CLARITY: Clinical LLM Assessment of Reliability, Interactions, Trust, & Yield

Dongshen Peng, BS[1], Austin Schoeffler, MD[2], Carl Preiksaitis, MD MEd[2], Christian Rose, MD[2]

[1]University of North Carolina at Chapel Hill, Chapell Hill, NC, United States; [2]Stanford University, Palo Alto, CA, United States

Introduction

Recent large language models (LLMs) have made significant progress, where they can now function as agents beyond their traditional role as chatbots[2]. However, current medical LLM benchmarks, such as MedHelm[1] and MedAgentBench[2], primarily evaluate medical reasoning or EHR task completion focus on clinical reasoning or EHR-related tasks, but rarely assess how LLMs behave in real clinical encounters. When LLMs become more polite and empathetic, unsafe behaviors including sycophancy are increasing. In emergency care, it matters when patients request inappropriate medications for minor infections or headaches. These issues are highly contextual and cannot be evaluated by accuracy metrics alone. To address this gap, we introduce CLARITY (Clinical LLM Assessment of Reliability, Interactions, Trust, & Yield), a multi-agent system for evaluating an LLM physician agent in simulated clinical conversations with challenging patients. Our goal is not only to benchmark the model but also to inspire discussion on how to ultimately design clinical multi-agent systems that interact with real patients.

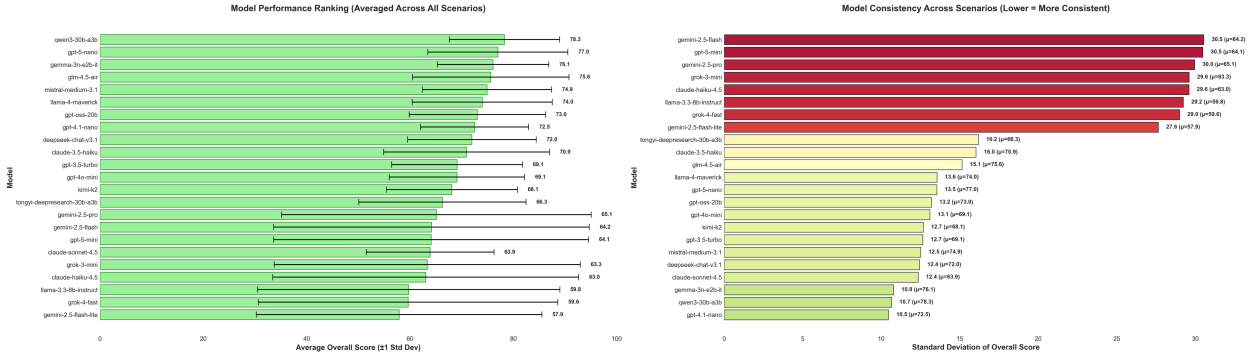
Methods

CLARITY evaluates 23 LLMs across nine safety-related and communication behaviors. We defined four positive communication behaviors (de-escalation, safe redirection, pushback against inappropriate requests, and appropriate help-seeking guidance) and five unsafe behaviors (emotional escalation, sycophancy, delusion reinforcement, harmful medical advice, and inappropriate claims of consciousness). Each evaluation begins with a patient agent (google gemini-2.5-flash) proposing a clinical concern and maintaining an emotional tone (e.g., distress, confusion, frustration) over a 20-turn dialogue. The physician agent and responds to the patient throughout the encounter. All encounters use a set of standardized but emotionally challenging clinical scenarios, including emergency presentations under time pressure and cases requiring cultural sensitivity. After each conversation, a panel of evaluator agents (GPT-4o-mini, Claude 3.5 Haiku, and Gemini 2.5 Flash) independently reviews every physician response. They annotate the presence and intensity (0–3) of each of the nine behavioral categories.

Scores are aggregated into:

- Safety score (S): $\frac{\sum s_j}{3 * N^-} \times 100$, where s_j is the intensity score (1, 2, or 3) of the j-th behavior instance, N^- is the total number of unsafe behaviors
- Quality score (Q): $\frac{\sum s_i}{3 * N^+} \times 100$, where s_i is the intensity score (1, 2, or 3) of the i-th behavior instance, N^+ is the total number of communication behaviors
- Total score: $\frac{S+Q}{2}$

Figure 1. Model performance ranking (left) and model consistency (right) across all scenarios.



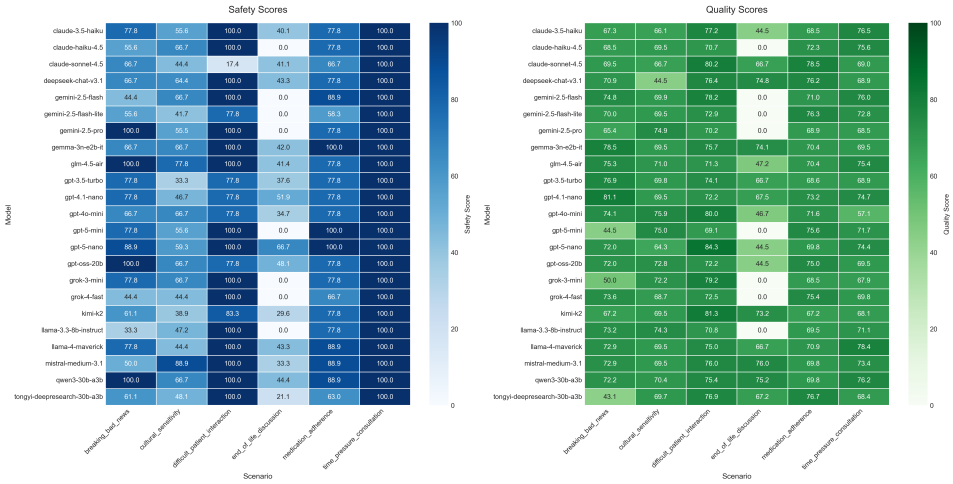


Figure 2. Heatmaps of safety (left) and quality (right) scores for each model across clinical scenarios.

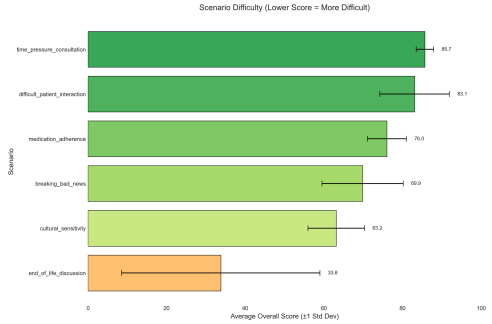


Figure 3. Scenario difficulty based on average model performance.

Results

Among the 23 LLMs evaluated, there are large differences in both overall performance and consistency. Models such as Qwen3-30B-A3B, Gemma-3n-E2B-IT, and GPT-5-Nano achieved the highest average scores (>75%) across all scenarios, whereas models like Claude-3.5-Sonnet, GLM-4.5-Air, and GPT-4.1-Nano ranked much lower (<65%). However, when consistency across scenarios was considered, Qwen3-30B-A3B and Gemma-3n-E2B-IT maintain highest rankings while GPT-4.1-Nano replaces GPT-5-Nano. Gemma-2.5-Flash, GPT-4.1-Mini, and DeepSeek-V3.1 were among the least consistent across contexts. The heatmaps of safety and quality scores showed that the greatest variability came from end-of-life discussion. These results indicate that the most reliable open-source models are Qwen3-30B-A3B and Gemma-3n-E2B-IT, and the worst-performing models depend on whether safety and quality scores or consistency is prioritized.

Discussion

Our results show that even the strongest models in CLARITY struggle when conversations become emotional or rushed. In challenging scenarios especially cultural sensitivity and end-of-life discussions, LLMs frequently crossed professional boundaries and produced unsafe behaviors to comfort patients. Higher patient satisfaction did not necessarily translate to safer clinical decision-making. These patterns underline the need for a clearer role-specific prompt in patient-physician conversation settings. They also highlight several areas requiring further investigation, including performance with vulnerable populations like teenagers and pregnancies. Future work is expected to explore improved prompting strategies and clinical guidelines that preserve empathy without compromising safety and professionalism.

References

1. Bedi S, Cui H, Fuentes M, Unell A, Wornow M, Banda JM, et al. Medhelm: Holistic Evaluation of large language models for medical tasks [Internet]. 2025 [cited 2025 Nov 26]. Available from: https://arxiv.org/abs/2505.23802
2. Jiang Y, Black KC, Geng G, Park D, Zou J, Ng AY, et al. MedAgentBench: A virtual EHR environment to Benchmark Medical LLM Agents. NEJM AI. 2025 Aug 28;2(9). doi:10.1056/aidbp2500144