FCEIA - UNR

XIX Jornadas de Ciencias, Tecnologías e Innovación UNR

Sistema de consultas académicas mediante agentes de IA implementado en un servidor local.

MARTIN, Raúl Lisandro[1]; CREPALDO, Daniel[2]; PACHER[3], Federico VARELA[4], Carlos; GHORGHOR[5], Javier; BAILON[6], Eduardo.

Laboratorio de Microelectrónica – FCEIA - UNR

[1] rlmartin@fceia.unr.edu.ar [3] fpacher@fceia.unr.edu.ar [5] javghor@fceia.unr.edu.ar

[2] crepaldo@fceia.unr.edu.ar [4] charlyv@fceia.unr.edu.ar [6] ebailon@fceia.unr.edu.ar

Palabras claves: Inteligencia Artificial. Educación. Consultas. Accesible..

INTRODUCCIÓN

Se presenta el desarrollo de un sistema de consultas académicas automatizado mediante una arquitectura agéntica de IA basada en Corrective RAG (CRAG) y Self-Reflective RAG (Self-RAG). Está diseñado para responder dudas teóricas a los estudiantes de un curso de grado basándose exclusivamente en la bibliografía oficial de una asignatura. La arquitectura propuesta supera las limitaciones de los pipelines de Generación Aumentada por Recuperación (RAG) tradicionales –que pueden sufrir de alucinaciones al procesar material no relevante de forma lineal– evaluando la relevancia del material recuperado y verificando la fundamentación de las respuestas generadas. Su objetivo principal es ofrecer soporte continuo 24/7 sin depender de infraestructura en la nube ni de hardware costoso con GPU dedicada, ejecutándose localmente en un servidor de prestaciones básicas.

METODOLOGÍA / MATERIALES Y MÉTODOS

Se aplicó para el desarrollo de este proyecto la metodología de investigación-acción. La misma implica métodos que incluyen el objetivo de investigar, y el de actuar/modificar algo de lo investigado. A partir de la información recopilada surge un nuevo plan de acción, configurando una estrategia recursiva.

DESARROLLO

La arquitectura del sistema combina CRAG y Self-RAG sobre un grafo de estados orquestado con LangGraph y LangChain. La interacción con los estudiantes se realiza de manera asincrónica vía correo electrónico o mediante una API construida con FastAPI que se conecta con bases de datos MySQL (para usuarios) y MongoDB (como cola FIFO para consultas). El flujo de procesamiento RAG sigue estos pasos clave: Recuperación y Evaluación: Se extraen los 4 fragmentos más similares del índice vectorial FAISS (con embeddings GPT4All). Un nodo evaluador clasifica el material como relevante, parcialmente relevante o irrelevante. Reformulación: Si la información es irrelevante, el agente reformula la pregunta y reintenta la búsqueda hasta dos veces; de persistir la insuficiencia, notifica explícitamente que no puede responder. Generación y Verificación: Las respuestas son generadas por el modelo Llama 3.1 8B, mientras que las decisiones de control y verificación son ejecutadas por Phi-3.5-mini (3.8B). Un nodo Self-RAG valida que el texto generado esté estrictamente anclado en la bibliografía recuperada antes de enviarlo por correo. El entorno de ejecución se montó sobre una PC de escritorio modesta (Intel i3 de 3.ª generación, 16 GB de RAM DDR3 y SSD de 400 GB) mediante Ollama, corriendo ambos modelos en forma local.

Figura 1.
Diagrama en bloques

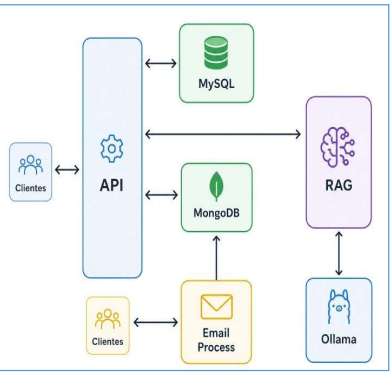
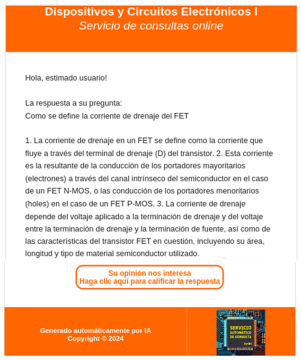


Figura 2.
Formato de respuesta


Dispositivos y Circuitos Electrónicos I
Servicio de consultas online

Hola, estimado usuario!

La respuesta a su pregunta:
Como se define la corriente de drenaje del FET

1. La corriente de drenaje en un FET se define como la corriente que fluye a través del terminal de drenaje (D) del transistor. 2. Esta corriente es la resultante de la conducción de los portadores mayoritarios (electrones) a través del canal intrínseco del semiconductor en el caso de un FET N-MOS, o las conducción de los portadores menoritarios (holes) en el caso de un FET P-MOS. 3. La corriente de drenaje depende del voltaje aplicado a la terminación de drenaje y del voltaje entre la terminación de drenaje y la terminación de fuente, así como de las características del transistor FET en cuestión, incluyendo su área, longitud y tipo de material semiconductor utilizado.

Su opinión nos interesa
Haga clic aquí para calificar la respuesta

Generado automáticamente por IA
Copyright © 2024

RESULTADOS y DISCUSIÓN

El sistema se encuentra en funcionamiento continuo y procesa las consultas satisfactoriamente. El tiempo medio de respuesta es de 48 a 52 minutos (frente a ~40 minutos de la versión lineal anterior) debido a las validaciones agénticas. Si se requieren iteraciones de reformulación, el tiempo alcanza entre 60 y 70 minutos, mientras que las consultas fuera de alcance se descartan tempranamente en 5 a 8 minutos. Esta variabilidad no impacta negativamente la experiencia del usuario dada la naturaleza asincrónica del correo electrónico. En cuanto a la calidad de los contenidos, las respuestas destacan por su precisión y apego al texto base, registrando únicamente detalles menores de terminología

CONCLUSIONES

Se logró implementar con éxito un asistente académico agéntico capaz de operar en hardware de bajo costo y sin conexión externa, garantizando respuestas precisas y drásticamente reducidas en alucinaciones gracias a los mecanismos de evaluación y auto-reflexión. El enfoque demuestra que es posible desplegar soluciones avanzadas de IA en instituciones educativas con presupuesto limitado. Como trabajo futuro, se contempla añadir capacidades multimodales para interpretar esquemas y diagramas, así como aplicar Adaptive RAG para ajustar la recuperación según la complejidad de cada pregunta.