Microsoft Elevate

CAPSTONE PROJECT

SmartDocs AI: Intelligent Document Understanding and Query Platform

PRESENTED BY

STUDENT NAME: Jonnala Parnika

COLLEGE NAME: Vignan's Foundation for Science, Technology & Research

DEPARTMENT: CSE - AIML

EMAIL ID: parnika2599@gmail.com

Microsoft Elevate

OUTLINE:

- **Problem Statement** (Should not include solution)
- **Proposed System/Solution**
- **System Development Approach** (Technology Used)
- **Algorithm & Deployment**
- **Result (Output Image)**
- **Conclusion**
- **Future Scope**
- **References**

Microsoft Elevate

PROBLEM STATEMENT:

In many organizations such as research institutions, legal firms, educational institutions, and enterprises, vast amounts of critical information are stored in PDF documents like manuals, policies, reports, and research papers. Retrieving specific information from these documents is difficult and time-consuming because users must manually search through lengthy files. Traditional keyword-based search methods fail to understand the actual context of user queries, often producing irrelevant results or missing important information.

Additionally, general AI chat systems may generate responses that are not grounded in the original documents, leading to incorrect or unverified answers. Therefore, there is a need for an intelligent document interaction system that can understand the content of multiple PDF documents, support natural language queries, and provide accurate, context-relevant information directly from source documents with reliable references.

PROPOSED SOLUTION:

- **The proposed system aims to address the challenge of efficient information retrieval from large PDF documents by developing an intelligent AI-based document understanding platform.**
- **The solution leverages Natural Language Processing, vector embeddings, and Retrieval-Augmented Generation (RAG) techniques to enable accurate semantic search and context-aware question answering from document content. The solution consists of the following components:**
- **Document Ingestion:**
  - **Upload and process multiple PDF documents including manuals, reports, and research papers.**
  - **Extract text and metadata from PDFs using advanced parsing techniques.**
- **Data Preprocessing:**
  - **Clean and normalize extracted text to remove noise and inconsistencies.**
  - **Split documents into meaningful text chunks with contextual overlap for semantic continuity.**
- **Embedding & Semantic Storage:**
  - **Convert text chunks into vector embeddings representing semantic meaning.**
  - **Store embeddings and metadata in a vector database for efficient similarity search.**
- **Intelligent Query Processing:**
  - **Convert user questions into semantic embeddings.**
  - **Retrieve the most relevant document sections using similarity search.**
- **AI Response Generation:**
  - **Generate context-aware answers using a Large Language Model based only on retrieved content.**
  - **Provide source citations including document name and page reference.**
  - Result:

PROPOSED SOLUTION:

- **Deployment:**
  - **Develop an interactive web interface for document upload and chat-based querying.**
  - **Deploy the system on a scalable cloud platform for reliable access and performance.**
- **Evaluation:**
  - **The system was evaluated based on retrieval accuracy, answer correctness, citation reliability, and response performance.**
  - **Verified that relevant document sections are retrieved for user queries using semantic similarity.**
  - **Checked that generated answers are accurate and grounded in document content.**
- **Result:**
  - **SmartDocs AI successfully enables intelligent querying over multiple PDF documents.**
  - **The system retrieves relevant content for natural language questions.**
  - **Documents are converted into a searchable semantic knowledge base.**
  - **The interface displays chat responses, document preview, and citation verification.**
- **Outcome:**
  - **Improved document search efficiency and reliable information access.**

Microsoft Elevate

SYSTEM APPROACH:

The "System Approach" section outlines the overall strategy and methodology used to develop and implement the SmartDocs AI intelligent document understanding system. Here's a suggested structure for this section:

- System requirements :
- Support uploading and processing multiple PDF documents with accurate text extraction and preprocessing.
- Convert document text into semantic embeddings and store in a vector database for efficient retrieval.
- Enable natural language query processing to retrieve relevant document content.
- Provide an interactive web interface for document management and querying.

Microsoft Elevate

SYSTEM APPROACH:

The "System Approach" section outlines the overall strategy and methodology used to develop and implement the SmartDocs AI intelligent document understanding system. Here's a suggested structure for this section:

- Library required to build the model :
- PyMuPDF, pdfplumber – PDF text extraction
- OpenAI API – Embedding generation and LLM responses
- ChromaDB – Vector database storage and retrieval
- LangChain – RAG pipeline orchestration
- FastAPI – Backend API development
- React.js – Frontend user interface
- Axios – Frontend–backend communication
- Python & JavaScript – Core development languages

ALGORITHM & DEPLOYMENT:

- In this system, a Retrieval-Augmented Generation (RAG) algorithm is used to enable intelligent querying over PDF documents by combining semantic search with AI-based response generation.
- **Algorithm Selection:**
  - The system uses a Retrieval-Augmented Generation (RAG) approach integrating vector embeddings with a Large Language Model.
  - RAG is selected because it enables context-aware answers grounded in document content and reduces hallucination.
- **Data Input:**
  - Input documents: PDF files containing text content (manuals, reports, research papers).
  - User input: Natural language queries entered through the chat interface.
  - Extracted text chunks and metadata from documents are used for processing.
- **Processing / Training Process:**
  - Extract and preprocess text from uploaded PDFs.
  - Split text into semantic chunks with overlap.
  - Generate vector embeddings for each chunk using embedding models.
  - Store embeddings and metadata in a vector database for similarity retrieval.

ALGORITHM & DEPLOYMENT:

- **Prediction Process :**
  - Convert user query into embedding representation.
  - Perform cosine similarity search to retrieve relevant document chunks.
  - Provide retrieved context to the language model.
  - Generate answer strictly from context with source citations.
- **Deployment :**
  - Backend API developed using FastAPI for document processing and retrieval.
  - Frontend interface built using React for document upload and chat interaction.
  - System deployed on cloud platform enabling scalable and reliable access.

Microsoft Elevate

RESULT:

The SmartDocs AI system demonstrates accurate and effective retrieval of information from multiple PDF documents using semantic search and Retrieval-Augmented Generation. Uploaded documents are transformed into a searchable knowledge base, enabling context-aware answers grounded in document content with precise source citations.

The system interface including document ingestion, chat-based querying, and citation verification confirms improved document search efficiency and reliable information access. The results obtained from the system are shown below:

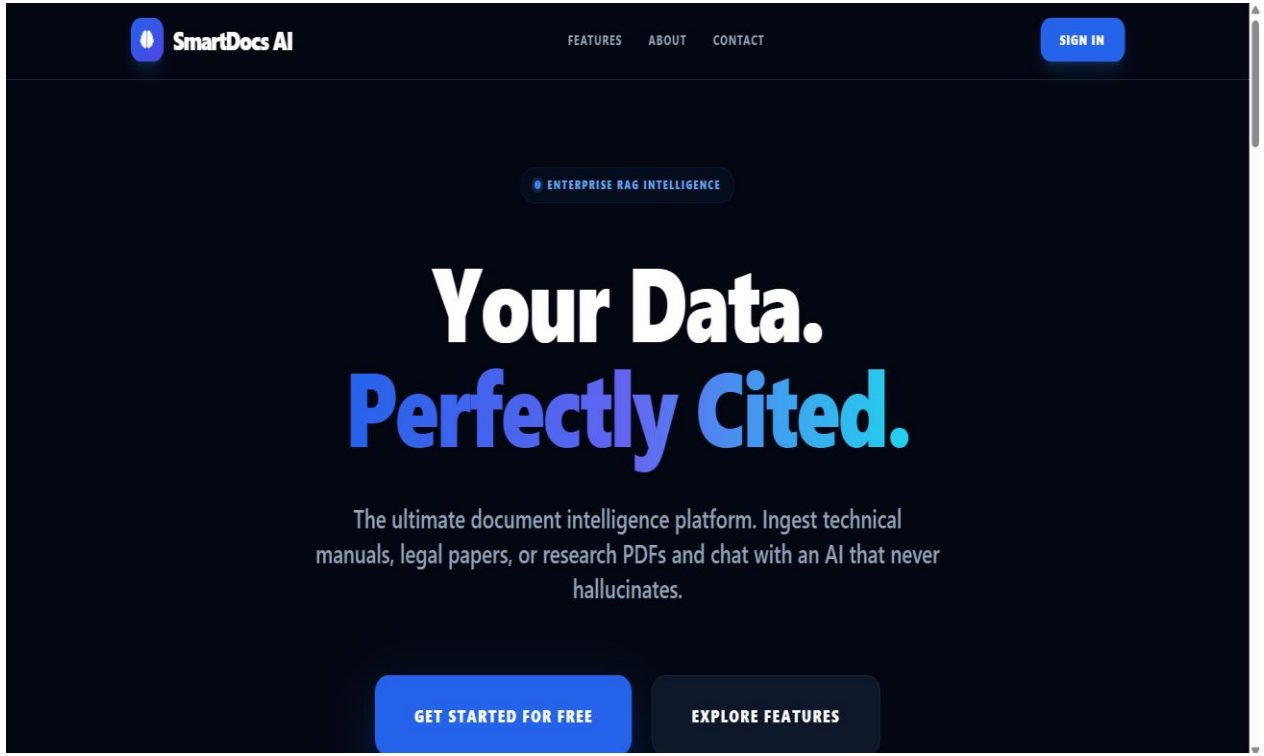
Deployment Link : https://pdf-ai-app-1.onrender.com

Microsoft Elevate

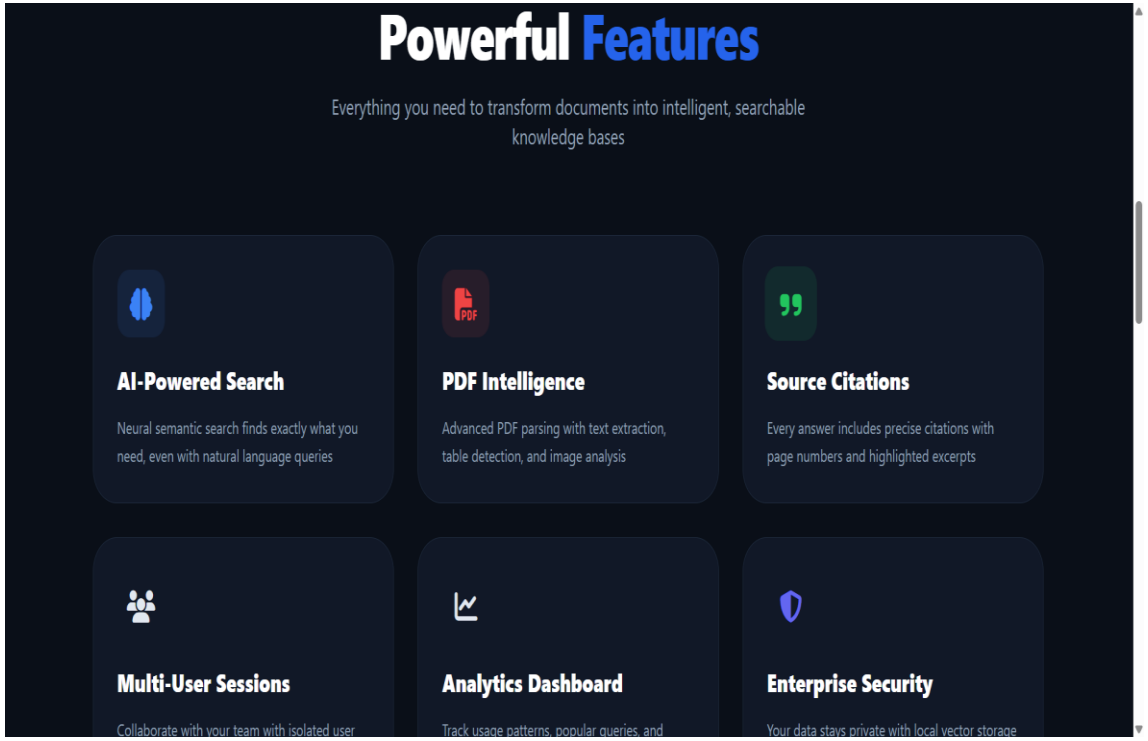
RESULT:

Screenshots:

Landing Page

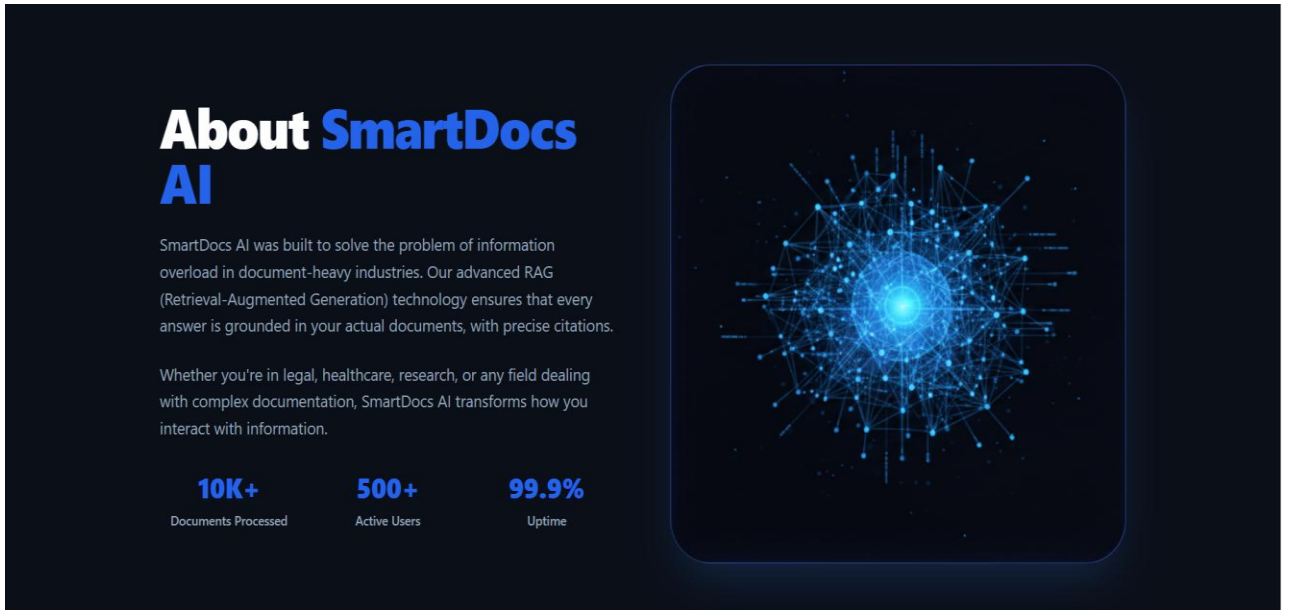


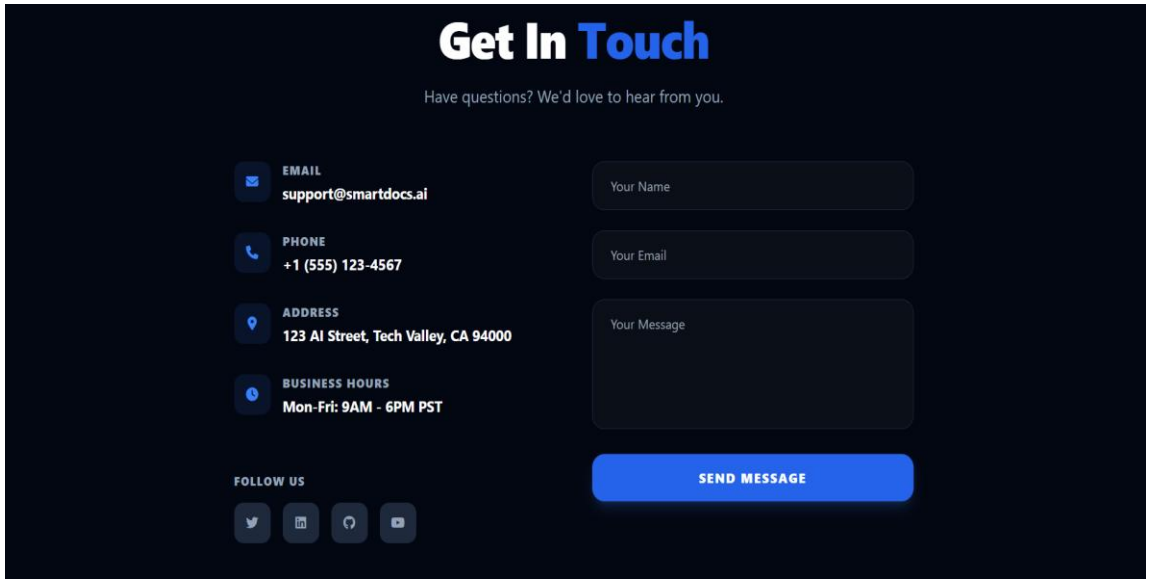
Features



RESULT:

Screenshots:



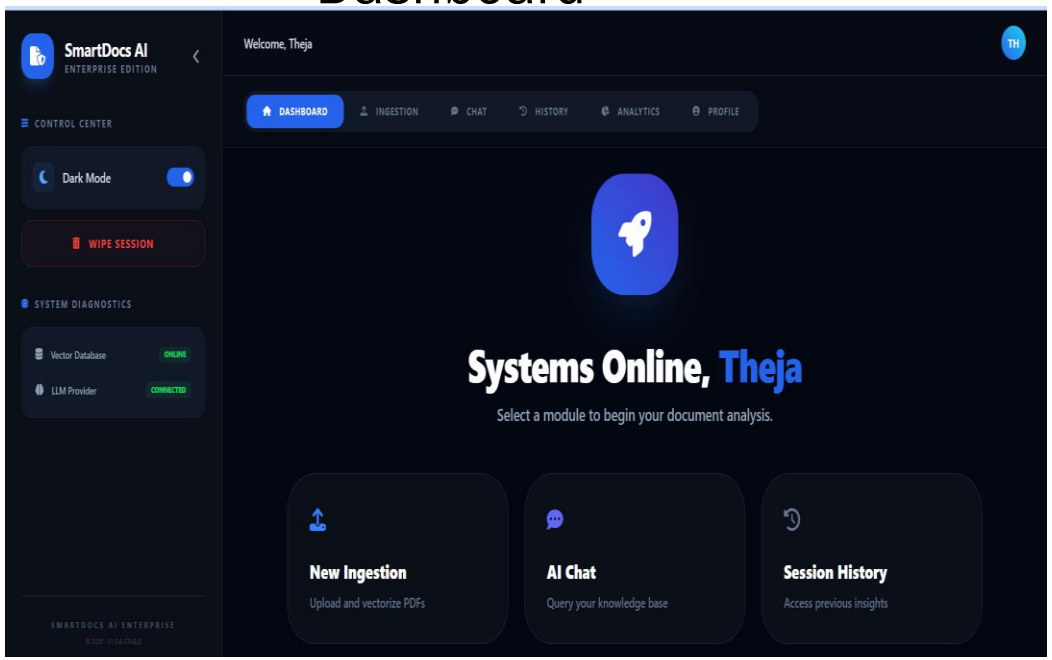


Microsoft Elevate

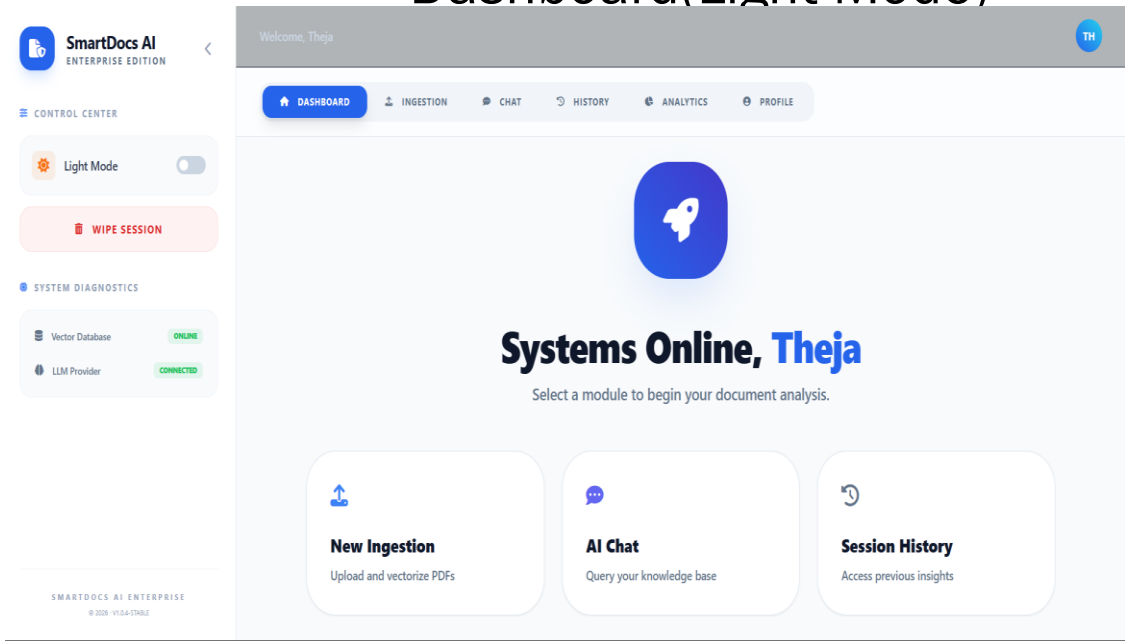
RESULT:

Screenshots:

Dashboard

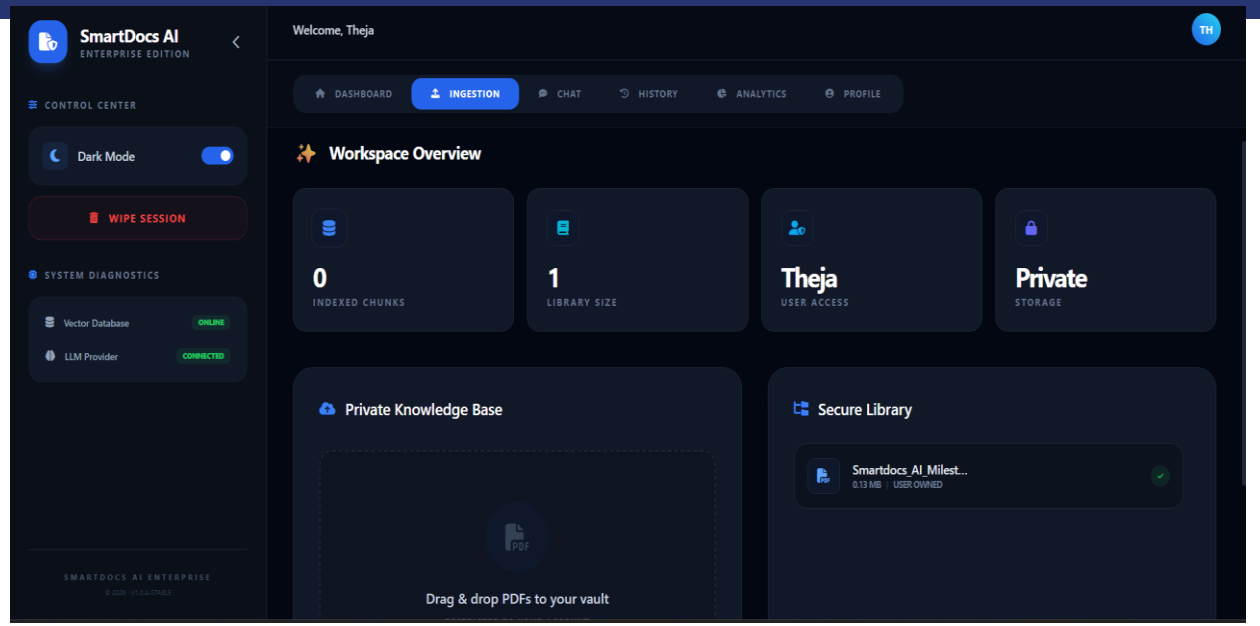


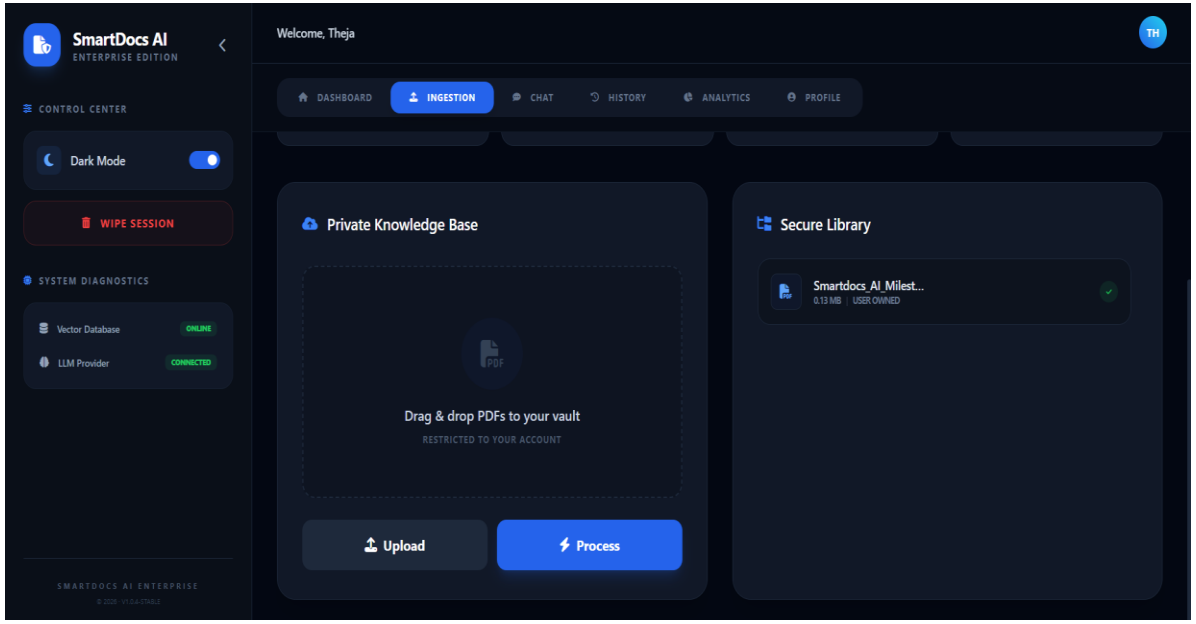
Dashboard(Light Mode)

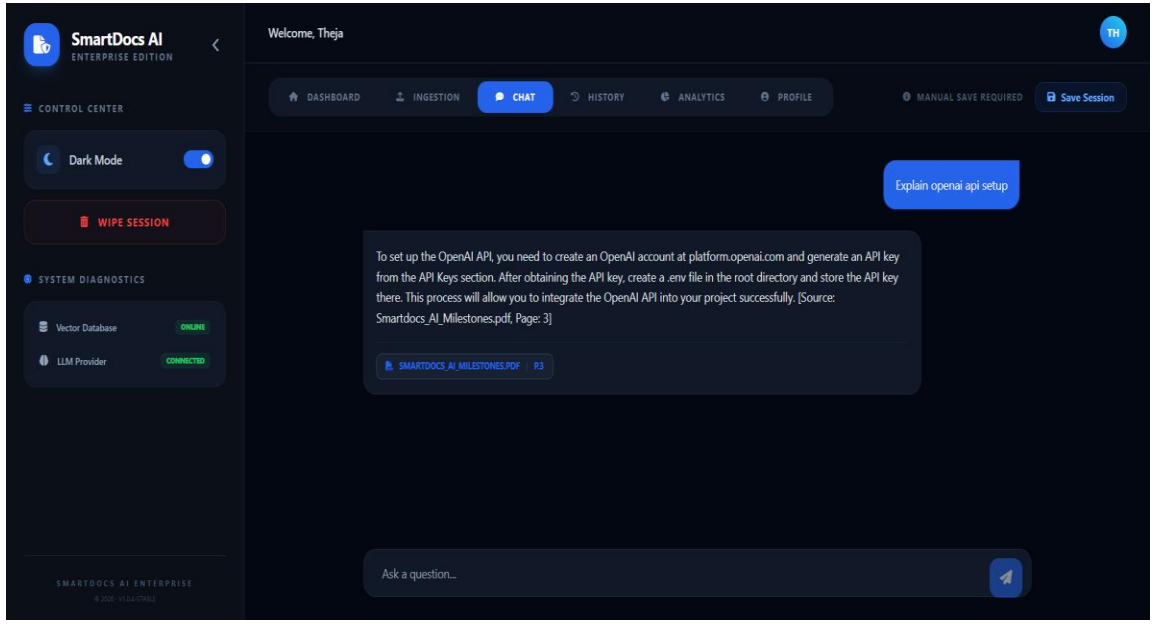


RESULT:

Screenshots:

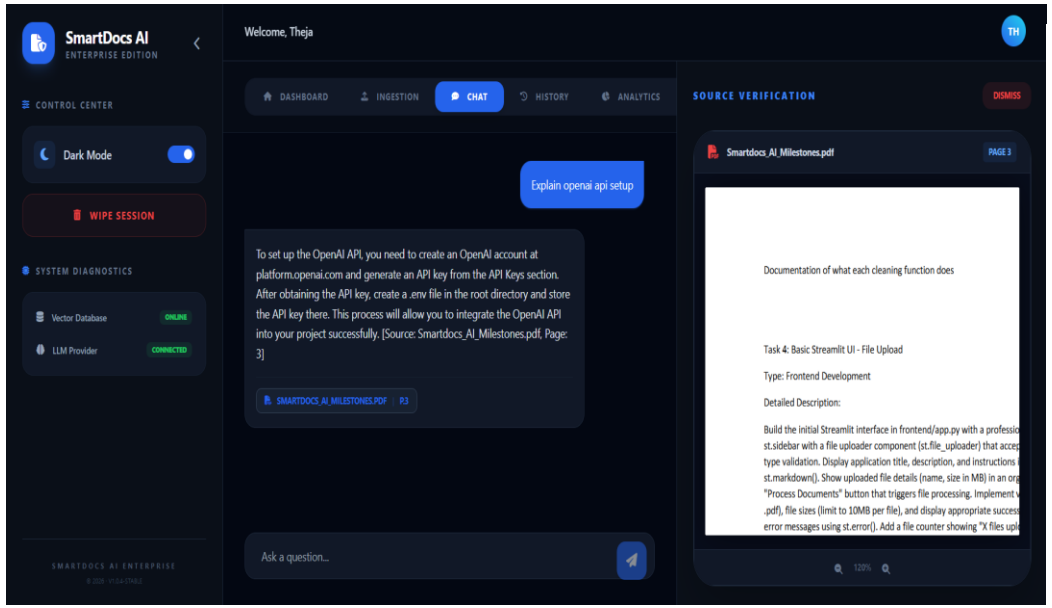






RESULT:

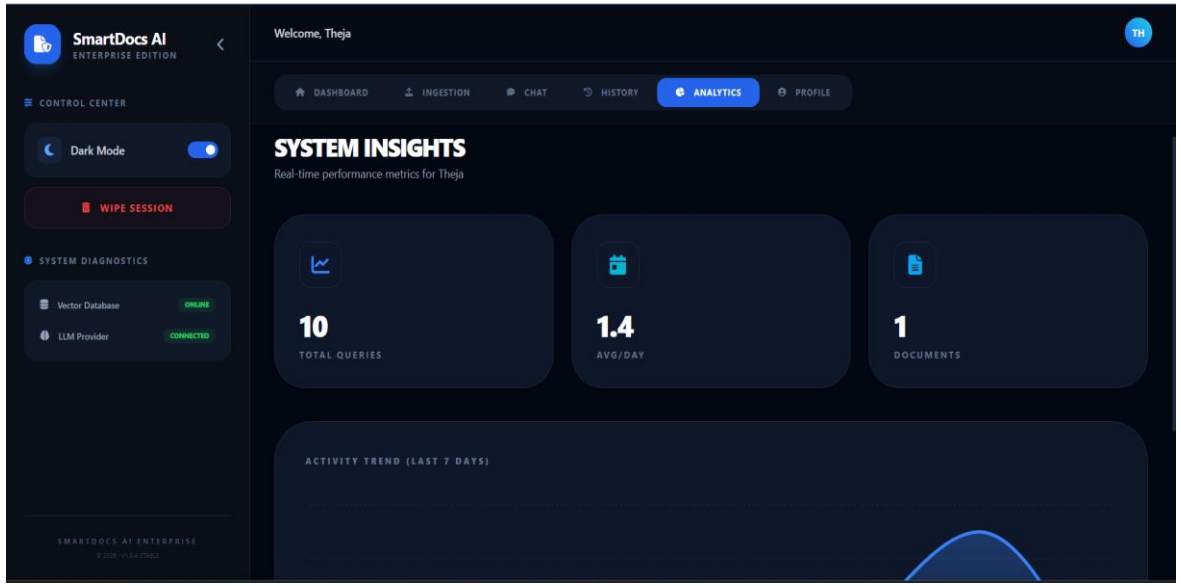
Screenshots:

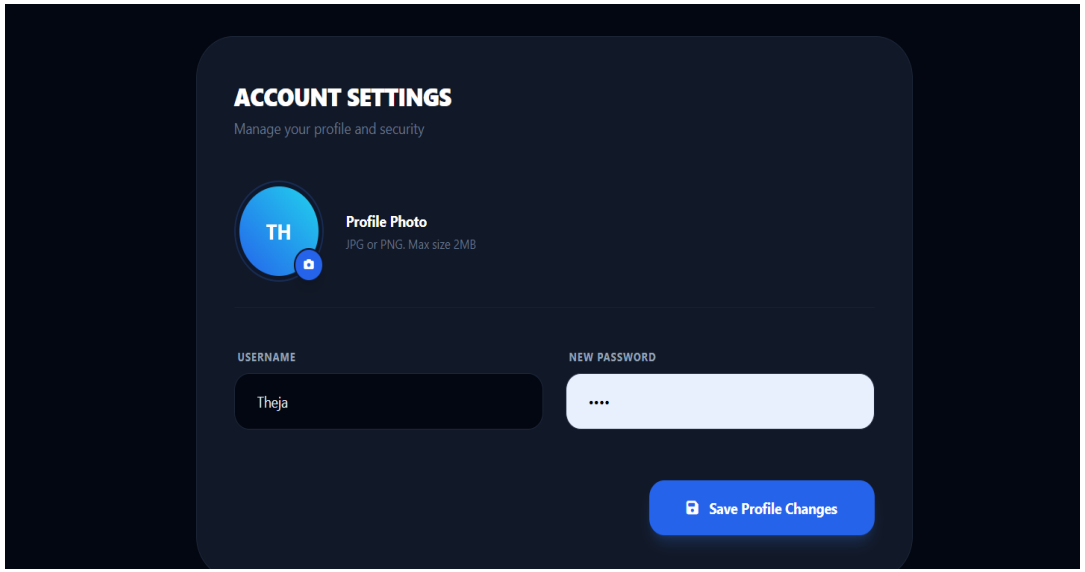


SmartDocs AI — Chat Session

Session ID: sess_20260211_061446
Created: 2/11/2026, 12:44:46 AM

Role	Content
USER	Can you provide a high-level summary of the uploaded documents?
AI	The uploaded documents detail the tasks and deliverables for an internship project involving the development of an AI application for processing PDF documents. The project is divided into two milestones: foundation (core infrastructure, PDF processing, API setup) and vector DB search (embeddings, database, search, QA functionality). The tasks include setting up the development environment, extracting text from PDFs, creating a Streamlit UI for file upload, and integrating the OpenAI API for testing. The deliverables encompass a complete test suite, sample PDFs, setup instructions, user and developer guides, deployment instructions, a demo video, presentation slides, and a deployed application on Streamlit Cloud with a public URL. [Source: Smartdocs_AI_Milestones.pdf, Page: 15]
SOURCE	Smartdocs_AI_Milestones.pdf (Page 15)





Microsoft Elevate

CONCLUSION:

The SmartDocs AI system successfully demonstrates an effective solution for intelligent retrieval and understanding of information from multiple PDF documents. The proposed approach using Retrieval-Augmented Generation enables context-aware question answering with accurate source citations, improving document search efficiency and reliability.

During implementation, challenges such as processing large documents, maintaining context continuity, and ensuring citation accuracy were addressed through chunking strategies, semantic retrieval, and prompt control. The system highlights the importance of intelligent document interaction tools for enhancing information access and supporting knowledge-driven decision-making in document-intensive domains.

Microsoft Elevate

FUTURE SCOPE:

The SmartDocs AI system can be further enhanced by incorporating support for additional document formats such as Word files, images, and scanned PDFs with OCR capabilities. Performance can be improved through advanced embedding models, hybrid search techniques, and optimized retrieval algorithms for handling large-scale document repositories.

The system can be expanded to support enterprise-level deployment with multi-user collaboration, role-based access control, and integration with organizational knowledge bases and cloud storage platforms. Future integration of emerging AI technologies such as multimodal understanding and advanced reasoning models can further enhance intelligent document interaction and automation capabilities.

Microsoft Elevate

REFERENCES:

1. Lewis, P. et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” NeurIPS 2020.
2. OpenAI, “OpenAI API Documentation – Embeddings and GPT Models,” https://platform.openai.com/docs
3. ChromaDB Documentation, “Vector Database for AI Applications,” https://docs.trychroma.com
4. FastAPI Documentation, “Modern Web API Framework,” https://fastapi.tiangolo.com

GitHub Link: Link

Microsoft Elevate

Thank You