Qeios

Peer Review

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

**Shrieraam Sathyanarayanan**[1]

1. Health Access Network, Oklahoma State University Oklahoma City, Oklahoma City, United States

## *Strengths*

- **Innovative approach**: Moves beyond static QA to dynamic, agentic workflows.
- **Practical implications**: Provides a pathway for integrating AI agents into clinical workflows.
- **Transparency**: Publicly available dataset and code (https://github.com/stanfordmlgroup/MedAgentBench).
- **Detailed error analysis**: Identifies common failure patterns (e.g., incorrect syntax, wrong output format).

## *Limitations*

- **Scope**: Focused on medical record tasks; excludes surgical or nursing workflows.
- **Bias**: Patient profiles derived from Stanford Hospital—may not generalize globally.
- **Reliability**: Current LLMs are far from meeting clinical safety standards.
- **Evaluation constraints**: Only pass@1 metric used; does not test robustness under repeated trials.

## *Overall Assessment*

The paper is **timely and impactful**, addressing a critical gap in evaluating LLMs for healthcare. It sets a strong foundation for future research on **agent reliability, multi-step reasoning, and integration into EMR systems**.



doi.org/10.32388/YMC3LG

## Declarations

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