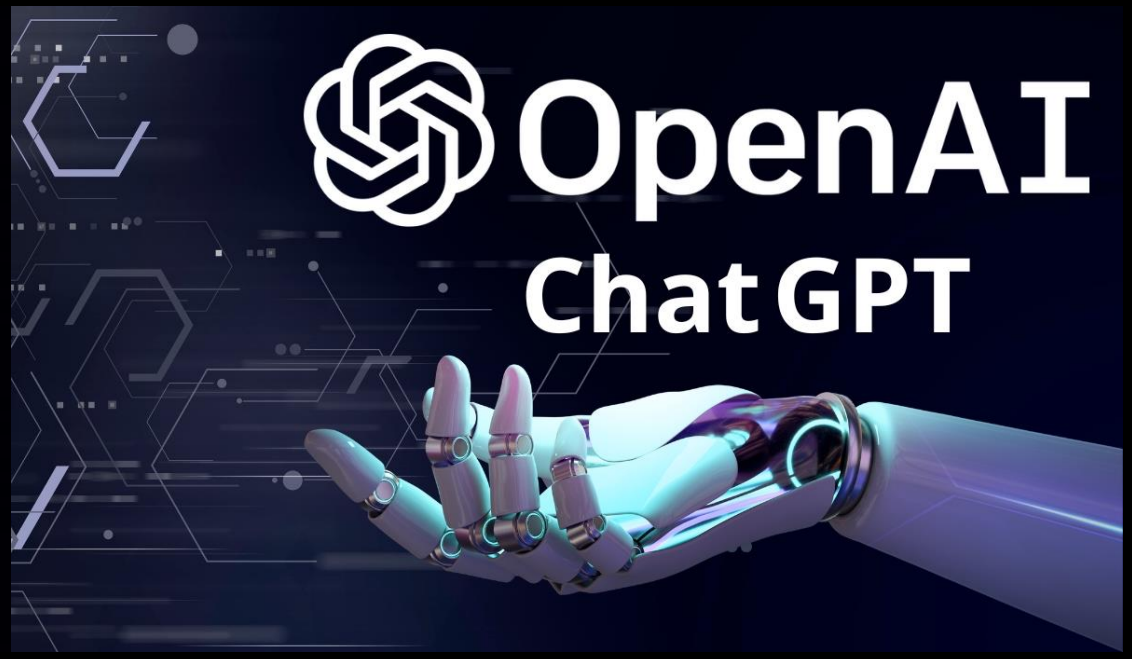
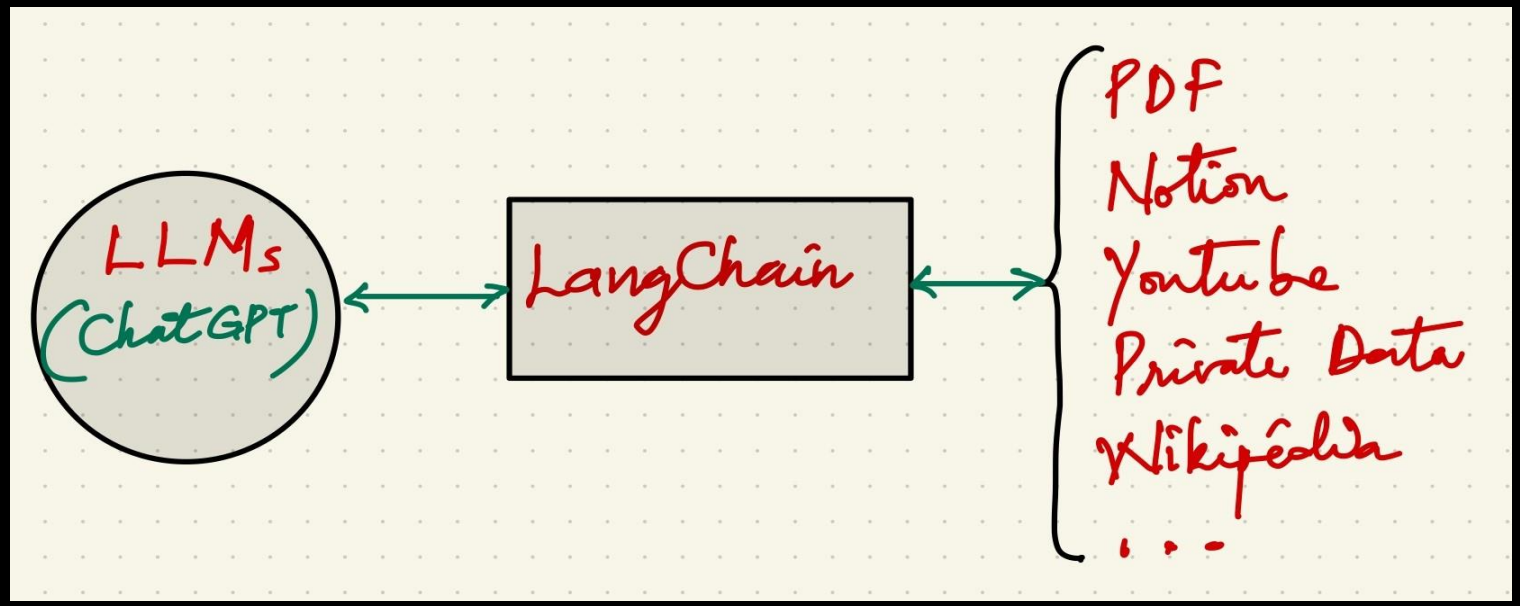


COMMUNIQUER AVEC NOS PROPRES DOCUMENTS VIA LANGCHAIN ET GPT

GUIDJIME ADINSI Ahouahounko

LANGCHAIN ?

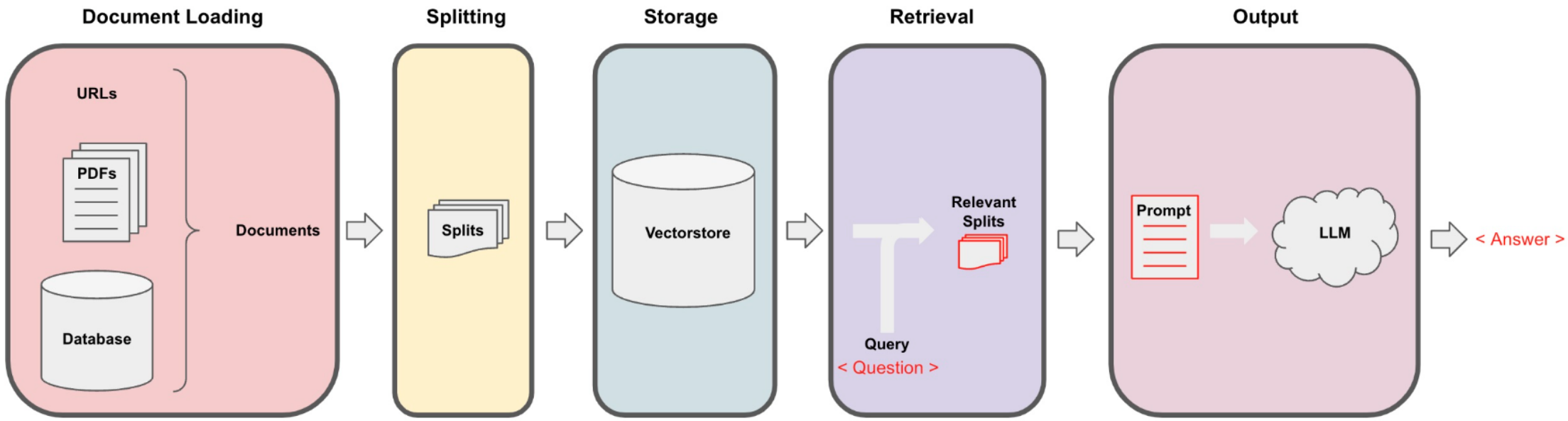
LangChain est un cadre de développement pour, créer des applications alimentées par des LLMs. Il permet de connecter principalement un LLM à des sources de contexte.



RÉCUPÉRATION AUGMENTÉE DE LA GÉNÉRATION (RAG) ?

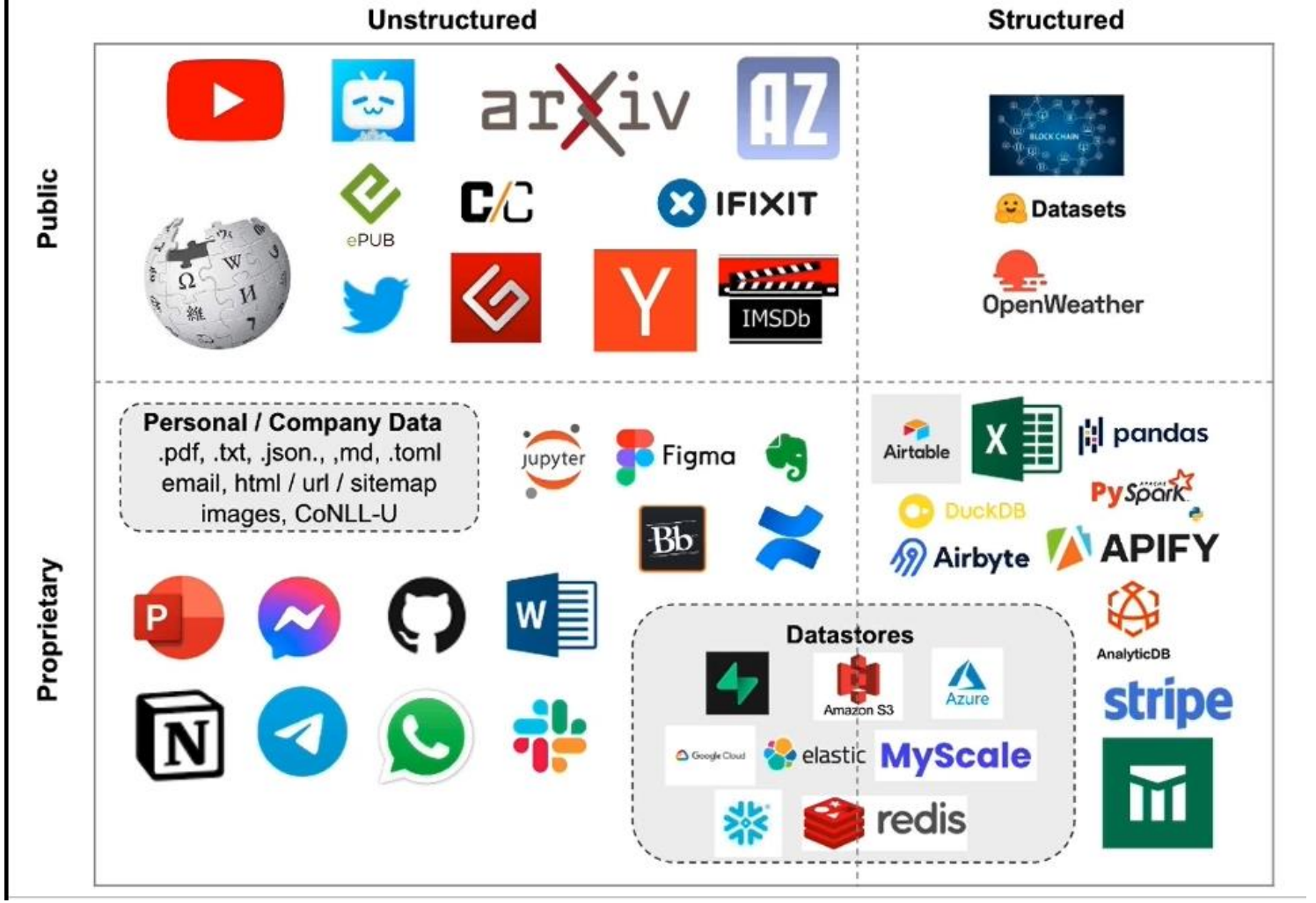
- **Retrieval Augmented Generation(RAG)** est un ensemble de techniques permettant d'augmenter les connaissances d'un LLM en incorporant des informations provenant de documents externes récupérés souvent privées ou en temps réel ;

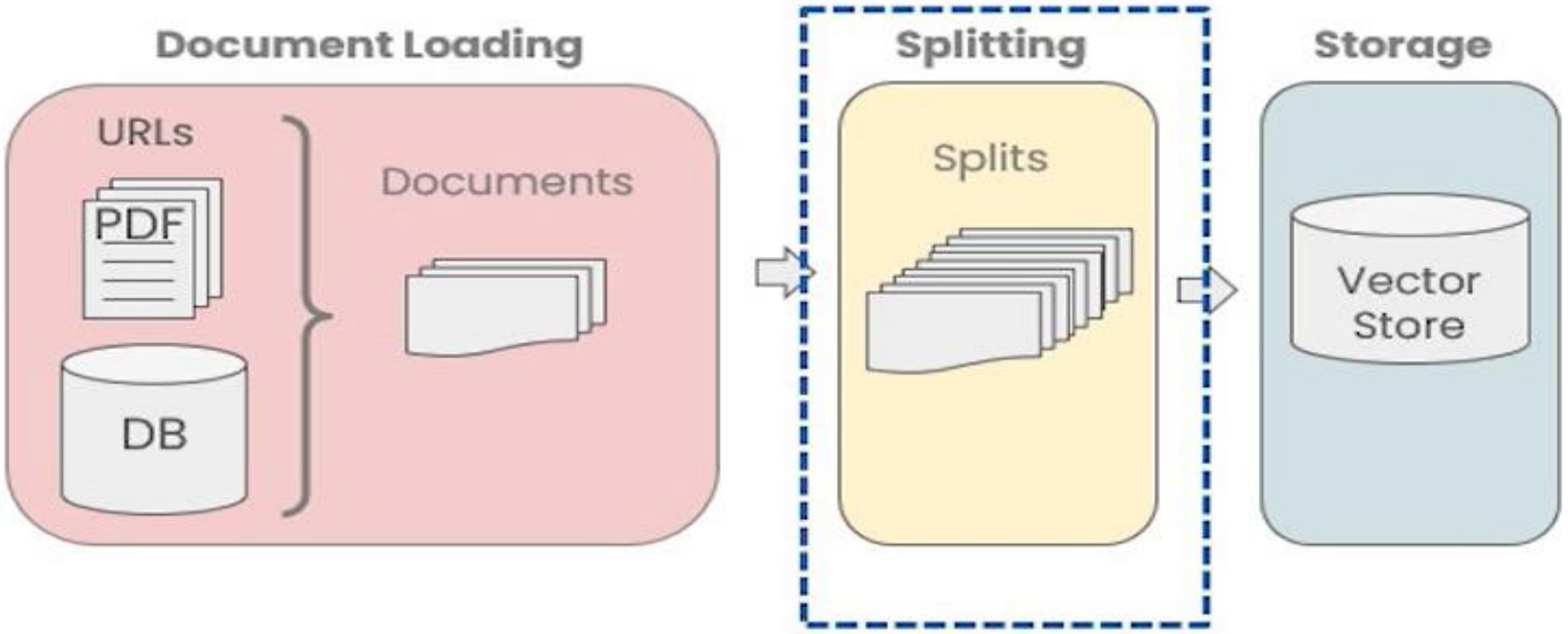
- Architecture basique d'un RAG :
  - **Indexing**(Indexation)
  - **Retrieval and generation(**Récupération et génération**)**



ARCHITECTURE D'UN RAG ?

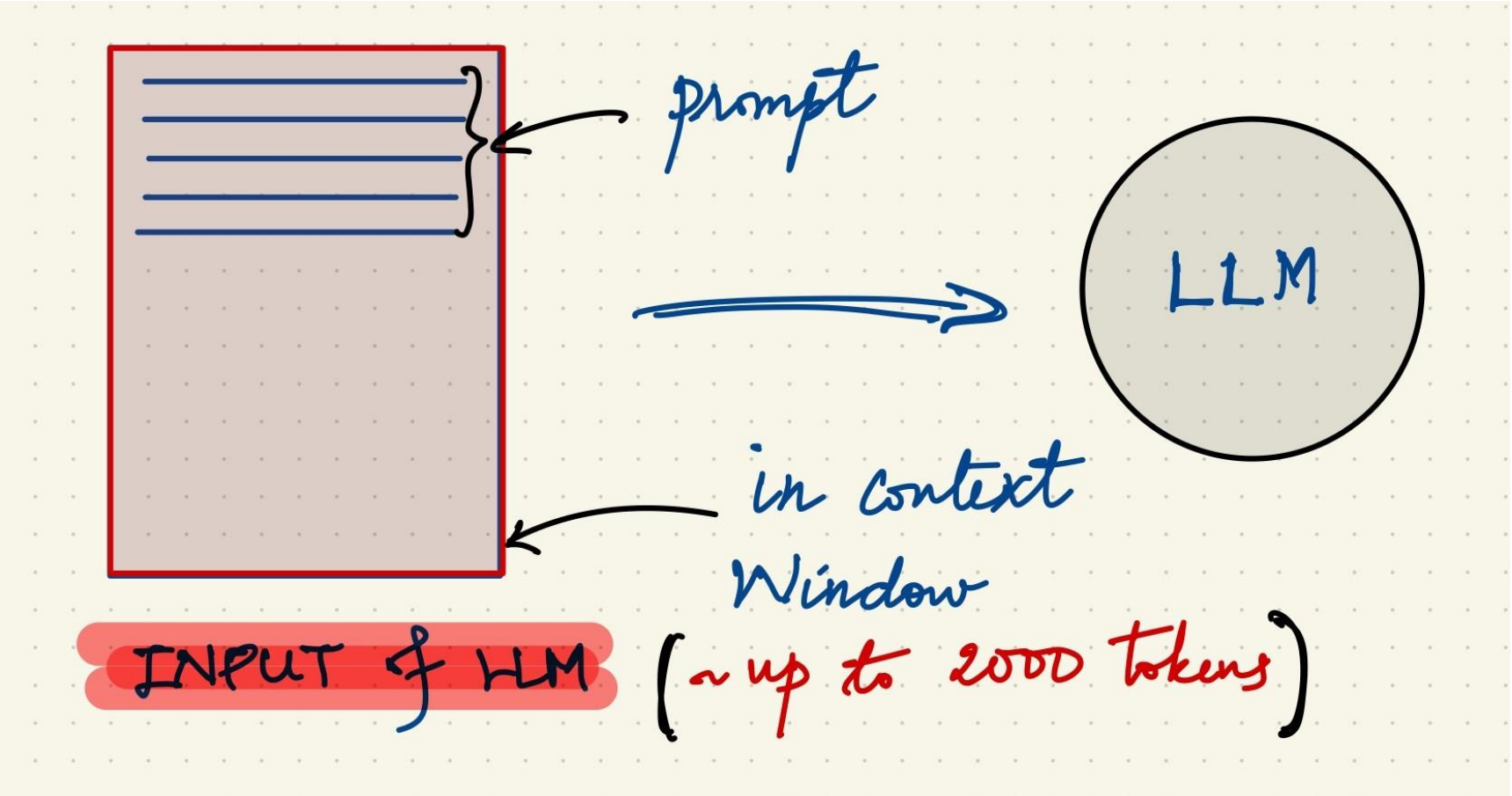
DOCUMENTS LOADING ?



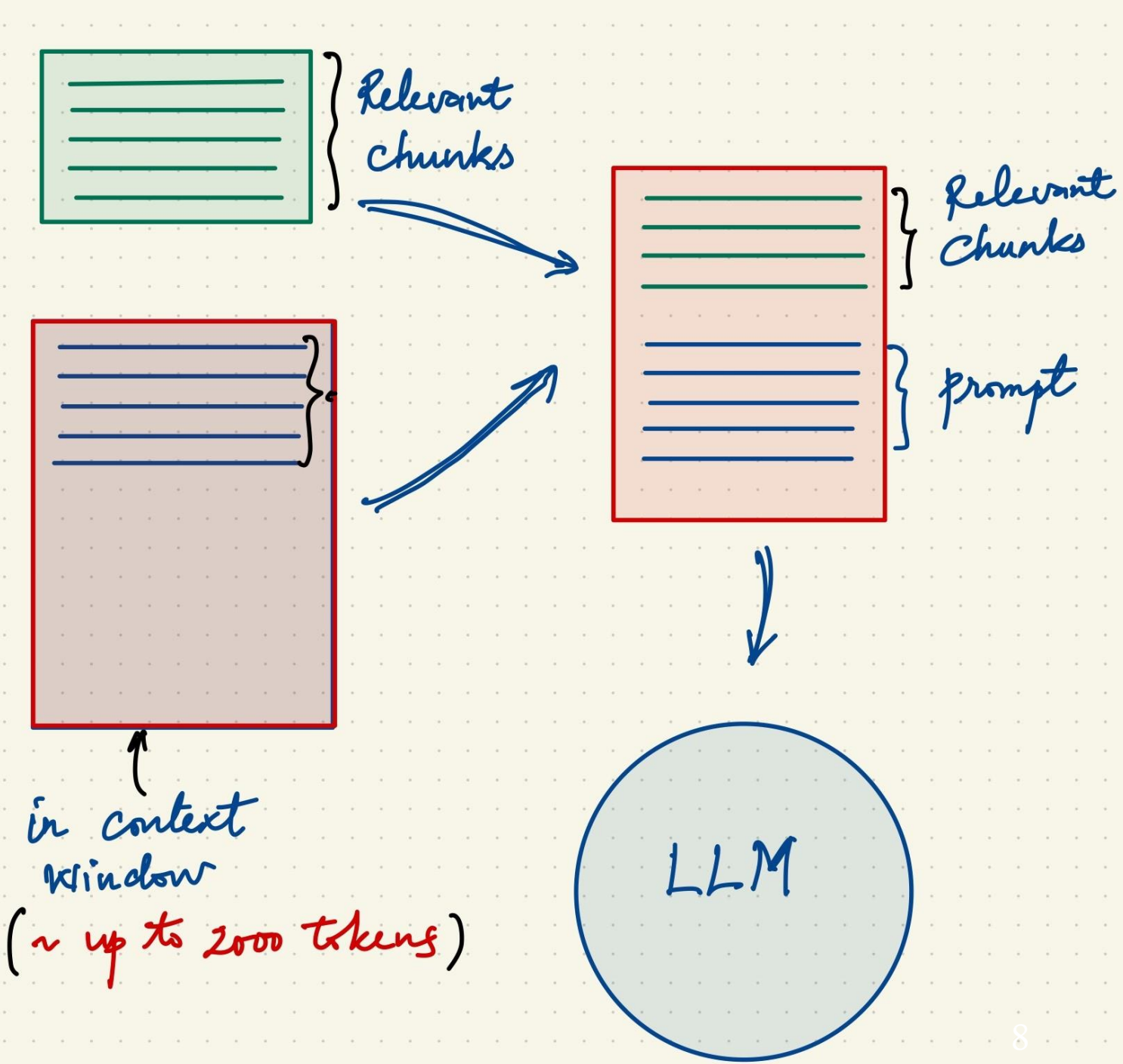


DOCUMENTS SPLITTING ?

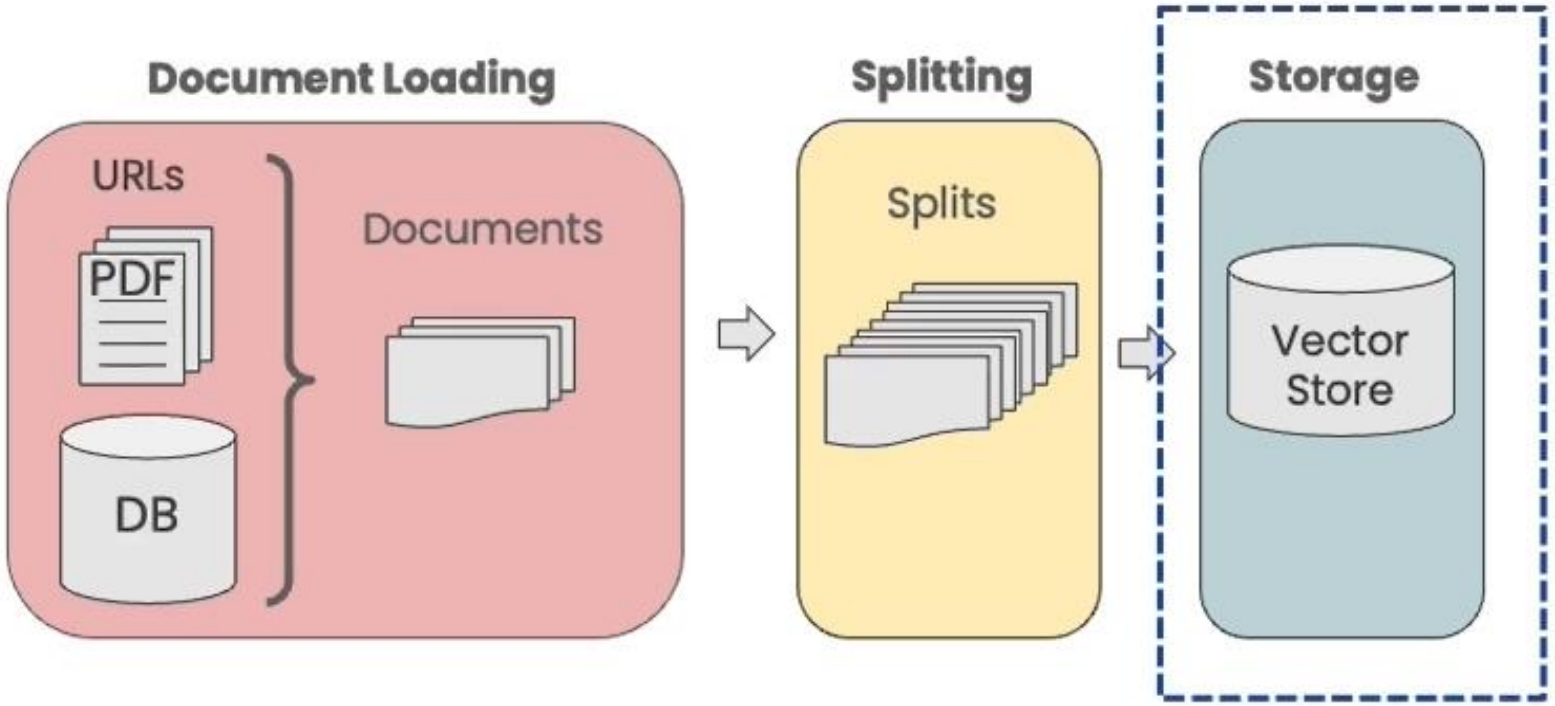
IN CONTEXT WINDOW ?



IN CONTEXT WINDOW AUGMENTED ?



EMBEDDING AND VECTORSTORES



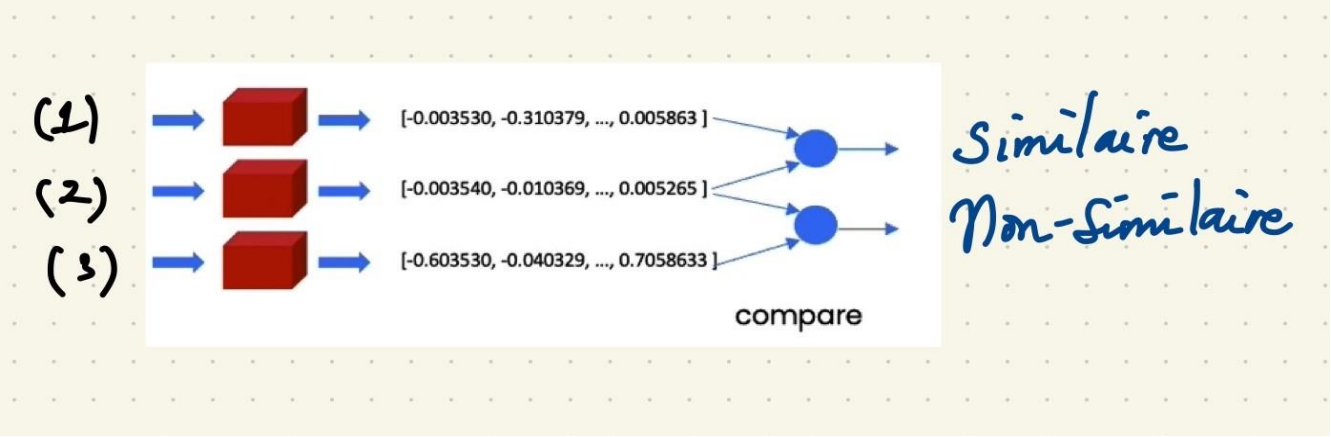
EMBEDDING(1)

- Embeddings: créer une représentation numérique d'un texte;

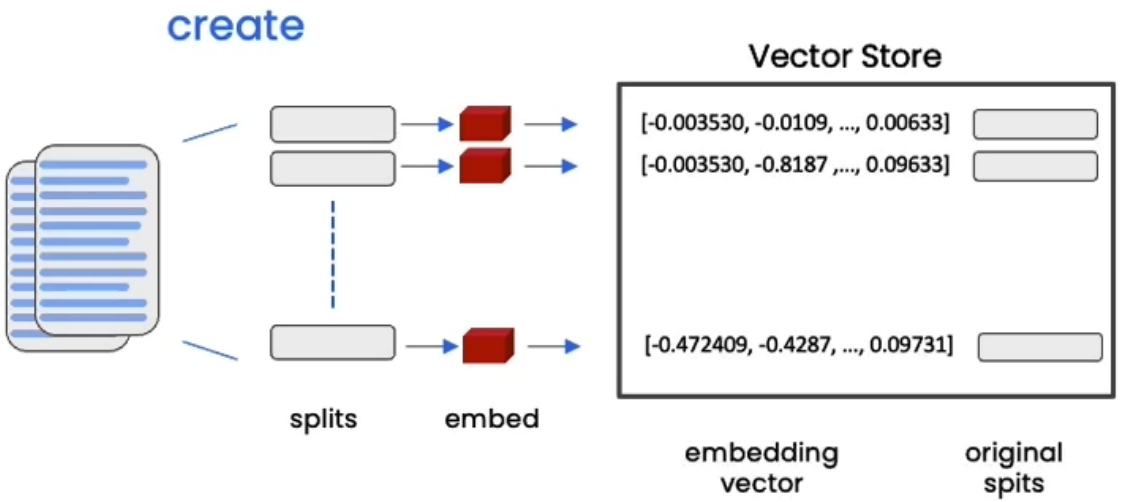
EMBEDDING(2)

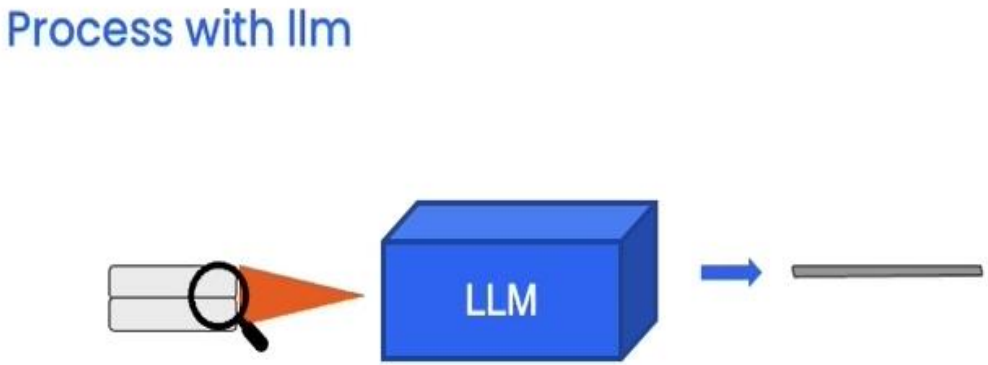
- Les textes dont le contenu est sémantiquement similaire auront des vecteurs similaires dans l'espace d'embedding.

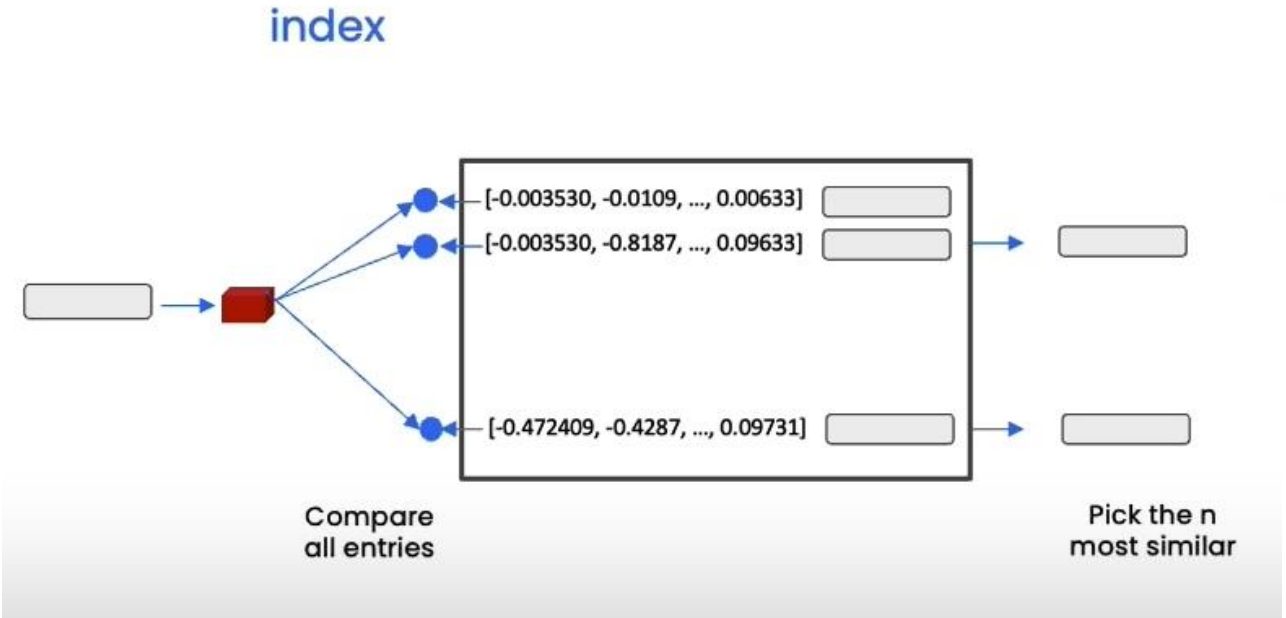
1) C'est un Chien;
2) C'est une canine;
3) Le temps est mauvais.



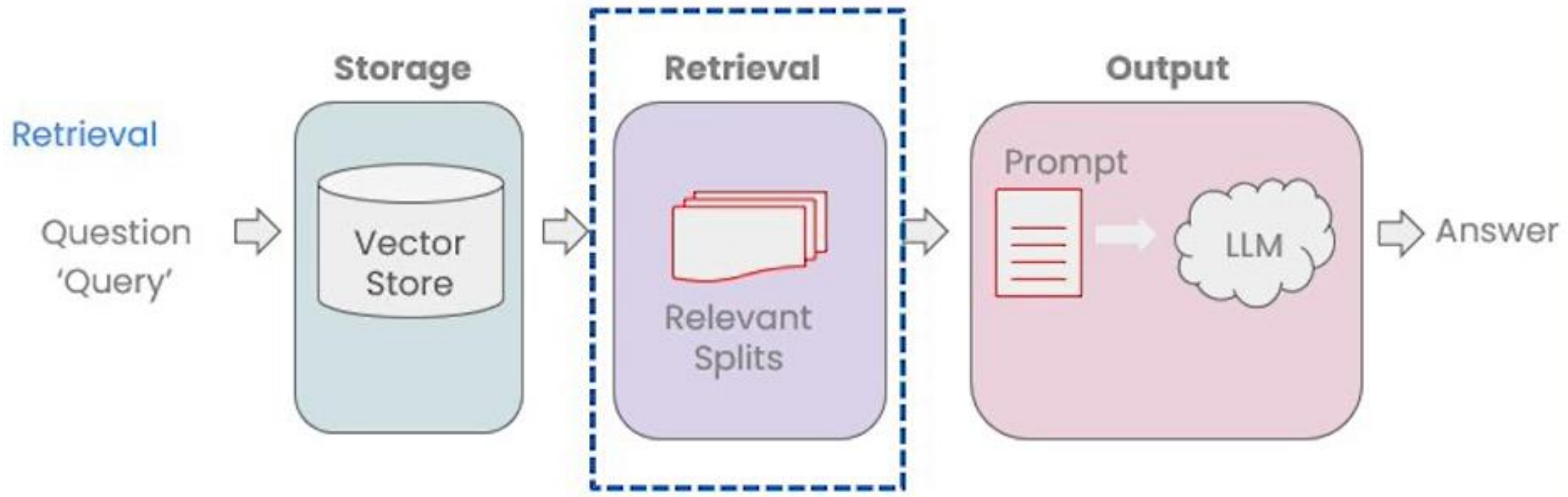
VECTOR STORES



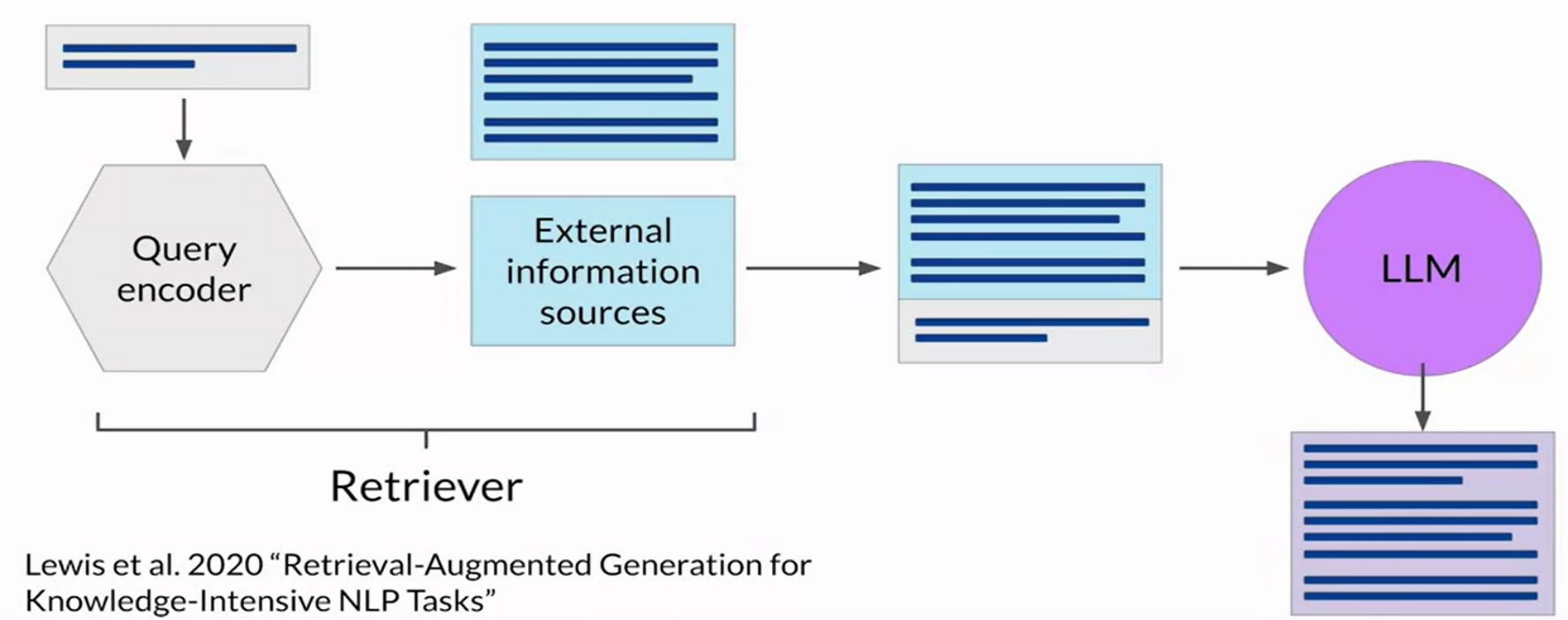




RETRIEVAL


Retrieval
Question
'Query'
Storage
Vector
Store
Retrieval
Relevant
Splits
Output
Prompt
LLM
Answer

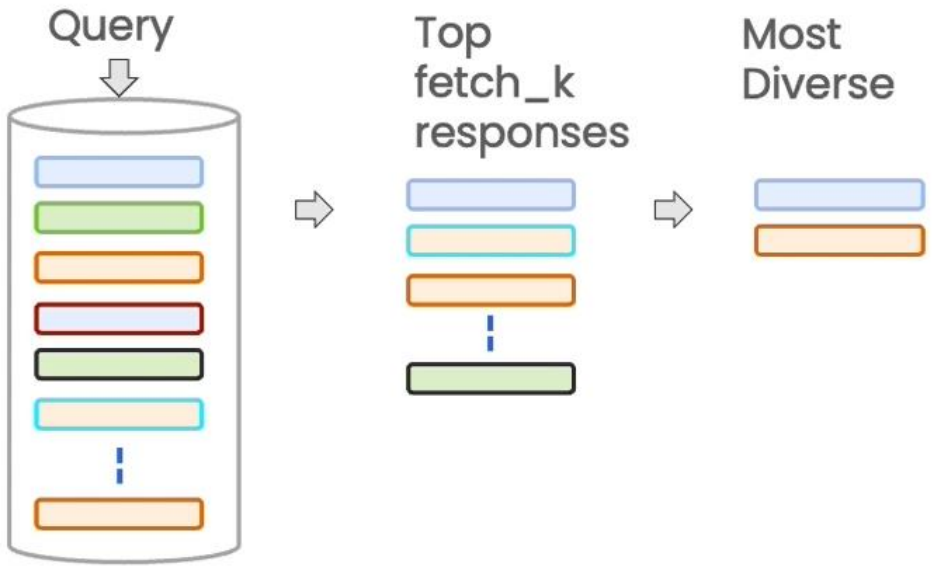
STRUCTURE DU RETRIEVAL



Lewis et al. 2020 "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks"

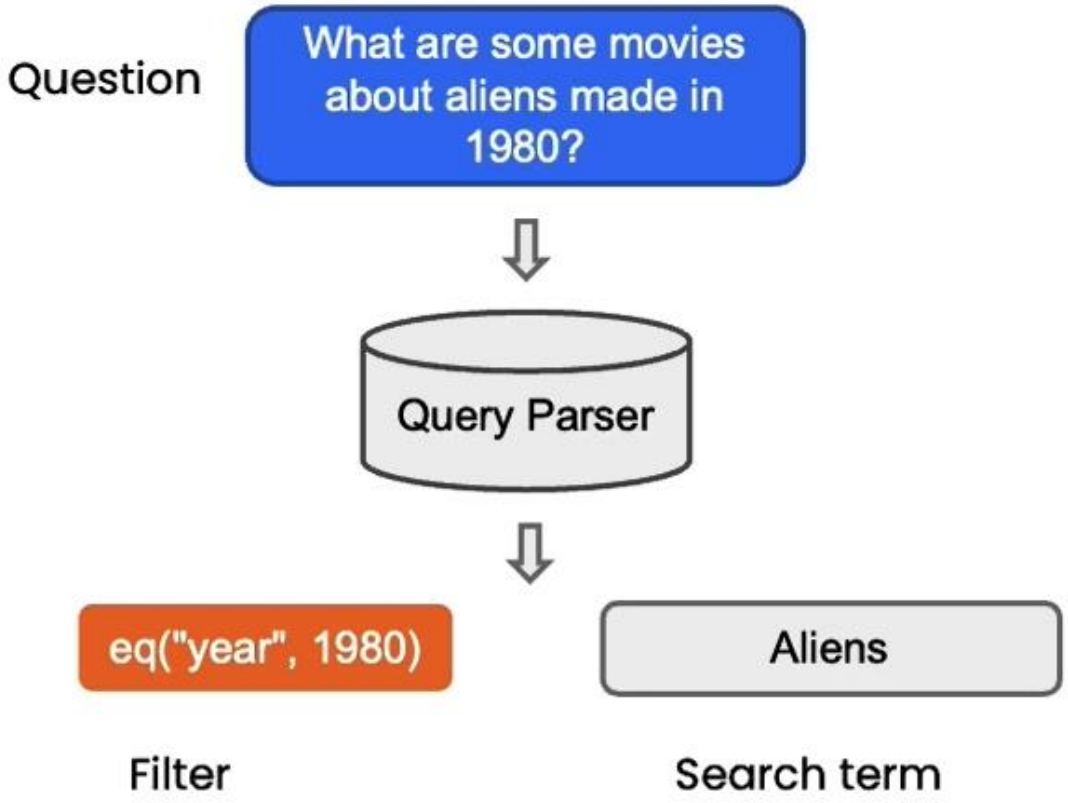
MAXIMUM MARGINAL RELEVANT (MMR)

- Récuperation des informations pertinantes basée sur la **similarité sémantique** (via fetch_k avec langchain);
- Optimisation de la **diversité** des informations (via k)



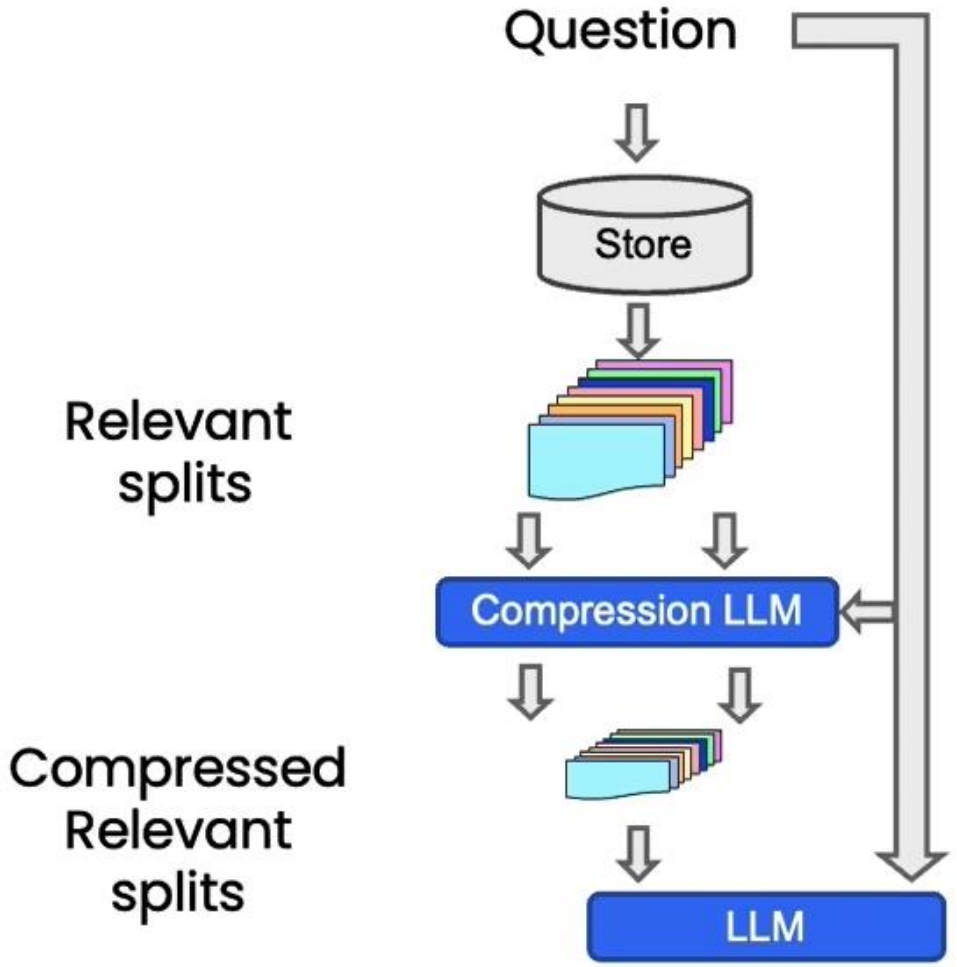
SELF-QUERY

- Usage sur des questions possédant des **similarités sémantiques** et des **filtres** sur des metadata ;
- Conversion de la question requête avec deux terme: le filtre(**filter**) et le terme de recherche(**search term**).



COMPRESSION

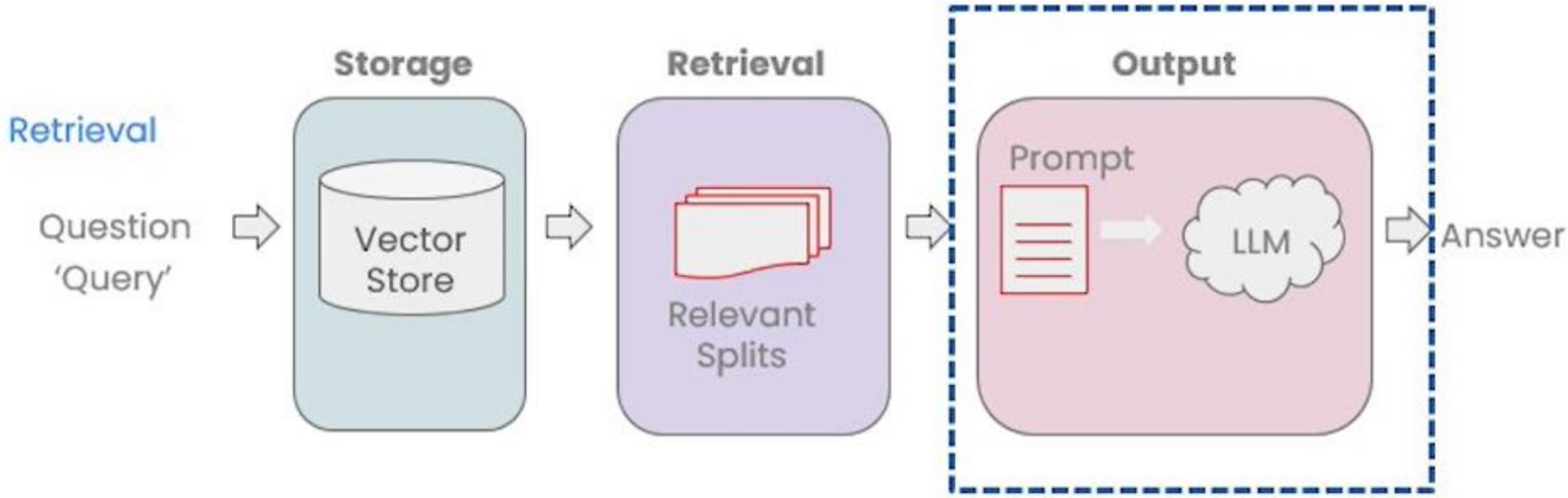
- Transition des documents pertinents via un **compression LLM** ;
- Extraction des bouts de segments pertinents vers LLM final.



AUTRES TYPES DE RECUPÉRATEURS

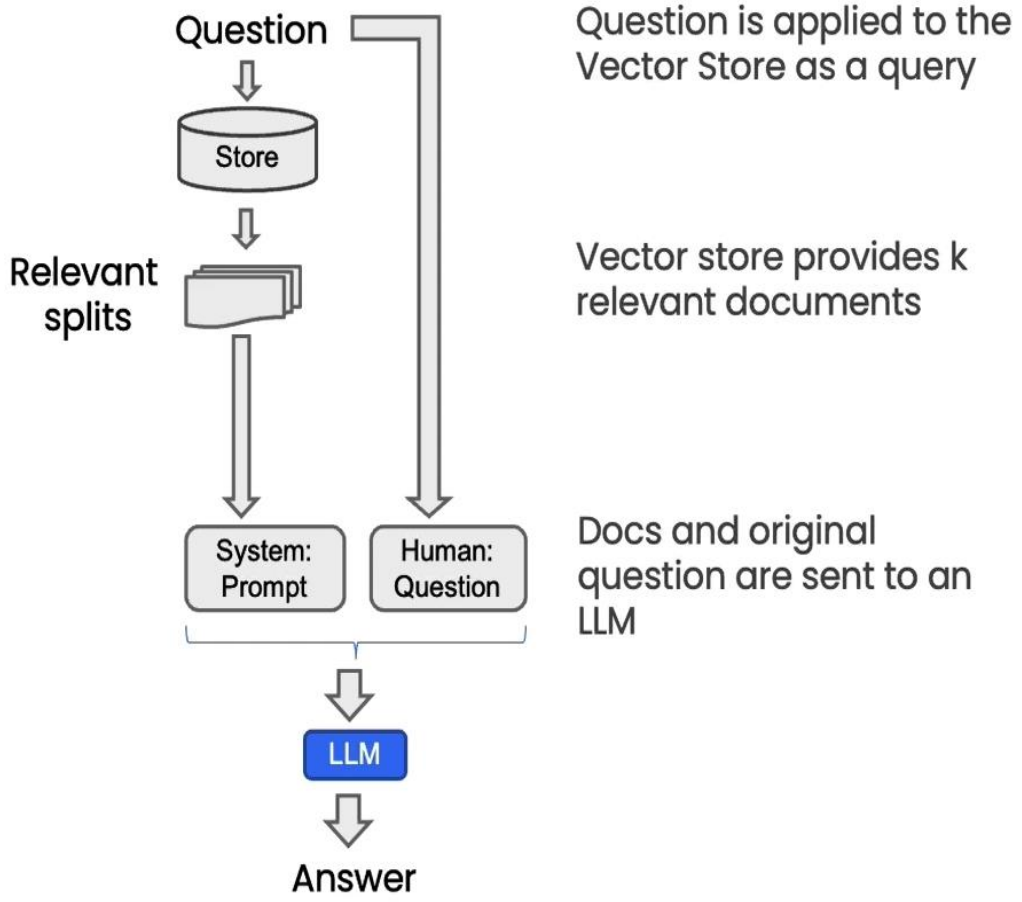
- SVMRetriever (Support Vector Machine)
- TFIDFRetriever (TF-IDF)
- …

QUESTION-ANSWER(QA)



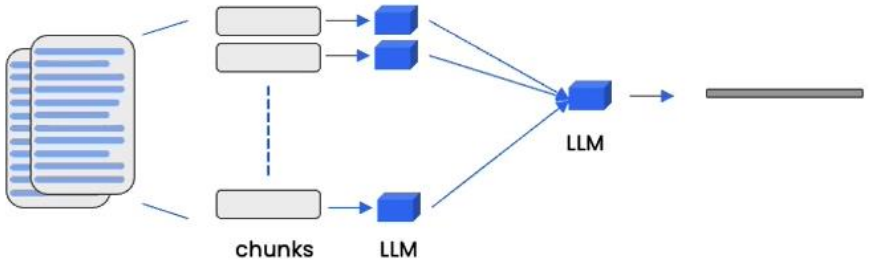
RETRIEVALQA CHAIN

RetrievalQA.from_chain_type(,
chain_type = "stuff", ...)

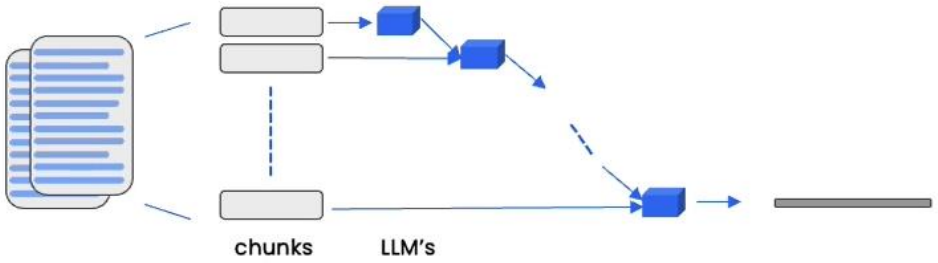


CHAINS TYPE

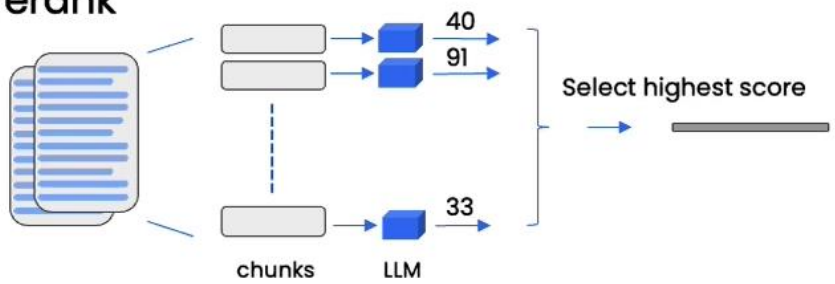
1. Map_reduce

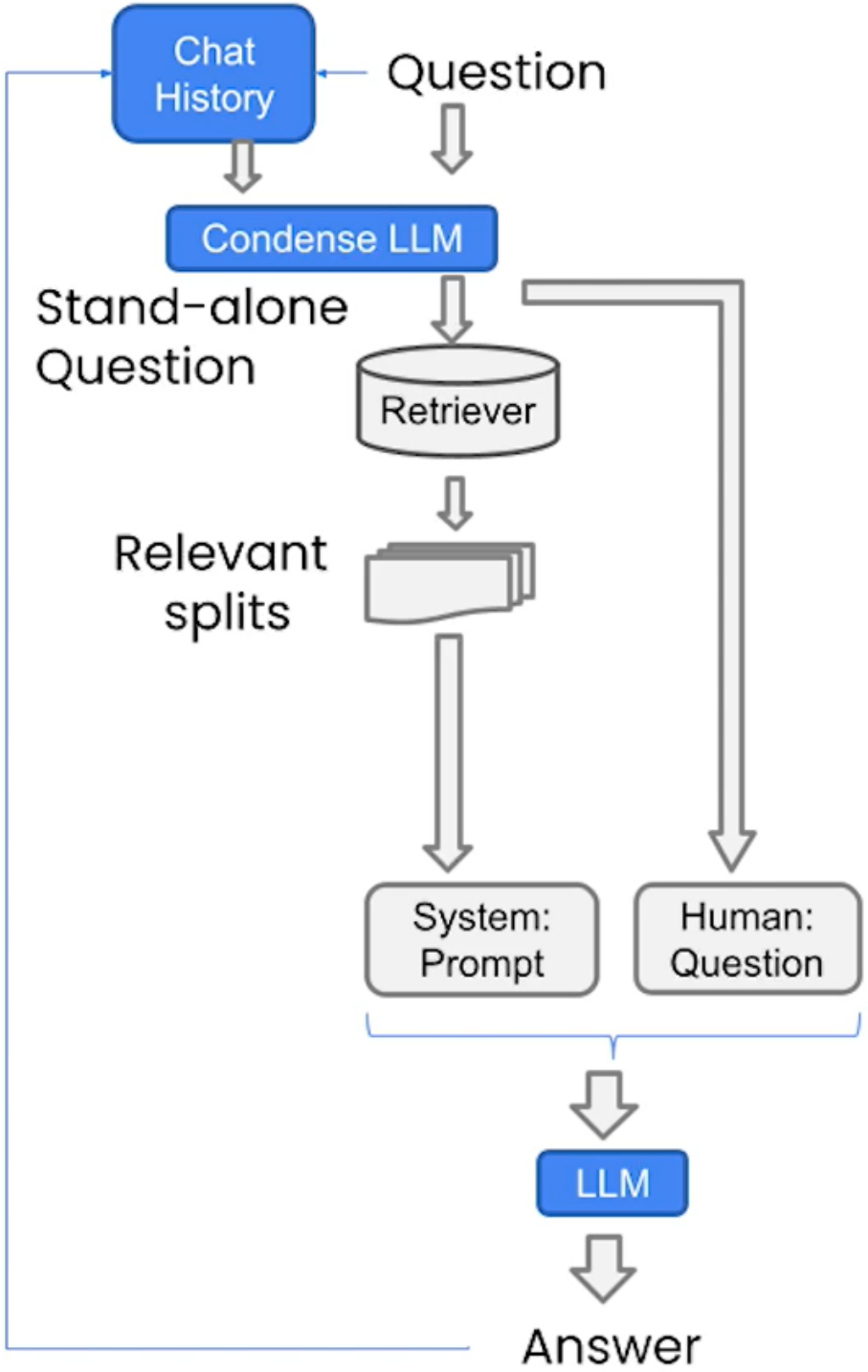


2. Refine



3. Map_rerank





CONVERSATIONAL RETRIEVAL CHAIN

- qa = **ConversationalRetrievalChain**.from_llm(ChatOpenAI(temperature = 0), vectorstore.as_retriever(), memory = memory)

ChatGPT API & LangChain

MERCI

GUIDJIME ADINSI Ahouahounko