ReflexBench: Evaluating Observer-Participant Failures and Counterfactual Trustworthiness in Agentic AI

Mian Zhang — KDD 2026 Workshop on Evaluation and Trustworthiness of Agentic AI

20
scenarios across six domains

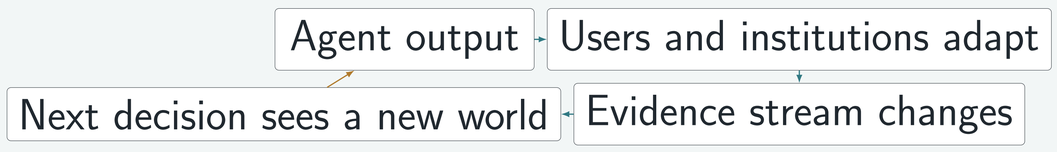
720
scored responses from nine models

≈ -0.44
mean shallow-to-deep degradation

The Failure Mode

Most benchmarks treat an AI agent as an outside observer of a fixed task. Deployed agents are different: recommendations change users, forecasts change incentives, interventions change future data, and benchmarks can change the systems being benchmarked.

Observer-participant failure: the agent gives a plausible answer while ignoring that its own answer changes the system being evaluated.



Observer Depth

OD-0	Fixed-world decision making.
OD-1	First-order self-impact.
OD-2	Other actors adapt to the agent.
OD-n	Recursive, rebound, or equilibrium effects.

$$G_m = \frac{S_m(2) + S_m(n)}{2} - \frac{S_m(0) + S_m(1)}{2}$$

The primary signal is degradation across observer depth, not a winner-only leaderboard.

ReflexBench Protocol

- 20 scenarios across six feedback-sensitive domains.
- Four observer-depth levels per scenario.
- A 0–1 rubric with partial credit for incomplete reflexive reasoning.
- Failure tags separate static mistakes from feedback-loop mistakes.
- Audit fields retain model, prompt, answer, score, tool state, memory state, and evidence provenance.

Scenario Families

Market impact; public signals; platform feedback; organizational adaptation; user-response loops; and coordination equilibria.

Evidence Snapshot

Layer	Registered signal
Model panel	Nine public models and 720 scored responses.
Main result	Rounded mean observer-depth degradation of about -0.44.
Interpretation	Surface competence does not imply recursive-feedback competence.
Boundary	Fixed audit snapshot, not a permanent ranking or safety certificate.

Counterfactual Trustworthiness

The audit question is not only whether an answer was correct in the observed trace. It asks whether the agent would behave the same way if it did not know it was being evaluated, if users had not adapted, or if prior outputs had not changed the evidence stream.

Evidence-integrity channels: prompt provenance, benchmark visibility, model version, tool state, memory state, user adaptation, and post-output environmental change.

Lifecycle Evaluation

- **Pre-deployment:** run the observer-depth profile before release.
- **Runtime:** tag static-world, adaptation, and contamination failures.
- **Post-deployment:** test whether users or data shifted after agent outputs.
- **Model update:** rerun gates after model, prompt, tool, or policy changes.
- **Incident review:** identify the feedback loop that caused the failure.

Scope and Reproducibility

- Diagnostic evaluation, not certification or universal deployment proof.
- Does not replace domain simulation, real-world causal evaluation, or human review.
- The artifact specifies scenario and prompt schemas, OD rubric, failure tags, audit fields, and a statistical hardening plan.
- Full confidence intervals, scorer agreement, rubric sensitivity, and domain breakdown require the retained row-level score table.

Three Questions for Deployment Teams

1. Who observes and reacts to the agent's output?
2. Which future evidence is changed by that reaction?
3. Which regression test will detect the resulting feedback failure?

For feedback-sensitive agents, evaluation is part of the system being evaluated.
ReflexBench turns that risk into an explicit, bounded, and repeatable audit target.

August 9, 2026 — Jeju, Republic of Korea — Non-archival workshop paper