

Evaluación del aislamiento de memoria de largo plazo en agentes conversacionales multi-tenant

Pablo Cardozo
Independent Applied AI Researcher
Buenos Aires, Argentina

Resumen

Este working paper define un protocolo reproducible para evaluar memoria conversacional con aislamiento entre tenants. El sistema combina estado de corto plazo del hilo, memoria vectorial y evidencia documental basada en grafos. El artefacto disponible crea dos tenants sintéticos, registra hechos distintos para cada uno, prueba recuperación, detecta filtraciones cruzadas y mide latencia. No se afirma ningún resultado cuantitativo hasta publicar una ejecución versionada.

Palabras clave: memoria conversacional; aislamiento multi-tenant; GraphRAG; evaluación; latencia

ESTADO DE LA EVIDENCIA	REVISIÓN FUENTE
Protocolo de evaluación disponible; falta publicar el artefacto de resultados	1470f8f17163

1. Pregunta de investigación

¿Puede un agente conversacional multi-tenant recuperar hechos persistentes de cada usuario sin filtrar información entre tenants y respetando un objetivo explícito de latencia?

2. Sistema bajo estudio

Un runtime conectado a WhatsApp construido con FastAPI, LangGraph, checkpointer en PostgreSQL, Mem0 sobre pgvector, Neo4j GraphRAG, coordinación de jobs en Convex, generación con Gemini o Vertex y trazas en Langfuse. La precedencia documentada es evidencia del grafo, luego memoria de largo plazo y finalmente estado del hilo.

3. Método

- Crear dos tenants sintéticos con hechos mutuamente excluyentes sobre color, alergia y empleador.
- Registrar tres memorias por tenant y ejecutar seis consultas de recuperación en hilos aislados.
- Usar términos obligatorios y prohibidos para puntuar recuperación y detectar filtraciones entre tenants.
- Reportar exactitud, latencia promedio, p50, p95, latencia máxima y proporción debajo del objetivo.
- Evaluar contra objetivos declarados de 0,86 de exactitud y 700 ms de latencia p95.

4. Evidencia actual

- Un runner versionado calcula verificaciones léxicas deterministas y percentiles de latencia.
- Un dataset JSON de seis casos separa recuperación correcta de filtración entre tenants.
- Los tests unitarios cubren respuestas esperadas, filtraciones, hechos faltantes y cálculo de percentiles.
- El README documenta el comando de evaluación desplegado y la arquitectura productiva bajo prueba.

5. Limitaciones

- El dataset contiene solo seis casos sintéticos y no permite una afirmación general de calidad.

2. La coincidencia léxica no mide corrección semántica más allá de los términos esperados.
3. El repositorio fuente es privado y todavía no se publicó el JSON público de resultados.
4. No existe aún un baseline ni una ablación que aísle estado del hilo, memoria vectorial y recuperación por grafo.

6. Próximo experimento

1. Ampliar el benchmark con paráfrasis, contradicciones, recuperación diferida, borrado y consultas adversariales de filtración.
2. Ejecutar ablaciones con solo hilo, solo vectores, solo grafo y recuperación combinada.
3. Publicar dataset anonimizado, manifiesto del entorno, resultados crudos y mediciones de costo.
4. Agregar revisión semántica humana junto con las verificaciones deterministas.

7. Fuentes y reproducibilidad

1. Repositorio privado de implementación, revisión registrada arriba

Revisión fuente: 1470f8f17163cd87538c5e12ae621658bd46afe3

Cita sugerida

Cardozo, Pablo. "Evaluación del aislamiento de memoria de largo plazo en agentes conversacionales multi-tenant." Working paper, version 0.1, 2026. <https://pablo.cardozo.com.ar/es/research/long-term-memory-isolation>.

Documento de trabajo independiente. Sin peer review.