DocTalk - System Architecture Overview

# Introduction

DocTalk is a retrieval-based Question Answering (QA) system developed using FastAPI, FAISS, SentenceTransformers, LangChain, and Docker. It allows users to upload documents, generate embeddings, store them in a vector database, and retrieve answers based on queries.

# Architecture and Components

* Frontend (HTML/CSS/JS): For file uploads and question submissions.
* Backend API (FastAPI): Handles uploads, serves files, and provides search results.
* Document Processing: PDF/TXT extraction using pdfplumber and Python.
* Text Chunking using RecursiveCharacterTextSplitter.
* Embedding generation using SentenceTransformers ('all-MiniLM-L6-v2').
* Vector storage with FAISS for efficient retrieval.
* Answer retrieval and formatting via LangChain.
* Docker for containerization and deployment.
* Pytest-based unit tests for API reliability.

# Conclusion

The DocTalk project integrates advanced NLP techniques and modern web development practices to deliver robust, efficient, and accurate question-answering capabilities directly from uploaded documents.