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

Evaluación con un juez automático de un agente conversacional transaccional

Pablo Cardozo
Independent Applied AI Researcher
Buenos Aires, Argentina

Resumen

Este technical report examina un harness de evaluación para un asistente de restaurantes que debe responder preguntas, gestionar pedidos de varios turnos, manejar pagos, resistir entradas inseguras y escalar a una persona. Un modelo juez separado puntúa exactitud, accionabilidad, completitud, relevancia y tono de 0 a 100, mientras el harness registra latencia, tokens, trazas y estadísticas por categoría. Un informe fechado del repositorio registra una mejora de 19 de 46 casos aprobados a 31 de 46 después de una implementación parcial. Esas cifras son las reportadas por el repositorio y no fueron reejecutadas de forma independiente para esta publicación.

Palabras clave: LLM como juez; agentes conversacionales; evaluación; handoff humano; observabilidad

ESTADO DE LA EVIDENCIA	REVISIÓN FUENTE
Implementación pública y resultados reportados por el repositorio	8467633188cd

1. Pregunta de investigación

¿Puede un juez basado en un LLM, organizado por categorías, revelar fallos concretos y orientar mejoras medibles en un agente conversacional transaccional?

2. Sistema bajo estudio

Un asistente conversacional de pedidos con estado persistente, catálogo, handlers de pagos y carrito, handoff humano, orquestación con LangGraph, persistencia en Convex y trazas de evaluación en Langfuse.

3. Método

1. Generar una batería en once categorías, incluyendo FAQ, menú, pedidos, pagos, workflow, handoff, seguridad y resiliencia.
2. Usar un identificador de conversación nuevo por test y ejecutar la secuencia completa de mensajes.
3. Puntuar la respuesta final con un modelo juez separado, cinco criterios explícitos y umbral de aprobación de 75.
4. Registrar tokens del sistema y del juez, latencia, trazas, tasa agregada de aprobación, score promedio y p95 por categoría.
5. Usar las categorías fallidas para priorizar implementación y volver a ejecutar la misma batería.

4. Evidencia actual

1. El código público define once categorías, resultados tipados del juez, contabilidad de tokens, métricas de latencia y agregación por categoría.
2. El runner separa el sistema evaluado del juez y registra ambas trazas en Langfuse.
3. Un informe del repositorio del 6 de marzo de 2026 registra un baseline de 19/46 y un resultado parcial de 31/46.
4. El mismo informe marca como incompletas la integración de workflow, la integración de handoff y la validación final con el juez.

5. Limitaciones

1. El cambio reportado en la tasa de aprobación no fue reproducido de forma independiente para este informe.
2. Un único modelo juez puede introducir sesgos de preferencia, verbosidad y autoconsistencia.
3. No existe una calibración publicada contra múltiples evaluadores humanos.
4. El informe del sprint mezcla comportamiento unitario terminado con integraciones que todavía estaban pendientes.

6. Próximo experimento

1. Versionar el dataset completo de 46 casos y las salidas crudas del juez.
2. Calibrar los umbrales contra evaluaciones humanas ciegas y reportar concordancia.
3. Fijar modelos, prompts, temperaturas, fixtures del catálogo y versiones de dependencias.
4. Reejecutar un baseline limpio y el sistema final en el mismo entorno.

7. Fuentes y reproducibilidad

1. Repositorio público
2. Tipos de resultados y umbral del juez
3. Runner de evaluación
4. Informe fechado de implementación

Revisión fuente: 8467633188cd89e4708b8011ee2a0fb39b7f1a24

Cita sugerida

Cardozo, Pablo. "Evaluación con un juez automático de un agente conversacional transaccional." Technical report, version 0.1, 2026. https://pablo.cardozo.com.ar/es/research/conversational-agent-evaluation.

Documento de trabajo independiente. Sin peer review.