Proyecto Final — R7 "Agentic Tools" (E1)

Agente explícito con herramientas (RAG híbrido + Calculator + Verification)

Jose Emiliano Pachas Mariluz
Curso: Procesamiento de Lenguaje Natural
Profesor: Cesar Lara
Repositorio: https://github.com/Aushzlten12/R7—Agentic-Tools-Development

13/12/2025

Resumen

Se implementó un sistema *agentic* que resuelve consultas sobre un plan de estudios (PDF) mediante herramientas (*tools*). El agente usa un **RAG** híbrido (BM25 + embeddings/FAISS), una **calculadora** y una herramienta de **verificación** para proporcionar respuestas. Se evaluó el RAG con Recall@k, MRR y latencia, además de medir el desempeño del agente completo con métricas de enrutamiento y precisión.

1. Objetivo

El objetivo del proyecto es diseñar un agente que utiliza herramientas como RAG, Calculator y Verification para responder consultas, con un flujo explícito. En este informe se presentan los resultados preliminares de evaluación del agente.

2. Arquitectura del sistema

El agente sigue el siguiente flujo:

- Preprocesamiento: analizar la consulta y seleccionar herramienta.
- Ejecución: invocar *RAG*, *Calculator* o *Verification*.
- Postprocesamiento: sintetizar la respuesta.
- Logging: registrar la ejecución.

El agente opera en modo explícito, tomando decisiones de enrutamiento de manera determinística según la consulta.

3. Herramientas implementadas

3.1. RAG Tool (híbrido)

RAG utiliza una combinación de BM25 para recuperación de texto basado en coincidencia exacta y embeddings con FAISS para recuperación semántica. Los resultados se combinan ponderando ambos métodos.

3.2. Calculator Tool

Herramienta para operaciones aritméticas básicas, como sumas, restas, multiplicaciones y divisiones.

3.3. Verification Tool

Herramienta de verificación que revisa los requisitos para matricularse en cursos. Utiliza un catálogo de cursos y un historial de estudiante simulado.

4. Evaluación del RAG (Recall@k, MRR y latencia)

Se evaluaron 11 preguntas utilizando Recall@k, MRR y latencia. Los resultados muestran que el modo híbrido (RAG) tuvo un rendimiento superior, con Recall@3 y MRR perfectos en la mayoría de consultas.

Modo	**Recall@3**	**MRR**	**Latencia prom. (s)**
Sparse (BM25)	100.00 %	1.0000	0.025
Dense (FAISS)	72.73 %	0.5758	0.013
Híbrido ($\alpha = 0{,}45$)	100.00 %	1.0000	0.012

Cuadro 1: Evaluación del RAG.

5. Evaluación del agente completo (flujo explícito)

Se evaluó el agente completo usando un LLM simulado (*MockLLM*). Las métricas de evaluación fueron: enrutamiento, precisión de salida y latencia. Los resultados muestran una latencia promedio en el rango de 12-32 segundos, dependiendo de la herramienta utilizada.

Tool	**N**	**Routing Acc.**	**Output Acc.**	**Latencia prom. (s)**
Calculator	5	100.00 %	100.00 %	11.654
Verification	5	100.00 %	0.00 %	12.156
RAG	5	100.00 %	100.00 %	31.915

Cuadro 2: Evaluación del agente completo.

6. Estado de E1 y pendientes

6.1. Completado

- Agente explícito en CPU con herramientas RAG, Calculator y Verification.
- Evaluación del RAG con métricas de Recall@k, MRR, latencia.
- Evaluación del agente con métricas de enrutamiento y precisión.

7. Próximos pasos (E2)

Para la entrega final se propone:

1. Agente autónomo con decisiones automáticas de enrutamiento.
2. Guardrails para validar inputs y limitar llamadas por consulta.
3. Observabilidad extendida: agregación de métricas y visualización.
4. Experimentación adicional: variar parámetros y ampliar el conjunto de pruebas.