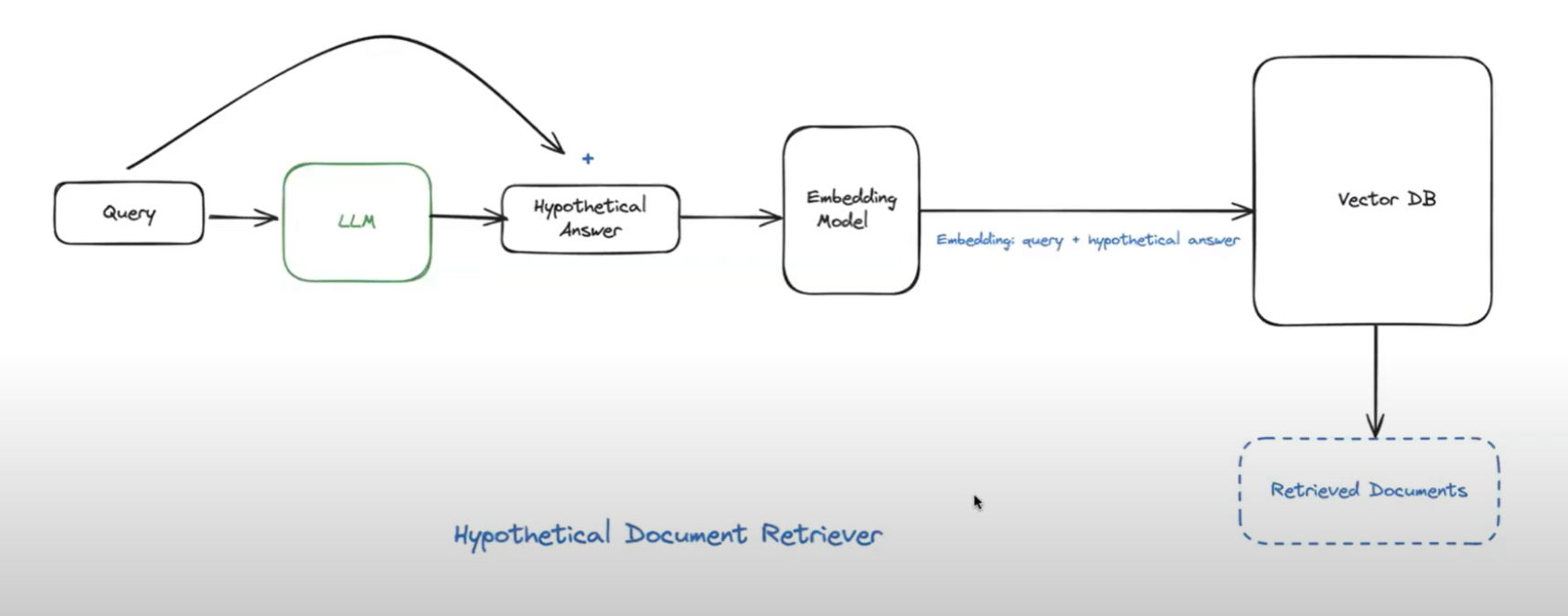
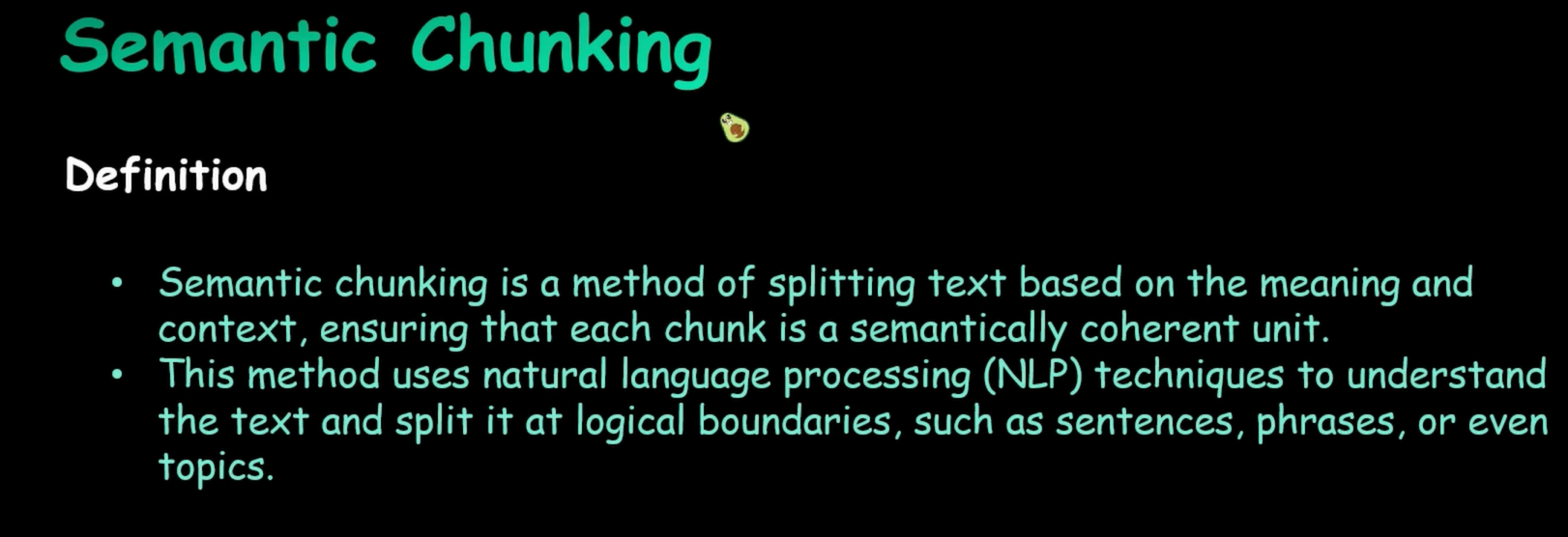
**HyDE - Hypothetical Document Embedding**



YouTube link: <https://www.youtube.com/watch?v=FHUX0hlYNZo>

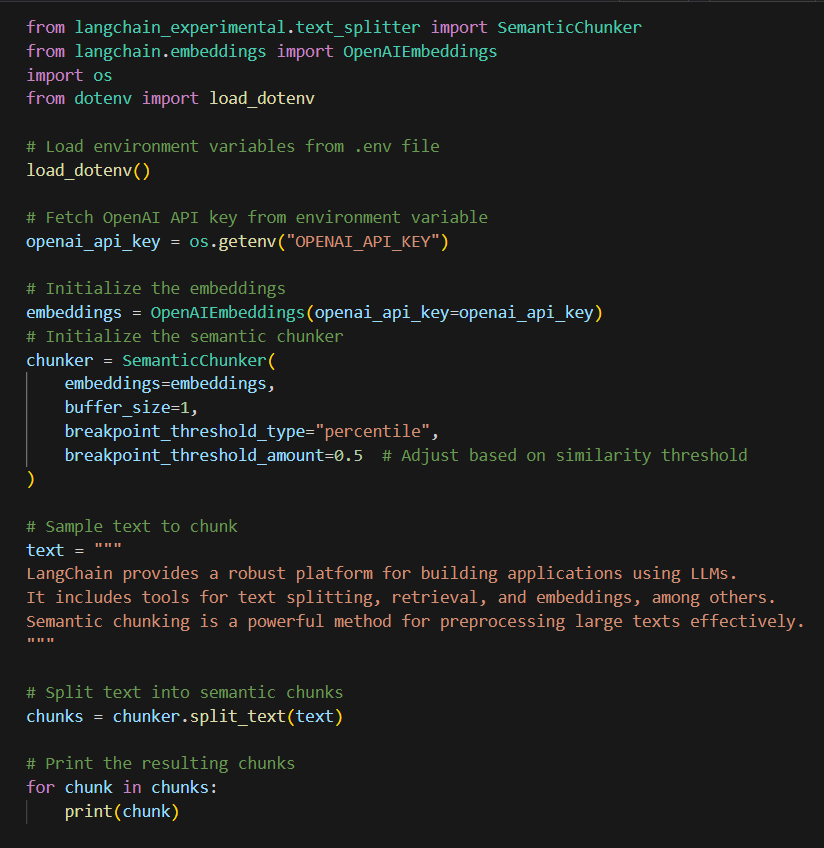
1. hypothetical document embedding to enhance query retrieval in language models, improving document relevance through semantic similarity
2. **Hypothetical Document Approach**: Using LLM to generate hypothetical answers for queries.
3. **Enhanced Retrieval**: Embedding both the query and hypothetical answer improves document search.
4. **Semantic Similarity**: Queries expanded in vector space lead to better document matches.
5. **Leveraging LLMs**: By generating hypothetical answers, we can significantly widen the search space for relevant documents, leading to more effective retrieval strategies.
6. **Vector Space Expansion**: Embedding both the user query and a generated hypothetical answer allows the model to tap into a larger, more relevant vector space, thereby increasing the likelihood of finding pertinent documents.

**Semantic chunking**

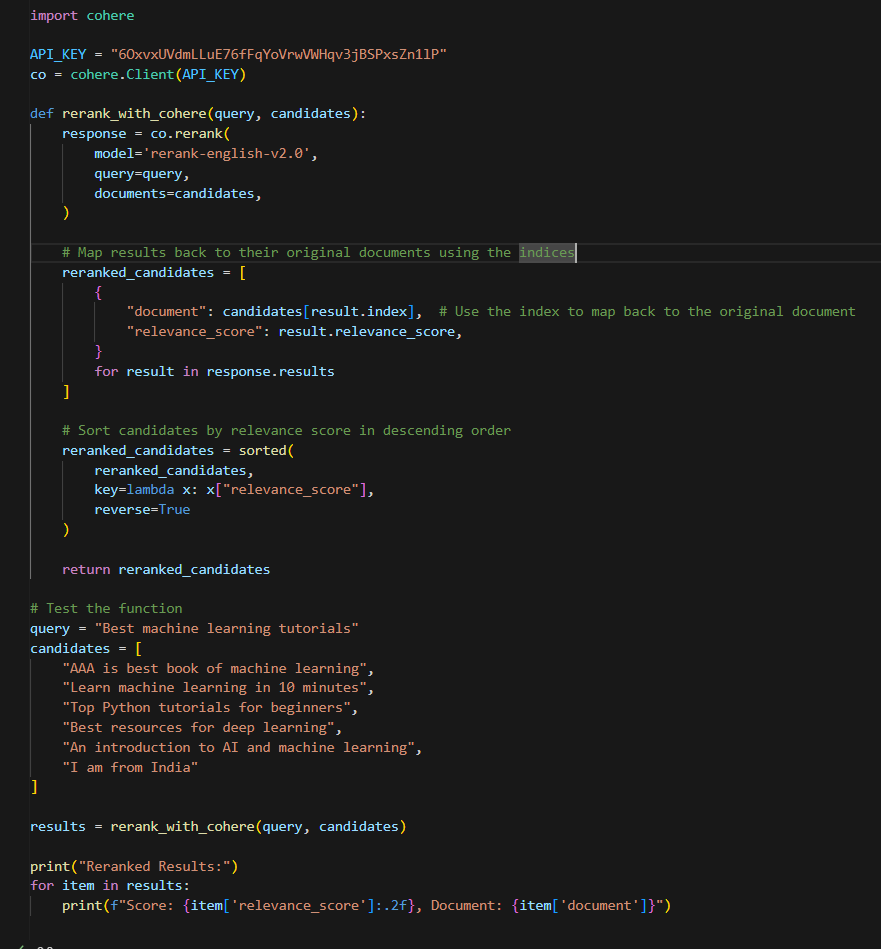


A black background with white text

Description automatically generated



Re ranking using cohere AI API



Output:

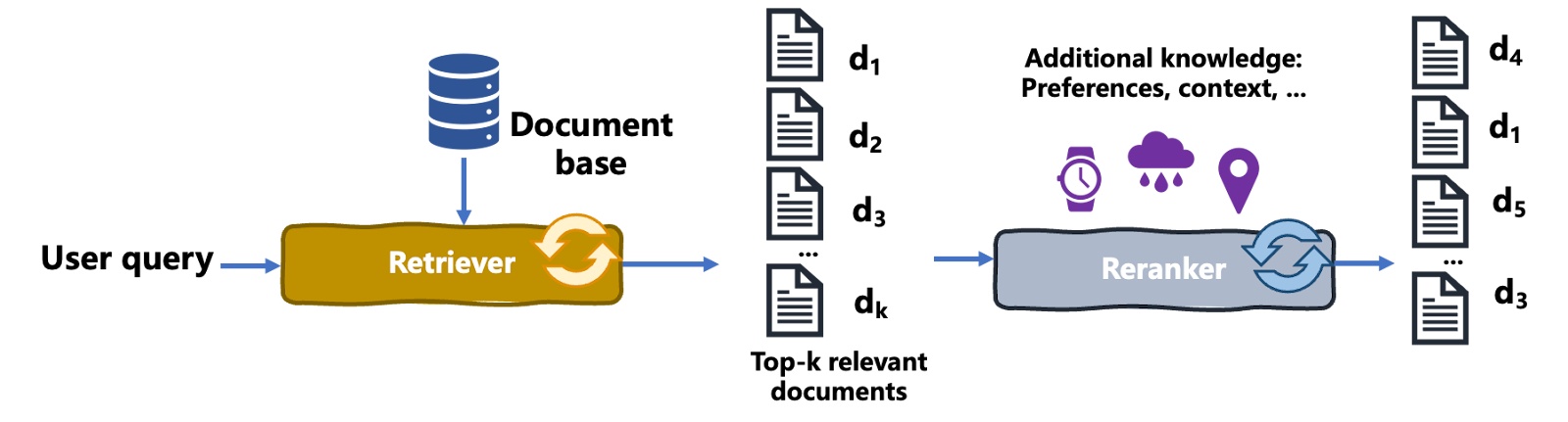
A screen shot of a computer

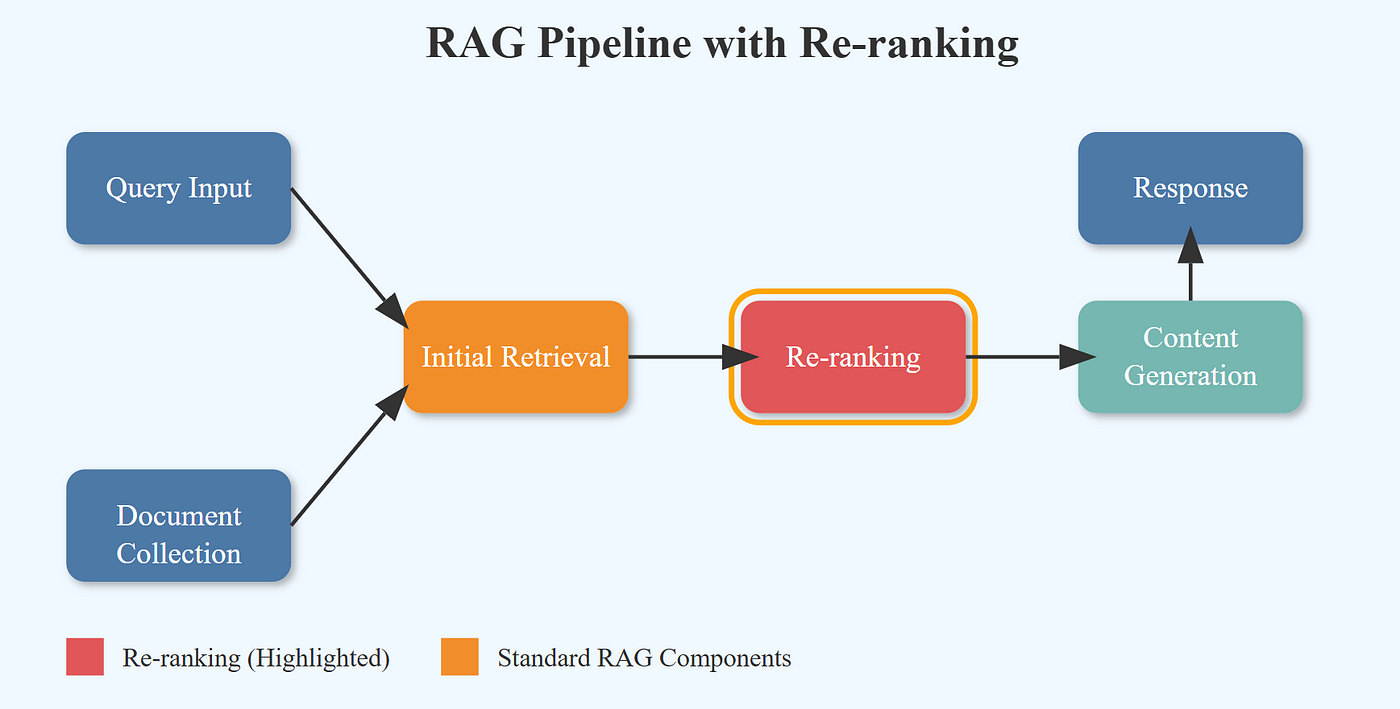
Description automatically generated

 **Concept of Reranking**: Reranking involves ordering a list of retrieved documents, answers, or candidates based on their relevance to a query, using a model that assigns scores to each item.

 **Use Case**: It ensures the most relevant information is prioritized, improving accuracy in search engines, recommendation systems, or conversational AI.

 **Stage in RAG (Retrieval-Augmented Generation)**: Reranking is typically used after the retrieval phase to optimize the quality of the retrieved context before passing it to the language model for response generation.





A diagram of a multi-vector retriever

Description automatically generated