Can LLM judges catch unsafe clinical actions by EHR agents?

27 November 2026 · 5 min read · blogs.iictp.org/llm-judges-unsafe-clinical-actions-ehr-agents

Short answer: LLM judges for clinical AI agents are language models asked to decide whether an agent's actions in an electronic health record were safe, and a fault-injection study by Neda Ghafouri, posted on arXiv on 24 November 2026, tested how reliable they are. Judges did far better when shown the real record responses than when shown only the final answer. It matters because an agent can appear to finish a task while taking a dangerous step.

TL;DR The short version

- A fault-injection study posted on 24 November 2026 tested whether LLM judges can spot unsafe actions by agents working in electronic health records.
- Judges did far better when shown the real record responses than when shown only the agent's final answer.
- An agent can appear to finish a task while taking a dangerous step, so evaluate agents on their full action trace, not only the result.

Read online, listen to it, or change the text size: https://blogs.iictp.org/llm-judges-unsafe-clinical-actions-ehr-agents

What are LLM judges in clinical agent testing?

An EHR agent reads and writes records to complete tasks such as checking a laboratory value or placing an order. Developers often use a second model, the judge, to grade the sequence of actions, called a trajectory. The study asks whether the judge notices clinically important mistakes.

How the fault injection worked

1. Start from verified correct trajectories in the MedAgentBench simulated EHR benchmark.
2. Inject controlled clinical errors across eight categories, producing 416 faulty trajectories.
3. Ask six open-weight models to judge them.
4. Compare each faulty run with its correct twin, so a judge that simply flags more things gets no credit.
5. Vary what the judge sees: final answer, full trajectory, record responses and the agent's reasoning.

What is new compared with grading by final answer

Paired comparison separates real error detection from general suspicion. The study also checks the benchmark's own grader and a rule-based checker, so it measures the evaluation tools as well as the models.

Results reported by the author

- With only the final answer, judges could not distinguish errors that left it unchanged, and they flagged 56% of correct trajectories.
- With the full trajectory, the ability to tell unsafe from correct runs reached a score of 0.75, and adding the actual EHR responses improved discrimination by 0.39.
- Showing the agent's reasoning did not help overall and reduced detection of wrongly reported clinical values by 0.18. These were the hardest errors to catch.
- MedGemma, a medically specialised model, did worse than the general-purpose comparison and was the only judge to miss potentially dangerous medication overdoses.
- The original benchmark grader accepted all 38 trajectories where a laboratory check had been deliberately skipped.
- LLM judges combined with a rule-based checker that had the full record detected 97% of injected faults.

This is an arXiv preprint or workshop paper. Treat its numbers as the authors' own results, not as independently confirmed findings.

Limitations and open questions

- One simulated EHR environment with relatively short tasks.
- One injected error per case, not combinations.
- Harm levels came from predefined rules, not from clinician review.
- Open-weight judges only, so results say nothing direct about other models.
- Injected errors may differ from the mistakes agents make naturally.

Sources were read on 27 November 2026. Products, rules and data change, so check the linked originals before you rely on any detail.

How to apply it to a clinical AI pilot

1. Require evaluation on the whole action log, including what the record system returned.
2. Write deterministic checks for things that must always happen, such as a laboratory check before an order.
3. Test the evaluator itself with deliberately planted errors before trusting its pass rate.
4. Do not assume a medical model is the better judge; measure it.
5. Keep clinician sign-off for orders, doses and messages to patients.
6. The author released code at github.com/neda86/clinical-judge.

What it means for clinicians and informatics teams

A vendor's success rate is only as good as its grader. Ask which errors the grader can see, how often it flags correct work, and whether anyone planted faults to test it. This complements the questions raised by longitudinal clinical agent benchmarks.

Frequently asked questions

- **What are LLM judges for clinical AI agents?** Models that grade whether an agent's actions in a health record were correct and safe.
- **What helped judges most?** Access to the actual EHR responses.
- **Did showing the agent's reasoning help?** Not overall, and it hurt detection of wrong clinical values.
- **Is this ready for clinical use?** No. It is a preprint based on a simulated environment.

Related reading

- WARD benchmark for clinical agents
- MedUIGym and EHR GUI agents
- SAFE: evaluating LLM triage
- AI Practitioner certification

> Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. Neda Ghafouri. Can LLM Judges Catch Unsafe Clinical Actions? A Fault-Injection Study. arXiv 2611.33012. 24 November 2026 (preprint). arxiv.org/abs/2611.33012

Published by IICTP, an independent technical credential body, on 27 November 2026. The current version of this post is at https://blogs.iictp.org/llm-judges-unsafe-clinical-actions-ehr-agents. More guides: blogs.iictp.org.