RAG (Retrieval Augmented Generation)

An innovative approach to searching your PDF application using the power of Langchain, ChromaDB, and Open Source LLM, all running on your CPU.

Langchain is a powerful library designed for generative AI tasks, providing a range of capabilities that enhance language generation and understanding.

ChromaDB, on the other hand, acts as a vector store and database, enabling us to store and retrieve vectors efficiently. By integrating ChromaDB into our search tool, we can create a robust and scalable solution for managing the vector representations of PDF documents, allowing for faster and more accurate searches

Finally, we utilise an Open Source LLM (Language Model) to enable question-answering capabilities within our search tool. With the power of LLM, we can process user queries and extract relevant information from PDF documents, providing precise and context-aware answers to the user's questions.

We'll start with Lang chain then we'll move to sentence Transformers for embedding and then we'll utilize chroma DB you know as a vector store and in end we'll use an open source language model called lamini LM.

<https://huggingface.co/MBZUAI/LaMini-T5-738M>

<https://www.trychroma.com/>

<https://python.langchain.com/v0.1/docs/get_started/introduction.html>

<https://github.com/AIAnytime>

They are not that big or that large when we call large language models but they are good enough to you know give you decent accuracy for whatever task you are trying to do when it comes to Natural languages right for example semantic search summarization etc.

This model is one of our LaMini-LM model series in paper "LaMini-LM: A Diverse Herd of Distilled Models from Large-Scale Instructions".

This model is a fine-tuned version of t5-large on LaMini-instruction dataset that contains 2.58M samples for instruction fine-tuning.

This model lamini t5 738m means 738 million parameters. It is an extremely underrated language model.

MBZUAI - The Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) in Abu Dhabi, is a graduate-level, research-based academic institution that offers specialized degree programs for local and international students in the field of Artificial Intelligence.

I think the last six to ten months after the origin of chat GPT, Langchain is the most used library because it does all the heavy lifting for us.