¿Qué aprendió Gemma?

Cambiamos la guía de recepción.

La biblioteca de consulta sigue siendo la misma.

Imagina que Gemma trabaja en recepción. Cuando un cliente hace una pregunta, Gemma usa la guía para identificar la tarea. El código del enrutador decide qué fuente consultar, si hace falta.

1 Leer el mensaje
Mensaje actual y anterior

2 Identificar la tarea
Gemma usa la guía

3 Proponer una fuente
El enrutador aplica una regla

Un mismo tema, tres pasos distintos

«Facturación», después de un menú de departamentos
Dirigir a Facturación. No hace falta consultar registros.

«¿Cuánto cuesta el plan de fibra?»
Consultar el catálogo público de ventas.

«¿Cuánto debo?»
Proponer una consulta de facturación. Verificar primero el permiso de acceso.

¿Qué papel tuvo OpenAI?

Durante el desarrollo, OpenAI aportó respuestas de ejemplo. Las comparamos con las etiquetas guardadas y usamos los errores para revisar la guía. Después, Gemma funcionó por sí sola.

¿Qué cambiaría la biblioteca de consulta (RAG)?

Agregar una lista de precios al índice de ventas cambia la información que el sistema puede recuperar. Cambiar cuándo debe consultar ese índice modifica el enrutador. Eso último fue lo que hicimos.

¿Qué demostró la prueba piloto?

En dos rondas, el modelo local Gemma 2 2B acertó en los mismos 12 mensajes practicados. Sus pesos no cambiaron y no se ejecutó ninguna consulta. Su desempeño en conversaciones nuevas sigue sin verificarse.