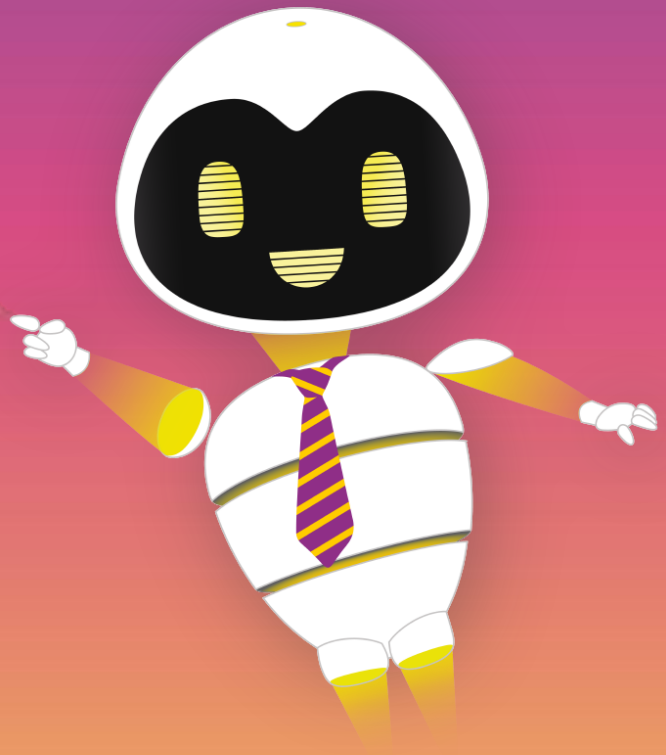
PI DATA STRATEGY & CONSULTING

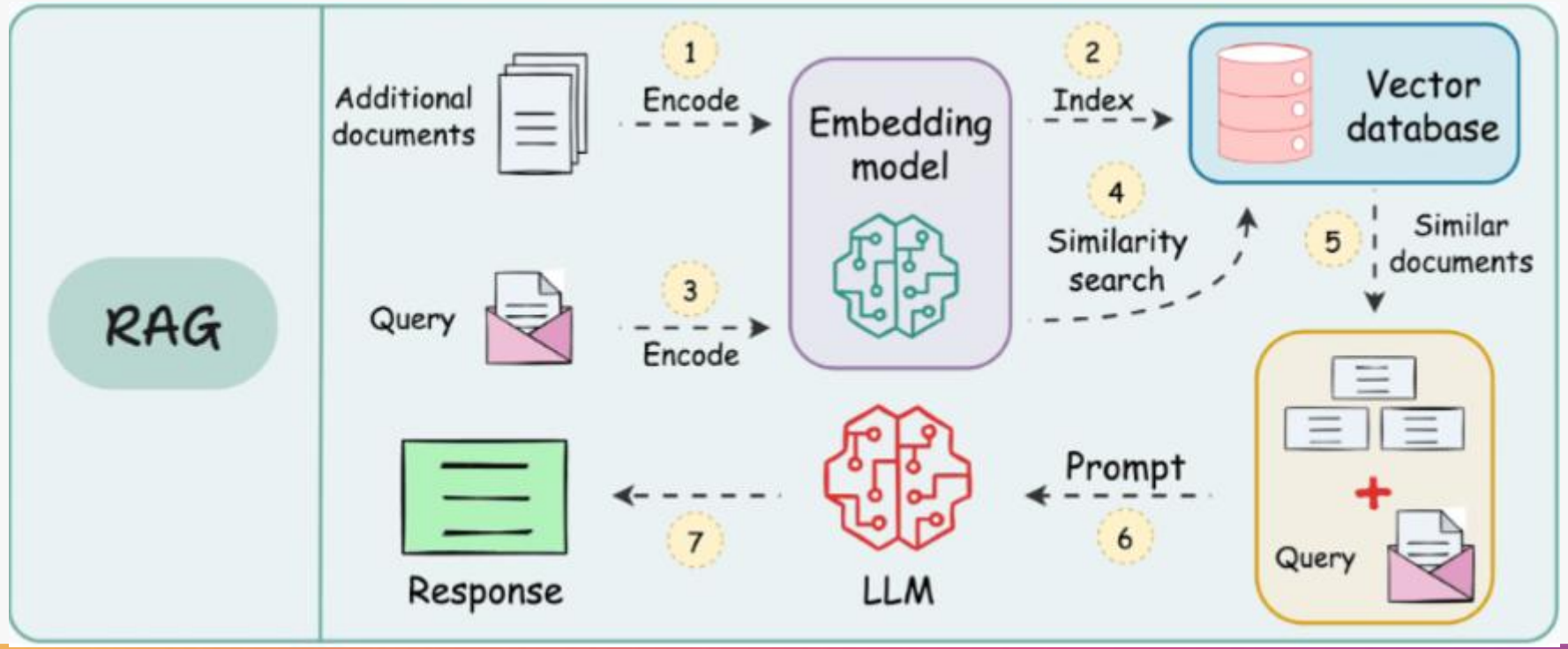
get Talent

AI EDITION

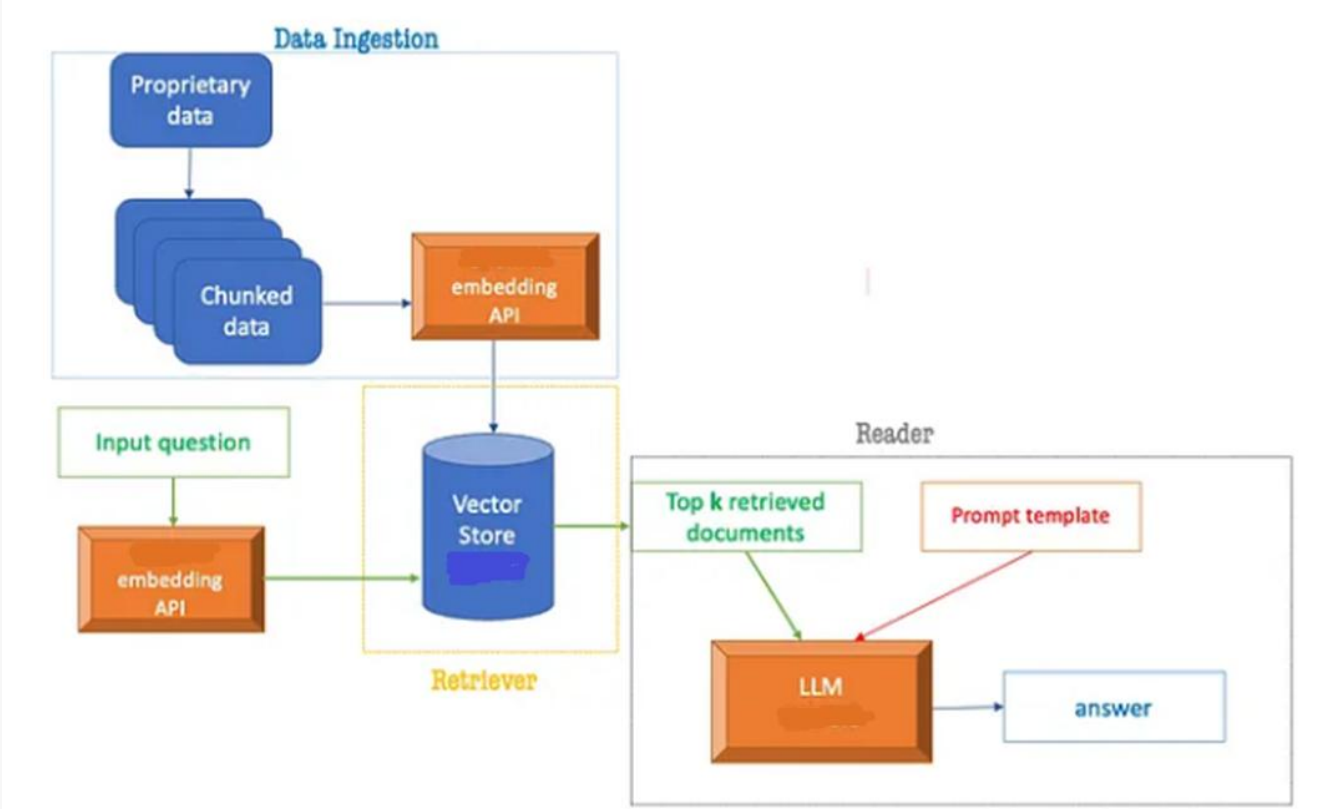


RAG

RAG (Retrieval-Augmented Generation) es un enfoque en inteligencia artificial donde un modelo generativo utiliza información recuperada de una base de datos o documentos externos para mejorar la precisión y relevancia de sus respuestas.



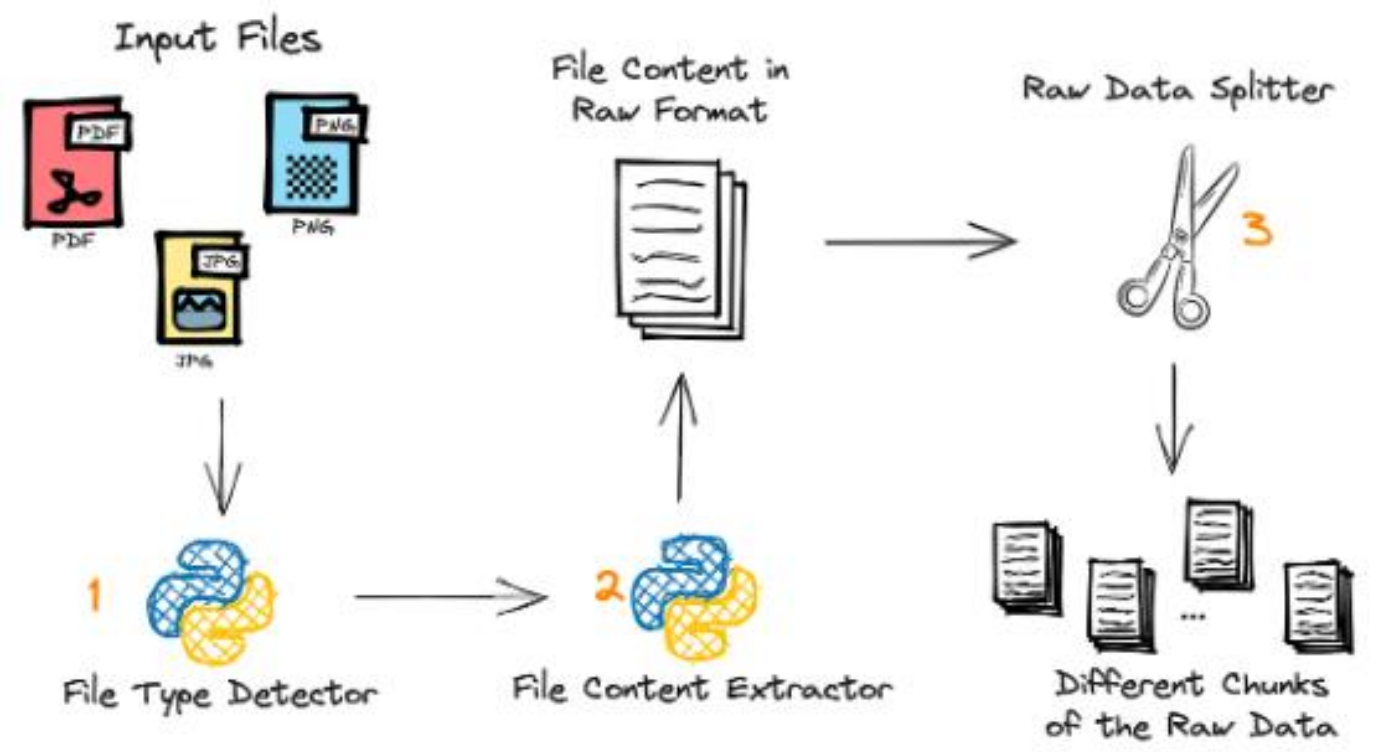
1- Data Ingestion
2- Retriever
3- Reader (generation)



A typical RAG pipeline

1- Data Ingestion

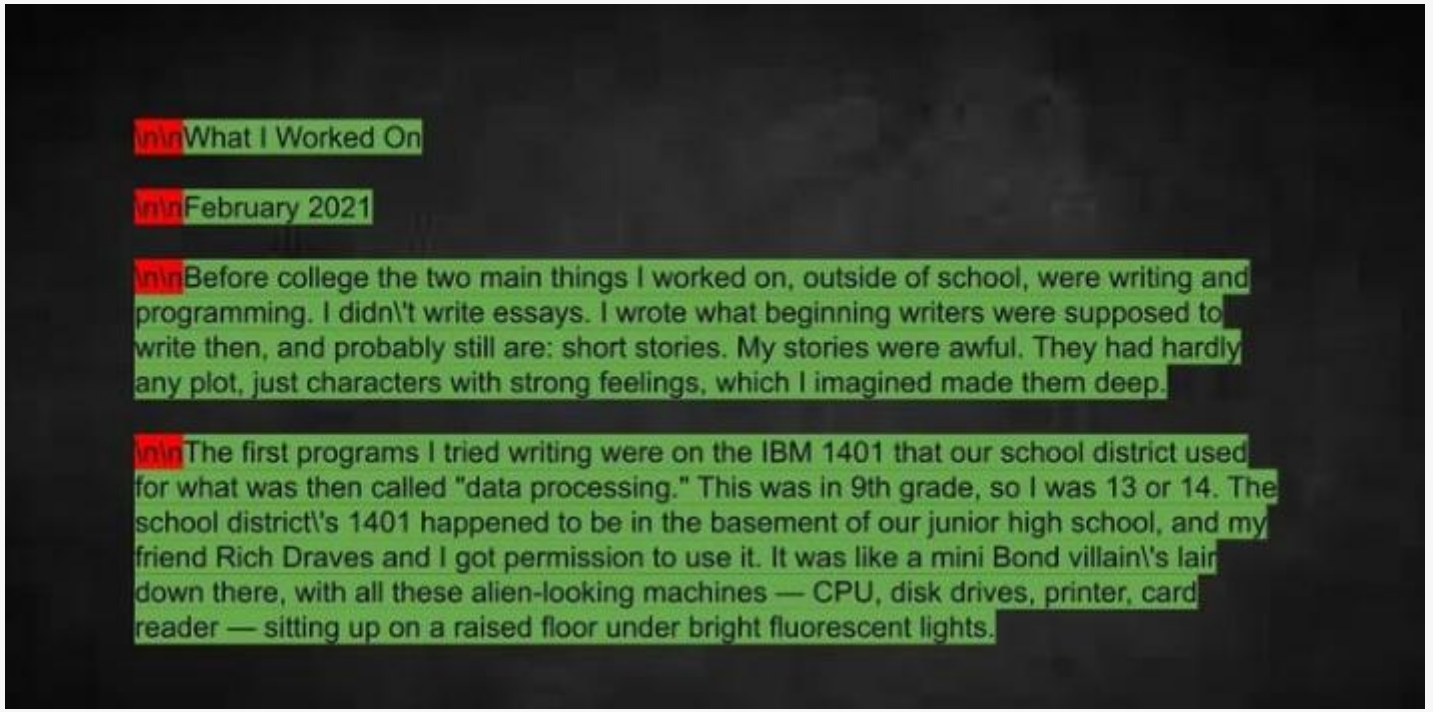
Data Loaders in LangChain



- https://www.geeksforgeeks.org/working-with-pdf-files-in-python/

Chunking

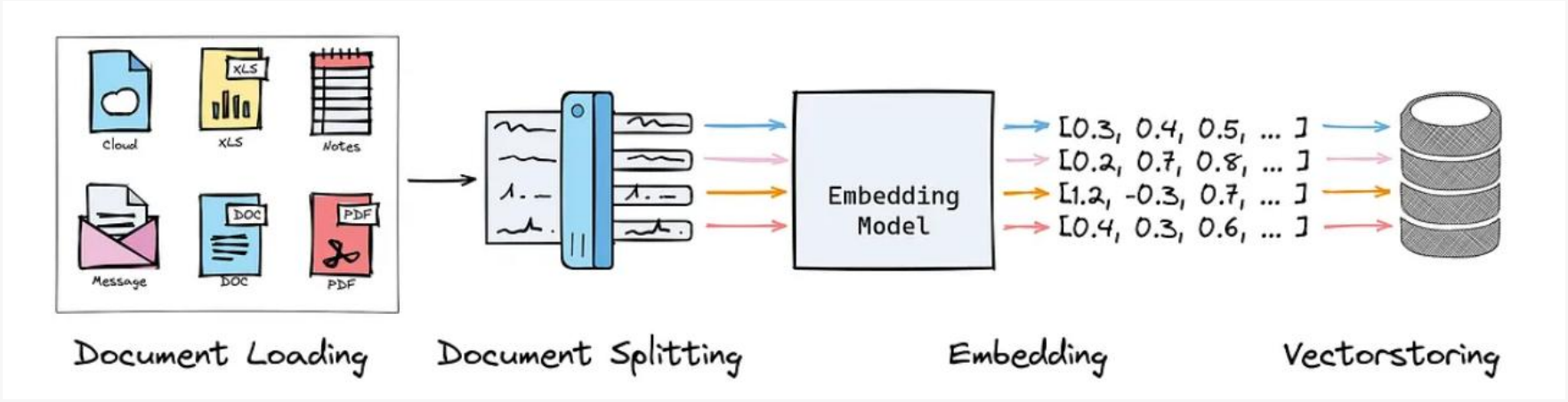
```
text_splitter = RecursiveCharacterTextSplitter(
    chunk_size = 100,
    chunk_overlap  = 0,
    length_function = len,
)
```



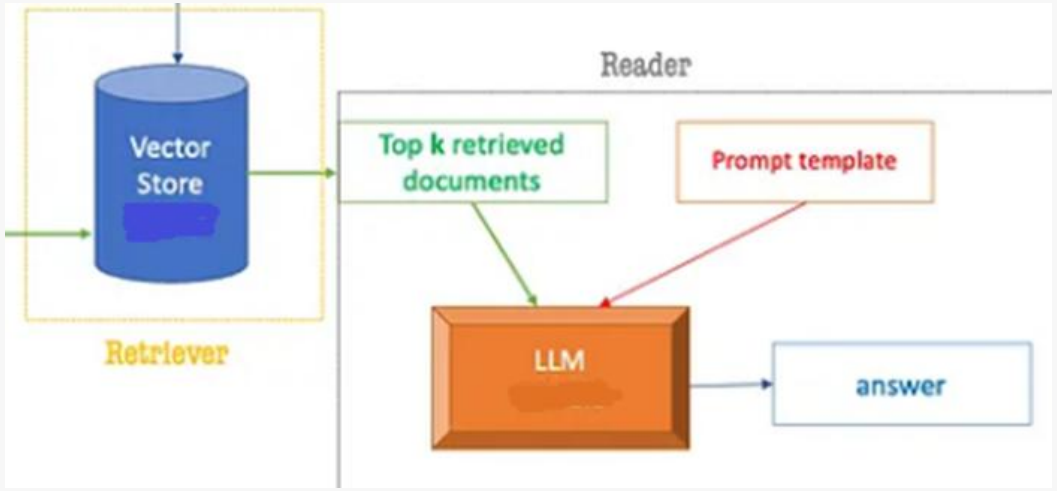
- https://chunkviz.up.railway.app/#explanation
- https://dev.to/eteimz/understanding-langchains-recursivecharactertextsplitter-2846
- https://python.langchain.com/docs/concepts/text_splitters/
- https://www.pinecone.io/learn/chunking-strategies/

2- Retriever

chroma



3- Reader (generation)

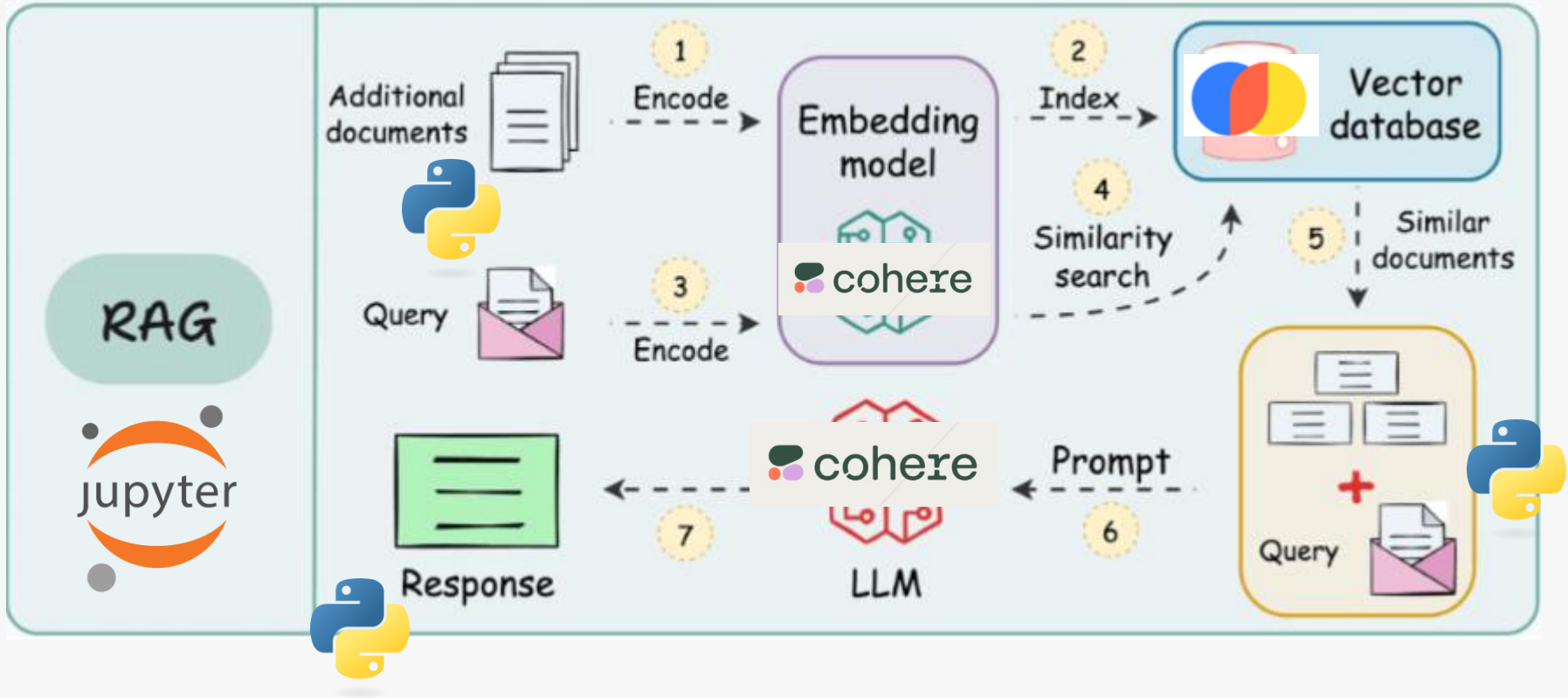



```
prompt_template = """Usa el siguiente contexto para responder la pregunta.

        {context}

        Pregunta: {question}
        """
```

Challenge:


RAG
jupyter
Additional documents
1
Encode
Embedding model
cohere
2
Index
Vector database
4
Similarity search
5
Similar documents
Query
3
Encode
cohere
Prompt
6
Query
7
Response
LLM

Muchas gracias.