

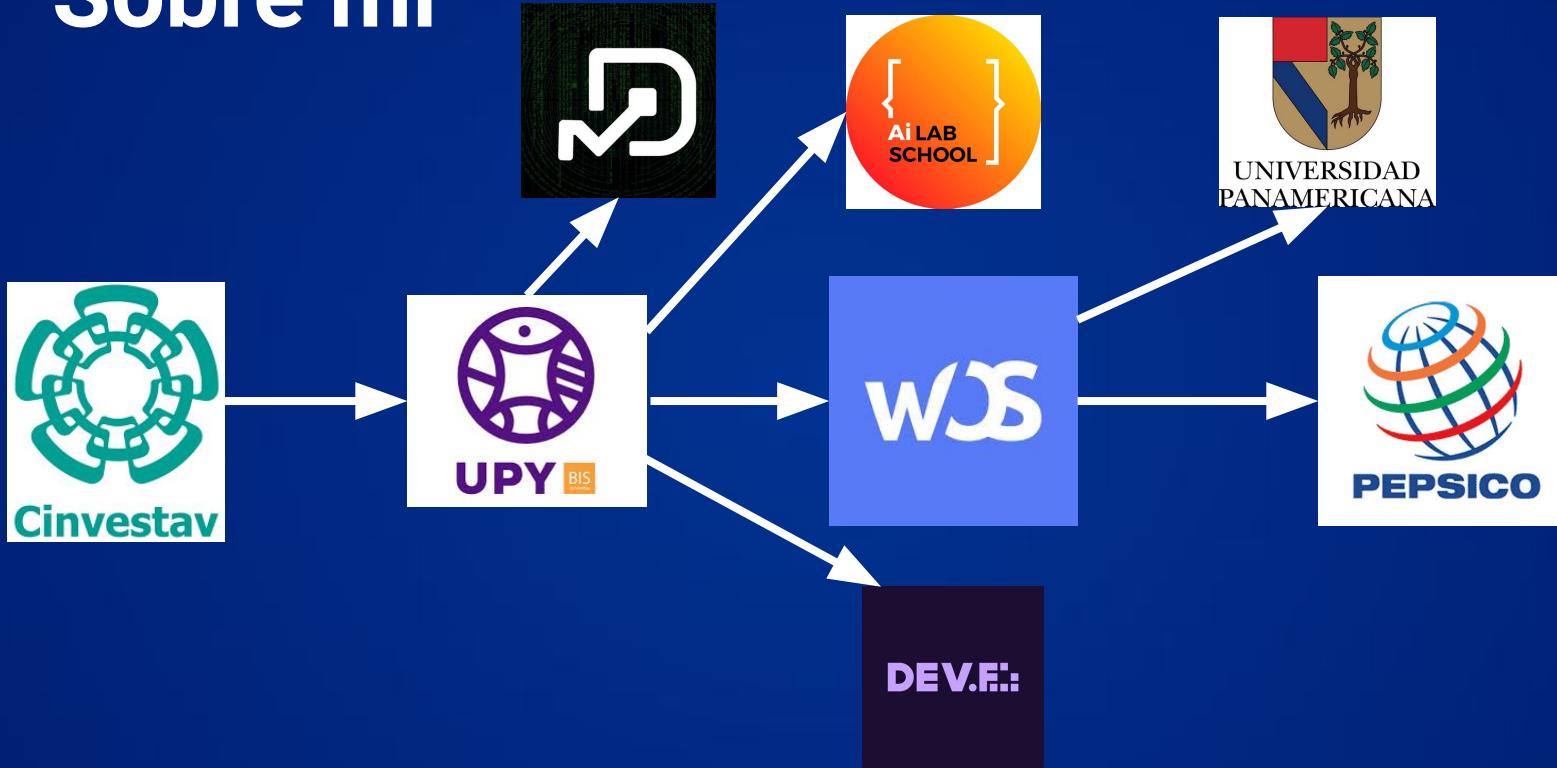
Introducción a Transformers y Large Language Models

Juan Vázquez Montejo

The background is a dark blue space filled with dynamic light trails and circuit-like patterns. Bright blue and red lines streak across the frame, creating a sense of motion and depth. Some lines form recognizable circuit board traces, while others are straight, radiating paths. Small, glowing blue and red dots are scattered throughout, resembling data points or nodes in a network. The overall effect is a high-tech, digital aesthetic.

<http://sg1.run/-i>

Sobre mi



Agenda del taller

Transformers y LLM

Desarrollo de código con ejemplo de RAG

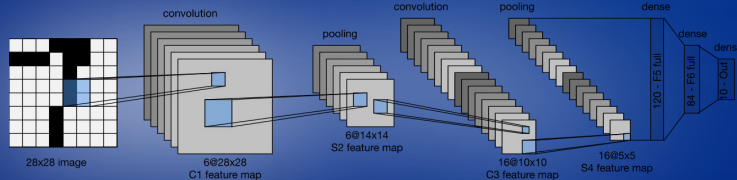
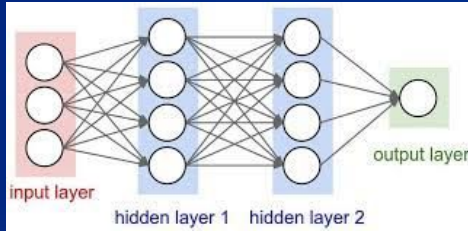
Comparación de langchain vs haystack vs llama index

preguntas

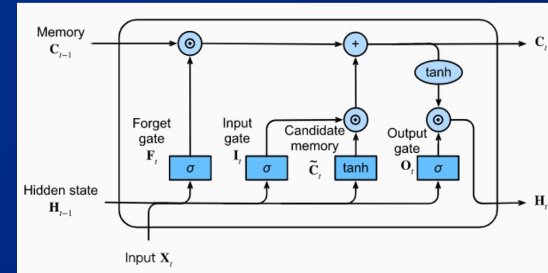
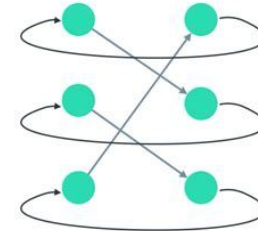
Cierre y recursos

¿Qué son los transformers?

Arquitecturas: DNN, CNN, RNN (LSTM GRU)

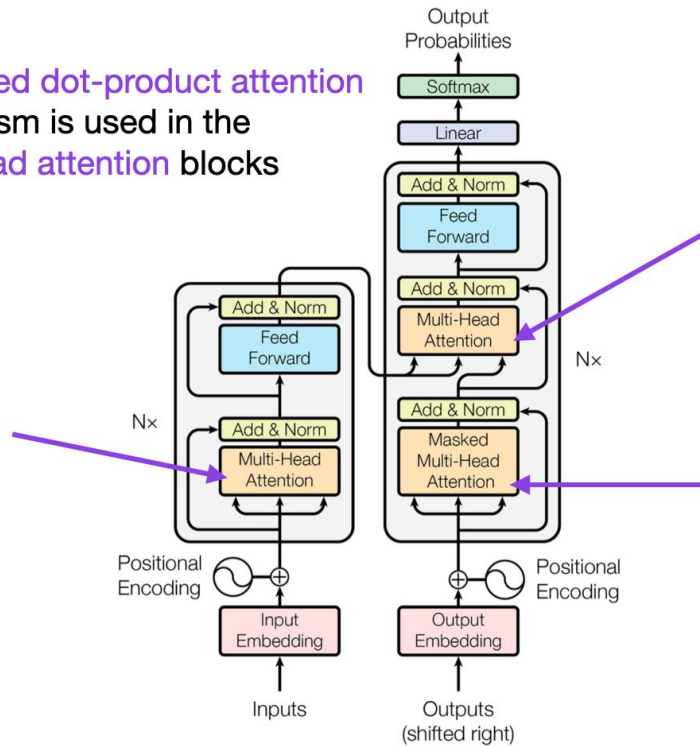


Simple (Recurrent) Neural Network



Mecanismo de auto-atención

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



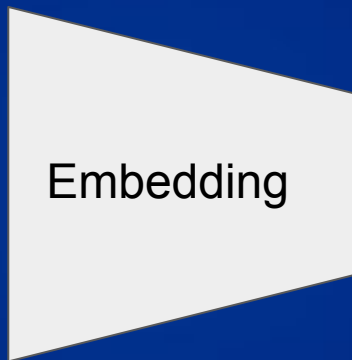
¿Qué son los LLM?

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

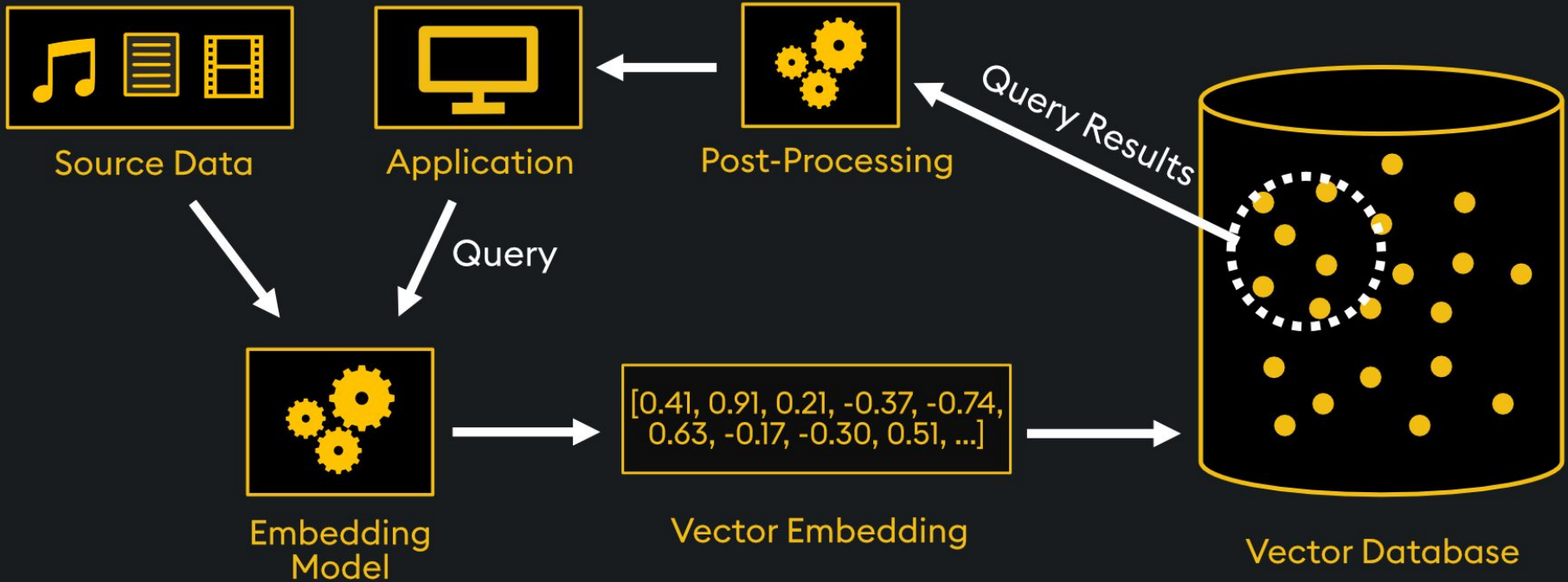
Textos -> vectores

“hoy estamos en un taller de RAG”



texto como vectores
[0.73, -1.23, 1.44, ...]

vector databases y embeddings



Ecosistema actual de LLM

LLMs y Embeddings



Vector DB



AI Frameworks



Cómo abordar la IA generativa y personalizar los transformers/LLM

conocimiento externo	<i>alta</i>	RAG	FINE TUNING + RAG
	<i>baja</i>	PROMPT ENGINEERING	FINE TUNING
		<i>bajo</i>	<i>alto</i>

especificidad de dominio

fuentes: [LinkedIn](#)

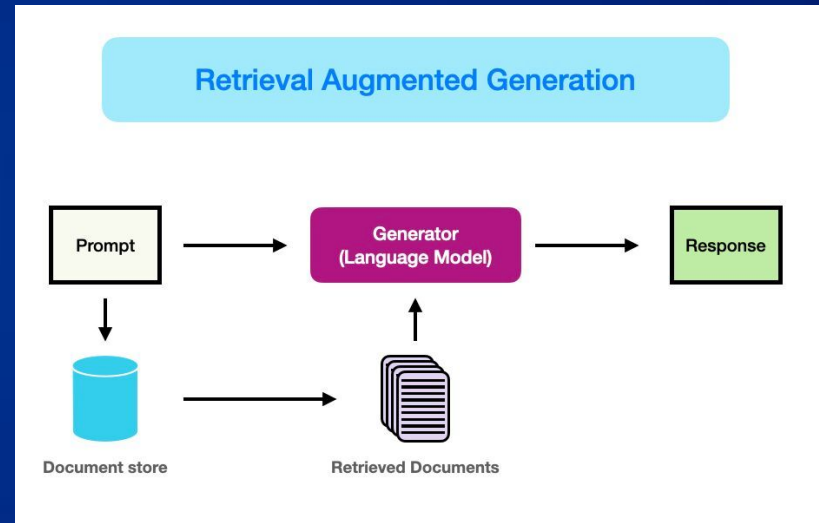
¿Qué es RAG?

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

