**Abstract**

The integration of Large Language Models (LLMs) into clinical diagnostics has the potential to transform patient-doctor interactions. However, the readiness of these models for real-world clinical application remains inadequately tested. This paper introduces the Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD), a novel approach for evaluating clinical LLMs. Unlike traditional methods that rely on structured medical exams, CRAFT-MD focuses on natural dialogues, using simulated AI agents to interact with LLMs in a controlled, ethical environment. We applied CRAFT-MD to assess the diagnostic capabilities of GPT-4 and GPT-3.5 in the context of skin diseases. Our experiments revealed critical insights into the limitations of current LLMs in terms of clinical conversational reasoning, history taking, and diagnostic accuracy, emphasizing the need to evaluate clinical LLMs beyond static exam-questions. The introduction of CRAFT-MD marks a significant advancement in LLM testing, aiming to ensure that these models augment medical practice effectively and ethically.

**Introduction**

Doctor-patient conversations enable physicians to uncover key details that guide their clinical decisions. However, the mounting pressure of escalating patient numbers, lack of access to care (Lasser, Himmelstein, and Woolhandler 2006), short consultation times (Irving et al. 2017; Wong, Vincent, and Al-Sharqi 2017), and the expedited adoption of telemedicine due to the COVID-19 pandemic (Shaver 2022) have presented formidable challenges to this conventional model of interaction. As these factors risk compromising the quality of history taking and thereby diagnostic accuracy (Bubeck et al. 2023), there is an urgent need for innovative solutions that can enhance the efficacy of these crucial conversations.

New advances in Large Language Models (LLMs), could present a potential solution to this problem (Nori et al. 2023; Singhal et al. 2023; Sarraju et al. 2023; Rajpurkar et al. 2022; Lee, Bubeck, and Petro 2023). These AI models have the ability to engage in nuanced and complex conversations, making them ideal candidates for extracting comprehensive patient histories and assisting physicians in generating differential diagnoses (Moor et al. 2023; Ayers et al. 2023; Au Yeung et al. 2023). However, a considerable gap remains in assessing these models’ readiness for application in real-world clinical scenarios (Wornow et al. 2023; Shah, Entwistle, and Pfeffer 2023; Ali et al. 2023). The predominant method for evaluating LLMs in the medical field involves medical exam-type questions, with a strong emphasis on multiple-choice formats (Fijačko et al. 2023; Kung et al. 2023; Han et al. 2023). Although there are instances where LLMs are tested on free-response and reasoning tasks (Strong et al. 2023; Nair et al. 2023; Lowell et al. 2001), or for medical conversation summarization and care plan generation (Shanahan, McDonell, and Reynolds 2023), these are less common. However, none of these assessments explore LLMs’ ability for engaging in interactive patient conversations, a crucial aspect of their potential role in revolutionizing healthcare delivery.

**Methods**

To address the evaluative shortfall, we propose a new framework for evaluation of clinical LLMs, called the Conversational Reasoning Assessment Framework for Testing in Medicine (CRAFT-MD). CRAFT-MD allows multi-faceted testing of clinical abilities of LLMs, including medical history gathering and open-ended diagnosis, by employing AI agents in simulations to represent patients or graders, rather than relying completely on human evaluators. This strategy significantly enhances the scalability of evaluations and allows for broader and quicker testing, keeping
pace with the rapid evolution of LLMs (Figure 1).

Results

We applied the CRAFT-MD framework on 140 case vignettes focused on skin diseases, sourced from both an online question bank1 (100 cases) and 40 newly created cases, encompassing a variety of skin conditions seen in both primary care and specialist settings. Our evaluations focused on the performance of GPT-4 and GPT-3.5 (versions “gpt-4-0314” and “gpt-3.5-turbo-0301”) across 10 simulations per case vignette, revealing several limitations in clinical LLMs’ conversational reasoning abilities (Appendix Figure 1, Table 1). In 4-choice multiple choice questions (MCQs), multi-turn conversations decreased accuracy versus vignettes. Notably, multi-turn conversations did not improve over single-turn, but summarizing conversations into concise paragraphs increased accuracy, indicating inability to synthesize across dialogues. Importantly, vignettes had the highest accuracy compared to all conversational setups, indicating limitations in medical history gathering skills. Replacing 4-choice MCQs with free response questions (FRQs), we observed a further decrease in accuracy across all experimental setups, with similar trends for inability to synthesize information and take medical histories. Removing physical exam details further decreased accuracy, indicating potential benefit of multimodal integration in LLMs. Code and data for reproducing experimental results is available online2.

<table>
<thead>
<tr>
<th>Type</th>
<th>GPT-4</th>
<th>GPT-3.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vignette</td>
<td>0.919</td>
<td>0.833</td>
</tr>
<tr>
<td>Multi-turn conversation</td>
<td>0.854</td>
<td>0.724</td>
</tr>
<tr>
<td>Single-turn conversation</td>
<td>0.868</td>
<td>0.745</td>
</tr>
<tr>
<td>Summarized conversation</td>
<td>0.856</td>
<td>0.810</td>
</tr>
<tr>
<td>Multi-turn conversation</td>
<td>0.774</td>
<td>0.642</td>
</tr>
<tr>
<td>(without physical exam)</td>
<td>0.324</td>
<td>0.318</td>
</tr>
</tbody>
</table>

Table 1: Experimental Results. MCQ = 4-choice Multiple Choice Questions; FRQ = Free Response Questions.

Conclusion

Recent studies showing high diagnostic accuracy on medical exam questions for LLMs such as GPT-4 may present an overly optimistic outlook for clinical use case, as these evaluations overlook crucial real-world complexities. CRAFT-MD reveals significant deficiencies in LLMs’ abilities to gather thorough patient histories, synthesize information over dialogues, and clinical reasoning for diagnosis without answer choices. This work emphasizes the need for responsible and comprehensive evaluation of clinical LLMs.

Acknowledgments

S.J. is supported by the 2023 Quad Fellowship. We thank the Microsoft Accelerating Foundation Models Research (AFMR) program for providing Azure credits.

---

1https://www.clinicaladvisor.com/
2https://github.com/rajpurkarlab/craft-md
References


Nair, V.; Schumacher, E.; Tso, G.; and Kannan, A. 2023. DERA: Enhancing large language model completions with dialog-enabled resolving agents.


Appendix

Figure Appendix 1: Schematic showing experimental setups for assessing GPT-4 and GPT-3.5 using CRAFT-MD using (a) vignettes, (b) multi-turn conversations, (c) single-turn conversations, and (d) summarized conversations. (e, f) Clinical LLM’s accuracy for all experimental setups on the 140 case vignettes.