WARD explained: why clinical AI agents see the evidence but fail to act on it

25 November 2026 · 8 min read · blogs.iictp.org/ward-benchmark-longitudinal-clinical-agents-evidence-to-action-gap-explained

Short answer: WARD is a benchmark, described in an arXiv preprint submitted on 22 November 2026 by Chao Ding, Yuyin Zhou and colleagues, that tests whether language-model agents can update their view of a patient and choose safe, timely actions as hospital information accumulates. Across 12 models the authors find that recognising evidence runs well ahead of updating patient state and choosing actions, and that only 11.46% of executed actions in the interactive track advance a reference transition. This is an arXiv preprint. It has not been peer reviewed, and its numbers are the authors' own claims.

TL;DR The short version

- WARD tests whether clinical AI agents update their view of a patient and choose safe, timely actions as hospital information accumulates.
- Across 12 models, recognising evidence consistently outpaced acting on it, and linked evidence went unanswered 74% to 91% of the time.
- It covers 95 hospitalisations in an arXiv preprint; score evidence, state and action separately when evaluating clinical agents.

Read online, listen to it, or change the text size: https://blogs.iictp.org/ward-benchmark-longitudinal-clinical-agents-evidence-to-action-gap-explained

What is WARD?

Most clinical LLM benchmarks score a single answer or a final outcome. According to the abstract, that hides intermediate failures in using evidence, updating the patient's state and selecting an action. WARD instead follows patients over time: 4,560 clinical tasks across seven clinical output domains, organised into 240 checkpoints from 40 longitudinal trajectories spanning 95 hospitalisations.

How it works

- **Longitudinal track.** Tasks are evaluated under controlled, cumulative evidence, so the model sees what a clinician would have known at each checkpoint.
- **Interactive track.** End-to-end, tool-mediated clinical management, scored against 271 verifiable evidence items and 141 reference transitions.
- **Three-step lens.** Evidence recognition, then patient-state judgment, then action. The benchmark separates them so you can see where an agent fails.

What is new compared with earlier approaches

Earlier benchmarks look at isolated responses. WARD’s contribution is process-level evaluation across time, which matches how care actually happens: assessments are revised as results arrive.

Results reported by the source

- Across 12 models with scorable checkpoint outputs, evidence recognition consistently outpaces downstream state and action decisions.
- Linked-evidence nonresponse reaches 74.47% to 90.90%, according to the authors.
- Models struggle to retain and update information over time.
- In the Interactive Track only 11.46% of executed actions advance a reference transition.

Limitations and open questions

- The data covers 95 hospitalisations, a small number compared with real-world variety.
- Models were tested as agents, not as part of a clinician-supervised workflow.
- The abstract does not say which models were weakest or strongest. Sources were read on 25 November 2026. Products, rules and standards change, so check the linked originals before you rely on any detail.

How to use the idea in an evaluation

1. Break any clinical agent task into evidence, state and action, and score each separately.
2. Test with accumulating information, not a single snapshot.
3. Count actions that move the patient closer to a reference path, not only answers that look right.
4. Keep clinician review of any recommended action.

What it means for clinicians and health-IT teams

The finding is a caution, not a verdict: a model that can quote the right lab result may still not change its plan. If you are evaluating a vendor’s clinical agent, ask for process-level results over time, not a single accuracy figure. See SAFE-Triage for another evaluation approach.

Frequently asked questions

- **What is WARD?** A benchmark of longitudinal clinical tasks that tests how agents turn accumulating evidence into patient-state judgments and actions.
- **How big is it?** 4,560 tasks, 240 checkpoints, 40 trajectories and 95 hospitalisations.
- **What is the main finding?** Evidence recognition is ahead of state and action decisions; only 11.46% of interactive actions advance a reference transition.
- **Is it peer reviewed?** No, it is a preprint.

Related reading

- SAFE-Triage: evaluating LLM emergency triage tools
- Gemini for healthcare and prior authorization agents
- AI Practitioner certification

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

Sources and further reading

1. Ding C, Jiang Y, Ning J, Wu J, Wang Y, Lu R, Yuan M, Wang G, Wang L, Chen P, Yin D, Wang M, Ding J, Zhou Y. Seeing Is Not Acting: WARD Diagnoses the Evidence-to-State-to-Action Gap in Longitudinal Clinical Agents. arXiv 2611.29243, 22 Nov 2026 (preprint). arxiv.org/abs/2611.29243

Published by IICTP, an independent technical credential body, on 25 November 2026. The current version of this post is at https://blogs.iictp.org/ward-benchmark-longitudinal-clinical-agents-evidence-to-action-gap-explained. More guides: blogs.iictp.org.