**AI-Powered Document Search and Summarization System**

An AI-driven document search and summarization tool that allows users to upload PDFs, Word files, or text documents, search for relevant information, and generate concise summaries.

**Key Components of the project:**

1. Document ingestion: Allow users to upload documents (PDFs, text files).

2. Use BERT-based embeddings to convert text into searchable vectors.

3. Store embeddings in a vector database (FAISS/Pinecone) for efficient retrieval.

4. Develop a search interface that allows users to input queries and retrieve the most relevant document sections.

5. Use LLMs (on your local computer or hosted on the Hugging-Face website) to answer questions regarding the document, and to summarize the document.

6. Use LLMs (local or hosted) to answer general questions (not related to the document).

**How to use it:**

1. First you have to **load** a document (option **L**).
2. Define a **question** you would like to ask regarding the document (option **A**).

Loading the document also creates:

**embeddings** (with “sentence-transformers/all-MiniLM-L6-v2”)

**Indexing** (with L2 distance-based FAISS index)

1. Now you can search the document, either by:

**FAISS** L2 distance (option **1**)

Or:

**Cosine similarity** (option **2**)

1. You can also search the document, not with embeddings, but directly using **local LLMs**:

"microsoft/**Phi-4**-mini-instruct" (option **3**)

Or:

**gpt2** (option **4**)

The **Microsoft** LLM would take a lot of **CPU power, and about a minute of calculation.**

The **gpt2** LLM is much quicker, but not as good as the Microsoft LLM.

1. You can **summarize** the document using a local LLM (Microsoft) or a hosted LLM. (options **S** and **M**)
2. You can use LLMs (local or hosted) to **answer general questions** (not related to the document). (options **G** and **H**)

When using a **hosted LLM**, you can always choose what LLM model you would like to use.

**Here are some options:**

TinyLlama/TinyLlama-1B

tiiuae/falcon-7b-instruct

cutycat2000/MeowGPT-2

EleutherAI/gpt-j-6B

facebook/bart-large-cnn

microsoft/Phi-4-mini-instruct

Notice that some hosted LLMs are not free unless you pay money to use them, some are blocked, some are sometimes very busy (because many users are currently using them), and so on.