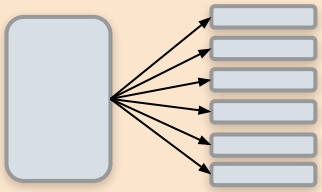
Query

Input: query from the user for which the domain are the documents.

E.g) ***What is the current buy price of the described stock?***

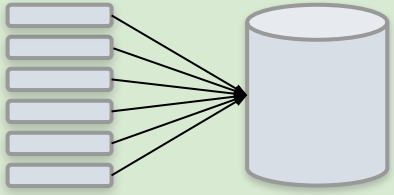
Indexing

Chunking



The uploaded document is split into ***chunks***. These chunks are then converted to **vectors** via *embedding*.

Embedding



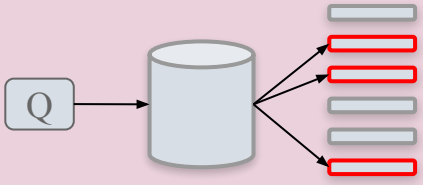
Each vector is then ***embedded*** into a **Vector Database.**

User

Generation

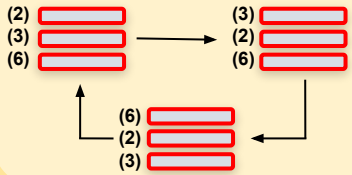
The **LLM** *combines* the final ordered relevant documents and ***generates*** an output.

Retrieval



Similarity Search is used on the Vector DB and embeddings within to ***retrieve*** chunks (split documents) that are **most relevant** to the inputted query.

Reranking



After relevant chunks are retrieved, they are ***re-ranked*** based on relevance to query

Output

The **LLM** returns **generated output** in the form of a comprehensible response based on the retrieval of relevant chunks, re-ranking, and the combination of those chunks.

E.g) ***The current buy price of the stock for XYZ Corp is $125.75. Please note that stock prices fluctuate constantly due to market conditions, so it's important to check a reliable financial news source or brokerage platform for the most up-to-date information.***