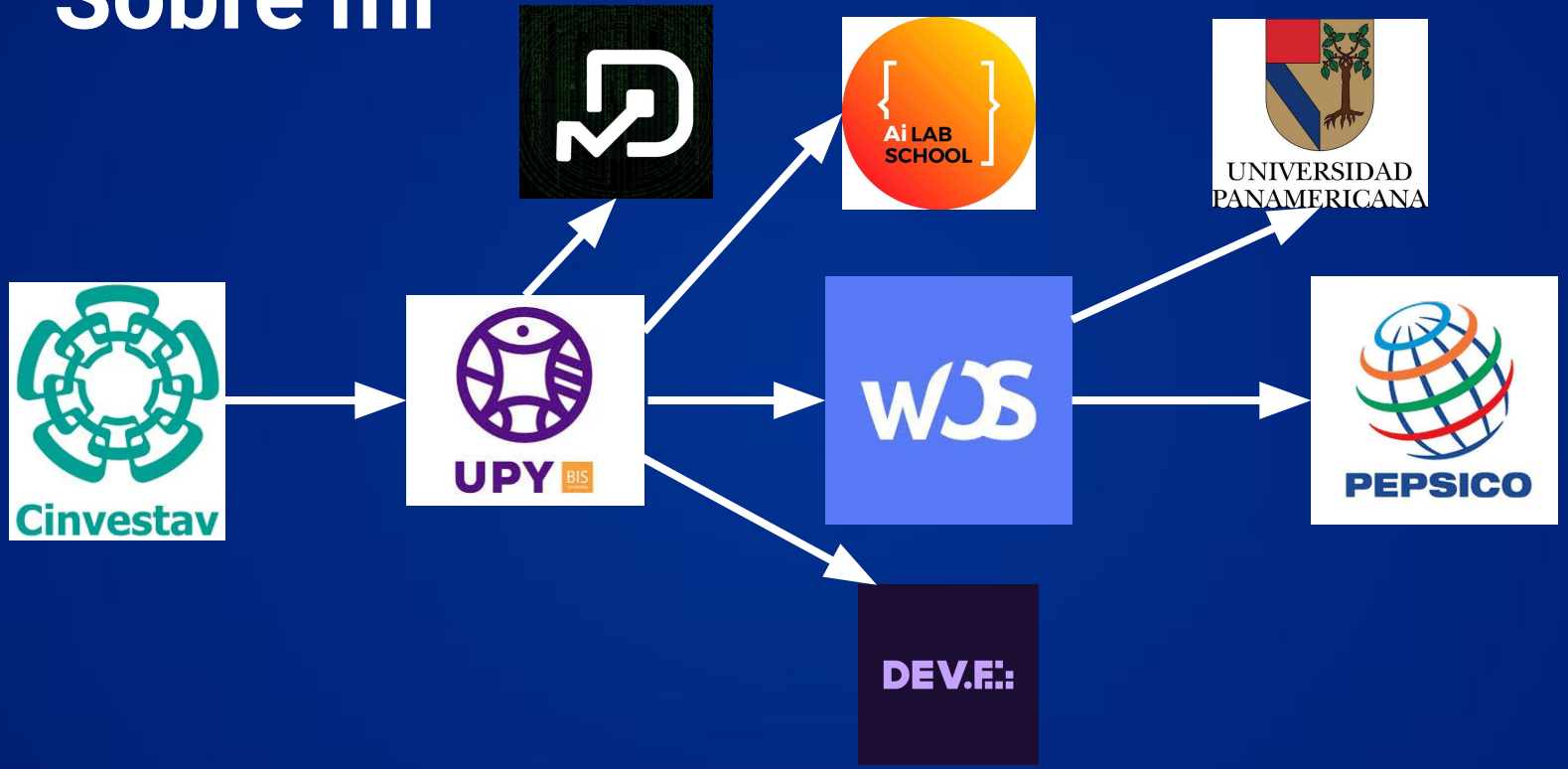
Introducción a Transformers y Large Language Models

Juan Vázquez Montejo


http://sg1.run/-i

Sobre mi


Ai LAB SCHOOL
UNIVERSIDAD PANAMERICANA
Cinvestav
UPY
PEPSICO
DEV.F::

Agenda del taller

Transformers y LLM

Desarrollo de código con ejemplo de RAG

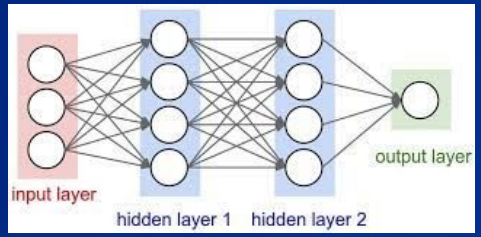
Comparación de langchain vs haystack vs llama index

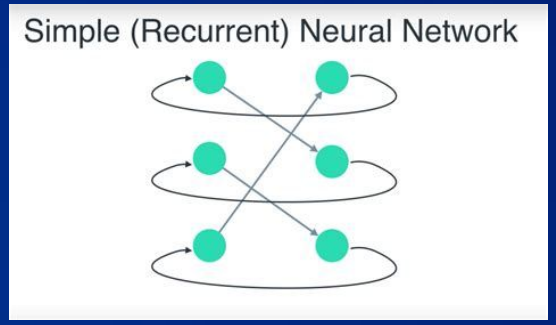
preguntas

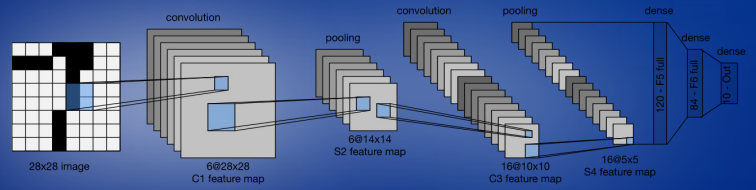
Cierre y recursos

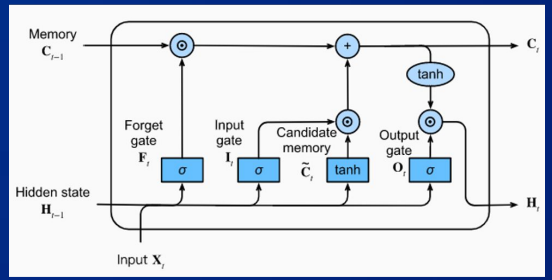
¿Qué son los transformers?

Arquitecturas: DNN, CNN, RNN (LSTM GRU)



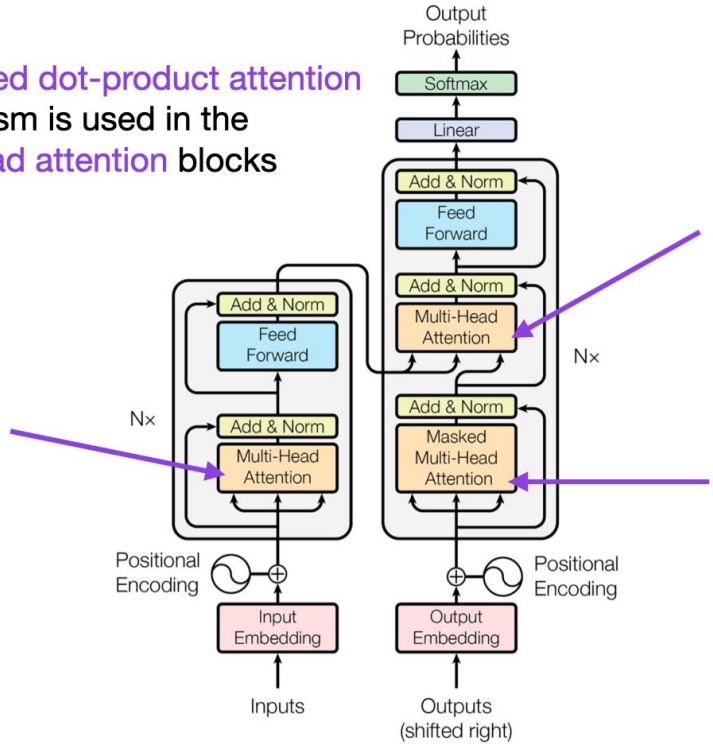






Mecanismo de auto-atención

The scaled dot-product attention mechanism is used in the multi-head attention blocks



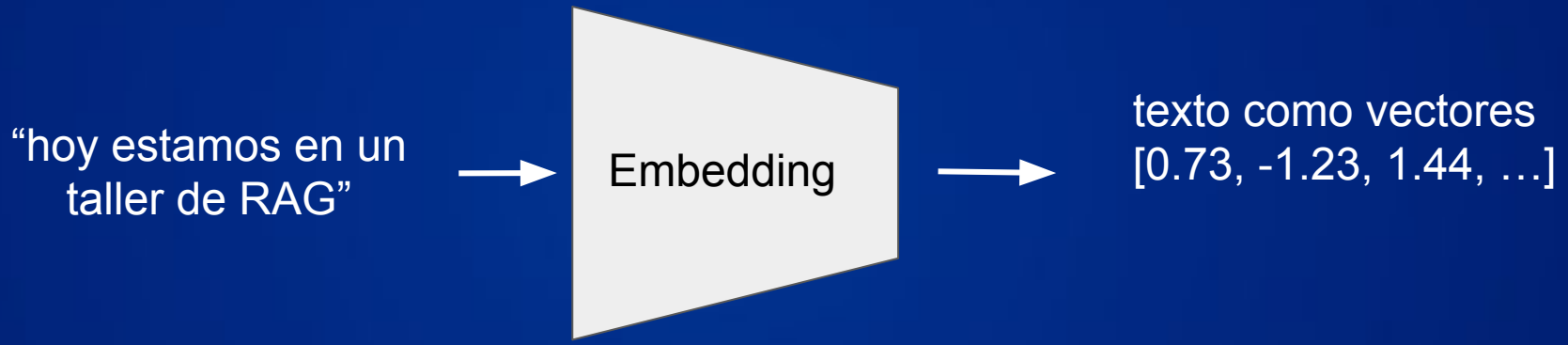
Source: “Attention Is All You Need” (https://arxiv.org/abs/1706.03762)

¿Qué son los LLM?

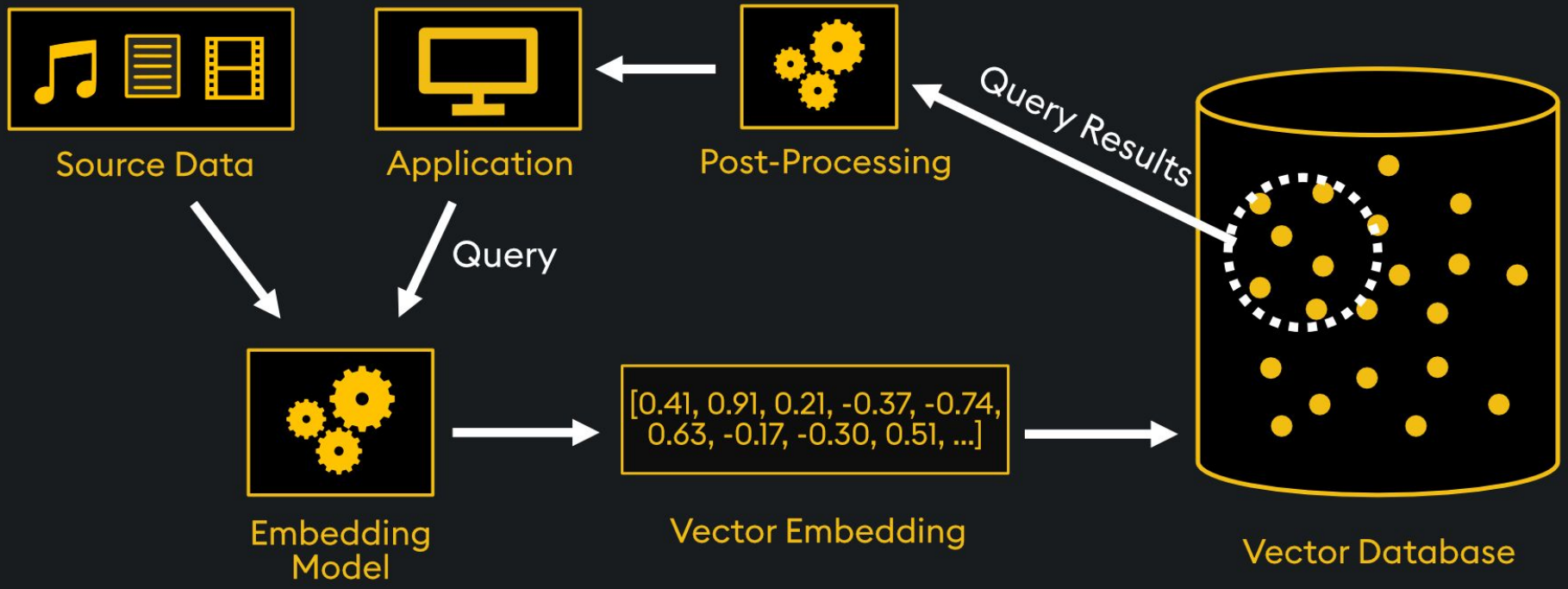
Large Language Models (LLMs)

- Distribución de probabilidad sobre cadenas de texto.
- Redes neuronales con miles de millones de parámetros.
- Casos de uso: responder preguntas, traducción, generación de resúmenes, generación de código, etc

Textos -> vectores



vector databases y embeddings

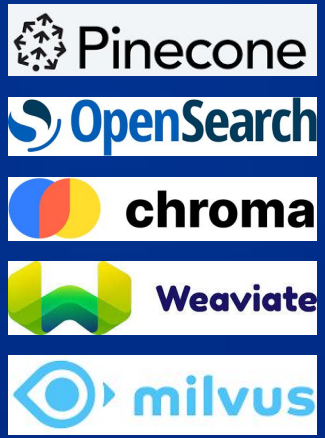


https://thedataquarry.com/posts/vector-db-1/

Ecosistema actual de LLM

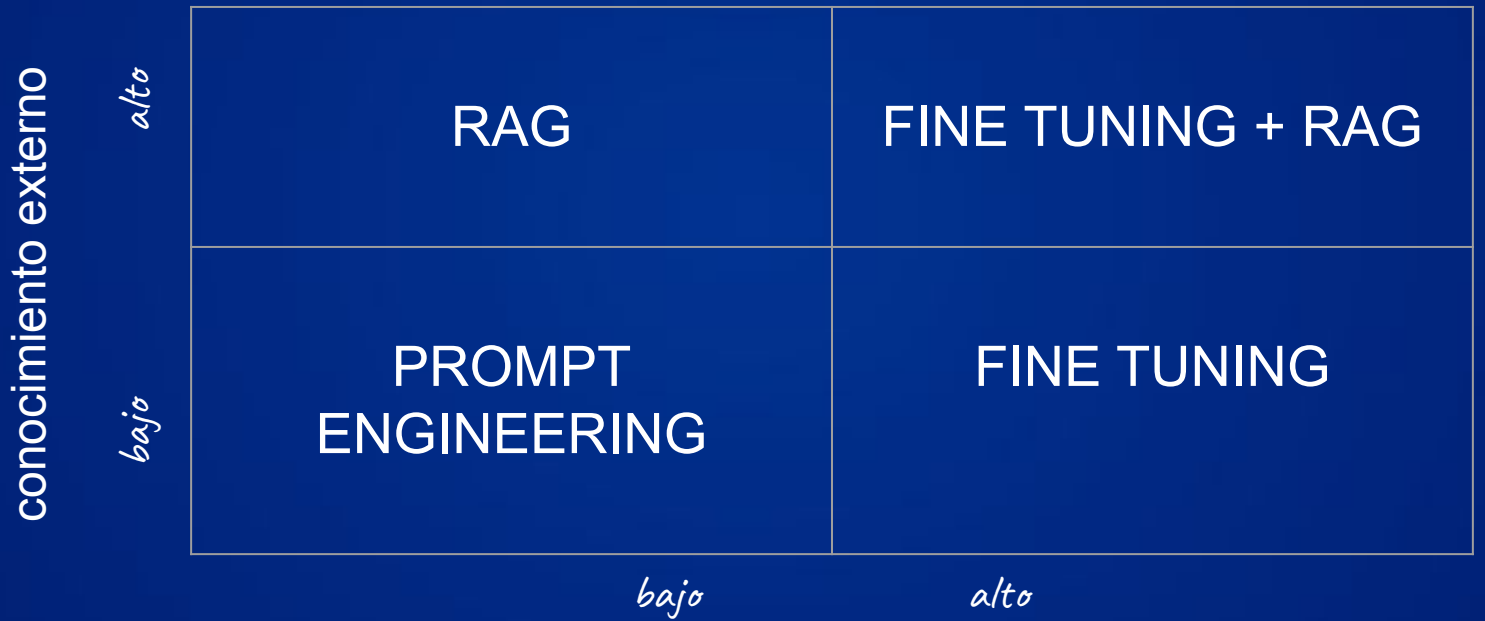
LLMs y Embeddings

Vector DB



AI Frameworks

Cómo abordar la IA generativa y personalizar los transformers/LLM



fuente: LinkedIn

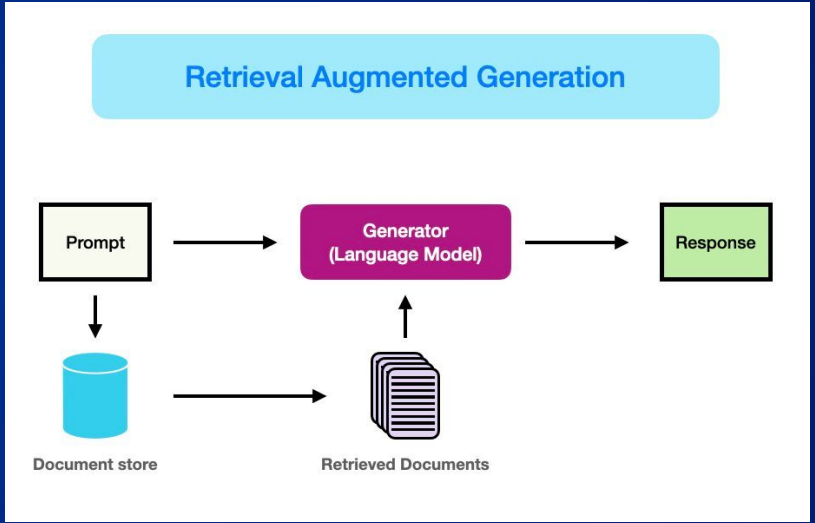
¿Qué es RAG?

Retrieval Augmented Generation (RAG)

Retrieval (Recuperador): una interfaz que devuelve documentos dada una petición.

Augmented, Prompt augmentation (o Aumento de mensajes): proporcionar contexto o información adicional en el mensaje para mejorar el rendimiento.

Generation: (la IA generativa, LLM, GPT, etc)



Notebook en colab

Contacto: LinkedIn y Twitter

Agradecimientos:

Muchas gracias por su atención

Data.dλy