

AI Study Assistant Using Vector Database and Large Language Models for Intelligent Question Answering from PDF Documents

Marisa. Umadevi

Department of computer science(M.sc)

Co author: V. Satish

Assistant professor in Dr.Lankapalli Bullayya college

Abstract - The rapid advancement of Artificial Intelligence (AI) and Large Language Models (LLMs) has introduced new possibilities in personalized education and intelligent learning systems. This paper proposes an AI Study Assistant using Vector Database and Large Language Models that enables students to interact with educational documents through intelligent question answering.

The proposed system implements a Retrieval-Augmented Generation (RAG) framework, where uploaded PDF documents are processed, converted into semantic embeddings, and stored in a vector database. When a user submits a query, the system retrieves the most relevant document sections using similarity search and provides accurate responses through an integrated Large Language Model.

Unlike traditional search-based learning platforms, the proposed approach understands the context of user queries and generates meaningful answers from the provided study materials. The system helps students reduce manual searching time, improve knowledge understanding, and achieve personalized learning experiences.

Keywords: Artificial Intelligence, Large Language Model, Vector Database, Retrieval-Augmented Generation, Natural Language Processing, Question Answering System, Educational Technology.

I. Introduction

Digital learning resources such as textbooks, research papers, lecture notes, and academic PDFs have become widely available. However, students often face difficulties in extracting important information and finding specific answers from large documents.

Traditional document search methods rely mainly on keyword matching, which fails to understand the semantic meaning of user queries. Recent developments in Large Language Models (LLMs) and Vector Databases provide an effective solution by enabling intelligent document understanding and context-aware responses.

This research presents an AI Study Assistant that allows students to upload PDF documents and interact with them through natural language conversations. The system combines document processing, embedding generation, vector similarity search, and LLM-based response generation to provide accurate answers.

II. Problem Statement

Students spend significant time manually searching through lengthy PDF documents to locate required information. Existing search systems provide keyword-based results but lack contextual understanding.

Major challenges include:

- Difficulty extracting information from large academic documents.
- Lack of personalized interaction with study materials.
- Time-consuming manual searching.
- Inability of traditional systems to understand natural language questions.
- Limited accessibility to intelligent learning support.

Therefore, an AI-powered document-based question answering system is required to provide efficient and accurate academic assistance.

III. Objectives

The objectives of the proposed system are:

- To develop an AI-based study assistant for PDF document interaction.
- To implement vector-based semantic search for document retrieval.
- To integrate Large Language Models for intelligent answer generation.
- To provide context-aware responses

- from uploaded study materials.

To reduce time required for information searching.

- To enhance personalized and interactive learning.

IV. Literature Survey

A. Traditional Document Search Systems

Conventional search engines use keyword matching techniques. These systems cannot understand semantic relationships between words.

B. Chatbot-Based Educational Systems

AI chatbots provide interactive learning support but may generate general responses without considering user-specific documents.

C. Large Language Models

LLMs such as transformer-based models standing and generation capabilities.

D. Retrieval-Augmented Generation

RAG combines information retrieval with LLM generation, improving response accuracy by providing external knowledge sources.

V. Proposed System Architecture

The proposed system follows a Retrieval-Augmented Generation architecture.

User

|

↓

Upload PDF / Ask Question

|

↓

PDF Processing Module

|

↓

Text Chunking & Embedding Generation

|

↓

Vector Database

|

↓

Similarity Search Retrieval

|

↓

Large Language Model

↓

Intelligent Answer General

VI. Methodology

1. PDF Document Processing

The uploaded PDF is converted into machine-readable text.

Processes include:

- Text extraction.
- Document cleaning.
- Sentence segmentation.
- Chunk creation.

2. Text Embedding Generation

The extracted text is converted into numerical vectors using embedding models

These vectors represent semantic meaning and allow similarity comparison.

3. Vector Database Storage

The generated embeddings are stored in a vector database.

Examples:

- ChromaDB
- FAISS
- Pinecone
- Weaviate

The database enables fast similarity-based document retrieval.

4. Query Processing

When a student asks a question:

1. User query is converted into an embedding.
2. Similar document chunks are retrieved.
3. Relevant context is provided to the LLM.
4. AI generates the final response.

VII. System Implementation

Technologies Used

Component	Technology
Programming Language	Python
Framework	Flask / FastAPI
Frontend	HTML, CSS, JavaScript
AI Model	LLM (GPT/Llama/Gemini)
Embedding Model	Sentence Transformers
Vector Database	FAISS/ChromaDB
PDF Processing	PyPDF/PDFPlumber
Database	SQLite/MySQL

VIII. Results and Discussion

The developed AI Study Assistant successfully performs:

- PDF-based question answering.
- Context-aware information retrieval.
- Automated explanation generation.
- Faster access to educational content.

The system provides more relevant responses compared with traditional keyword-based searching because it understands the semantic relationship between queries and documents.

IX. Advantages

- Intelligent document understanding.
- Reduces study time.
- Provides instant answers.
- Supports self-learning.
- Handles large educational documents.
- Improves accessibility.

X. Limitations

- Performance depends on the quality of the LLM.
- Large documents require higher computational resources.

- Incorrect document information may affect responses.
- Vector search accuracy depends on embedding quality.

XI. Future Scope

Future enhancements include:

- Voice-based AI learning assistant.
- Multilingual PDF understanding.
- Automatic note generation.
- AI-generated quizzes from documents.
- Integration with Learning Management Systems.
- Personalized learning analytics.

XII. Conclusion

This research presents an AI Study Assistant based on Vector Database technology and Large Language Models for intelligent question answering from PDF

documents. By implementing Retrieval-Augmented Generation, the system effectively combines document retrieval with AI-based response generation. The proposed approach provides students with a powerful learning companion that improves knowledge accessibility, reduces information searching time, and supports personalized education.

References

1. Vaswani et al., "Attention Is All You Need", 2017.
2. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks", 2020.
3. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2019.
4. Brown et al., "Language Models are Few-Shot Learners", 2020.