XXX CONGRESO ARGENTINO
DE CIENCIAS DE LA COMPUTACIÓN
CACIC2024

08
OCTUBRE
La Plata | Bs. As. | Argentina

Uso de LLMs con RAG

Curso 2: Deep Learning

Marcelo Errecalde[1] , **Horacio Thompson**[1,2]

[1]Universidad Nacional de San Luis (UNSL), Argentina

[2]Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina

¿Qué son los LLMs?

LLM Redes neuronales enormes capaces de procesar y generar textos

¿Qué son los LLMs?

LLM Redes neuronales enormes capaces de procesar y generar textos

OpenAI

Google

LLAMA 2

MISTRAL AI_

¿Qué son los LLMs?

LLM

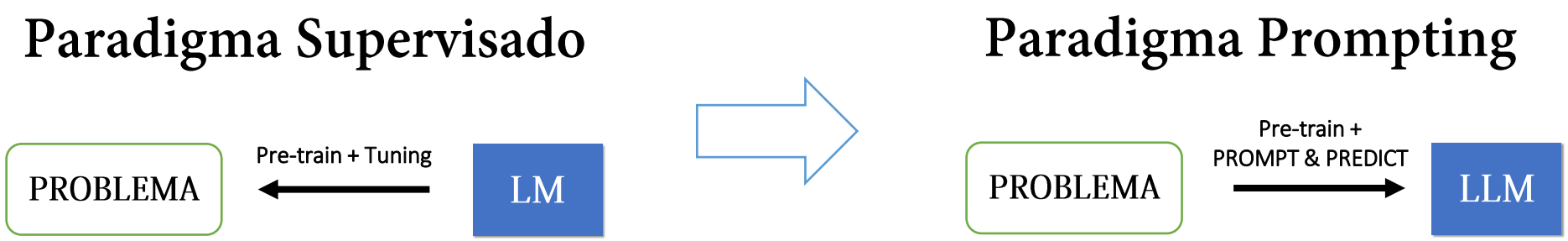
Redes neuronales enormes capaces de procesar y generar textos

Transformers

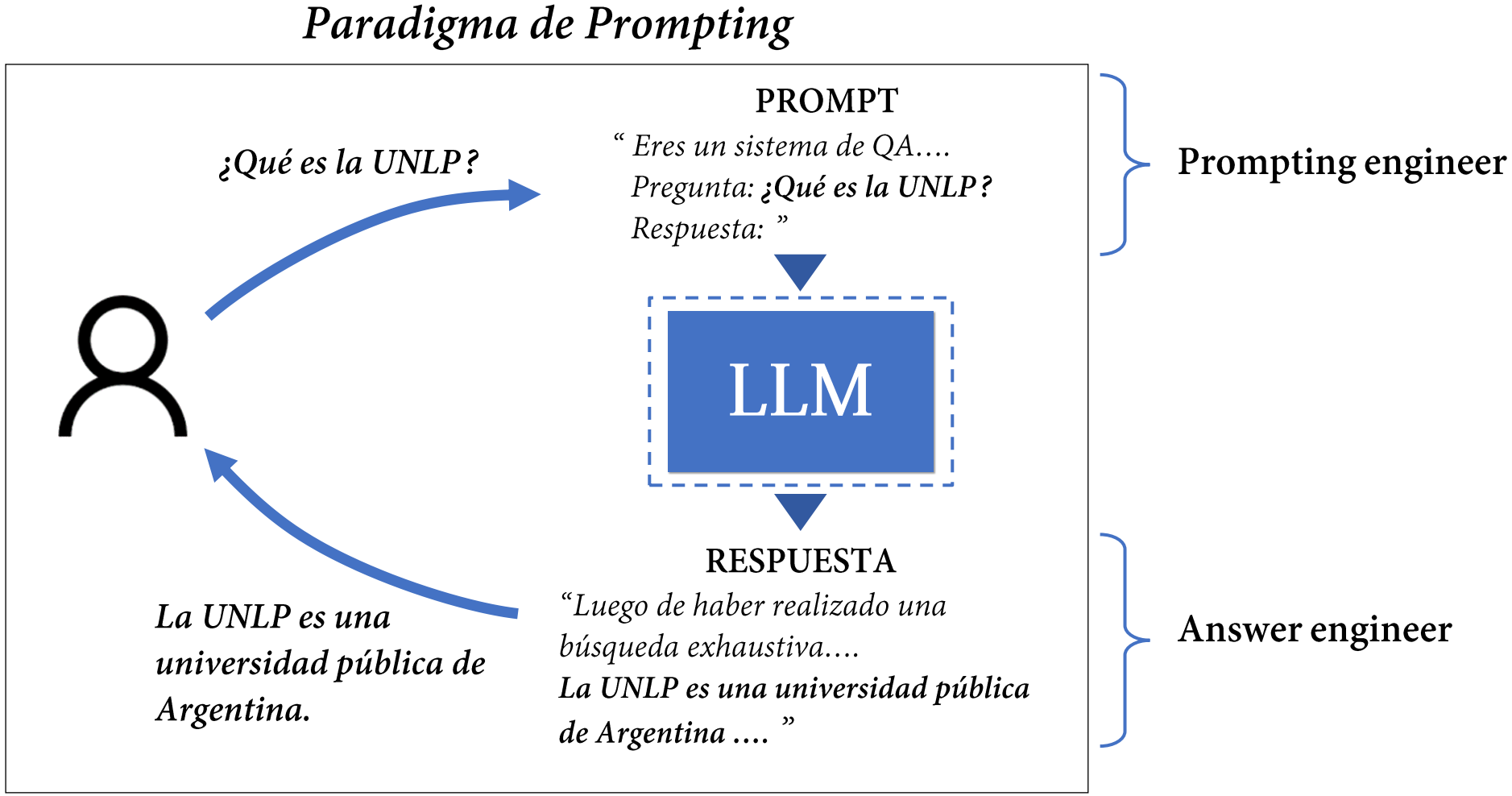
Attention Is All You Need

Vaswani et al. (2017)

¿Cómo se utilizan los LLMs?



¿Cómo se utilizan los LLMs?



LLMs: Ventajas y limitaciones

LLM

- ✓ *Procesan y generan contenido en lenguaje natural*
- ✓ *Eficaces para resolver diversas tareas lingüísticas*
- ✗ *Falta de transparencia*
- ✗ *Alucinaciones y sesgos*
- ✗ *Alto costo computacional*

LLMs: Ventajas y limitaciones

LLM

- ✓ *Procesan y generan contenido en lenguaje natural*
- ✓ *Eficaces para resolver diversas tareas lingüísticas*
- ✗ *Falta de transparencia*
- ✗ *Alucinaciones y sesgos*
- ✗ *Alto costo computacional*

→

RAG

Retrieval-Augmented Generation

¿Qué es RAG?

- **Generación Aumentada por Recuperación de Información**
- Combinar el poder generativo de los LLMs con RI, permitiendo obtener **respuestas precisas** y **actualizadas** en **dominios específicos**.

¿Qué es RAG?

- **Generación Aumentada por Recuperación de Información**
- Combinar el poder generativo de los LLMs con RI, permitiendo obtener **respuestas precisas** y **actualizadas** en **dominios específicos**.

✓ ***Adaptar*** *un LLM a un dominio específico*

✓ ***Mejorar la calidad*** *de las respuestas con conocimiento externo*

✓ *Reducir* ***alucinaciones***

✓ *Contribuye a la* ***interpretabilidad*** *y el* ***razonamiento***

✗ *Complejidad en la implementación*

✗ *Tiempo de respuesta y recursos adicionales*

RAG – Arquitectura

- Estructura de un sistema que aplica RAG se organiza en 3 componentes:

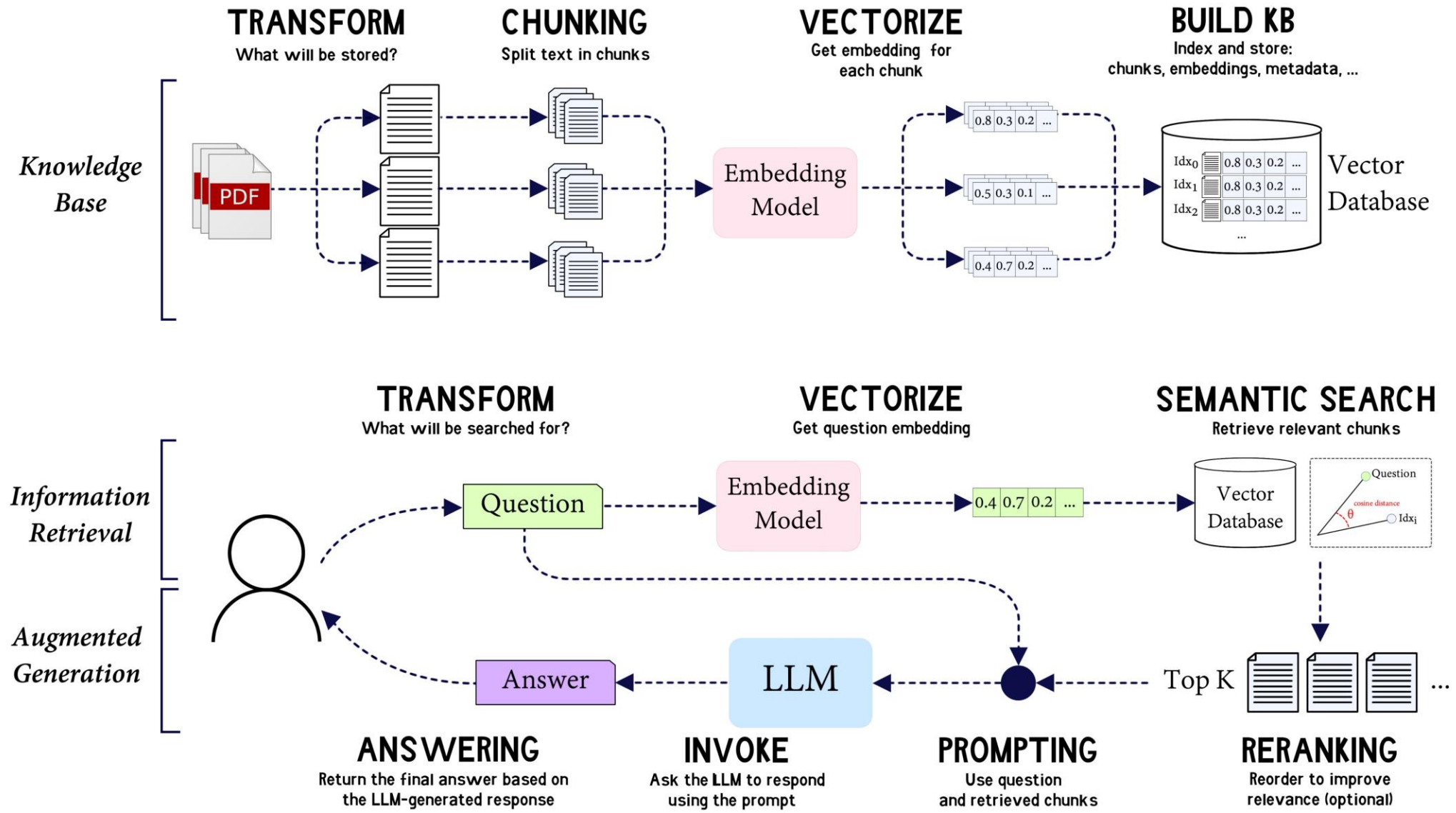
1) Base de conocimiento (Knowledge Base): construir **BD vectorial** que permita almacenar y gestionar la información disponible.

2) Recuperación de Información (Information Retrieval): obtener **conocimiento relevante** que sea útil para responder una pregunta.

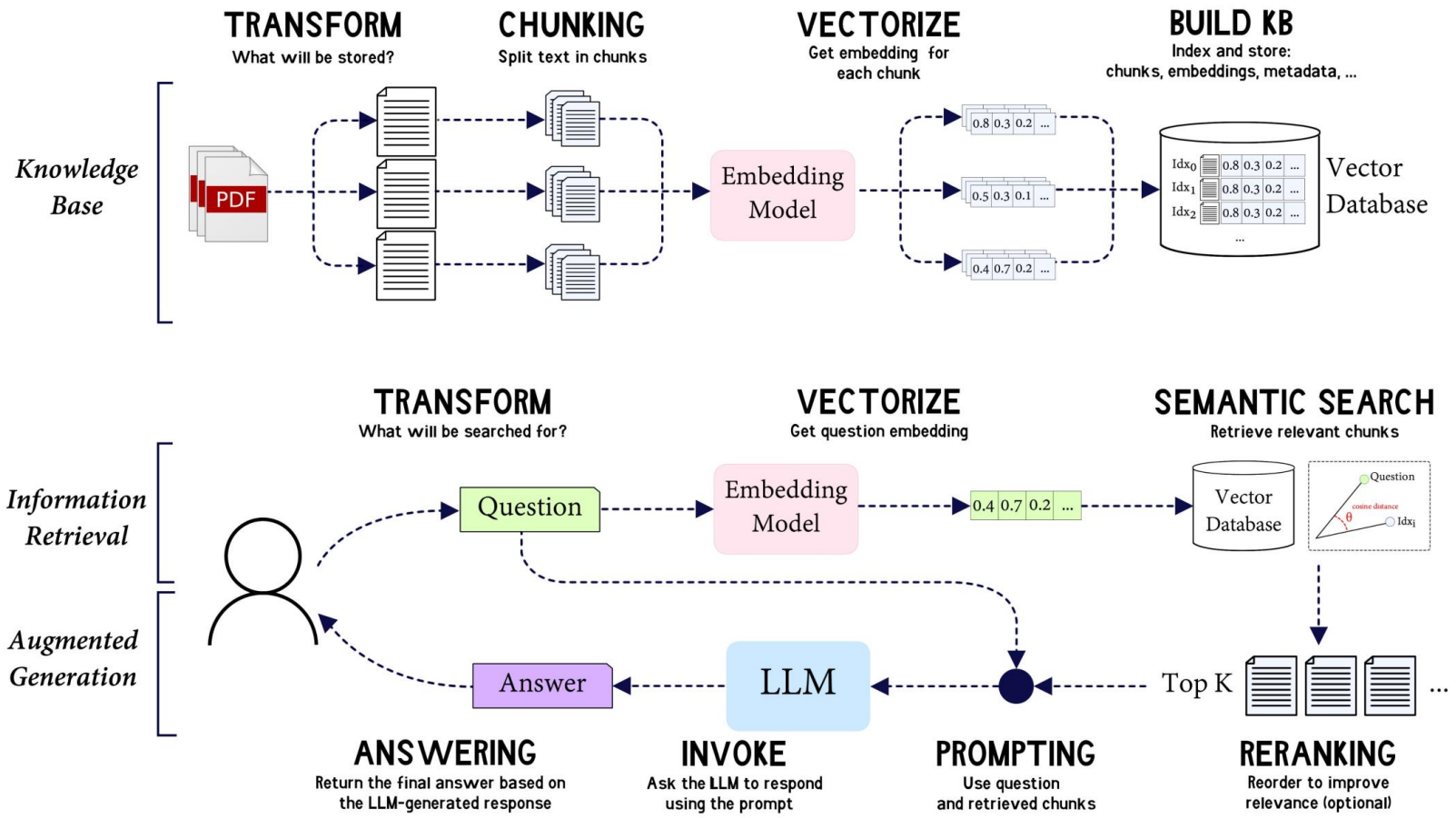
3) Generación aumentada (Augmented Generation): solicitar al LLM que responda la pregunta original usando únicamente la información recuperada.

Sistema QA con LLMs y RAG sobre documentos PDFs

Sistema QA con LLMs y RAG sobre documentos PDFs



Sistema QA con LLMs y RAG sobre documentos PDFs



TECNOLOGÍAS

Python

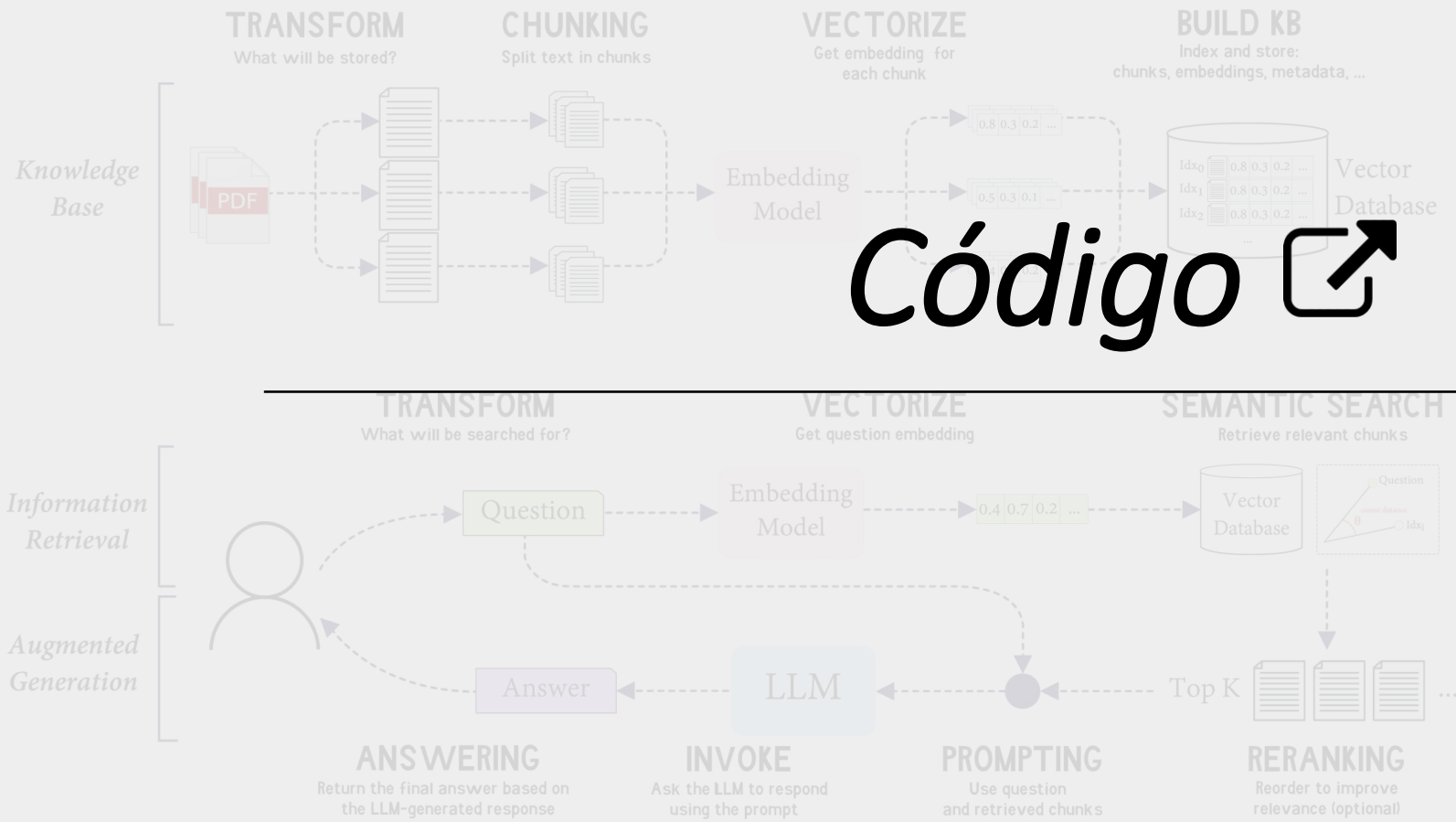
Chroma

SBERT.net

Gemini

LangChain

Sistema Question-Answering usando RAG



TECNOLOGÍAS

Chroma

SBERT.net

Gemini

LangChain

XXX CONGRESO ARGENTINO
DE CIENCIAS DE LA COMPUTACIÓN

CACIC2024

GRACIAS

CONICET

LIDIC Laboratorio de Investigación y Desarrollo en Inteligencia Computacional

Lic. Horacio J. Thompson – hjthompson@unsl.edu.ar

Dr. Marcelo L. Errecalde – merreca@email.unsl.edu.ar