

Proyecto de Generación de IA

Chatbot simple en ambiente local usando un modelo de generación de texto

Por: Fausto Morales

ffmogbaj@gmail.com

1. Propuesta

La propuesta de proyecto final consistió en realizar un Chatbot ejecutado en un entorno local empleando mi computadora personal MAC M2 usando un modelo de generación de texto como Deepseek.



Se propuso integrar por lo menos la arquitectura de front, APIS y el modelo. Y se comentó como opcional realizar un Fine Tuning, sin embargo se observó que los recursos para realizar esta actividad son elevados.



2. Propuesta final

Derivado al comentario de dificultad para un Fine Tuning se descartó la opción y en su lugar se exploró la generación aumentada por recuperación (RAG) que extiende las capacidades de los LLM a una base de conocimientos adicional.

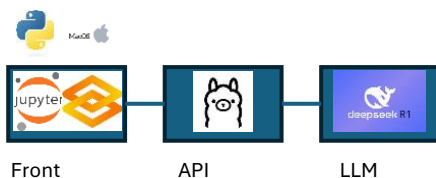
Para ello se realizaron 3 casos:

1. Chat bot incluyendo sólo LLM cumpliendo el requerimiento mínimo de la propuesta.
2. Chat bot incluyendo LLM + RAG con un archivo pdf.
3. Chat bot incluyendo LLM + RAG con una carpeta de archivos pdf.

El diagrama de esta solución quedó de la siguiente forma:

1. Chat bot incluyendo sólo LLM.

Lenguaje y Sistema Operativo

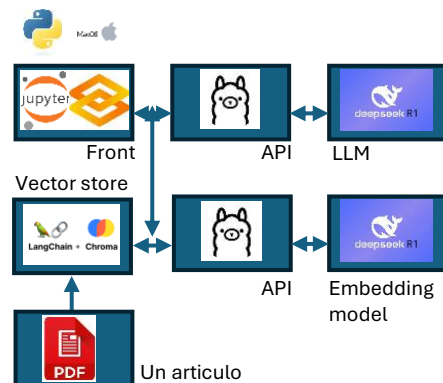


Se empleó una Macbook Air M2 con python3 dentro de jupyter. Este prototipo se corrió completamente en jupyter y el notebook sirve de front para su interacción con el usuario empleando el paquete: gradio.

Para poder interactuar con el modelo se empleó el software y APIS locales instalando Ollama que permitieron descargar e interactuar con el modelo más pequeño de Deepseek: "deepseek-r1:1.5b"

2. Chat bot incluyendo LLM + RAG con un archivo pdf.

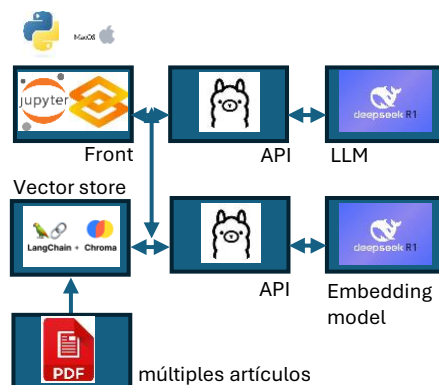
Lenguaje y Sistema Operativo



Para este nuevo modelo se integró el RAG a partir de un artículo de Wikipedia .pdf el cual se transformó y almacenó temporalmente en embeddings (o almacenamiento de vectores) a través de LangChain y Chroma, en conjunto con el API de Ollama y nuevamente DeepSeek: "deepseek-r1:1.5b" como modelo de Embeddings.

3. Chat bot incluyendo LLM + RAG con una carpeta de archivos.

Lenguaje y Sistema Operativo



Para esta última versión del modelo se creó una carpeta de archivos para nutrir el almacenamiento de vectores y al mismo tiempo se almacenó este elemento Chroma en la carpeta local: ./chroma

Nota: Como pre-requisito para el funcionamiento correcto de estos componentes fue necesario instalar con brew: poppler y tesseract, así como diferentes dependencias Python.

3. Resultados

Para poder testear los chatbots se les preguntó a los 3 modelos la misma pregunta: “¿Cuál es la diferencia entre la comisión nacional para la protección y defensa de los usuarios de servicios financieros y el banco nacional de México?”.

1. Para el modelo 1, la respuesta es incorrecta y se encuentra que el modelo original no tenía información correcta de esta pregunta:

prompt

¿Cuál es la diferencia entre la comisión nacional para la protección y defensa de los Usuarios de servicios financieros y el banco nacional de México?

Clear Submit

output

The differences between the national cooperation for protection and defense of service users of financial institutions in the USA (referred to as "comision nacional" or "national cooperation") and the bank cooperators of Mexico (the "beta" cooperators) can be summarized as follows:

1. **Scope of Protection**:
 - **US National Cooperation**: This organization focuses on protecting services from external risks such as fraud, threats, and regulatory issues. It works with other national Cooperaciones in the U.S., often through a commission system, to ensure financial institutions operate safely.
 - **Mexican Bank Cooperators**: These focus specifically on their own banks, serving public needs. They concentrate on internal operations like transparency, fair lending practices, and compliance with Mexican regulations.

2. Para el modelo 2, se empleó la página de la [wikipedia de condusef](#). Esté fue su respuesta, también errónea:
- 3.

question

¿Cuál es la diferencia entre la comisión nacional para la protección y defensa de los usuarios de servicios financieros y el banco nacional de México?

Clear Submit

output

La diferencia principal entre la **Comisión Nacional para la Protección y Defensa de los usuarios de Servicios Financieros (CNPF)** y el **Banco Nacional de México (CNBV)** está en su focus y función central:

1. **Comisión Nacional para la Protección y Defensa de los usuarios de Servicios Financieros (CNPF)**:
 - Un organismo descentralizado, activo desde el 1999.
 - Opera como financiera, analítica, hola, duda y ahorro, asegurando la protección y defensa de los usuarios de servicios financieros.
 - Tiene funciones específicas en el contexto de la finanzas, incluyendo something to do with social media and online presence.
2. **Banco Nacional de México (CNBV)**:
 - Centraliza en la monitorización y regulación de las operaciones financieras.
 - Científica en financiera, analítica, hola y duda, pero también puede tener functions related to something like social media presence.

4. Por último para el modelo 3, se empleó la página de la [wikipedia de condusef](#), la página de [wikipedia de banamex](#) y una página aleatoria del banco de México sobre [el Artículo 48 Bis 2 de la Ley de Instituciones de Crédito](#). Siendo correcta su respuesta aunque en inglés:

query_text

¿Cual es la diferencia entre la comisión nacional para la protección y defensa de los usuarios de servicios financieros y el banco nacional de México?

Clear Submit

output

The national protection and defense commission for financial service users in Mexico focuses on regulating and overseeing financial services, ensuring legal compliance and protecting users from illegal activities. In contrast, the national bank of Mexico is a leading financial institution that operates by managing day-to-day financial tasks such as banking, loan provision, public debt management, tax management, and economic stability. The protection commission emphasizes legal oversight, whereas the bank handles operational responsibilities.

Flag

Use via API · Construido con Gradio

Nota: e a veces la respuesta era incorrecta ya que existían variaciones si se volvía a preguntar

4. Conclusiones

Fue posible realizar una demo de: un chatbot simple en ambiente local empleando un modelo de generación de texto por LLM (Large Language Model o Modelo de Lenguaje de Gran Escala) a partir de tecnologías que uso por primera vez como oLlama y LangChain, sin embargo, requiere mejoras para poder aplicarlo en entornos productivos y empresariales cómo, por ejemplo:

- Lograr tener una base de conocimientos especializada más amplia y optimizaciones para escoger información relevante.
- Mejorar el entorno de ejecución del modelo, implementando un módulo óptimo, así como sistemas frontales y de APIS especializados y dedicados con seguridad productiva.
- Realizar monitores sobre las respuestas automáticas generadas, así como una gestión de las respuestas incorrectas.

Puedes consultar el notebook completo en: [chatbot.ipynb](#)

5. Referencias

Referencias para este proyecto:

<https://aws.amazon.com/es/what-is/retrieval-augmented-generation/>

<https://www.datacamp.com/es/tutorial/deepseek-r1-ollama>

https://github.com/tonykipkemboi/ollama_pdf_rag/blob/main/notebooks/experiments/updated_rag_notebook.ipynb

<https://github.com/pixegami/rag-tutorial-v2/tree/main>

https://github.com/fcori47/rag_basico/blob/master/Clase%206%20-%20Final.ipynb

<https://aws.amazon.com/es/what-is/retrieval-augmented-generation/>

DeepSeek-AI (2025) presenta *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* ([arXiv:2501.12948])(<https://arxiv.org/abs/2501.12948>)).