Document Summarization using RAG – Report

Objective

To build a summarization system that combines retrieval-based context selection with large language model generation. The system ingests long documents and returns short, accurate summaries.

Components

1. Document Ingestion

- Used PyMuPDF (fitz) to extract text
- Split into chunks of 500 words with 100-word overlap

2. Embedding + Vector DB

- Used all-MiniLM-L6-v2 from sentence-transformers
- Stored chunk embeddings in FAISS index

3. Semantic Retrieval

- Used cosine distance to retrieve top-5 relevant chunks based on the query: *"Summarize this document"*

4. LLM Summarization

- Used facebook/bart-large-cnn via transformers.pipeline
- Limited input text to ~3000 characters for efficiency

5. Output

- Clean summary printed in terminal

Sample Result

Summary Output:

Artificial Intelligence (AI) is transforming industries with its capabilities in natural language processing, computer vision, and decision-making. From healthcare diagnostics to financial forecasting and personalized education, AI applications are expanding rapidly. Ethical concerns remain, including data privacy, algorithmic bias, and the potential displacement of jobs.

Deliverables

- Python Code (main.py)
- PDF Test File (sample.pdf)

- README.md
- requirements.txt
- Sample Output File
- This Report

Conclusion

This project demonstrates a complete, functional pipeline for retrieval-augmented summarization of long documents using modern NLP tools.