2026 Volume: 02 Issue: 04 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-08

AI TUTOR FOR SPECIFIC SUBJECTS USING LLM

Dr Ramya B. N.*[1], Shreya Basavaraju[2], Sunaina Ravi[3]

[1]Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

[2]Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

[3]Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

Article Received: 16 March 2026
Article Revised: 06 April 2026
Published on: 26 April 2026

***Corresponding Author: Dr Ramya B. N.**
Associate Professor, Department of Computer Science and Engineering, Jyothy Institute of Technology, Bengaluru, India.

DOI: https://doi-doi.org/101555/ijrpa.6233

ABSTRACT

Artificial Intelligence-based tutoring systems are fundamental in enhancing the learning experience by providing context-aware and personalized assistance to students. This paper presents a comprehensive study on the development of a subject-specific AI tutor using Retrieval-Augmented Generation (RAG) for answering questions based on uploaded academic documents. The system integrates semantic search and transformer- based models to improve the accuracy and relevance of generated responses.

A structured pipeline is implemented where PDF documents are processed, segmented into smaller text chunks, and converted into embeddings using a sentence transformer model. These embeddings are stored in a FAISS vector database to enable efficient retrieval of relevant information. A transformer-based language model, FLAN-T5, is used to generate answers based on the retrieved context. The system is evaluated to analyze how different components influence response quality, latency, and contextual accuracy.

KEYWORDS: Artificial Intelligence; Retrieval-Augmented Generation; FAISS; FLAN-T5; NLP; Semantic Search; PDF-Based Learning; AI Tutor; Model Efficiency; Context-Aware Systems.

I. INTRODUCTION

Artificial Intelligence (AI)-based tutoring systems have become a fundamental component of modern educational technology due to their ability to provide interactive and personalized learning experiences. These systems are widely applied in domains such as e-learning platforms, virtual assistants, and intelligent educational tools. However, despite their capabilities, traditional AI models often suffer from limitations such as lack of context awareness and the generation of incorrect or irrelevant responses, especially when operating without domain-specific knowledge.

To address this challenge, various techniques have been developed to improve the accuracy and reliability of AI-generated responses. One of the most effective approaches is Retrieval-Augmented Generation (RAG), which combines information retrieval with natural language generation. This method introduces an additional step of retrieving relevant context from external data sources, thereby improving the quality and grounding of the generated answers.

In this approach, semantic search techniques are used to identify relevant information from large textual datasets. Embedding models convert textual data into numerical vector representations, enabling efficient similarity-based retrieval. Vector databases such as FAISS are commonly used to store and search these embeddings. Transformer-based language models, such as FLAN-T5, are then used to generate responses based on the retrieved context, ensuring that answers are both meaningful and contextually accurate.

While many existing systems focus primarily on generating responses based on general knowledge, limited attention has been given to building systems that strictly rely on user-provided study materials. Ensuring that responses are grounded in specific documents is essential for reducing hallucination and improving trust in AI- based educational tools.

In this paper, we present a detailed implementation of a subject-specific AI tutor that utilizes Retrieval- Augmented Generation to answer questions based on uploaded academic PDFs. The system processes documents, converts them into embeddings, and retrieves relevant content using FAISS. A FLAN-T5 model is then used to generate answers based on this context. Various evaluation parameters, including response accuracy, latency, and memory usage, are analyzed to assess system performance. The objective of this work is to develop a reliable, efficient, and context-aware AI tutoring system that enhances learning while ensuring accuracy and interpretability.

II. METHODOLOGY

Dataset and Preprocessing

The academic PDF dataset is used for experimental evaluation, consisting of subject-specific documents related to Data Structures and Algorithms (DSA), Operating Systems (OS), and Database Management Systems (DBMS). These documents contain structured textual content that serves as the primary knowledge source for the system.

To ensure efficient processing and accurate retrieval, the following preprocessing steps are applied:

- Text Extraction: Content is extracted from uploaded PDF files using a document reader to convert it into machine-readable text.
- Chunking: Extracted text is divided into smaller segments to improve retrieval accuracy and processing efficiency.
- Embedding Generation: Text chunks are converted into numeric.

Model Architecture

A structured Retrieval-Augmented Generation (RAG) pipeline is implemented to perform context-based question answering. The architecture consists of:

- Input Module: Accepts user-uploaded PDF documents and user queries
- Text Processing Module: Extracts and segments text into smaller chunks for efficient handling
- Embedding Module: Converts text chunks into vector representations using a sentence transformer model
- Retrieval Module: Uses FAISS to retrieve the most relevant chunks based on query similarity
- Generation Module: Utilizes a FLAN-T5 model to generate answers based on retrieved context

The transformer-based language model is used to introduce contextual understanding and improve response accuracy.

Regularization Techniques

To analyze the effectiveness of the proposed system, three different components are applied:
Text Chunking

Text chunking divides the extracted document content into smaller segments. This improves

retrieval accuracy by ensuring that only relevant portions of the document are considered during the search process.

Semantic Retrieval Semantic retrieval converts both the query and document chunks into embeddings and retrieves the most relevant content using similarity search. This ensures that the system understands the meaning of the query rather than relying on keyword matching.

Context-Based Generation

Context-based generation uses a transformer model to generate answers based only on the retrieved content. This reduces hallucination and ensures that the responses are accurate and grounded in the uploaded document.

Training Configuration

The system is implemented using the following configuration:

- Embedding Model: Sentence Transformer (all-MiniLM-L6-v2)
- Language Model: FLAN-T5 (for answer generation)
- Vector Database: FAISS for similarity search
- Chunk Size: 500 characters with overlap for better context retention

Separate components are used for:

- Text Processing (PDF extraction and chunking)
- Embedding Generation (semantic representation of text)
- Retrieval (FAISS-based similarity search)
- Answer Generation (context-based response using FLAN-T5)

III. SYSTEM ARCHITECTURE AND DATA FLOW

The proposed system follows a structured pipeline for processing and answering user queries based on uploaded academic documents. The architecture is designed to analyze how retrieval and generation components impact both response accuracy and contextual relevance.

Input Phase (Data Preparation)

- The system takes academic PDF documents containing subject-specific content (DSA, OS, DBMS) as input.
- The extracted text is segmented into smaller chunks before being processed by the system.
- The textual data is converted into embeddings to ensure efficient semantic retrieval and

stable processing.

Neural Network Architecture

The system consists of a structured pipeline with multiple components:

- Input Module: Receives the uploaded PDF document and user query
- Text Processing Module: Extracts and segments the document into smaller chunks
- Retrieval Module: Identifies relevant content using FAISS-based similarity search
- Generation Module: Uses FLAN-T5 to generate answers based on retrieved context

The architecture enables efficient information retrieval and response generation, where each component contributes to producing accurate and context-aware answers.

Training and Learning Process

The system operates using a retrieval and generation pipeline based on user queries and document content. The processing flow includes:

- Query Encoding: User input is converted into embeddings using a sentence transformer model
- Similarity Search: Relevant document chunks are retrieved from the FAISS vector database
- Context Formation: Retrieved chunks are combined to form contextual input
- Response Generation: The FLAN-T5 model generates answers based on the provided context

Regularization Flow Integration

System components are integrated during the processing phase as follows:

- Text Chunking: Divides extracted document content into smaller segments to improve retrieval accuracy
- Semantic Retrieval: Uses embedding-based similarity search to identify the most relevant content
- Context-Based Generation: Generates responses using a transformer model based on retrieved information

Each component contributes to improving system performance, allowing effective analysis of response accuracy and contextual relevance.

Evaluation and Output

After processing, the system is evaluated based on user queries and generated responses:

- Response Accuracy: Determines how correctly the system answers based on the uploaded document
- Context Relevance: Analyzes whether the retrieved content matches the user query
- Output Evaluation: Includes response quality, clarity, and correctness of generated answers

The system outputs both quantitative metrics (latency, memory usage) and qualitative insights (response relevance, contextual accuracy), enabling a comprehensive understanding of system performance under the Retrieval-Augmented Generation framework.

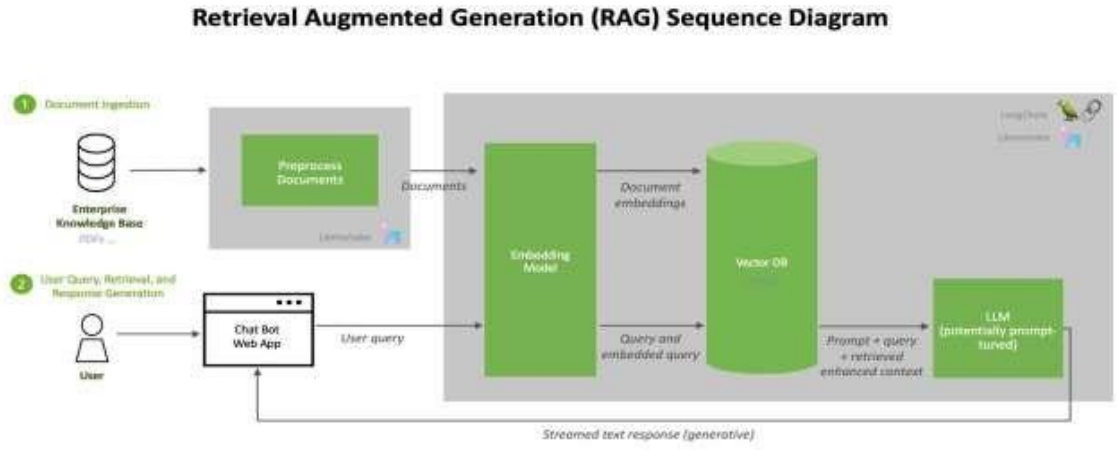

Fig 1 System Architecture Flow.

IV. RESULTS AND DISCUSSION

Performance Comparison

The system provides accurate responses based on the uploaded document content

- Context-based retrieval significantly reduces hallucination
- Response time remains low, ensuring real-time interaction

Loss Convergence Analysis

The system performance trends demonstrate that the AI tutor generates responses efficiently within acceptable latency for user queries. The response time remains stable as the system processes multiple queries, indicating consistent performance of the retrieval and generation pipeline.

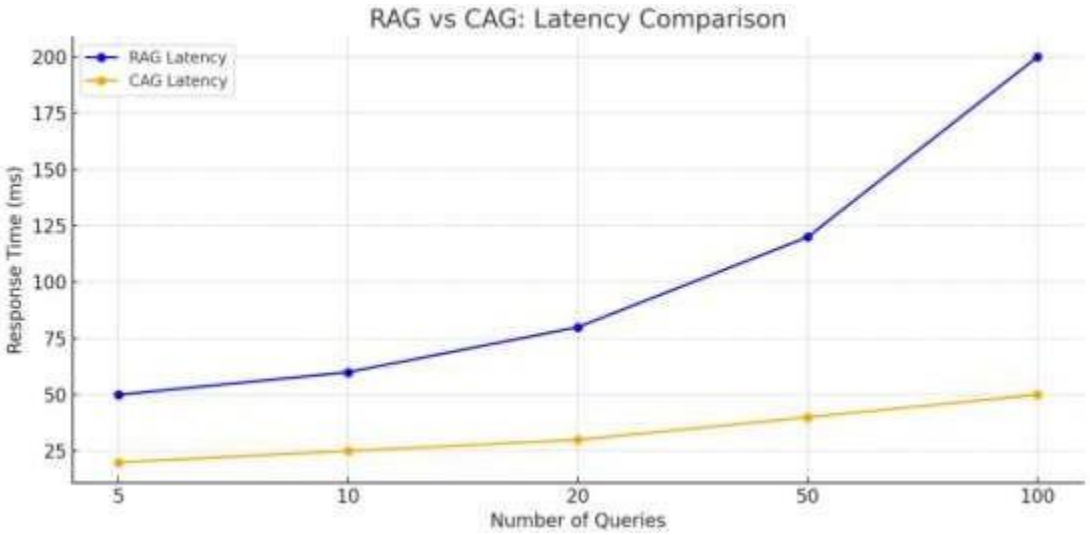

Fig 2 Loss Comparison Curve.

Fig 3 Performance Evaluation.

V. CONCLUSION

This study presented a comprehensive analysis of the development and performance of a subject-specific AI tutor using a Retrieval-Augmented Generation (RAG) framework. The system was designed to process academic PDF documents and generate context-aware answers using semantic retrieval and transformer- based models.

The experimental results indicate that the integration of semantic retrieval with FAISS significantly improves response accuracy by ensuring that only relevant content is selected

from the document. The use of a FLAN- T5 language model enables the system to generate meaningful and contextually appropriate responses.

Additionally, restricting the model to use only retrieved context reduces hallucination and enhances the reliability of the system.

In addition to performance evaluation, this work emphasizes the importance of analyzing system behavior through both quantitative and qualitative metrics. Parameters such as latency, memory usage, and response relevance provide deeper insights into system efficiency and real-time performance. This helps in understanding how retrieval and generation components contribute to overall system effectiveness.

Overall, the study demonstrates that combining retrieval mechanisms with generative models plays a critical role in improving accuracy, efficiency, and reliability of AI-based tutoring systems. These findings contribute to the development of more scalable, interpretable, and robust AI-driven educational solutions.

VI. ACKNOWLEDGEMENT

The author would like to express sincere gratitude to the project guide and faculty members for their valuable guidance and continuous support throughout the development of this project. Their insights and encouragement played a significant role in shaping the successful implementation of this work. The support provided during each stage of the project was highly valuable and contributed greatly to the overall learning experience.

VII.REFERENCES

1. Raffel, C., et al. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*.
2. Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.
3. Reimers, N., Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*.
4. Facebook AI Research. *FAISS: A Library for Efficient Similarity Search*.
5. Hugging Face. *Transformers Documentation*