable enough for complex medical decision-making. This underscores the ongoing need for further development, optimization, and real-world validation.

MedAgentBench serves as a valuable tool for model developers, offering a standardized benchmark to track progress and drive continuous improvements in AI-driven healthcare solutions. Its focus on interactive, agent-based tasks represents a shift from traditional static assessments, aligning with the evolving needs of AI integration in clinical practice. Furthermore, its public availability promotes transparency and facilitates broad adoption and collaborative advancement within the research community.

However, several limitations should be considered:

The patient profiles are derived exclusively from Stanford Hospital records, which may not be representative of diverse populations. This could introduce bias and limit the benchmark's

 doi.org/10.32388/FNYE02 

applicability to other healthcare settings.

Although MedAgentBench simulates real-world scenarios, it does not fully capture the complexity of multi-disciplinary team coordination often required in clinical practice.

The benchmark primarily focuses on inpatient and outpatient medical scenarios, potentially overlooking critical areas such as surgical specialties, emergency medicine, and nursing.

Longitudinal aspects of patient care, including follow-up visits and evolving treatment plans, are not incorporated, limiting the benchmark's ability to evaluate AI performance over extended clinical timelines.

AI models optimized specifically for MedAgentBench tasks may suffer from overfitting, reducing their generalizability to real-world medical environments.

While the benchmark uses standard APIs and communication protocols similar to modern EMR systems, differences between the simulated environment and actual hospital EMRs may limit direct applicability.

## Declarations

**Potential competing interests:** No potential competing interests to declare.