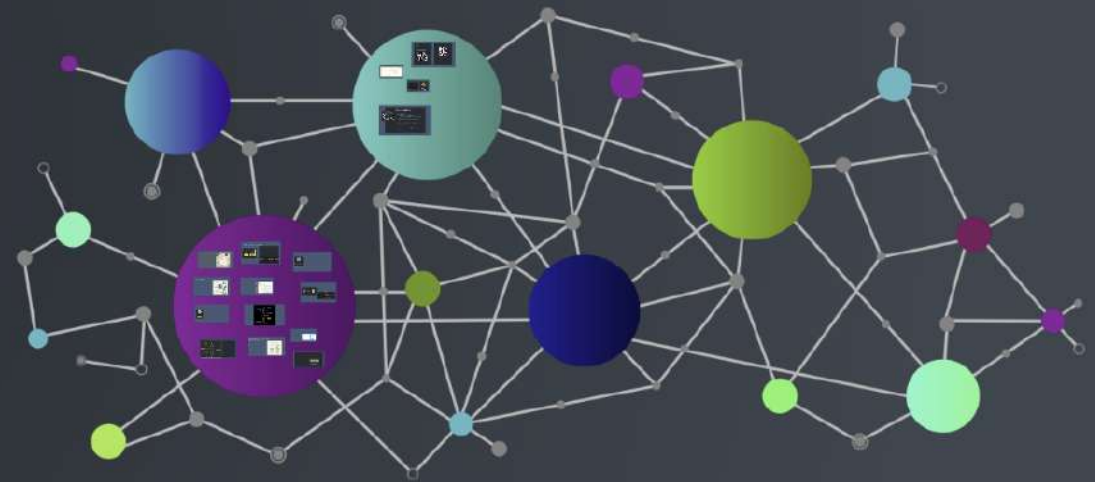


AI-Powered Academic Research Assistant

Intelligent Research Paper Summarization
and Citation Network System

- **Nivedita Nair**
- **Shanmukha Manoj Kakani**
- **Kalyani Chitre**



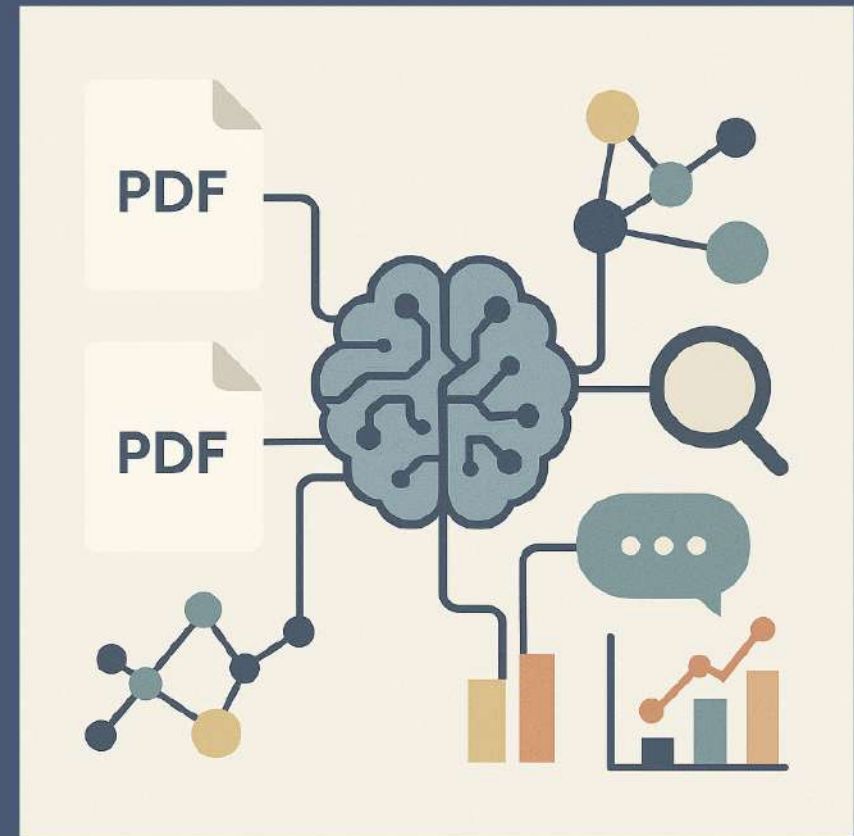
Agenda

- Project Summary
- Goals & Objectives
- Core Feature Set
- System Architecture
- Technology Stack
- Demo
- Future Roadmap



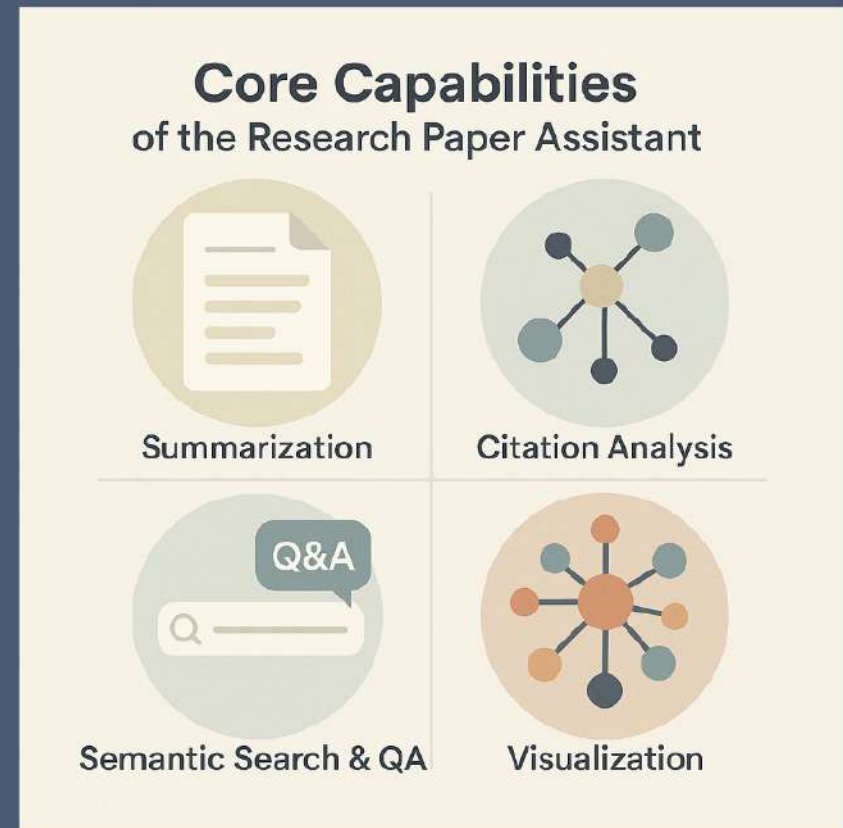
Project Summary

Objective: Develop an AI driven platform to automatically summarize research papers, extract and map citations, enable natural language semantic search & Q&A, and visualize research connections—all from large PDF corpora.

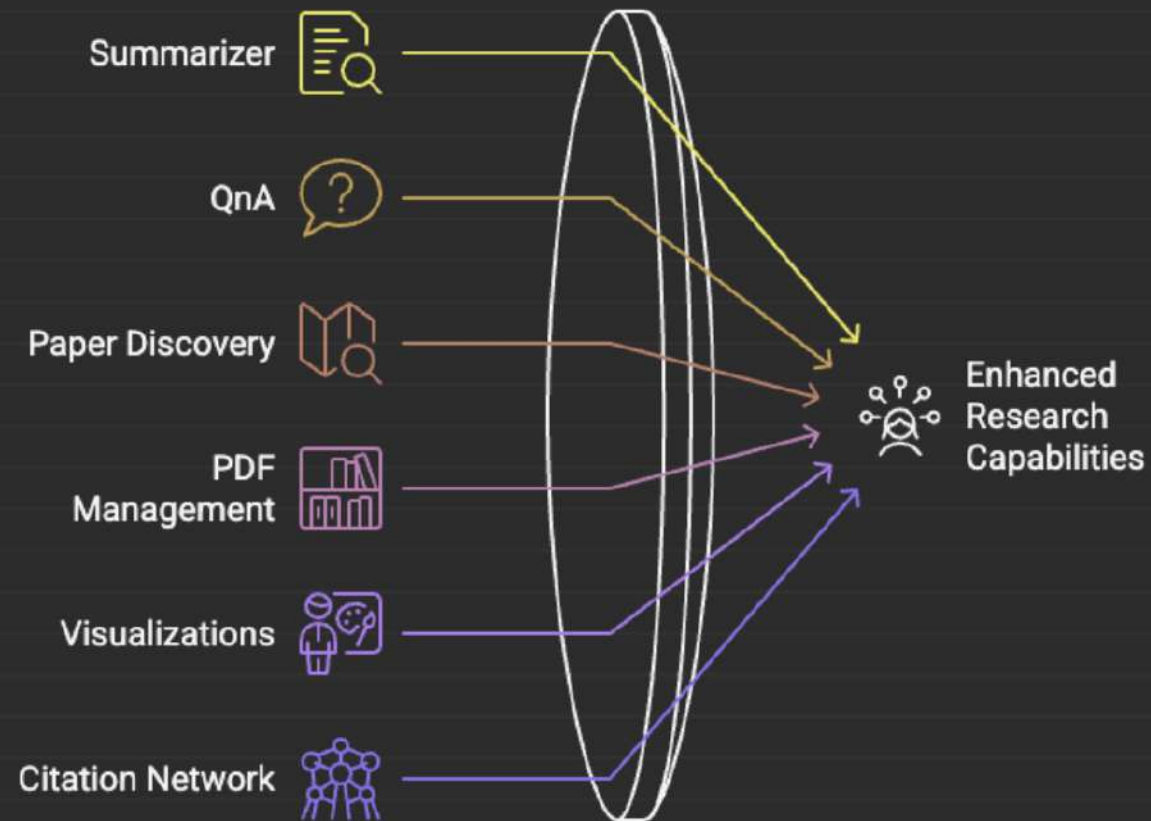


Core Capabilities of the Research Paper Assistant

- Summarization of Research Papers
- Citation & Reference Analysis
- Smart Semantic Search & QA
- Visualization of Research Connections



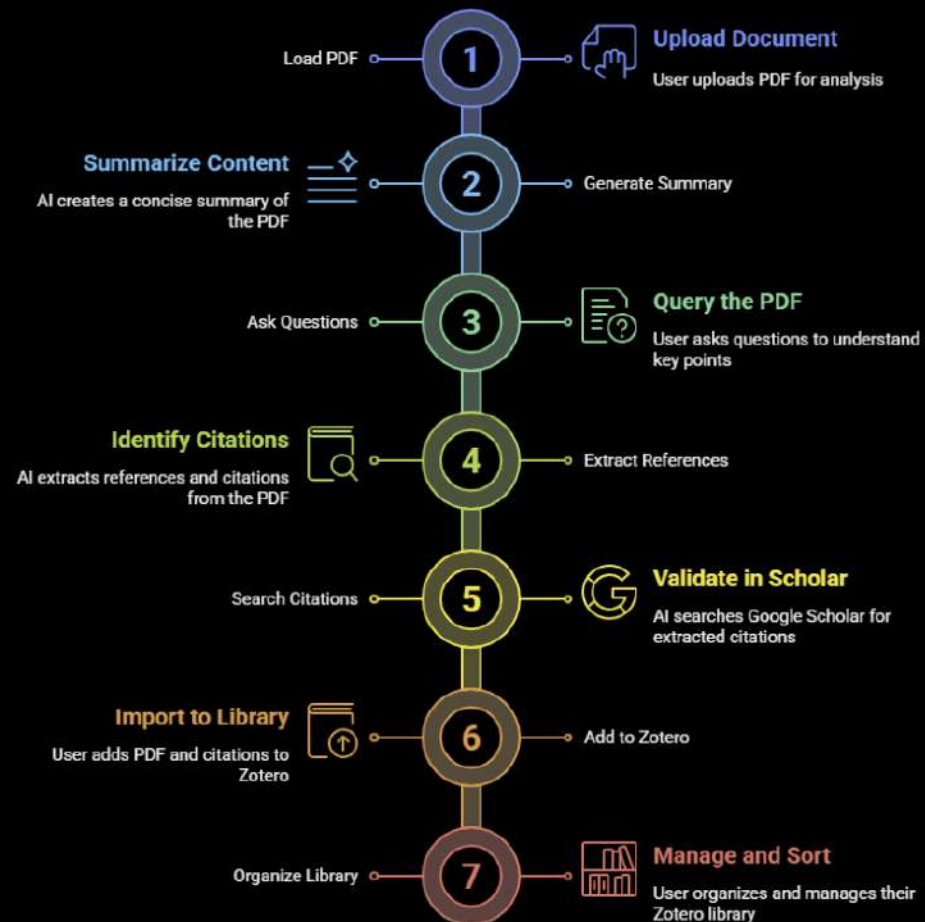
Unified Research Toolkit



Unified Research Toolkit

- QnA - across multiple doc
- Paper Discovery
- Managing PDFs (Zotero)
- Visualizations
- Citation network

PDF Research Workflow Steps



Features – PDF Analysis



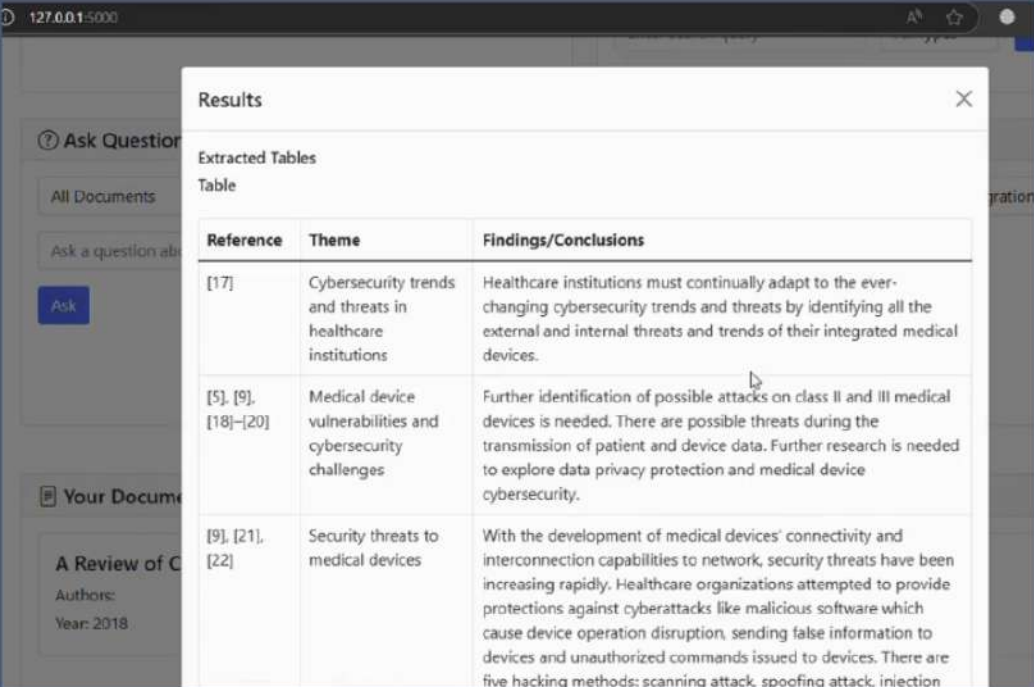
PDF/DOC Analysis

**Efficiently extract and
summarize text**

- PyPDF2/pdfplumber: PDF text extraction
 - pytesseract: OCR for scanned documents
 - Pillow: Image processing
 - transformers: For NLP tasks
 - sentence-transformers: For text embeddings
-
- Split text into semantic chunks (typically 500–1000 characters)
 - Uses RecursiveCharacterTextSplitter from LangChain
 - Generate embeddings using sentence-transformers
 - Store in FAISS vector database for efficient similarity search

Features – Table & Figures extraction

- camelot-py: Table extraction
- pdf2image: Convert PDF pages to images
- opencv-python: Image processing



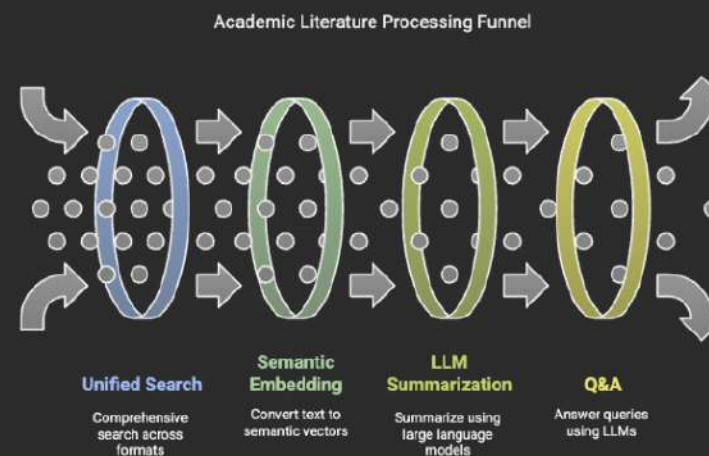
The screenshot shows a web application interface with a 'Results' modal window. The modal window displays 'Extracted Tables' and a table with three columns: 'Reference', 'Theme', and 'Findings/Conclusions'. The table contains three rows of extracted data from a PDF document.

| Reference | Theme | Findings/Conclusions |
|---------------------|---|---|
| [17] | Cybersecurity trends and threats in healthcare institutions | Healthcare institutions must continually adapt to the ever-changing cybersecurity trends and threats by identifying all the external and internal threats and trends of their integrated medical devices. |
| [5], [9], [18]–[20] | Medical device vulnerabilities and cybersecurity challenges | Further identification of possible attacks on class II and III medical devices is needed. There are possible threats during the transmission of patient and device data. Further research is needed to explore data privacy protection and medical device cybersecurity. |
| [9], [21], [22] | Security threats to medical devices | With the development of medical devices' connectivity and interconnection capabilities to network, security threats have been increasing rapidly. Healthcare organizations attempted to provide protections against cyberattacks like malicious software which cause device operation disruption, sending false information to devices and unauthorized commands issued to devices. There are five hacking methods: scanning attack, spoofing attack, injection |

Features – Question Answering

RAG

1. Encode question into embedding
2. Find most relevant document chunks using FAISS
3. Pass context + question to LLM
4. Uses LangChain with various LLM backends
5. Carefully crafted prompts for accurate answers
6. Handles token limits and context truncation



Features – Summarizer System

Hybrid Summarization

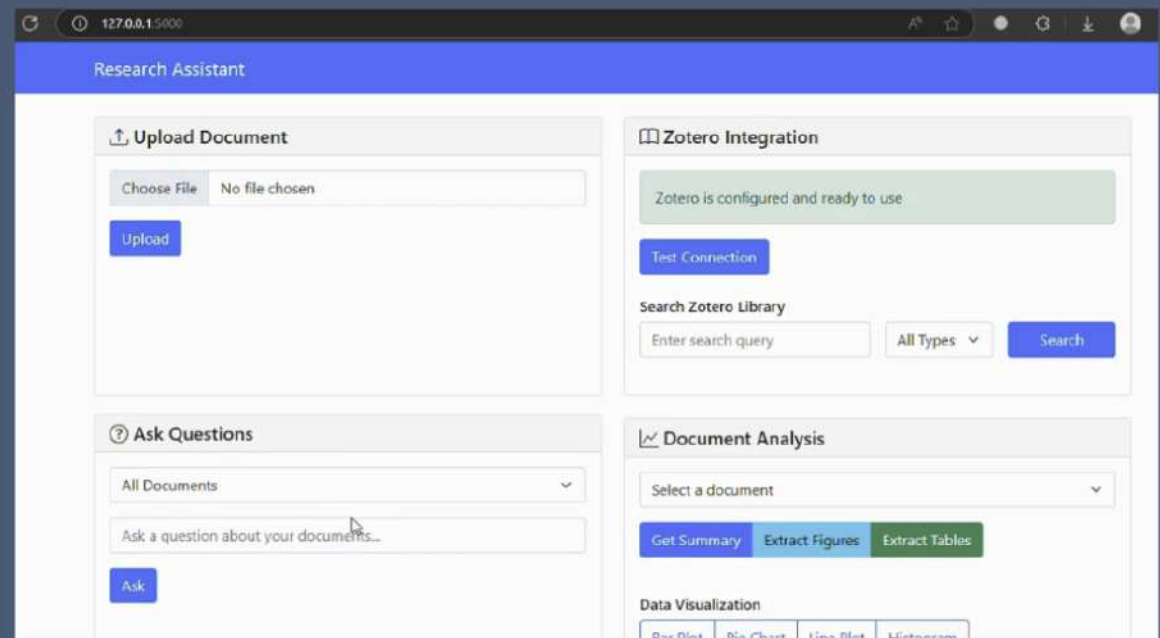
Extractive Summarization:

- Quick Overview Needed
- Preserves technical terms and key findings accurately

Abstractive Summarization:

- summary more concise and readable

Handles multiple related documents
Identifies common themes and unique points

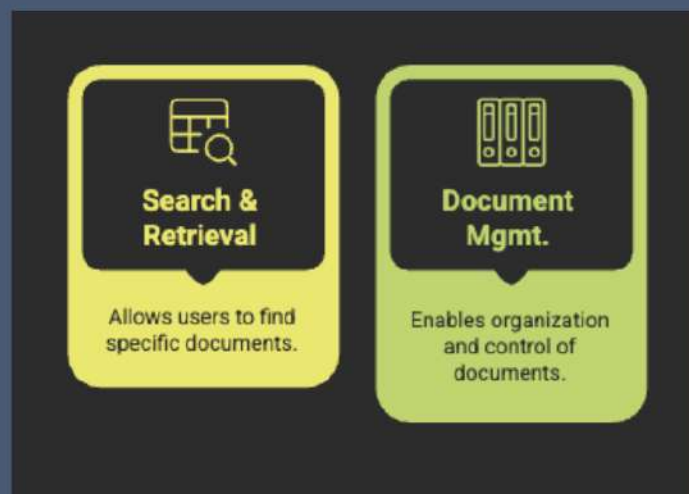


Features – Citation Analysis



- Citation search via Scholarly API, network graph visualization
- Unified Search & Retrieval – One UI across multiple sources
- Filter by year, relevance, citation count
- Paginated & sortable results
- Rule-based with regex patterns
- Citation pattern matching

Features – Document Management



- Zotero API – Open-source reference management software
- Collect, organize, cite, and share research
- Two-way Zotero sync (import/export)

| Domain | Features |
|--------------------|--|
| Search & Retrieval | Multi-source queries, filters, pagination |
| Document Mgmt. | Secure upload (PDF/DOCX/TXT), metadata catalog, tags/folders |

AI & References

Choose the best approach for reference management.



Traditional Methods

Time-consuming and complex

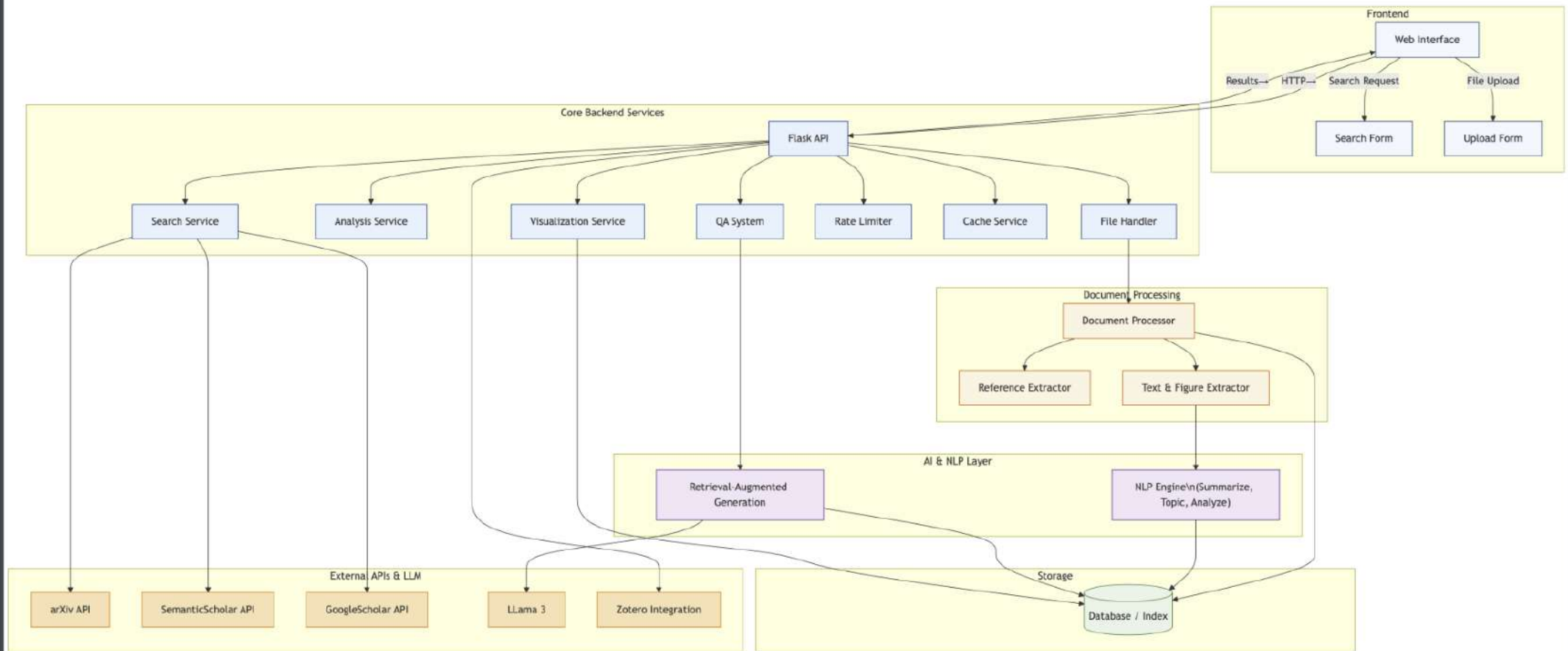


AI-Powered Tools

Efficient and streamlined

| Domain | Features |
|------------------|--|
| AI-Powered Tools | Extractive/abstractive summarization (LangChain/Ollama), contextual Q&A, translation |
| Reference Mgmt. | Zotero sync, BibTeX/RIS import-export, "Generate citation" widget |

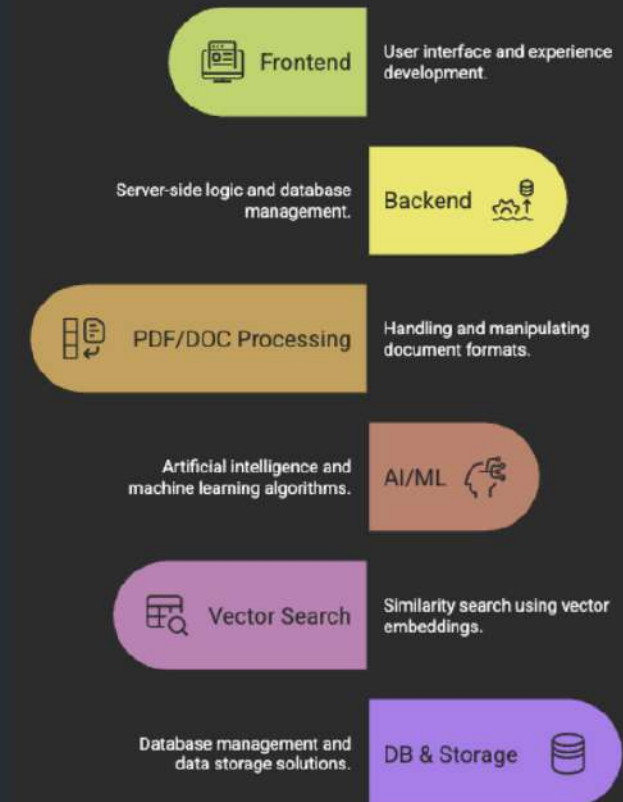
System Architecture



Tech Stack

| Category | Technology |
|----------------------------------|---|
| Vector Search & Vector DB | FAISS – Vector similarity search & DB |
| LLM Integration & Doc Processing | LangChain – LLM orchestration & pipelines |
| PDF Text Extraction | PyPDF2 |
| Word Doc Processing | python-docx |
| Text Embeddings | sentence-transformers |
| NLP Models | transformers |
| Deep Learning Framework | torch |
| Text Processing | NLTK, spaCy |
| Numerical Operations | numpy |
| Data Manipulation | pandas |
| Storage | Local filesystem & Vector database |

Software Layer Technologies



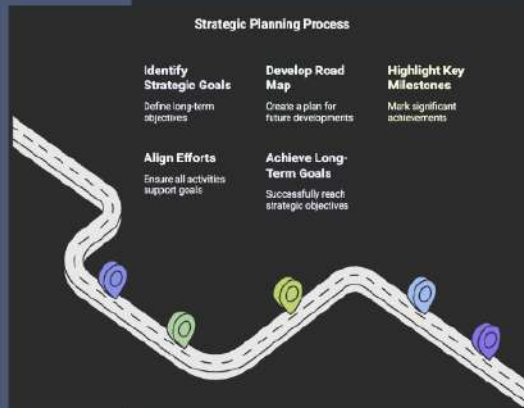
Demo

https://drive.google.com/file/d/1czsePL-WWL__pMbaEXeMawd34ueImqhx/view?usp=sharing



Presentation

Future Roadmap



- Advanced Analytics
 - Topic modeling, trend detection
 - Find connections and inferences across cited paper
 - web scraping of websites under the same topic
- Additional Database Connectors – PubMed, IEEE Xplore, Springer
- Cloud Storage & Serverless – AWS S3, Lambda
- Real-Time Collaboration – WebSockets for live edits