

Retrieval-Augmented Generation for LLMs

Fernando Gonzalez Aauto
Sr Machine Learning Engineer
at Kavak



@fer_adauto



fernando-adauto



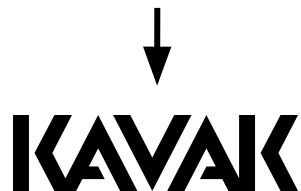
Ingeniería Telemática



Data Scientist



Master in Data Science
Research Assistant (NLP)



Sr. Machine Learning Engineer



Agenda

- ¿Qué son los LLMs?
- Limitaciones
- Retrieval-Augmented Generation (RAG)
- Herramientas existentes
- RAG Avanzado

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.

Escribe una función en Python
para imprimir la sucesión de
Fibonacci



```
python

def fibonacci(n):
    a, b = 0, 1
    for _ in range(n):
        print(a, end=" ")
        a, b = b, a + b

n = 10
fibonacci(n)
```

Limitaciones

- No se sabe cuál es la fuente de los datos.

Limitaciones

- No se sabe cuál es la fuente de los datos.
- Conocimiento limitado a los datos con los que fue entrenado.
- No pueden responder sobre eventos que pasan en la actualidad.

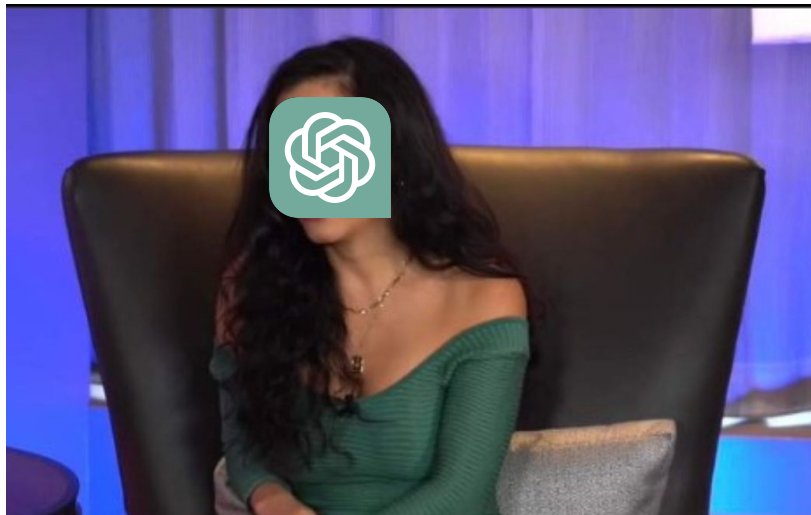


¿Cuáles fueron los eventos mundiales más importantes en septiembre de 2023?



Lo siento, pero mi conocimiento se detiene en septiembre de 2021 y no puedo proporcionar información actualizada sobre eventos de septiembre de 2023.

Alucinaciones



YR

¿Qué sabes de mis datos?



Yordi, yo soy experto en tus datos

Alucinaciones

¿Quien es Fernando Gonzalez Aauto?



Fernando González Aauto (nacido el 16 de noviembre de 1970 en Madrid) es un empresario español y exdirector ejecutivo del Grupo Santander, uno de los mayores grupos financieros del mundo.

Aauto se licenció en Derecho por la Universidad Carlos III de Madrid y comenzó su carrera profesional como abogado. En 2003 fue nombrado director general de Banco Santander Central Hispano, cargo que ocupó hasta 2008 cuando pasó a ser vicepresidente de la entidad.

Alucinaciones

¿Quién es Fernando Gonzalez Adatao (de Ecatepec, Estado de Mexico)?



Fernando González Adatao es un político mexicano nacido en la localidad de Ecatepec, estado de México. Es miembro del Partido Revolucionario Institucional (PRI) y ha ocupado diversos cargos públicos en el gobierno estatal y federal.

En 2018 fue elegido como gobernador del estado de México por el PRI, sucediendo a Alfredo Del Mazo Maza. Durante su mandato se enfrentó a críticas por sus decisiones sobre los temas de seguridad y educación. En particular, se criticó su decisión de dismantelar las escuelas secundarias que habían sido construidas durante el gobierno anterior para reemplazarlo con escuelas primarias.

Por otro lado, también se le acusó de ser responsable de una serie de asesinatos cometidos en el estado de México durante su mandato. Sin embargo, no se encontraron pruebas concluyentes que lo implicaran directamente en dichos crímenes.

Limitaciones

- No se sabe cuál es la fuente de los datos.
- Conocimiento limitado a los datos con los que fue entrenado.
- No pueden responder sobre eventos que pasan en la actualidad.
- Alucinaciones
- Sesgos políticos, género, religión.
- Malos en tareas que involucren matemáticas.

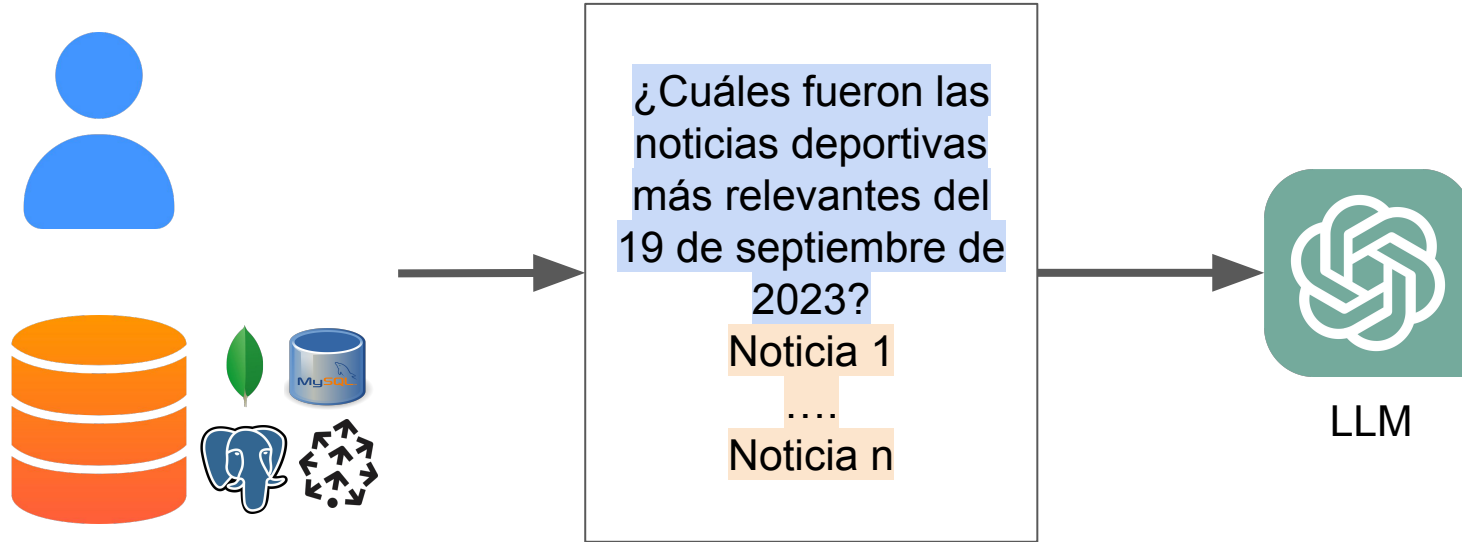
Soluciones

Fine-tuning:

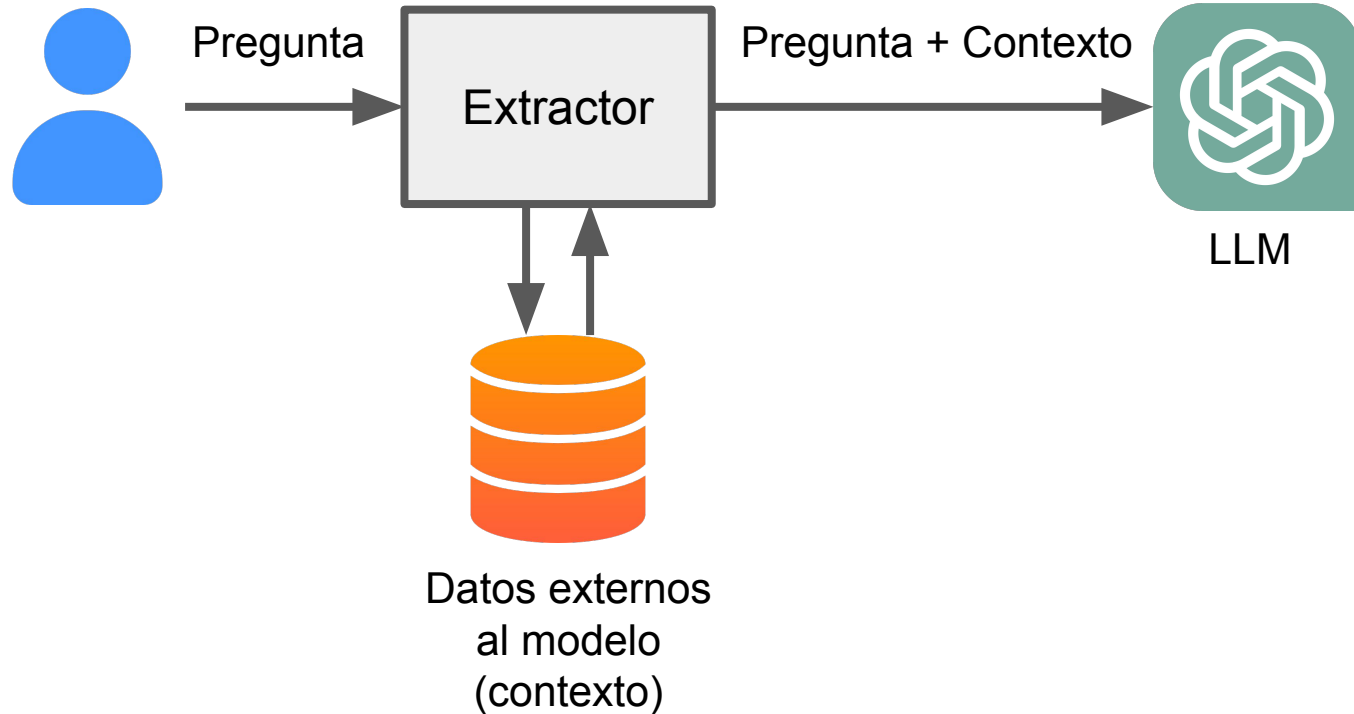
- Para que se mantenga actualizado se necesitaría entrenar periódicamente.
- Sigue siendo costoso.
- No se sabe la fuente de datos de las respuestas.

Retrieval-Augmented Generation (RAG)

Con base en el prompt, buscar información relevante y añadirla como contexto al input del LLM.



Arquitectura RAG

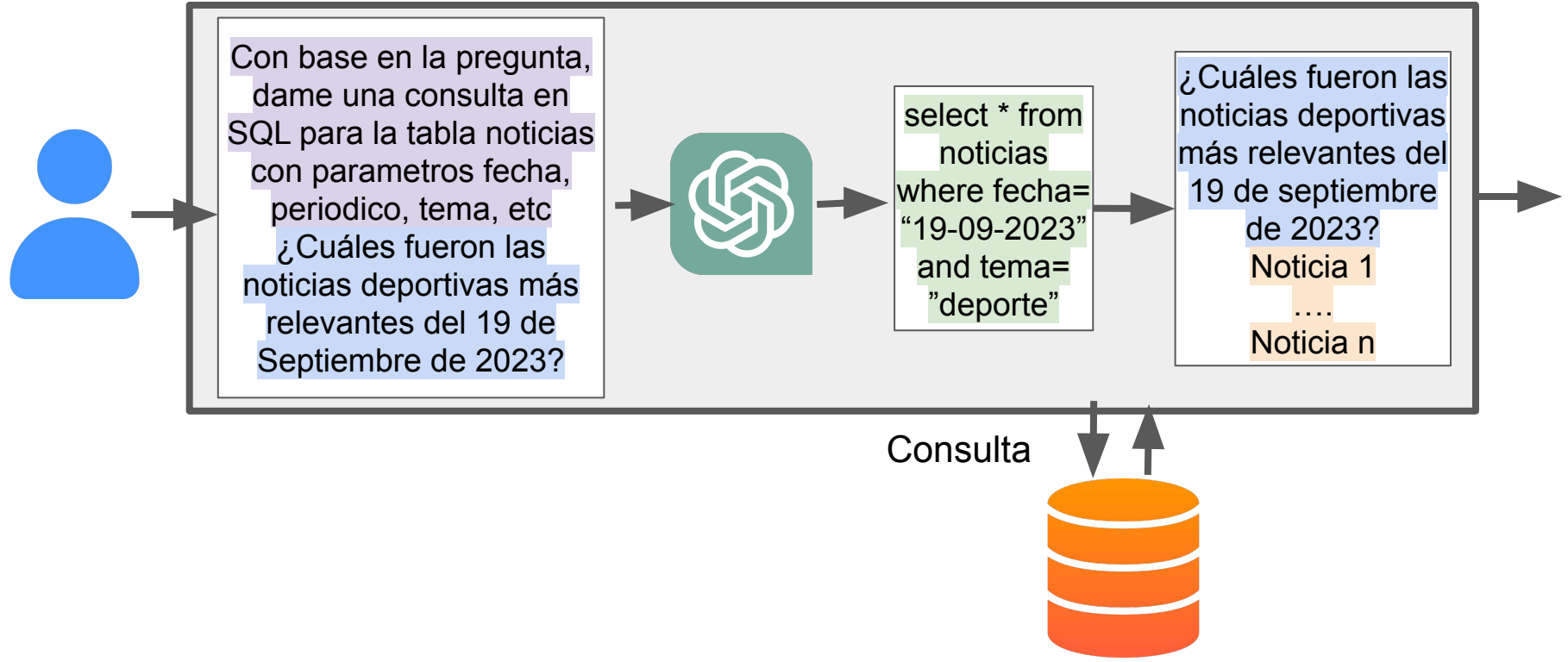


Extractor

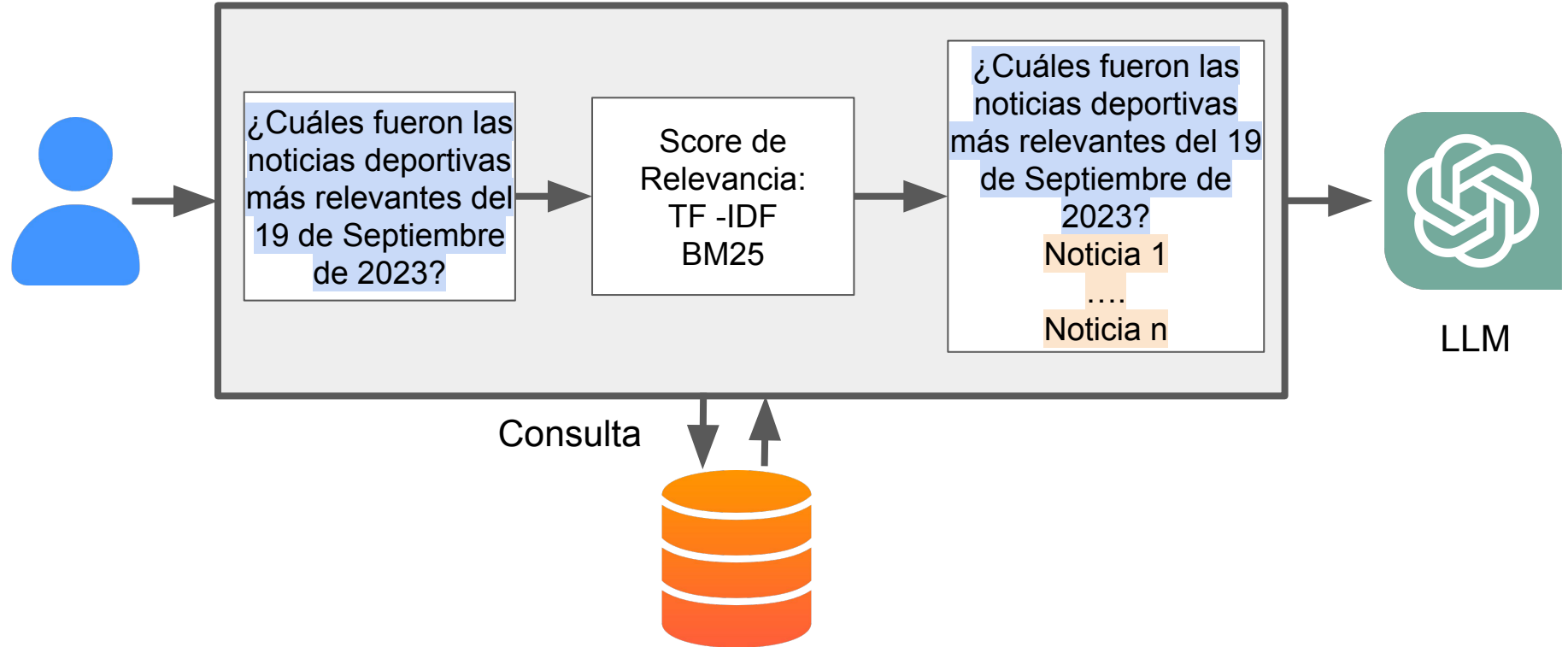
- Búsqueda por atributos
- Búsqueda de texto tradicional
- Búsqueda vectorial

Búsqueda por atributos

Usar el LLM como generación de parámetros de consulta.

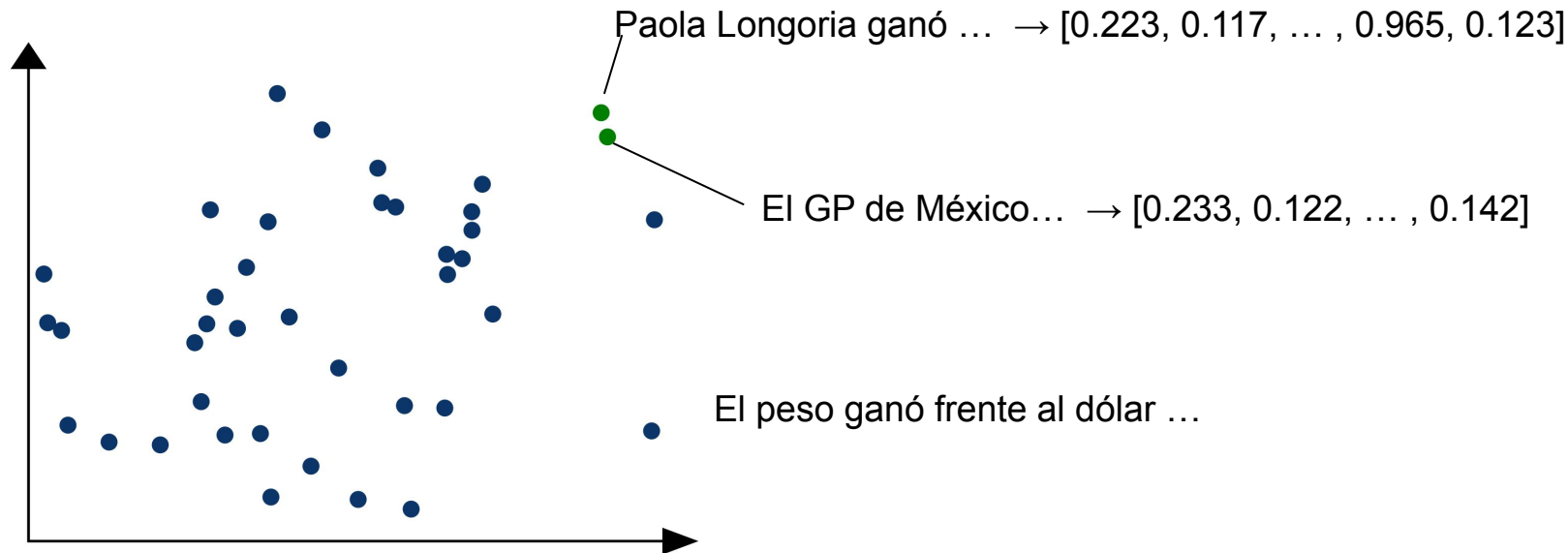


Búsqueda de texto tradicional



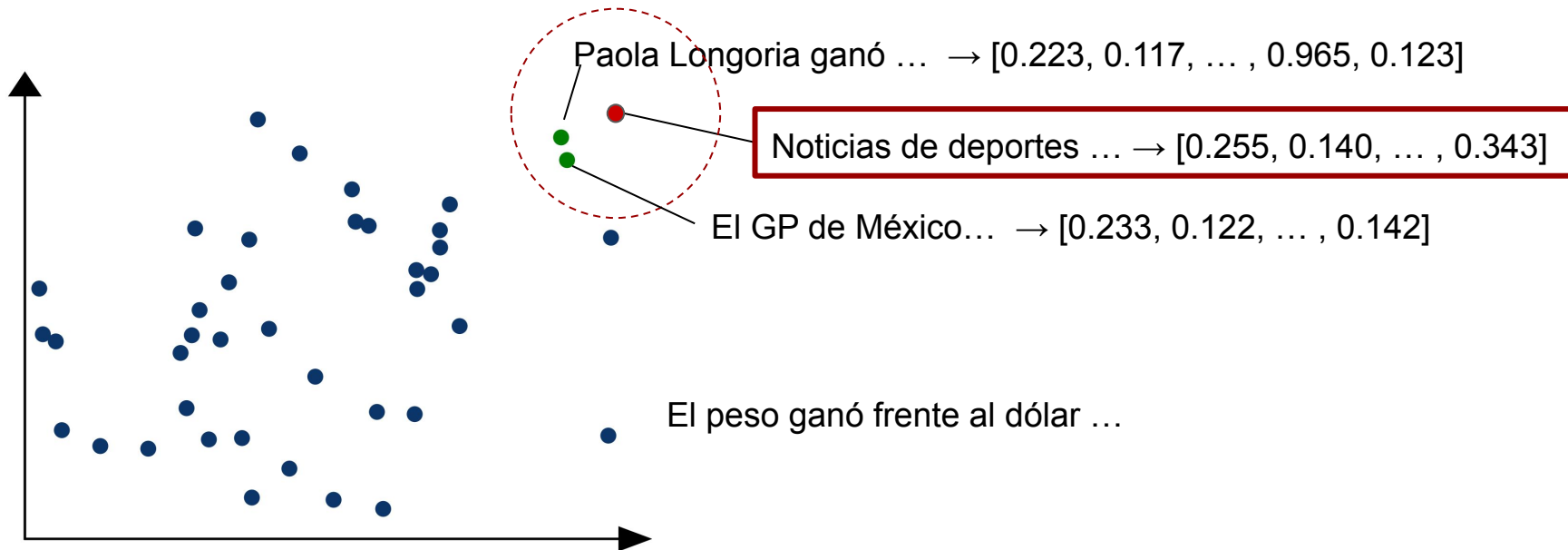
RAG Usando Búsqueda Vectorial

Representar la información en un vector denso para hacer búsquedas semánticas.

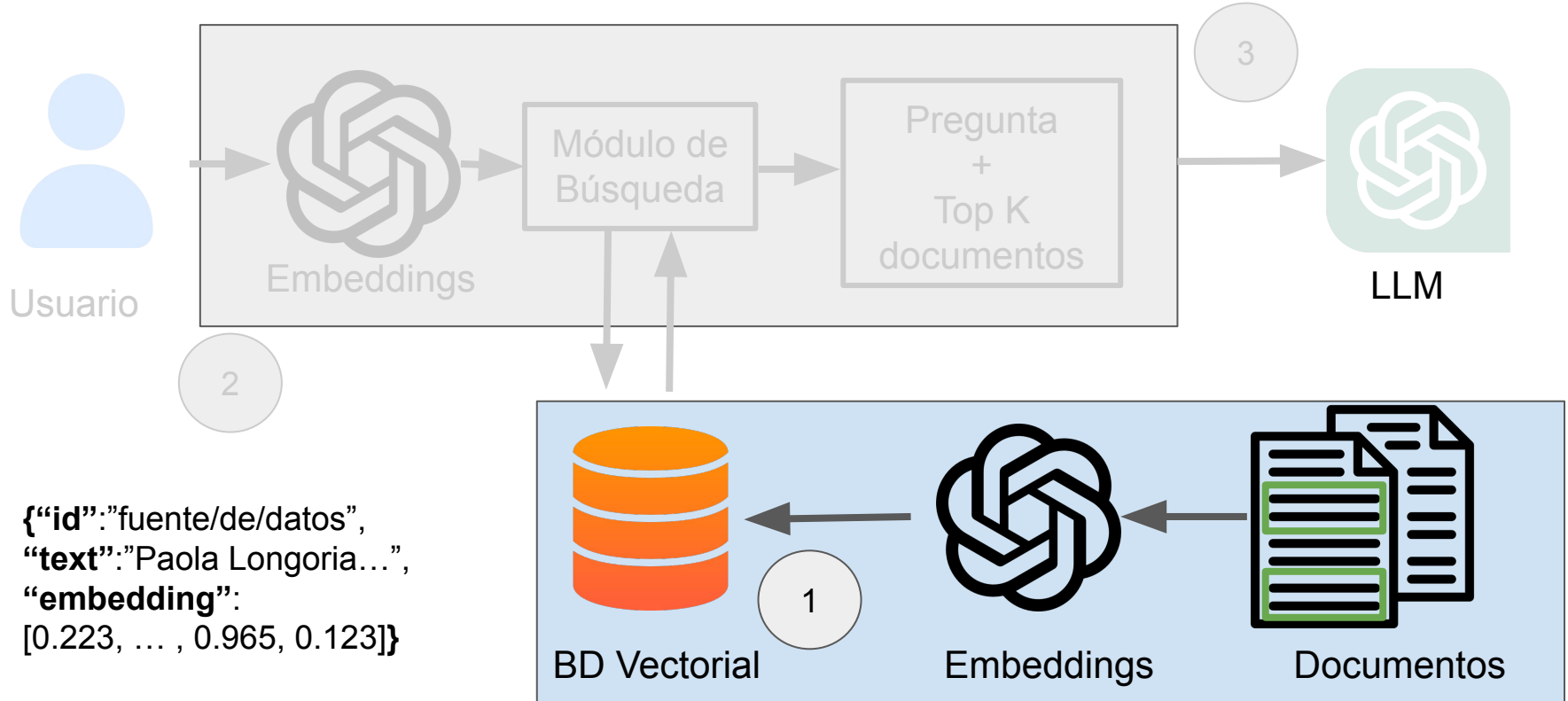


RAG Usando Búsqueda Vectorial

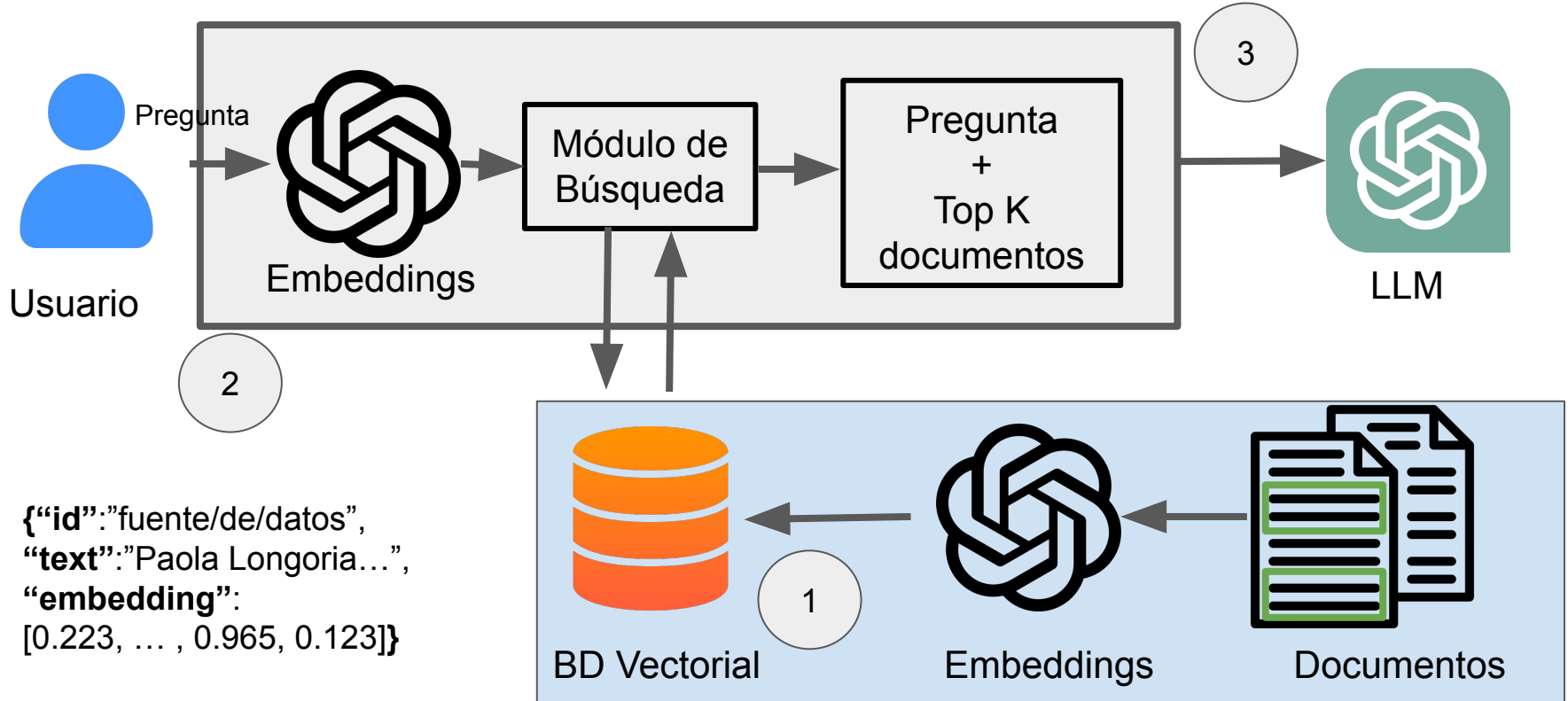
Representar la información en un vector denso para hacer búsquedas semánticas.



RAG Usando Búsqueda Vectorial



RAG Usando Búsqueda Vectorial



Parámetros

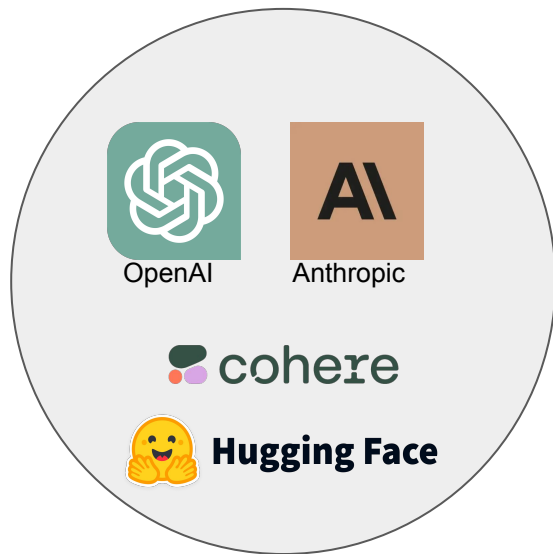
- Número de fragmentos a extraer.
- Límite de score de similitud.
- Filtrar documentos muy similares.
- Modelo de embeddings.

Evaluación

- Un proceso a la vez:
 - Extracción
 - Métricas de ranking: Recall at N, Precision at N
 - Generación
 - Métricas según el caso de uso (Accuracy, BERT Score)
 - Evaluar con un LLM de mayor capacidad.

Herramientas

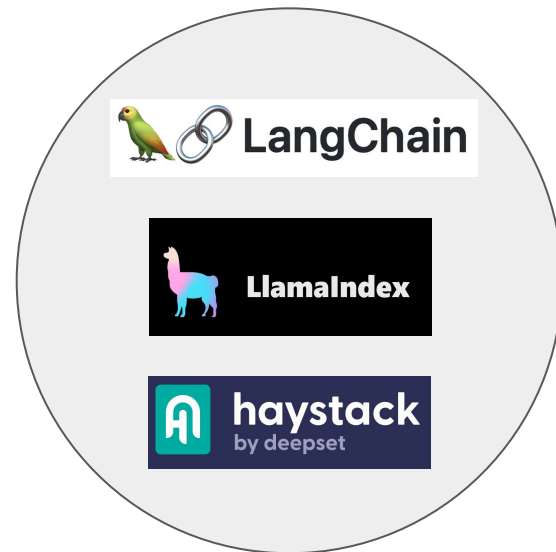
LLMs y Embeddings



Bases de Datos Vectoriales



AI Frameworks



```
import os
from haystack.document_stores import FAISSDocumentStore
from haystack.nodes import EmbeddingRetriever, PromptNode
from haystack.pipelines import Pipeline
from haystack.utils import convert_files_to_docs

## Cargar archivos en DocumentStore
docs = convert_files_to_docs(dir_path="", split_paragraphs=True)
document_store = FAISSDocumentStore(faiss_index_factory_str="Flat")
document_store.write_documents(docs)
retriever = EmbeddingRetriever(
    document_store=document_store, embedding_model="sentence-transformers/multi-qa-mpnet-base-dot-v1"
)
document_store.update_embeddings(retriever)

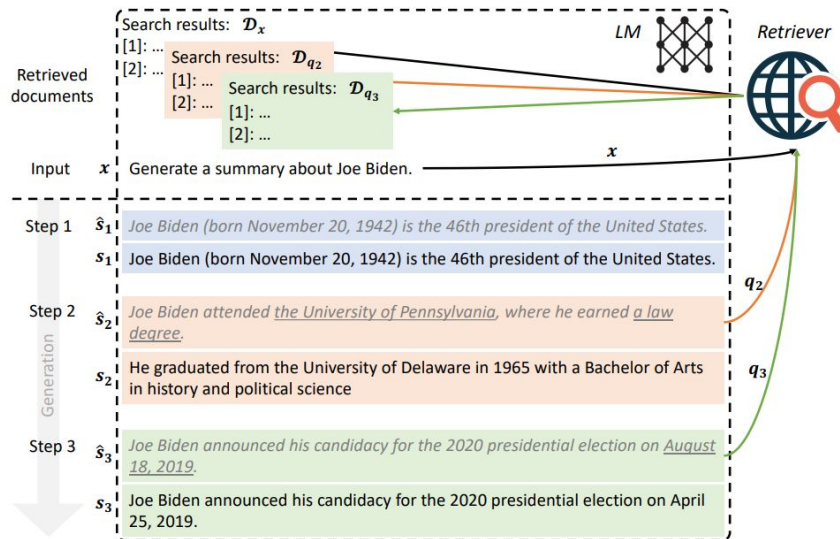
## Crear el Pipeline
prompt_node = PromptNode(
    model_name_or_path="gpt-3.5", api_key=os.getenv("OPENAI_API_KEY", None)
)

pipe = Pipeline()
pipe.add_node(component=retriever, name="retriever", inputs=["Query"])
pipe.add_node(component=prompt_node, name="prompt_node", inputs=["retriever"])

output = pipe.run(query="¿Cuales son las noticias mas relevantes sobre deportes en septiembre de 2023?")
```


RAG Avanzado

- ¿Cuándo extraer?
 - Cuando la probabilidad del próximo token sea baja.



(FLARE) Active Retrieval Augmented Generation (Jiang et al., 2023)

RAG Avanzado

- ¿Cómo extraer?
 - Entrenar generador y extractor end-to-end.
 - Parametrizados con una red neuronal.

RAG Avanzado

- ¿Cómo extraer?
 - Entrenar generador y extractor end-to-end.
 - Parametrizados con una red neuronal.

$$p(y|x) = \sum_{z \in Z} \underbrace{p(y|z, x)}_{\text{Generador}} \underbrace{p(z|x)}_{\text{Extractor}}$$

x : *Input*
 y : *Output*
 Z : *Documents*

REPLUG: Retrieval-Augmented Black-Box Language Models. Shi, W. et al. (2023).

REALM: Retrieval-Augmented Language Model Pre-Training Guu, Kevin et al. (2020)

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Lewis, W et al. (2020).


Beneficios RAG

- Fácil integración con fuentes de datos internas y actuales.
- Reduce alucinaciones.
- Información sobre la fuente de los datos.

Mas Información

- ACL 2023 Tutorial: Retrieval-based Language Models and Applications
<https://acl2023-retrieval-lm.github.io/>
- Documentación de Frameworks de AI

¡Muchas Gracias!

 @fer_adauto

 fernando-adauto

