Chat with Your PDF

DEEPSEEK 7B POWERED INTELLIGENT PDF Q&A SYSTEM
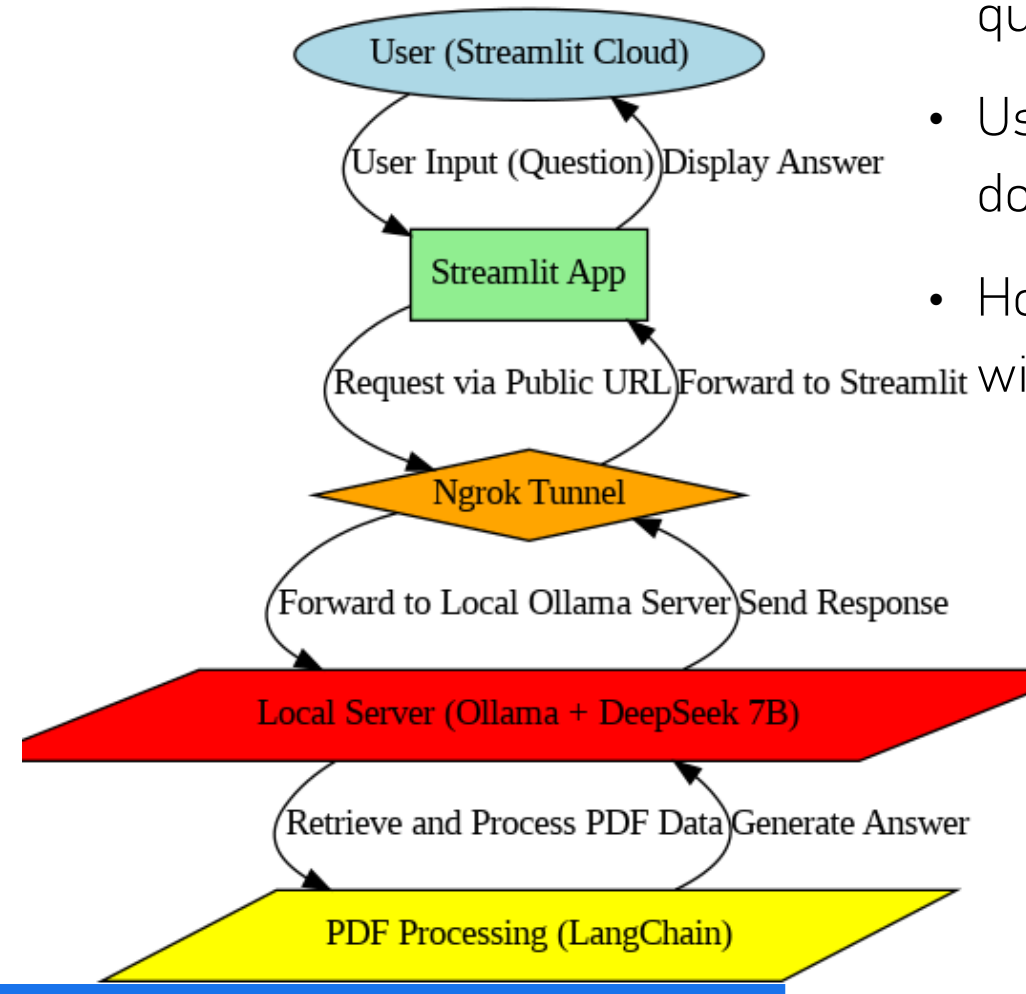
SYED ABDALLAH ALBEEZ
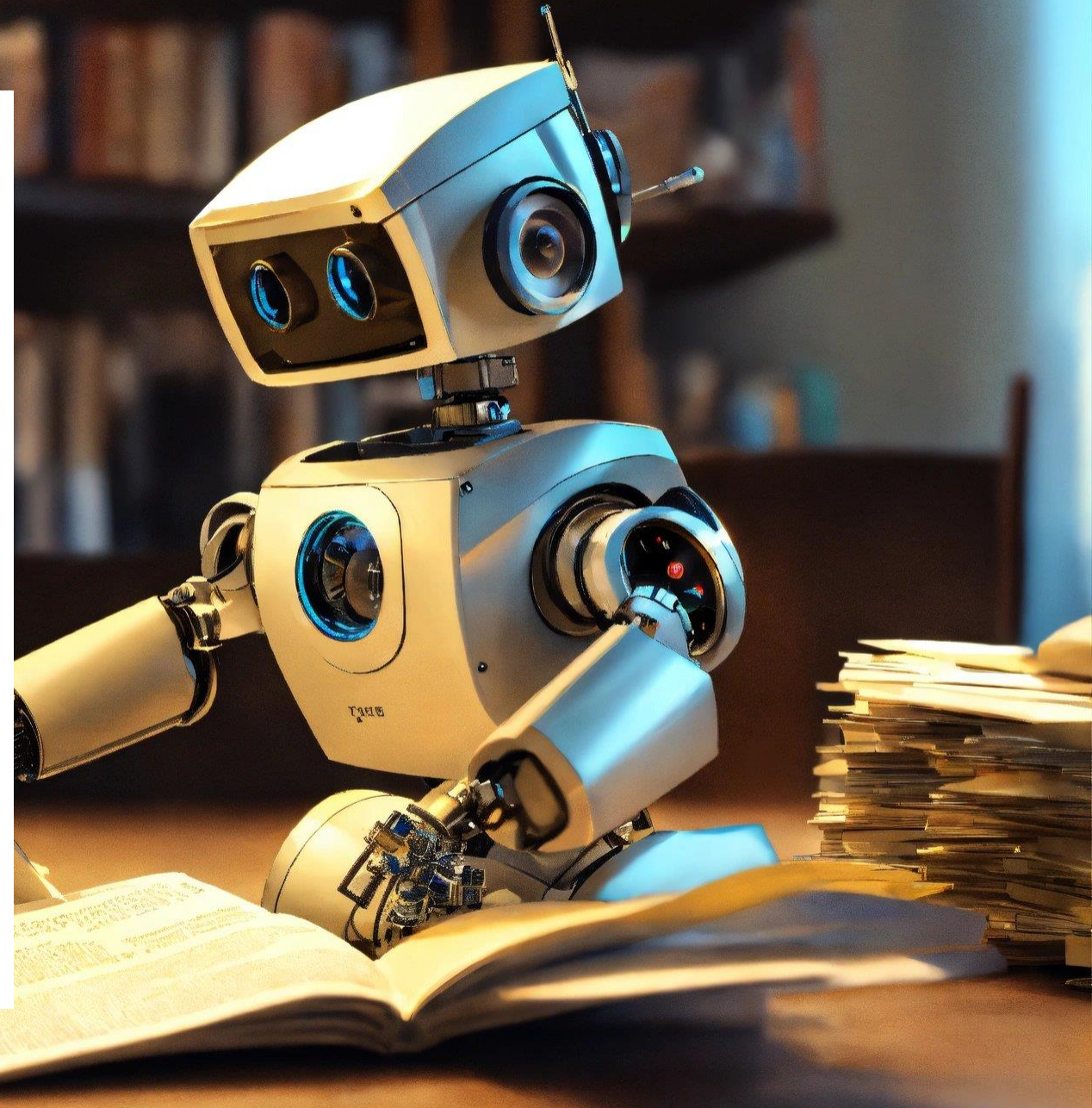
SYED7862001@GMAIL.COM

MST02-0082

Introduction

- The project enables users to chat with PDFs by asking questions and receiving AI-generated responses.

- Uses Ollama's DeepSeek 7B for NLP and LangChain for document processing.

- Hosted on Streamlit Cloud with Ngrok tunneling to connect with a local Ollama server

Problem Statement

- PDFs contain valuable information but are not easily searchable.

- Manually searching for information in long documents is time-consuming.

- Need for an AI-powered solution that enables efficient document interaction.

Technology stack

Ollama: Runs DeepSeek 7B for generating responses.

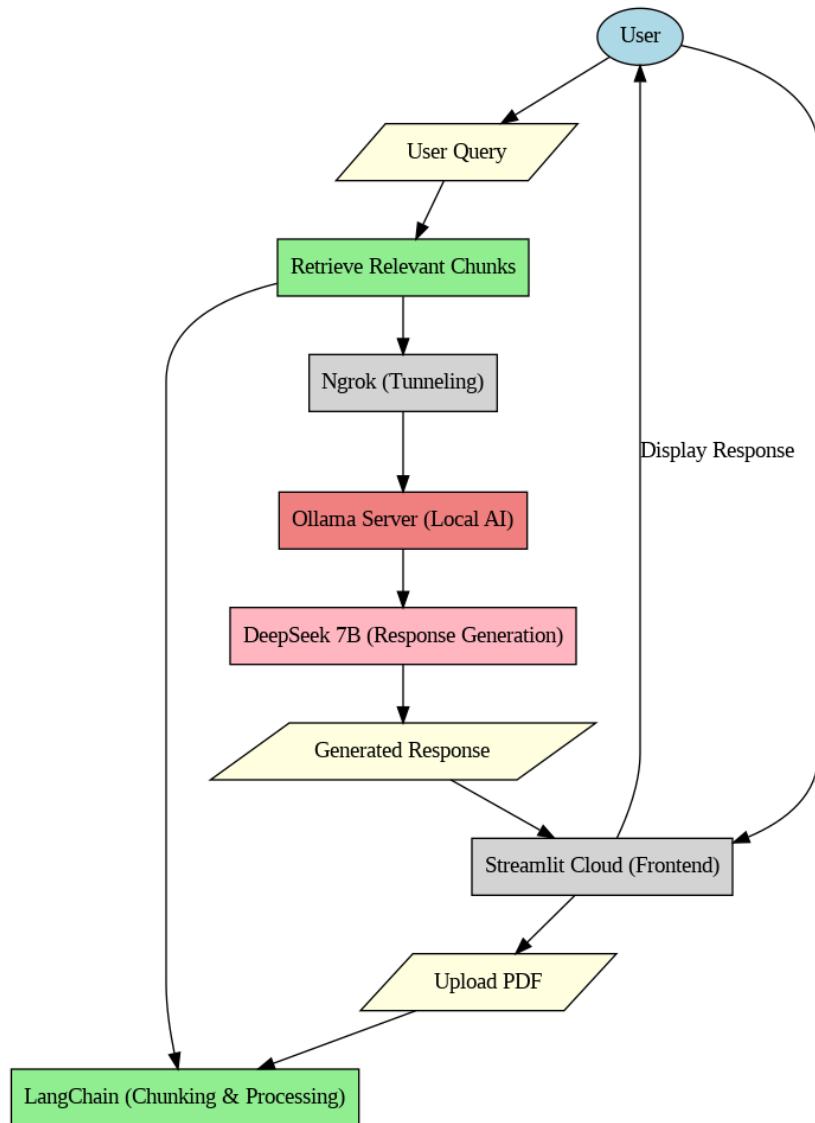**LangChain:** Manages document retrieval and processing.

Streamlit: Provides an interactive cloud-based UI.

Ngrok: Exposes the local Ollama server to the cloud.

PyTorch & CUDA: Enables GPU acceleration for AI computations.

System Architecture

- User uploads a PDF via Streamlit Cloud.

- PDF is processed using LangChain and split into chunks.

- User asks a question; relevant PDF chunks are retrieved.

- The query is sent to the local Ollama server via Ngrok.

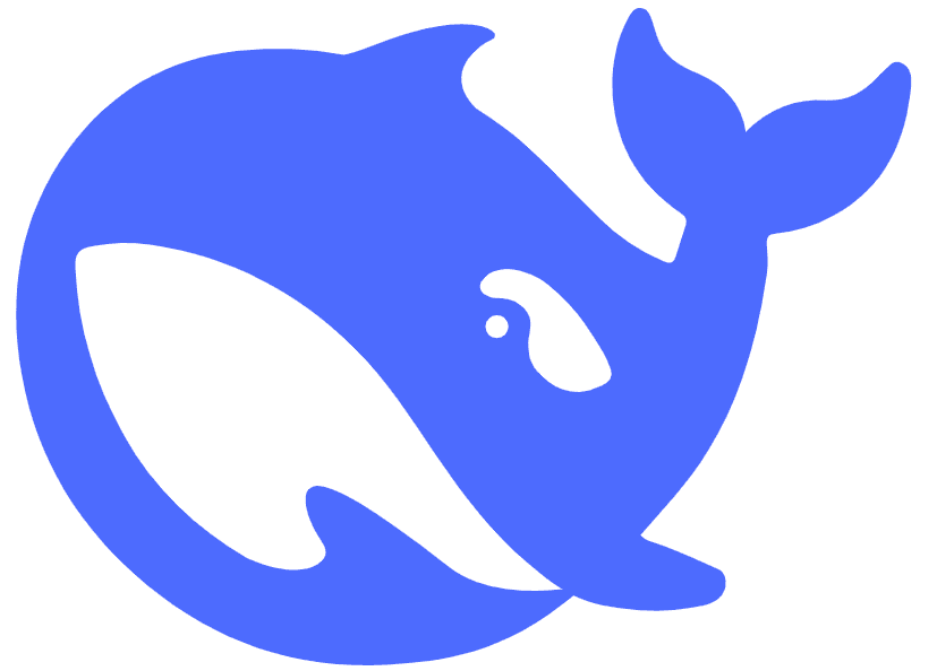- DeepSeek 7B generates a response, which is sent back to Streamlit.

Ngrok Tunnel Explanation

- Streamlit Cloud cannot directly communicate with local servers.

- Ngrok provides a secure public URL that tunnels requests to the local machine.

- Ensures seamless communication between cloud and local resources.

DeepSeek 7b

- **What is DeepSeek 7B?** – A powerful open-source large language model (LLM) optimized for natural language understanding and generation.

- **Why DeepSeek 7B?** – It balances performance and efficiency, making it ideal for handling PDF-based queries.

- **How is it used?** – It processes questions and generates relevant answers using context from the document.

- **Runs on Your Local Machine** – The model is hosted locally and accessed via **Ngrok** to serve the

Implementation Details

- **PDF Handling**: Extracts text using PDFPlumber.

- **Text Chunking**: Splits content into manageable sections.

- **Vector Store**: Uses embeddings to index and retrieve relevant content.

- **LLM Querying**: Ollama processes the question and returns an AI-generated response.

Results & Benefits

Provides quick and accurate answers from PDF content.

Enhances accessibility of information in large documents.

Enables seamless cloud-to-local AI integration via Ngrok.

Future scope: Add multi-document support and fine-tune models for better accuracy.

Output:

- Streamlit

Conclusion

- Successfully implemented an AI chatbot for PDF interaction.

- Combined LangChain, DeepSeek 7B, and Streamlit for an efficient solution.

- Ngrok enables secure cloud-to-local communication.

- Can be extended with more powerful models and integrations.

THANK YOU!