LLM y Prompt Engineer

Aprendizaje de Máquinas
Primer Semestre 2025

Instrucciones de Vuelo

El contenido que estás a punto de ver representa en gran medida la cúspide de la evolución humana en su búsqueda por IA accesible al público global.

Bastantes partes de la época de IA generativa en que nos encontramos toma provecho de conceptos que verás reflejados en esta presentación

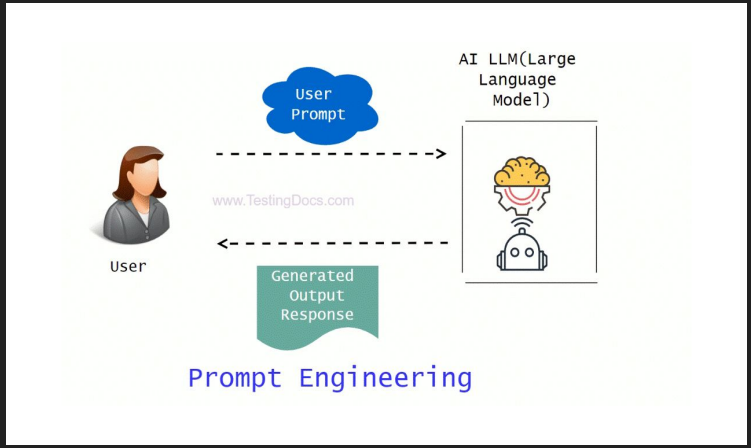
Por tanto pon mucha atención la material que revisaremos el día de hoy, dentro de esta diapositiva se encuentra la segunda parte del hito 1 parte 2 del curso que es evaluado, además se encuentra un laboratorio especial que aporta al 10% de la nota de se hito 1 parte 2 .

Que es Prompt Engineer?

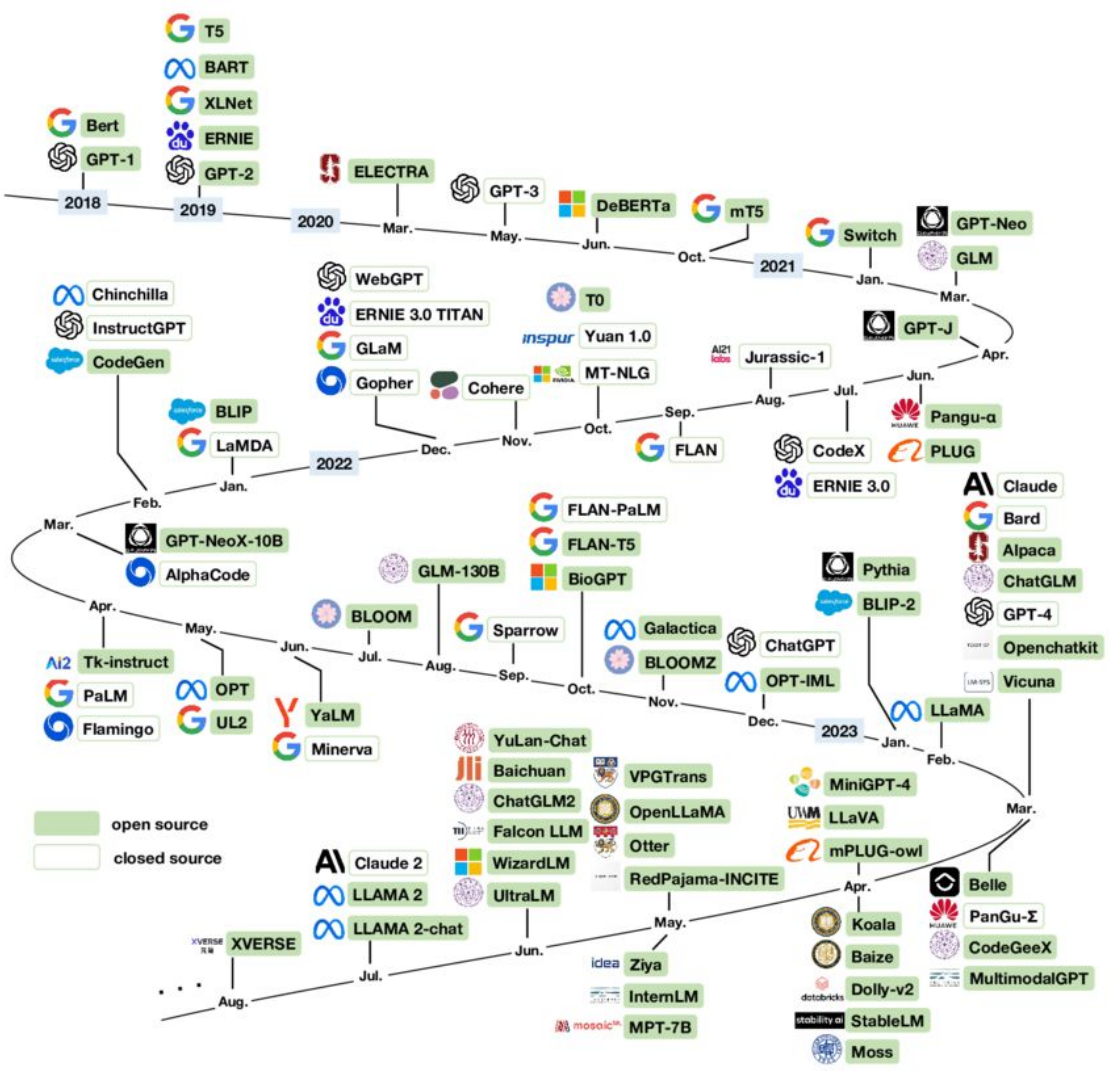
Prompt Engineering es el proceso de diseñar, estructurar y optimizar instrucciones (prompts) que se entregan a un modelo de lenguaje como GPT, Claude o LLaMA, para obtener respuestas útiles, precisas y alineadas con un objetivo específico.

Sin embargo llegar hasta el prompt engineer poch ha sido un proceso de cambios constantes.

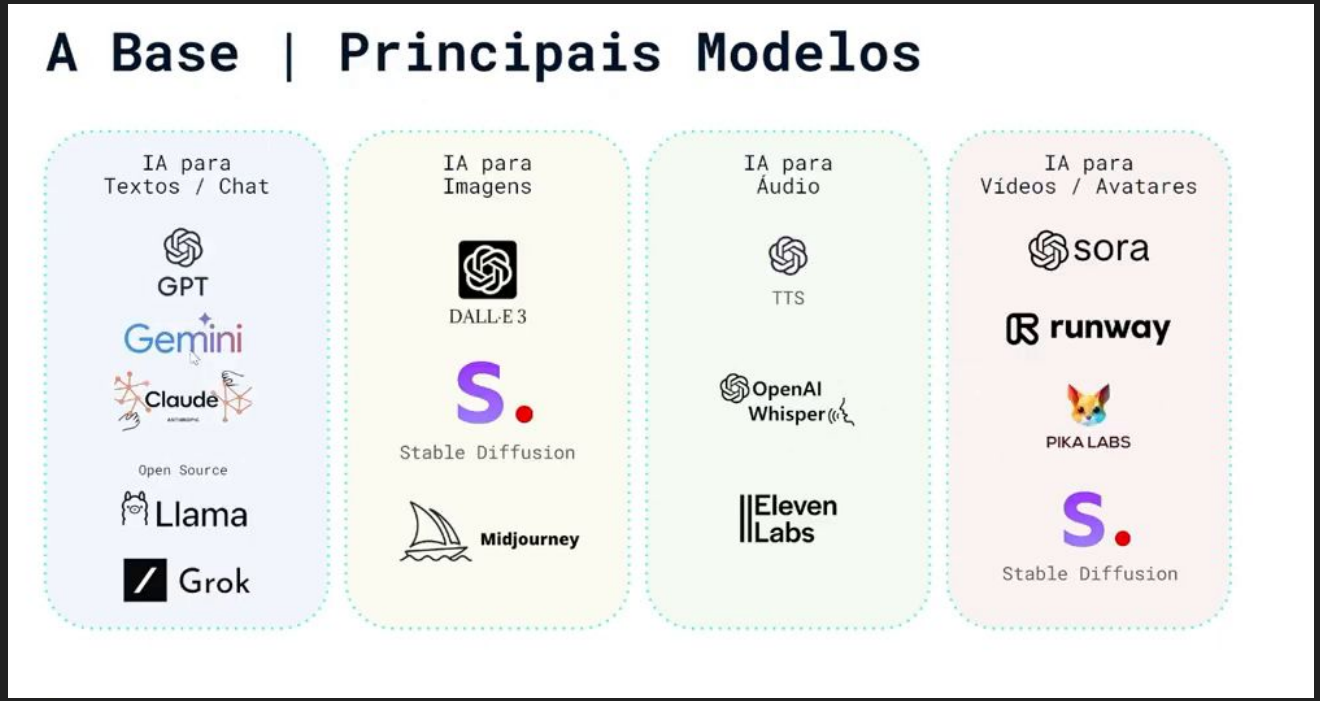
Los modelos tipo GPT, utilizan varias tecnologías en su arquitectura interna para lograr las respuestas que vemos expresadas en la actualidad.



TODO SE TRATA DEL LENGUAJE



Resumen actual modelos



El núcleo de las capacidades de todas estas IA es el NLP desde una perspectiva general

¿ Es suficiente solo con el modelo de lenguaje ?

¿ Es suficiente solo con el modelo de lenguaje ?

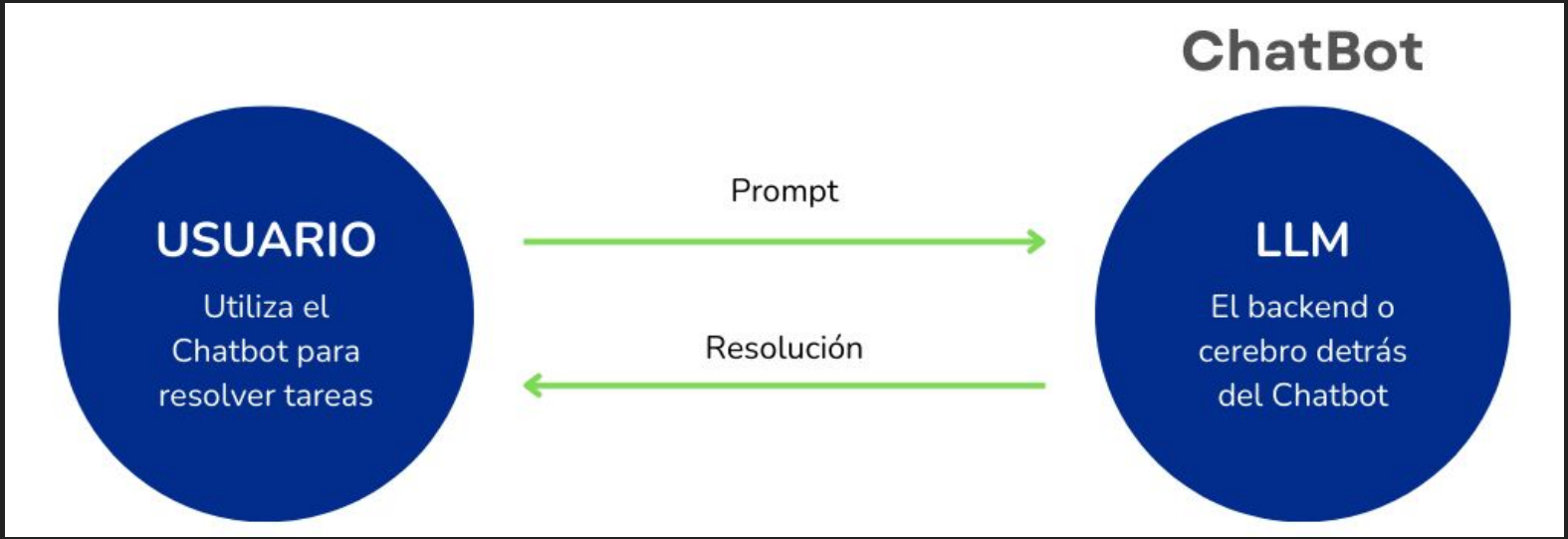
Hablemos de agentes en IA

Los AI Agents son programas creados para **percibir su entorno y tomar decisiones automáticas** utilizando modelos de inteligencia artificial. Por lo tanto, no es una IA con la que tienes que interactuar como Chat GPT, sino programas diseñados para realizar tareas basándose en su entorno.

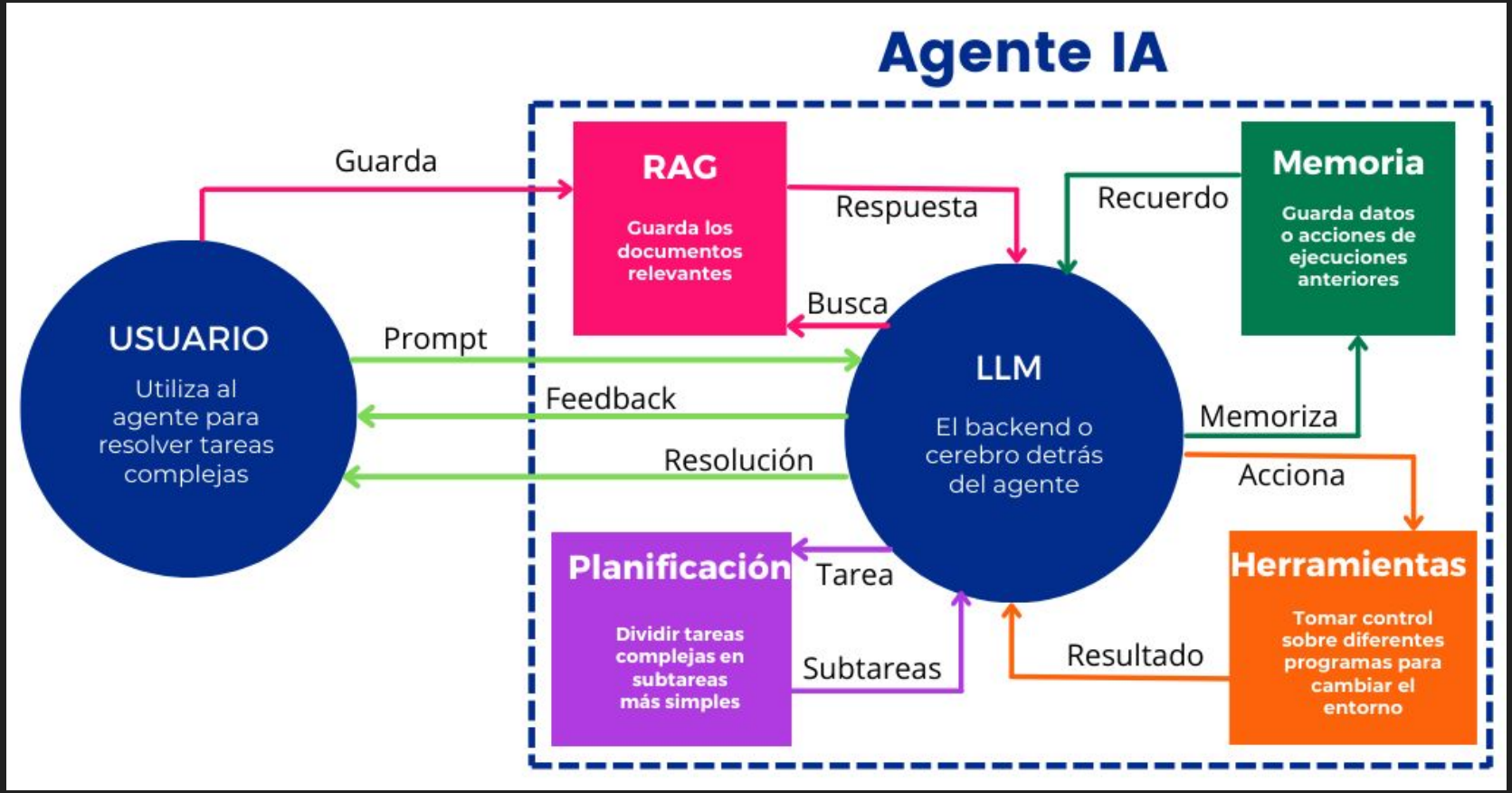
Hablemos de agentes en IA

Sin embargo también hay una expresión de agentes en IA tipo LLM como llama o Gpt, en donde podemos especializar su funcionamiento para una tarea o tema en particular

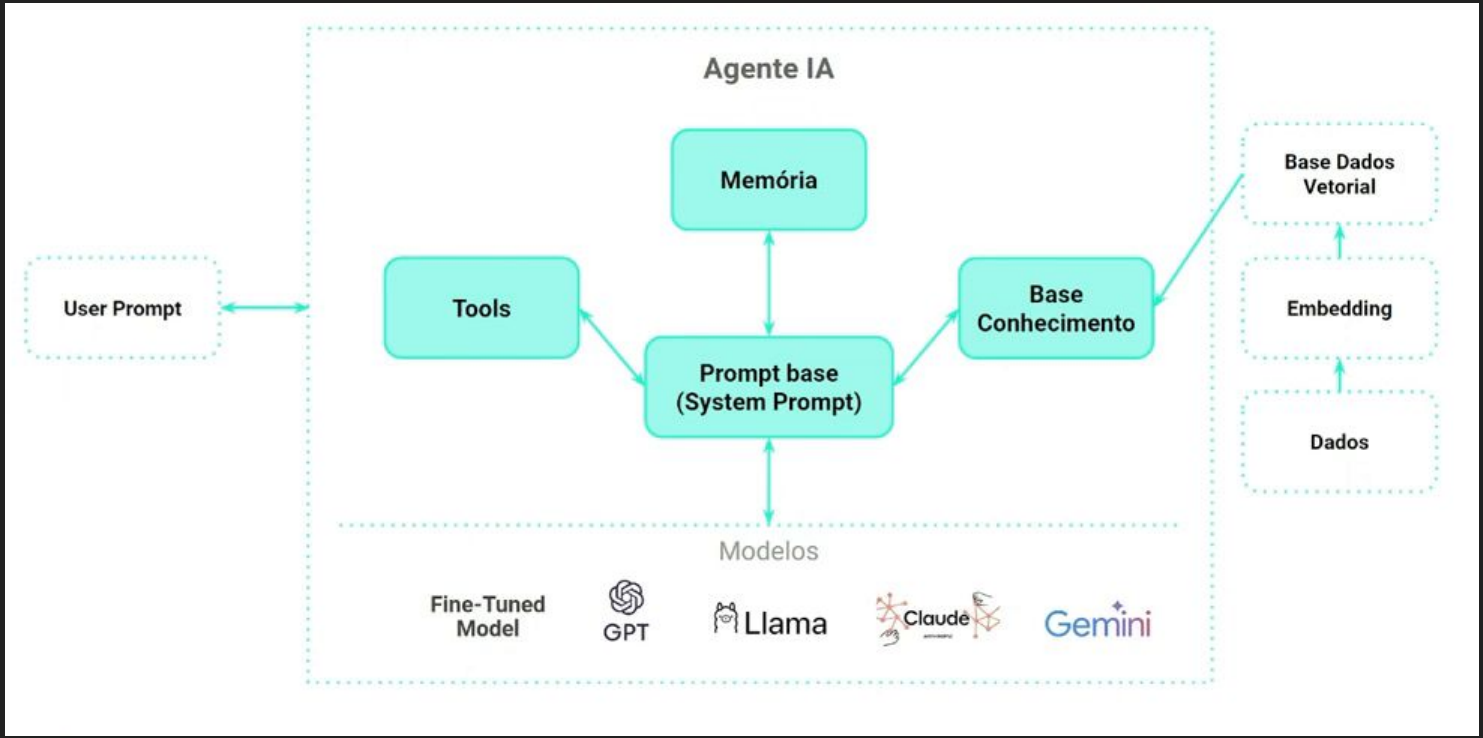
Tradicionalmente tenemos esto



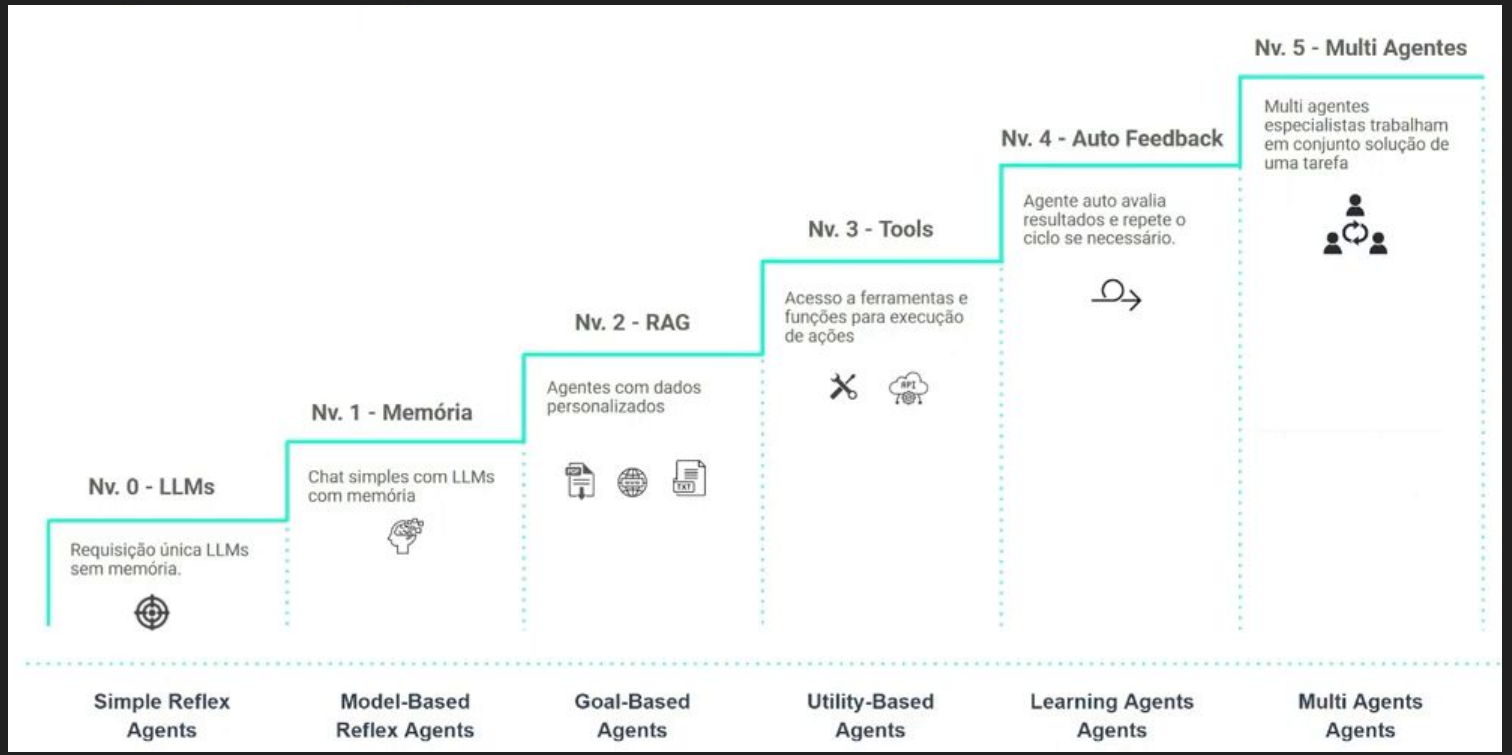
Pero al construir un agente con LLM …



Agente


Agente IA
Memória
User Prompt
Tools
Prompt base
(System Prompt)
Base
Conhecimento
Base Dados
Vetorial
Embedding
Dados
Modelos
Fine-Tuned
Model
GPT
Llama
Claude
Gemini

Herramientas y técnicas complementarias



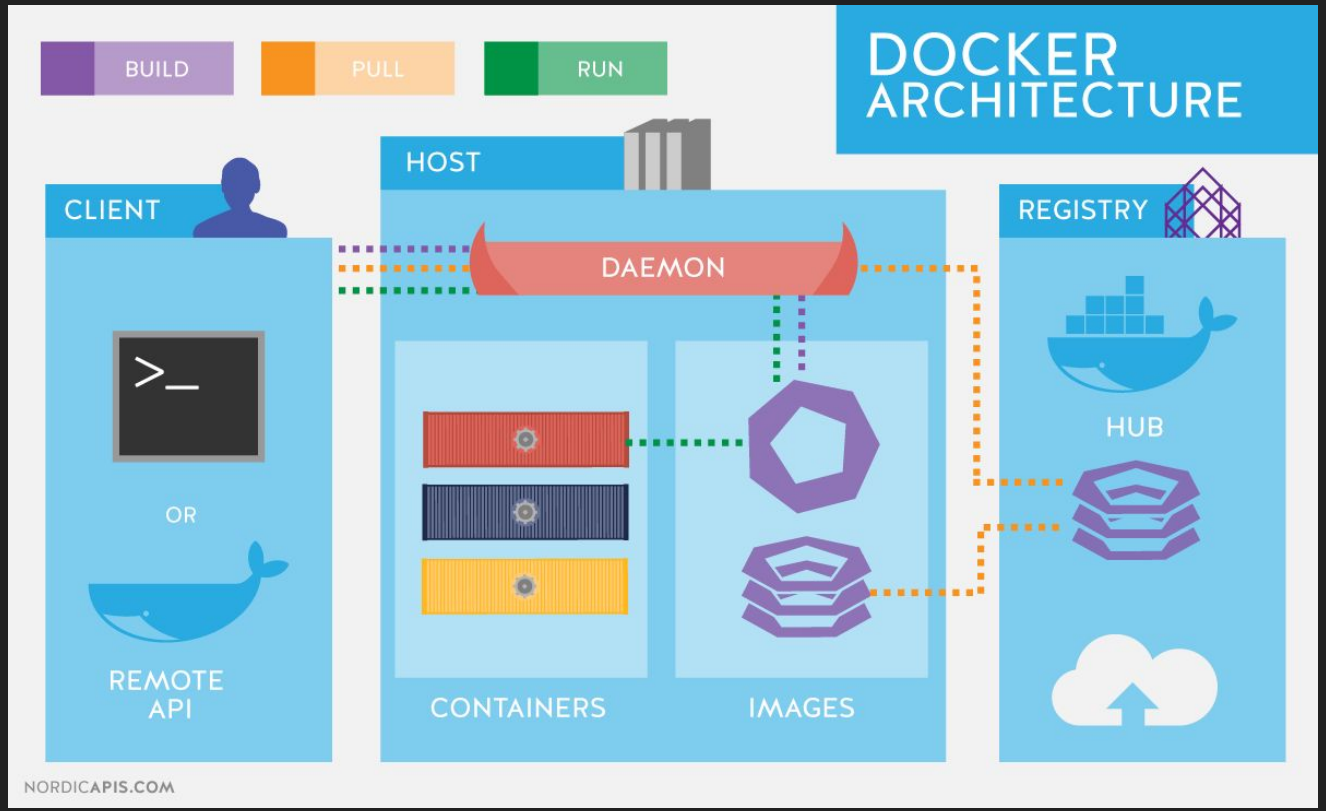
Como puedo utilizar LLM ?

Yo creo que ya lo sabes pero como usuario final … qué sucede si queremos ser un usuario especializado o desarrollar software utilizando e integrando funcionalidades basadas en este tipo de modelos?

Transfer Learning, LLM y modelos locales

Ollama permite correr de forma local diversos modelos de LLM, la gran diferencia en términos de arquitectura general es que ya cuenta con un implementaciones para consumir el modelo desde un endpoint local como api o desde terminal

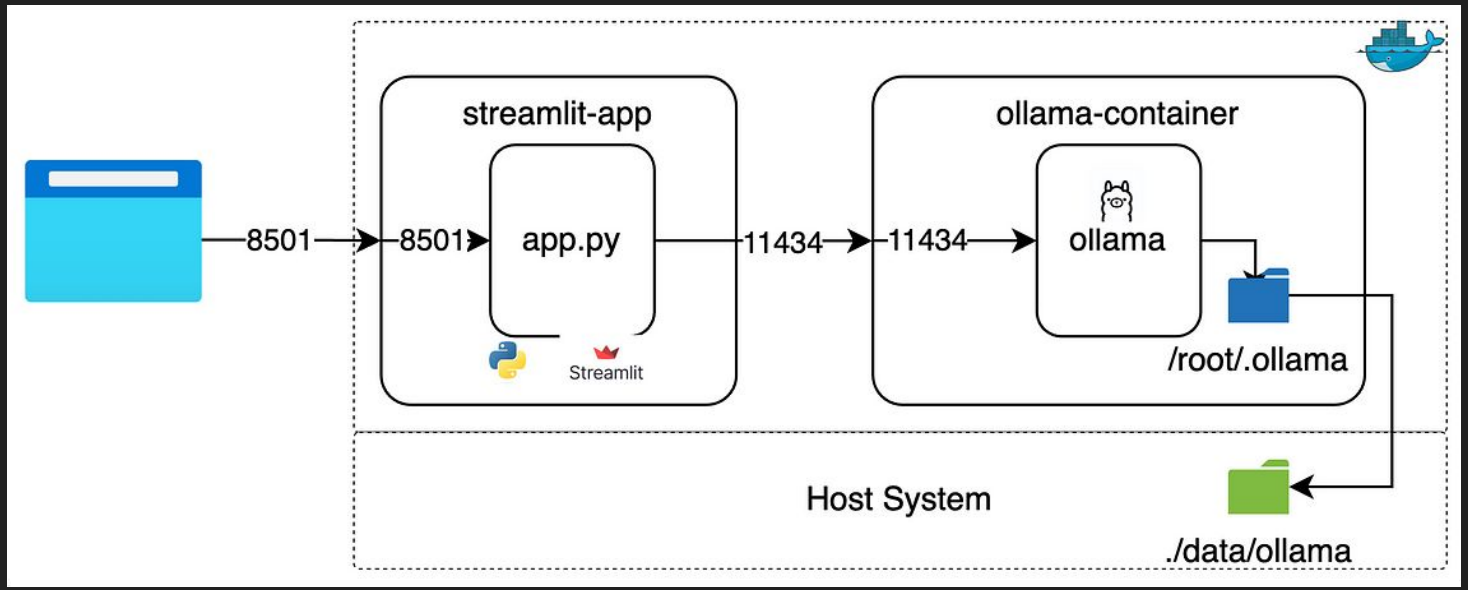
Hablemos brevemente de Docker



Hablemos brevemente de Streamlit

https://streamlit.io/

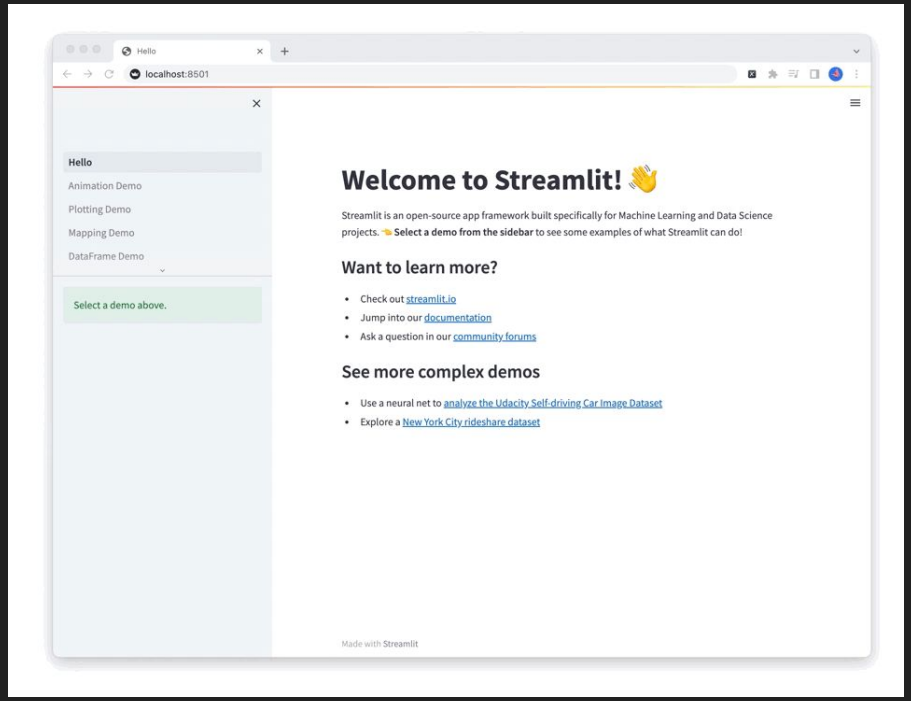
Arquitectura utilizando modelo Local - LAB



https://abvijaykumar.medium.com/ollama-build-a-chatbot-with-langchain-ollama-deploy-on-docker-5dfcfd140363

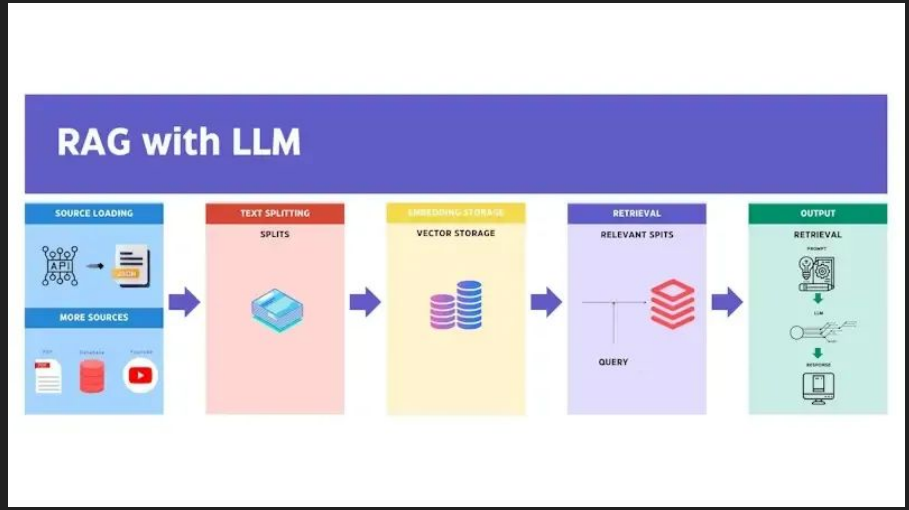
Arquitectura utilizando modelo Local

La arquitectura anterior está bien pero para pruebas locales e intentar integrar LLM como un experimento interesante, sin embargo escalaría pesimo si pensaramos en desplegar esto a usuarios a gran escala.



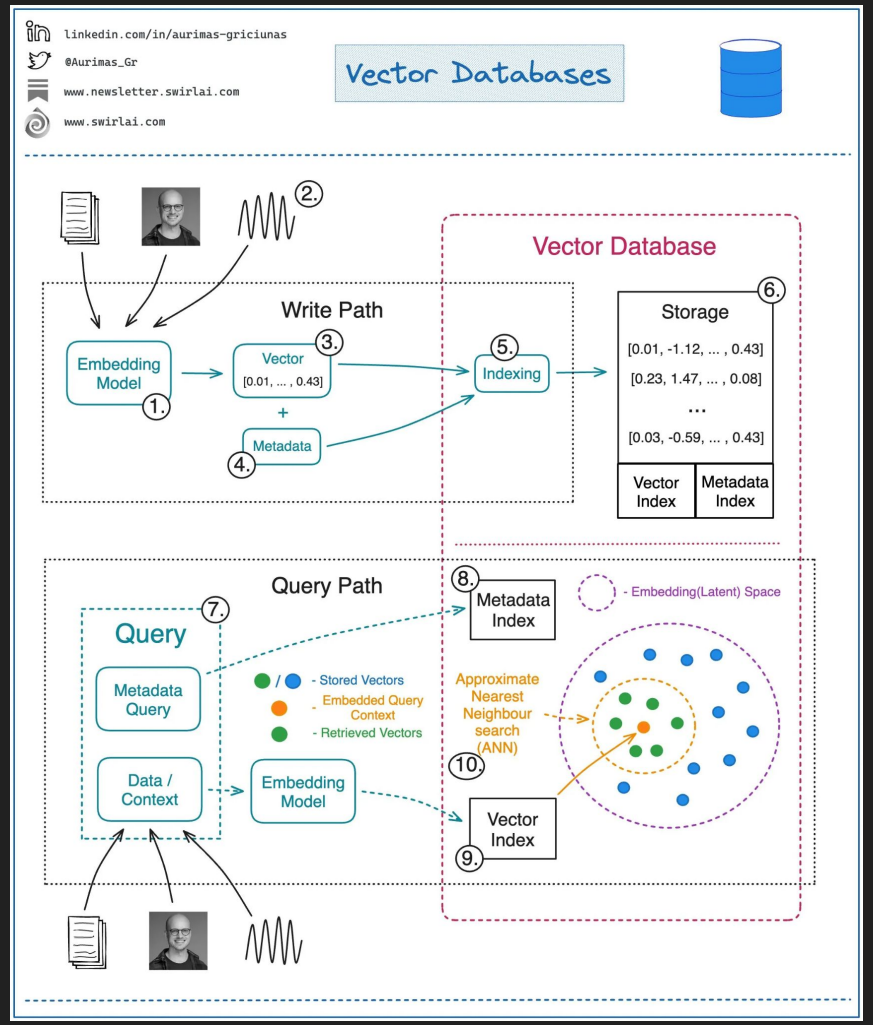
Como podemos especializar una LLM ?

La generación mejorada por recuperación (RAG) es el proceso de optimización de la salida de un modelo lingüístico de gran tamaño, de modo que haga referencia a una base de conocimientos autorizada fuera de los orígenes de datos de entrenamiento antes de generar una respuesta.



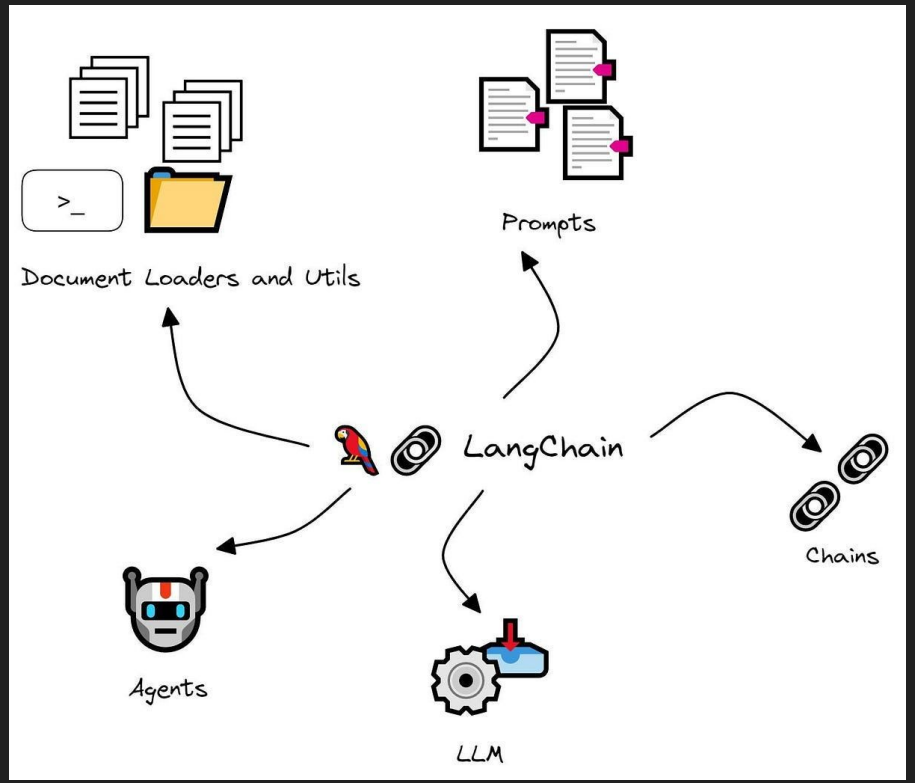
Base de datos vectoriales - Chroma

Las bases de datos vectoriales proporcionan la capacidad de almacenar y recuperar vectores como puntos de alta dimensión. Añaden capacidades adicionales para la búsqueda eficiente y rápida de los vecinos más cercanos en el espacio N-dimensional. Por lo general, se basan en índices de k vecinos más cercanos (k-nearest neighbor, k-NN) y se crean con algoritmos como los algoritmos Hierarchical Navigable Small World (HNSW) e Inverted File Index (IVF). Las bases de datos vectoriales proporcionan capacidades adicionales (como la gestión de datos, la tolerancia a errores, la autenticación y el control de acceso) y un motor de consultas.

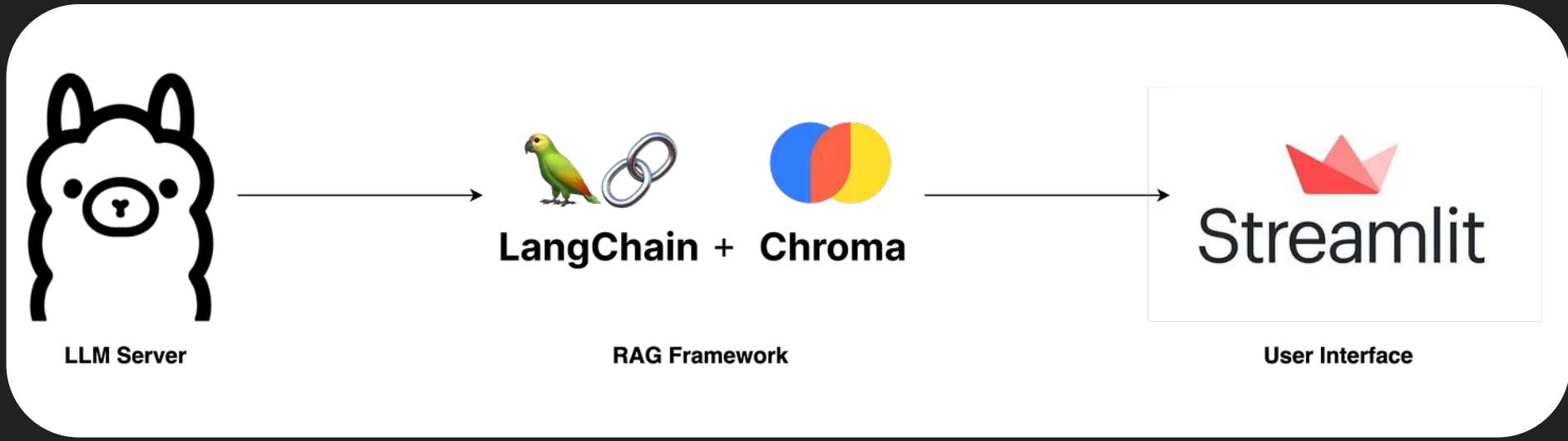


Langchain 🦜

LangChain es un marco de trabajo de código abierto para crear aplicaciones basadas en modelos de lenguaje de gran tamaño (LLM). Los LLM son grandes modelos de aprendizaje profundo entrenados previamente con grandes cantidades de datos que pueden generar respuestas a las consultas de los usuarios, por ejemplo, responder preguntas o crear imágenes a partir de peticiones basadas en texto.

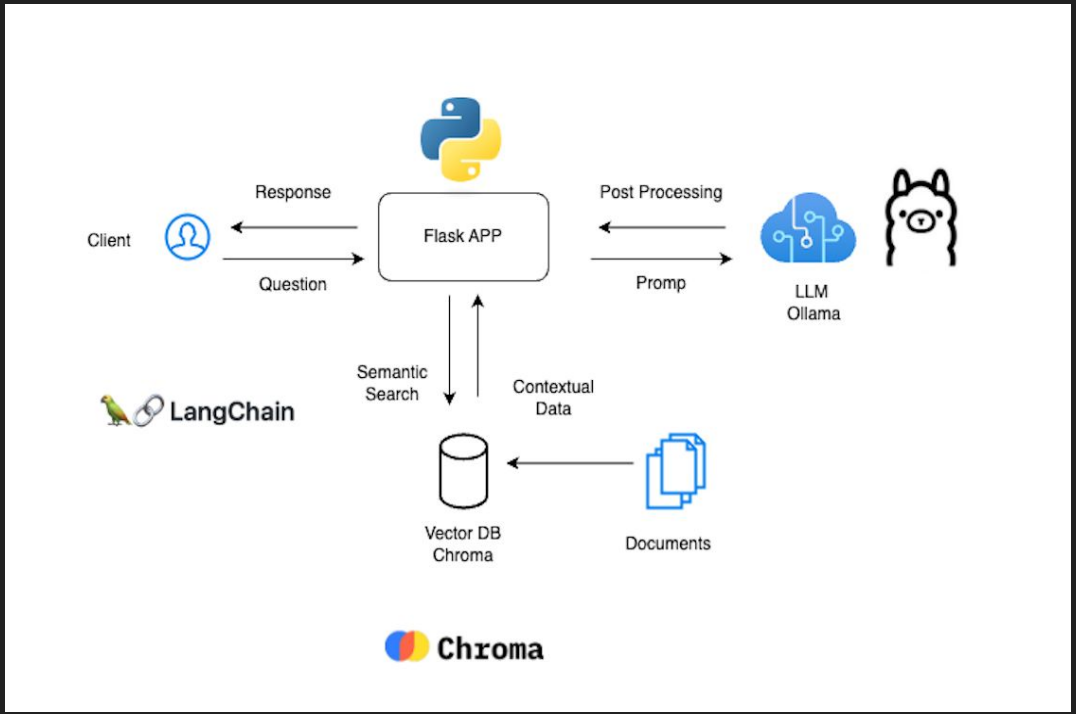


Como podemos especializar una LLM ?



https://thetechnician.substack.com/p/build-your-own-rag-and-run-it-locally

Hagamos una api :)



https://imosmanyilmaz.medium.com/build-your-free-ai-assistant-rag-python-ollama-45ba8e8fd847

Ahora Te Toca a Ti

Identifica un contexto en el que podamos utilizar las conceptos presentados:

- Profundiza sobre los conceptos vistos y responde a los siguientes puntos:
- Para que se utiliza el concepto de **`yiel`** en las api, que diferencias existen con una respuesta http común?
- Qué rol cumple el uso de GPU para procesamiento en LLM, qué valor tiene como empresa Nvidia en este punto ?
- Que formatos funcionan mejor para hacer Rag? porqué?, **pdf, md, docx, etc.**

- Propón casos de uso realistas para las temáticas vistas, ejemplo: análisis y consulta de documentos legales, documentos médicos, documentos gimnasio, etc.
- Genera una presentación donde podamos ver la definición del caso a trabajar y el funcionamiento de una app usando **streamlit**, **rag y ollama**, para el caso de uso propuesto.
    - **El código y datos debe estar disponible en un repositorio github publico.**
- Utiliza ollama para probar distintos modelos de llm al menos 2.
    - ej: Mistral, Llama, phi, etc.
    - Realiza un cuadro comparativo sobre los modelos utilizados que fortalezas tienen, que debilidades tienen, porque?
    - Considera que para los modelos más grandes la memoria Ram es un factor importante, con menos de 8 gb será muy difícil realizar pruebas, en los laboratorios de la **UNAB** contamos con 16 gb :) .
- Presenta tus resultados al curso la próxima semana, Lunes 21, desde las 10:20, el primer bloque de clases será para ajustes.