AI Research Assistant: Cross-Document Knowledge Synthesis

Group 10

Rashid Hamza
Jaleel Usman
Raja Wajahat Ali
Maqsood Asim
Zai Zohaib Sultan Yousuf

Course: Generative AI

Abstract

Students and researchers frequently work with large collections of academic documents such as lecture slides and research papers. While traditional PDF search or question-answering tools allow querying individual documents, they often fail to support cross-document understanding and synthesis. This project aims to address this gap by developing an AI-based research assistant that enables users to upload multiple academic PDFs, retrieve relevant information across documents, and generate answers grounded strictly in the provided material. The system targets students and researchers and is implemented as a Jupyter Notebook prototype with an additional demo web application. The main technical challenge explored is document-grounded retrieval and synthesis using classical text representations combined with a local large language model.

Description

Ideal User Journey

The intended users of the system are students and researchers working with multiple academic PDFs. The user begins by opening the system in a Jupyter Notebook or demo web application and uploading one or more PDF documents. The system automatically extracts text from the PDFs and splits it into smaller overlapping chunks. The user then asks a natural-language question related to the uploaded material. Relevant passages are retrieved across all documents, and an answer is generated strictly based on these passages. Retrieved chunks and document sources are shown to the user, enabling transparency and verification of results.

Technical Approach and Reasoning

The system follows a retrieval-augmented generation inspired pipeline, designed to work without reliance on paid APIs.

First, academic PDFs are parsed using a lightweight PDF reader. Each document is stored together with its filename to preserve minimal provenance information. Since academic documents are long and cover multiple topics, a chunking strategy is applied to split text into overlapping segments, enabling fine-grained retrieval.

For semantic retrieval, TF-IDF vectorization is used as an offline embedding method. Although simpler than neural embeddings, TF-IDF provides transparency, reproducibility, and sufficient performance for the project scope. Cosine similarity is used to retrieve the most relevant chunks for a given query.

Retrieved passages from multiple documents are then combined to support cross-document knowledge synthesis. A locally hosted large language model (LLaMA 3 via Ollama) generates answers using only the retrieved context. The prompt explicitly restricts the model from using external knowledge, ensuring document-grounded responses.

The primary user interface is a Jupyter Notebook, which supports step-by-step inspection of the pipeline. Additionally, a Streamlit-based demo application was developed to demonstrate a more user-friendly upload-and-query workflow.

Challenges and Technical Obstacles

Several challenges were encountered during development. PDF text extraction occasionally produced noisy output due to formatting and visual artifacts. Selecting an appropriate chunk size required experimentation to balance contextual completeness and retrieval accuracy. Early versions of the system produced verbose or repetitive answers, which was addressed through prompt refinement. Running a local LLM on Windows introduced encoding issues that required explicit UTF-8 handling. Finally, integrating Streamlit alongside Jupyter required clear separation between the notebook prototype and the demo application.

Evaluation

The system was evaluated using three academic PDFs, including lecture material on Transformers and Generative AI. Evaluation focused on retrieval relevance, answer grounding, and cross-document behavior. A set of predefined test questions was used, and retrieved passages were manually inspected for relevance. Generated answers were assessed for correctness, faithfulness to the source documents, and absence of hallucinated information.

The results showed that the system reliably retrieved relevant passages from the correct documents and produced concise, coherent answers grounded in the PDFs. Cross-document queries successfully combined information from multiple sources. While retrieval depth is limited by the use of TF-IDF, the system met its design goals for transparency and reproducibility.

Reflection

The system increases human agency by supporting understanding and synthesis rather than replacing human judgment. Users remain fully in control of the data and interpretation of results. Further development is desirable, particularly through the integration of neural embeddings, improved PDF preprocessing, structured citations, and more advanced evaluation metrics. The current prototype provides a strong foundation for future extensions.