TR-01 | Technical report | VERSION 0.1 | AUGUST 2026

Judge-Driven Evaluation of a Transactional Conversational Agent

Pablo Cardozo
Independent Applied AI Researcher
Buenos Aires, Argentina

Abstract

This technical report examines an evaluation harness for a restaurant assistant that must answer questions, manage multi-turn orders, handle payments, resist unsafe inputs, and escalate to a human. A separate judge model scores accuracy, actionability, completeness, relevance, and tone on a 0 to 100 scale, while the harness records latency, tokens, traces, and category-level statistics. A dated repository report records an increase from 19 of 46 passing cases to 31 of 46 after a partial implementation. These figures are repository-reported and have not been independently rerun for this publication.

Keywords: LLM-as-judge; conversational agents; evaluation; human handoff; observability

EVIDENCE STATUS	SOURCE REVISION
Public implementation and repository-reported results	8467633188cd

1. Research question

Can a category-based LLM judge reveal concrete failure modes and guide measurable improvements in a transactional conversational agent?

2. System under study

A conversational ordering assistant with persistent state, a catalog, payment and cart handlers, human handoff, LangGraph orchestration, Convex persistence, and Langfuse evaluation traces.

3. Method

1. Generate a test battery across eleven categories including FAQ, menu, orders, payment, workflow, handoff, security, and resilience.
2. Use a fresh conversation identifier for every test and execute the full multi-turn message sequence.
3. Score the final answer with a separate judge model using five explicit criteria and a pass threshold of 75.
4. Capture system and judge tokens, latency, traces, aggregate pass rate, average score, and category-level p95 latency.
5. Use failing categories to prioritize implementation work, then rerun the same battery.

4. Current evidence

1. The public source defines eleven test categories, typed judge results, token accounting, latency metrics, and category aggregation.
2. The runner separates the system under test from the judge and records both traces in Langfuse.
3. A repository report dated March 6, 2026 records a 19/46 baseline and a 31/46 partial result.
4. The same report marks workflow integration, handoff integration, and final judge validation as incomplete.

5. Limitations

1. The reported pass-rate change has not been independently reproduced for this report.
2. A single model judge can introduce preference, verbosity, and self-consistency biases.
3. There is no published calibration against multiple human raters.
4. The dated sprint report mixes completed unit-level behavior with integrations that were still pending.

6. Next experiment

1. Version the full 46-case dataset and raw judge outputs.
2. Calibrate pass thresholds against blinded human ratings and report agreement.
3. Pin models, prompts, temperatures, catalog fixtures, and dependency versions.
4. Rerun a clean baseline and final system under the same environment.

7. Sources and reproducibility

1. Public repository
2. Judge result types and threshold
3. Evaluation runner
4. Dated implementation report

Source revision: 8467633188cd89e4708b8011ee2a0fb39b7f1a24

Suggested citation

Cardozo, Pablo. "Judge-Driven Evaluation of a Transactional Conversational Agent." Technical report, version 0.1, 2026. https://pablo.cardozo.com.ar/research/conversational-agent-evaluation.
