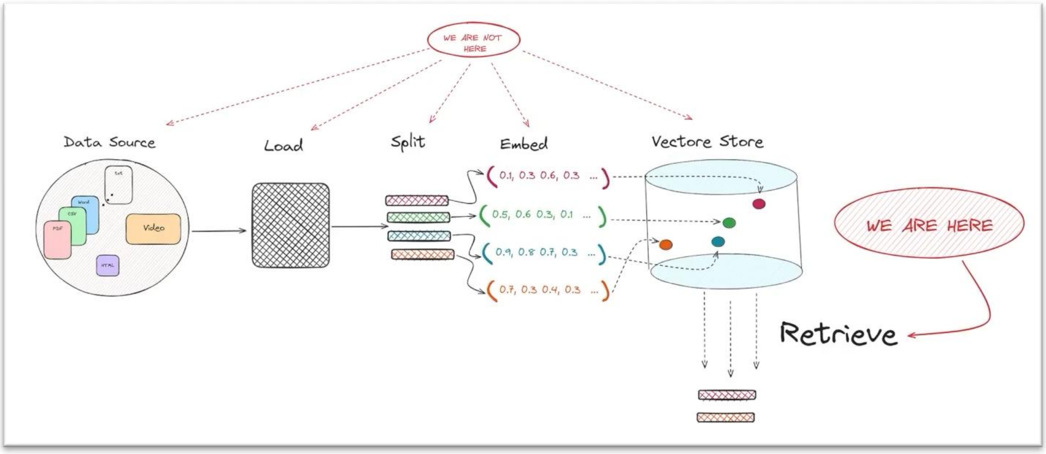
Retrievers:

Retrievers

Retrievers are responsible for finding the most relevant chunks of information from a vector store based on a user's query. They ensure that only the most semantically similar and meaningful data is passed to the language model for accurate response generation.

Retrievers

Retrievers are components responsible for **fetching the most relevant chunks of information** from a vector database or knowledge source based on a user query.
They act as a bridge between the **query** and the **relevant context** used by the LLM.



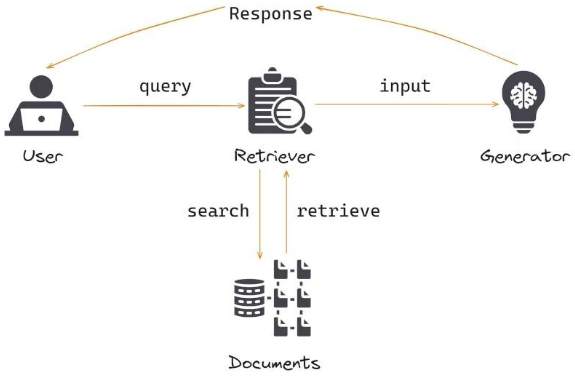
Where Retrievers Fit in a RAG Pipeline

User Query → Embed → 🔎 Retriever → Top Relevant Chunks → LLM → Answer

- ❑ Without retrievers, an LLM must rely on its pre-trained knowledge.
- ❑ With retrievers, we **ground** the model with up-to-date, accurate, and domain-specific information.

How Retrievers Work

- ❑ **Embed the Query** using the same model used to embed documents.
- ❑ **Search the Vector Store** (FAISS, Pinecone, ChromaDB, etc.)
- ❑ **Return Top-K Chunks** based on similarity metrics (cosine, dot product, etc.).



Common Retriever Types

Retriever Type	Description
Vector Store Retriever	Uses dense vector similarity to return most similar documents.
BM25 Retriever	Keyword-based retriever using classic information retrieval.
Hybrid Retriever	Combines vector search and keyword search (semantic + keyword).
Multi-query Retriever	Uses multiple query rewrites for better recall.
Parent Document Retriever	Retrieves large docs and chunks them later for better context.

Key Parameters

Parameter	Purpose
search_type	Type of retrieval (e.g., similarity, MMR, score)
k	Number of documents to retrieve
filter	Optional metadata filtering
threshold	Minimum similarity score (if available)

Benefits of Using a Retriever

Advantage	Why It Matters
Focused Context	LLMs get only relevant context → better answers
Lower Token Usage	Avoids passing huge documents into prompt
Plug in External Data	Enables LLMs to answer from private knowledge
Dynamic Updating	Vector DBs can be updated anytime without re-training

Use Cases

❑ Chat with PDF / Docs / Websites

❑ AI Customer Support Agents

❑ Medical, Legal, or Financial Search Tools

❑ AI Tutors with Curriculum Knowledge