24 February 2025 · CC-BY 4.0

Peer Review

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

Ângelo Fonseca[1]

1. Neurology, Hospital Pedro Hispano, Matosinhos Municipality, Portugal

This paper presents MedAgentBench, a promising framework for evaluating the agent capabilities of large language models (LLMs) within medical contexts. The authors propose an interactive, FHIR-compliant environment that enables benchmarking of LLMs through 300 clinically relevant tasks. While the paper provides an interesting and valuable contribution to AI research in healthcare, several areas require clarification and refinement.

General Comments:

1. **Author Affiliations**:

   The author affiliations are incomplete. To improve the clarity and professionalism of the paper, ensure that full affiliations, including department names and institutions, are provided for all authors.

2. **Abstract**:
   - The term "FHIR-compliant environment" is mentioned without explanation. It would benefit from a brief explanation of FHIR (Fast Healthcare Interoperability Resources) as a standard for health information exchange.
   - The acronyms "API" (Application Programming Interface) and "EMR" (Electronic Medical Record) are used without being defined. Consider expanding these terms the first time they appear.
   - The phrase "Agent-based task frameworks are the necessary next step to advance..." may come across as presumptive. Rewording this to a less definitive statement like, "Agent-based task frameworks could play a role in advancing..." would make the tone more neutral.

3. **Introduction**:

doi.org/10.32388/WVKIWE

- The term "traditional QA-based AI benchmarks" is used but not explained. Briefly define what these benchmarks are and how they differ from the proposed agent-based framework.

4. **Methods**:
    - Table 2.2.1, detailing the cohort characteristics, is placed in the Methods section but should be moved to the Results section. The Methods section should only mention the recruitment goal (100 patients).
5. **Results**:
    - The Results section should be a separate section (not a subsection) to improve clarity and readability.
    - Further analysis of the variations in performance across task categories would provide a more nuanced understanding of model capabilities and limitations.
6. **Security Considerations**:

    The authors mention that the FHIR-compliant environment is not secure for production settings. It would be valuable to briefly discuss how security concerns should be addressed in future iterations, especially given the sensitivity of healthcare data.
7. **Generalizability**:

    The paper acknowledges that the patient profiles are derived from a single institution (Stanford). This limits the generalizability of the results. It would be beneficial to mention how future work might involve more diverse datasets or adapt the framework for different healthcare settings.

Additional Strengths:

1. **Model Variability**:

    The evaluation of 12 state-of-the-art LLMs reveals promising results, although there is still significant room for improvement. The results emphasize the potential of LLMs in medical applications but also highlight that they are not yet reliable enough for deployment in clinical settings.
2. **Potential for Reducing Healthcare Burden**:

    The paper discusses how AI agents could alleviate administrative burdens and improve the quality of clinical care. This idea is particularly relevant given current healthcare workforce shortages.

Conclusion:

This is an interesting and promising study that introduces an important tool for benchmarking AI models in healthcare settings. With minor revisions, the paper would be a stronger and more comprehensive contribution. The framework has the potential to drive significant advancements in AI integration into clinical workflows, but further development is needed to address reliability concerns and extend its applicability across different medical domains.

Declarations

Potential competing interests: No potential competing interests to declare.