ClarifAI: A Desktop Assistant for Intelligent PDF Question Answering

1. Introduction:

With the increasing volume of textual data stored in PDFs across the bank, retrieving relevant information quickly becomes a challenge. This project proposes a desktop-based AI assistant that enables users to ask natural language questions and receive answers extracted directly from uploaded PDF documents.

2. Objectives:

- Allow employees to log in securely.
- Enable PDF document upload and manage file processing.
- Automatically extract and embed document content.
- Allow users to ask natural language questions.
- Use semantic search and an LLM to generate answers.
- Trace the source file where the answer was found.
- Provide a user-friendly local GUI (Tkinter-based, no browser required).

3. Scope:

This system focuses on PDF document understanding, embedding-based semantic search, and LLM-based question answering. It does not cover real-time OCR, handwritten content, or documents in languages other than English (in this version).

4. System Architecture:

- **Login System**: Uses employee database to verify credentials.
- **PDF Processor**: Extracts text from uploaded PDFs.
- **Chunker & Embedder**: Splits text and embeds it using a pretrained SentenceTransformer.
- **Semantic Search (FAISS)**: Finds relevant text based on user question.
- **LLM Answer Generator**: Uses a local LLM (e.g., Mistral via Ollama) to generate the final answer.

- **GUI Interface**: Tkinter-based desktop interface for upload, QA, and interaction.

5. Languages used:

- GUI: Tkinter (Python)
- Search index: FAISS
- PDF Parsing: PyMuPDF
- LLM: Mistral via Ollama
- Database: SQL
- File handling: shutil, os
- NLP Embeddings: sentence-transformers

6. Expected results:

Employees will be able to: login using their credentials, upload any number of PDF documents, view uploaded files, ask questions and receive LLM-generated answers and see which file the answer was found in.

7. Conclusion:

ClarifAI facilitates retrieving knowledge from PDFs using natural language, improving productivity and reducing time spent manually searching for information.