What is Gen AI?

- Generative AI refers to machine learning algorithms that enable computers to use existing content like text, audio and video files, images, and even code to analyze and generate new content.

Key activities can be performed using Gen AI

- **Written content augmentation and creation:** Producing a "draft" output of text in a desired style and length
- **Question answering and discovery:** Enabling users to locate answers to input, based on data and prompt information
- **Tone:** Text manipulation, to soften language or professionalize text
- **Summarization:** Offering shortened versions of conversations, articles, emails and webpages
- **Simplification:** Breaking down titles, creating outlines and extracting key content
- **Classification of content for specific use cases:** Sorting by sentiment, topic, etc.
- **Chatbot performance improvement:** Bettering "sentity" extraction, whole-conversation sentiment classification and generation of journey flows from general descriptions
- **Software coding:** Code generation, translation, explanation and verification

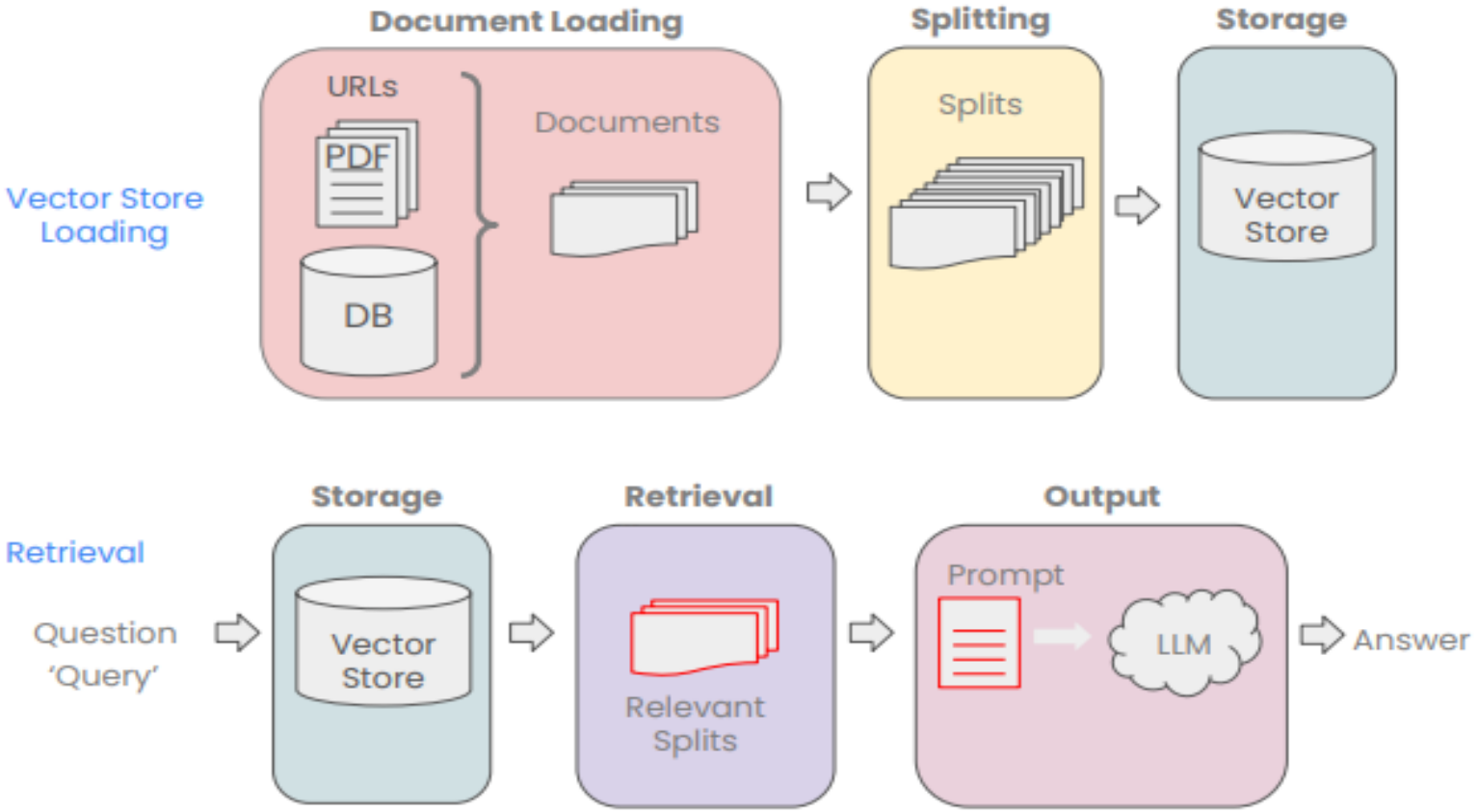
Key players in Gen AI

- **Open AI**
- **AWS**
- **Anthropic**
- **Google etc.,**

Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) is a very popular paradigm.

- Retrieve relevant documents and load into “working memory” / context window.



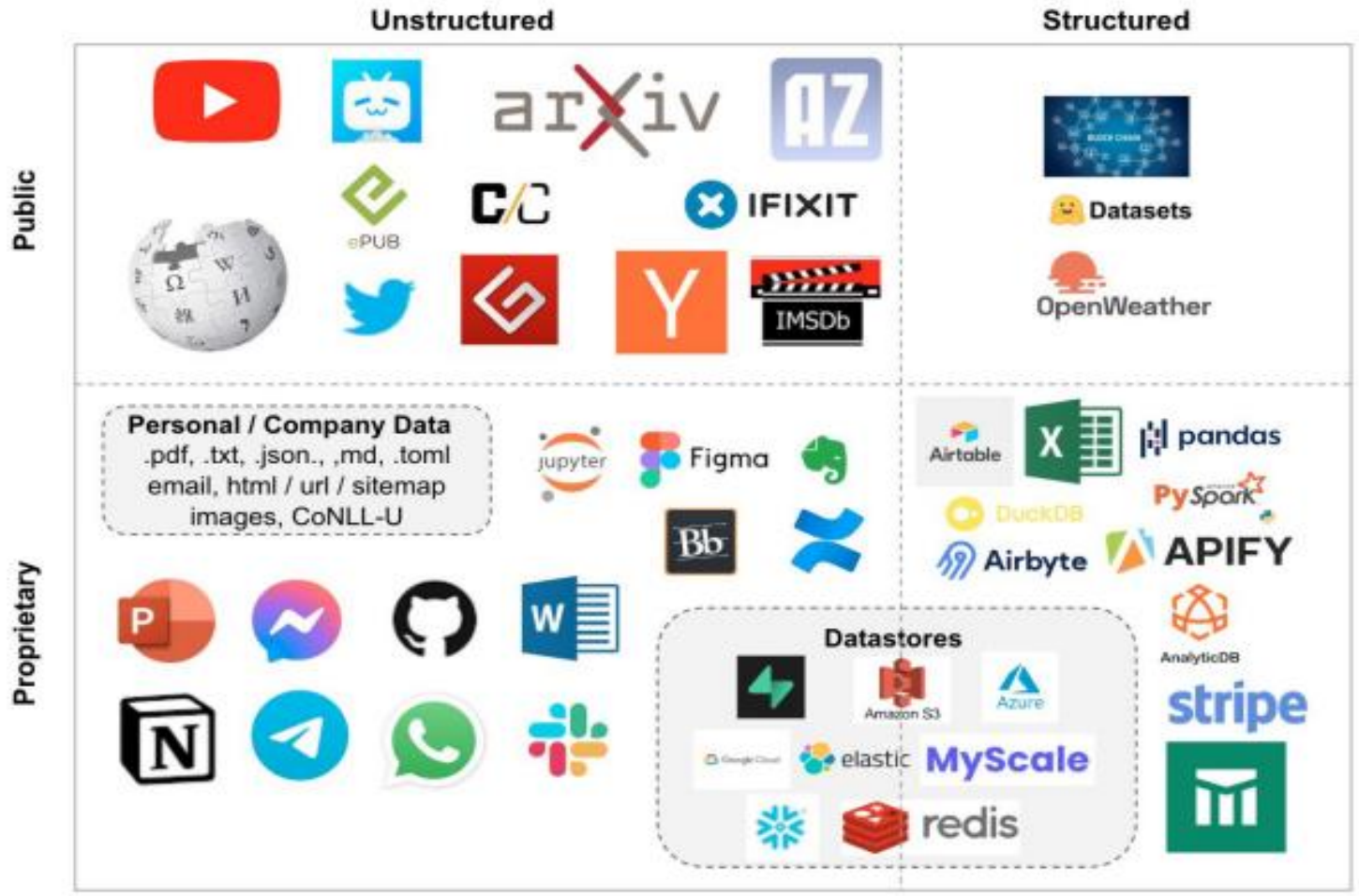
Chat with your data

- Document Loading

Loaders

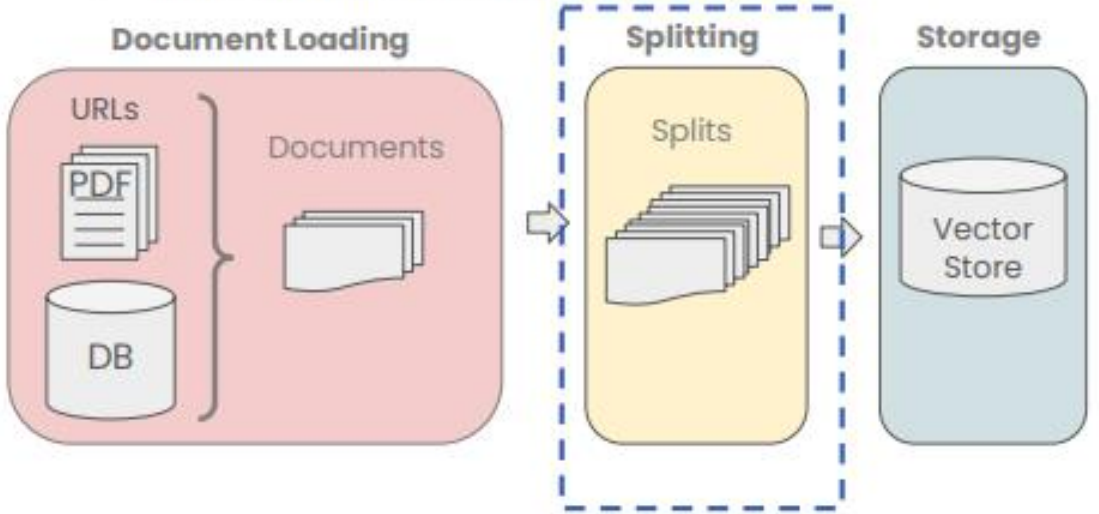
- Loaders deal with the specifics of accessing and converting data
  - Accessing
    - Web Sites
    - Data Bases
    - YouTube
    - arXiv
    - ...
  - Data Types
    - PDF
    - HTML
    - JSON
    - Word, PowerPoint...
- Returns a list of `Document` objects:

Document Loaders



Document Splitting

- Splitting Documents into smaller chunks
  - Retaining meaningful relationships!



...
on this model. The Toyota Camry has a head-snapping
80 HP and an eight-speed automatic transmission that will
...

Chunk 1: on this model. The Toyota Camry has a head-snapping

Chunk 2: 80 HP and an eight-speed automatic transmission that will

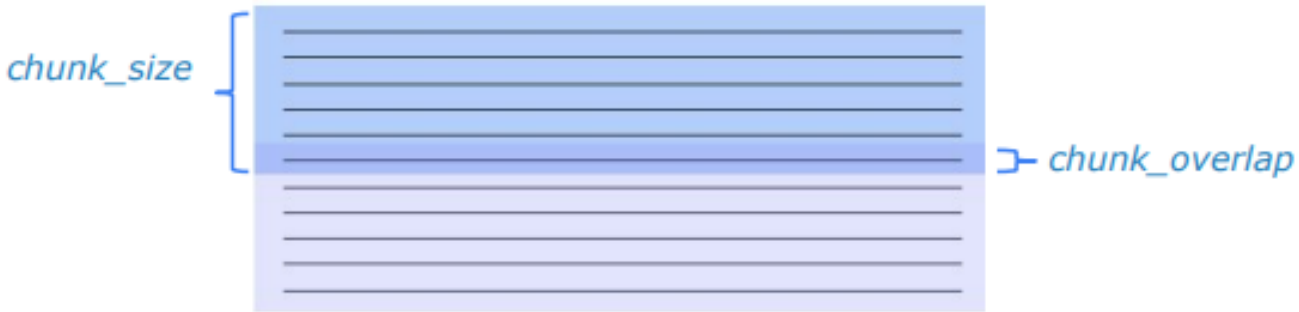
```
langchain.text_splitter.CharacterTextSplitter(
    separator: str = "\n\n"
    chunk_size=4000,
    chunk_overlap=200,
    length_function=<builtin function len>,
)
```

Methods:

create_documents() - Create documents from a list of texts.

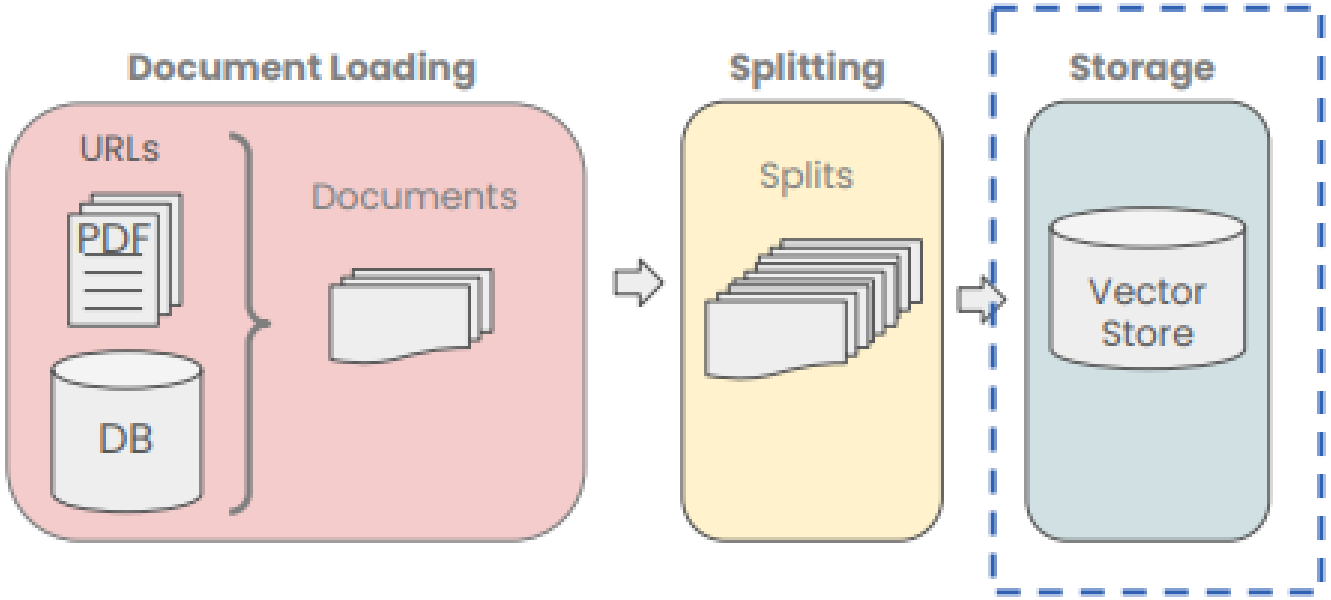
split_documents() - Split documents.

Sample Splitter Example



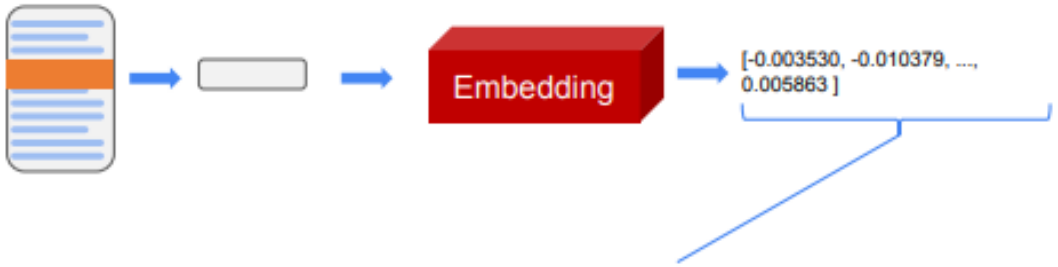
URL to refer : Text Splitters | 🦜🔗 Langchain

Vector Stores and Embeddings



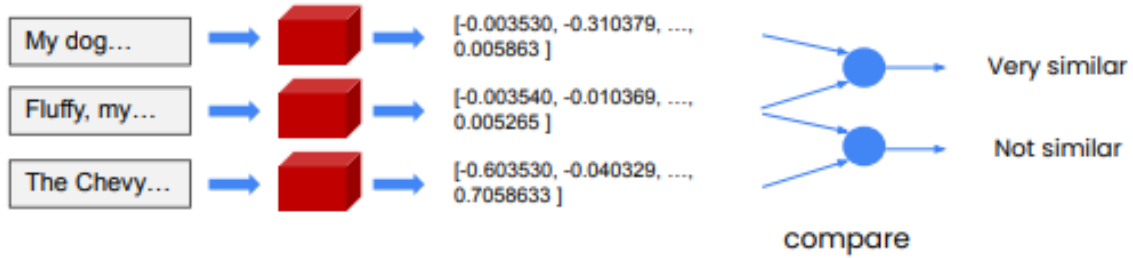
Vector stores | 🦜🔗 Langchain

Embeddings

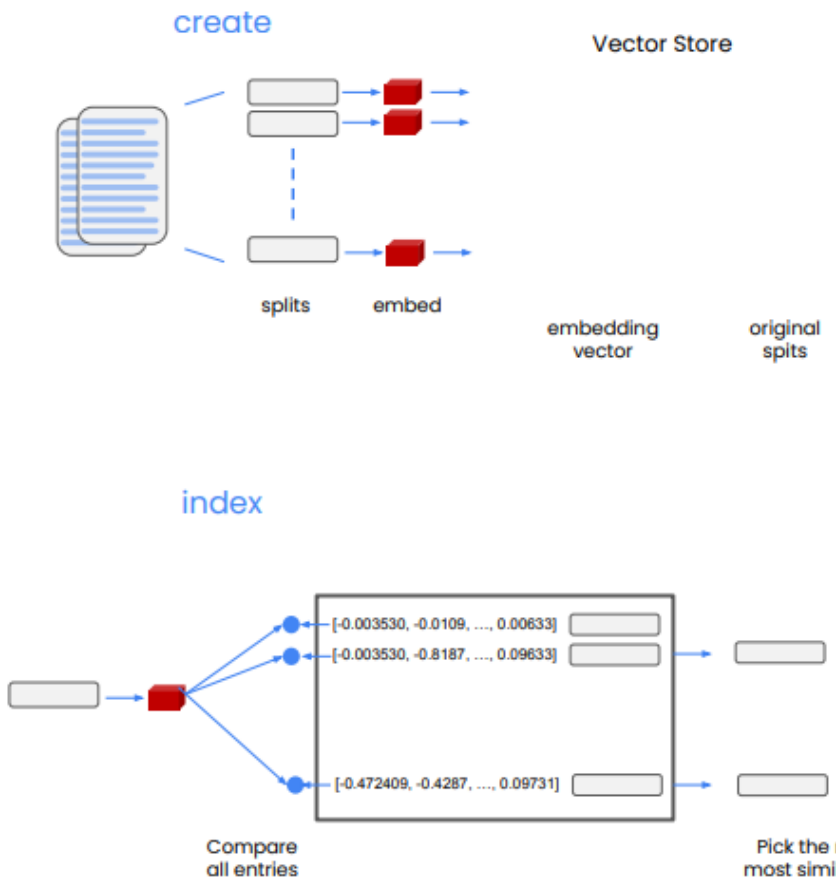


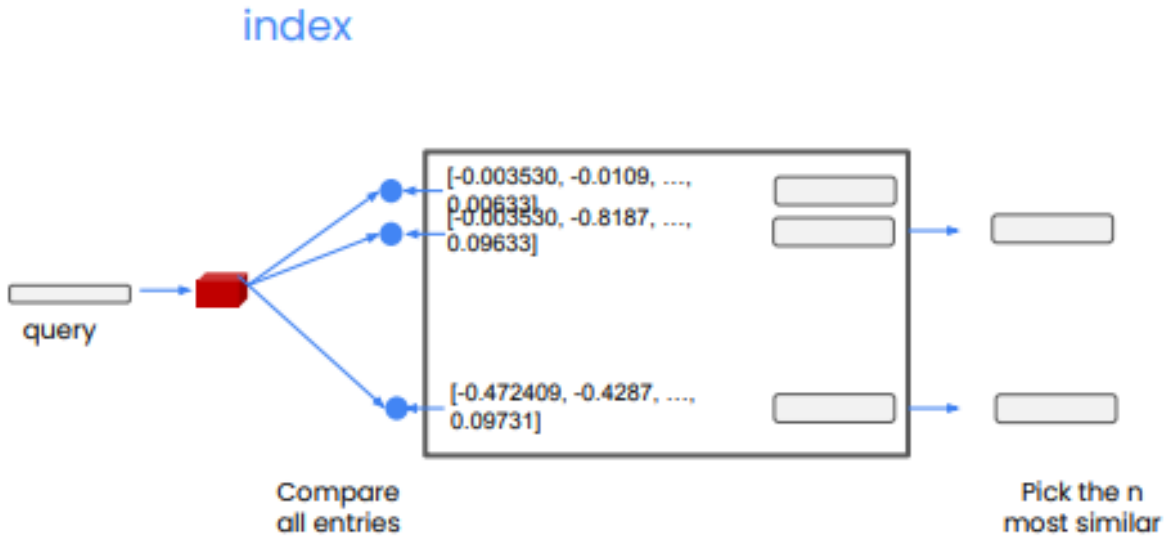
- Embedding vector captures content/meaning
- Text with similar content will have similar vectors

1) My dog Rover likes to chase squirrels.
2) Fluffy, my cat, refuses to eat from a can.
3) The Chevy Bolt accelerates to 60 mph in 6.7 seconds.

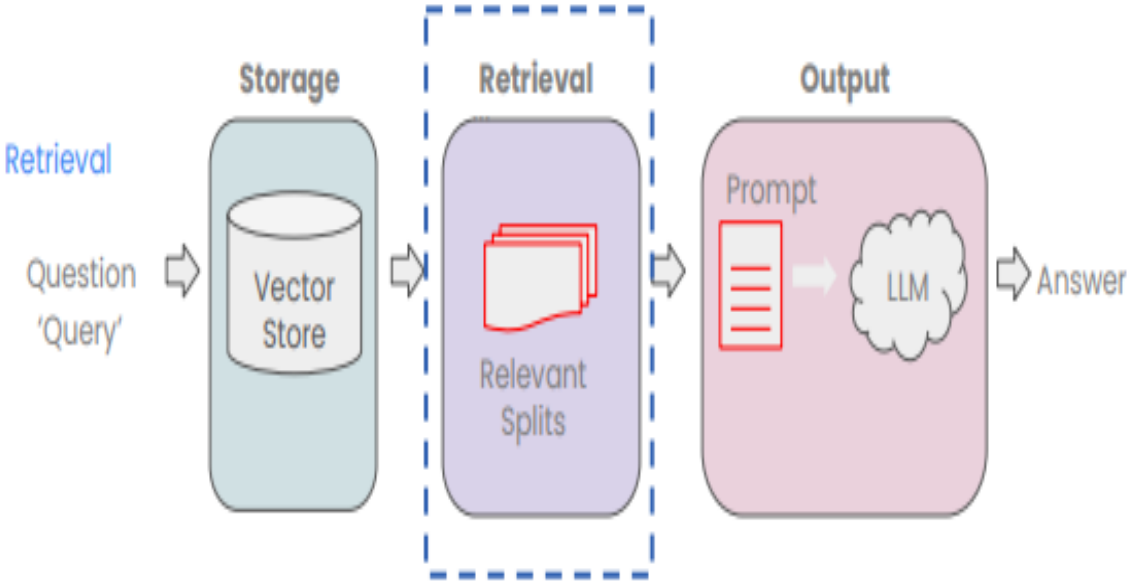


Vector Store / Data base





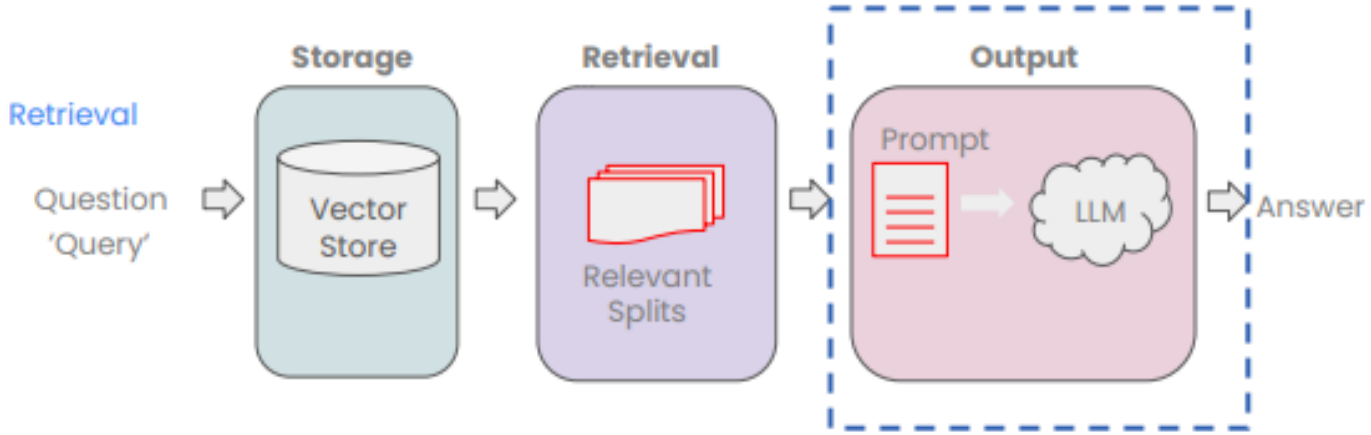
Retrieval



Process with llm

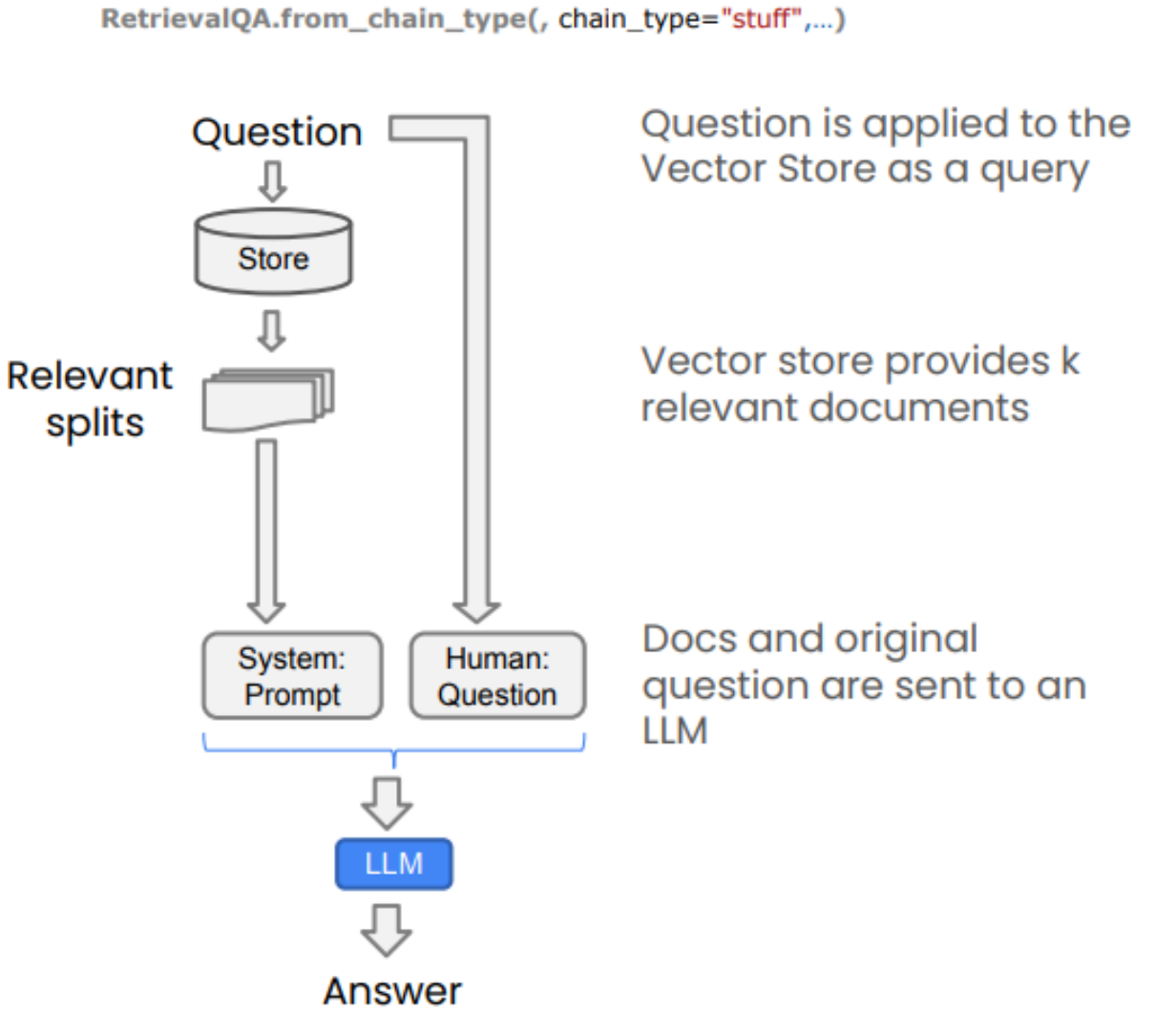
The returned values can now fit in the LLM context

Question Answering



- Multiple relevant documents have been retrieved from the vector store
- Potentially compress the relevant splits to fit into the LLM context
- Send the information along with our question to an LLM to select and format an answer

RetrievalQA chain

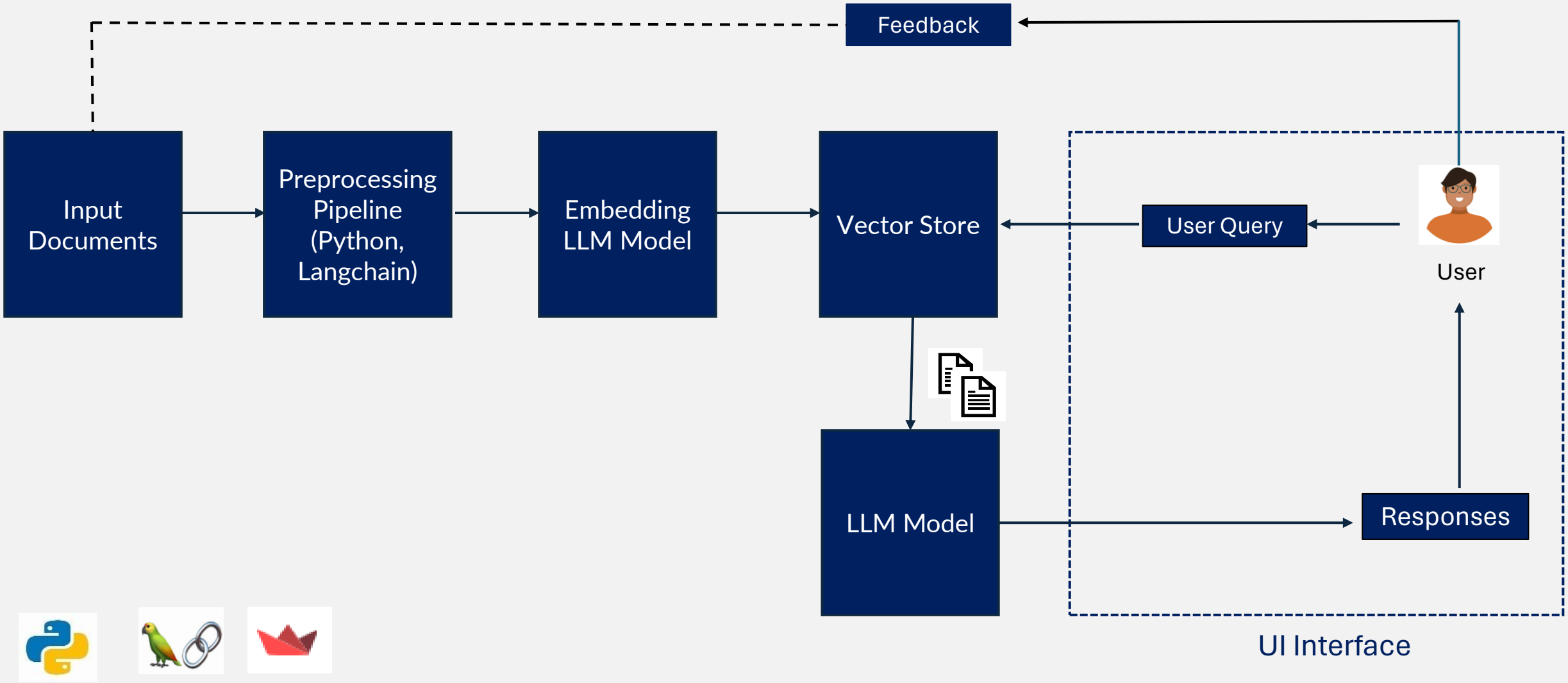


Chain types : [Mastering Document Chains in LangChain - Comet]

Things to Explore..!

- Agents
- Grounding vs Finetuning
- Involving Feedback Based Learning (RLHF)
- Serving the Gen AI Application
- Monitoring etc.,

Demo Architecture - Advanced



Thank you