

RAG-Based AI Teaching Assistant

Vaishnavi Yuvraj Samane¹, Sejal Bapuso Patil², Ankit Kalyan Ambure³, Prof. Dr Kiransing Pardeshi⁴

^{1,2,3}Department of Electronics and Computer Science Engineering Padma bhooshan Vasant Rao Dada Patil Institute of Technology, Maharashtra, India

Abstract - The increase in digital learning platforms has led to extensive educational content in various formats, including video lectures and audio recordings. Students frequently struggle to find specific explanations in lengthy lectures or large documents, leading to inefficient learning and wasted time searching for relevant information. Many traditional platforms only provide access to raw content and basic keyword search, lacking intelligent systems that understand concepts or retrieve precise explanations from a student's study materials. Retrieval-Augmented Generation (RAG) addresses these challenges by combining semantic retrieval with large language models to generate responses based on relevant data. This paper presents an AI-powered teaching assistant that processes educational content, such as MP4 videos and MP3 audio lectures, to create a unified knowledge base. The system uses retrieved context to generate clear explanations, accompanied by video timestamps or document references. This approach enhances learning efficiency, reduces search time, and delivers accurate, context-aware academic support to students.

Keywords: Retrieval Augmented Generation, Educational AI, Semantic Search, Large Language Models, AI Teaching Assistant, Vector Embedding's.

1. INTRODUCTION

The widespread growth of digital education platforms has reshaped the landscape of modern learning, granting students access to an extensive array of multimedia content. Resources such as recorded lectures, instructional videos, textbooks, scholarly articles, and digital notes are now commonplace in universities and online courses. Despite their abundance, these learning materials are often stored in distinct formats and systems, compelling students to sift through lengthy videos or documents to find specific information. Most conventional learning platforms offer only basic playback features and simple keyword searches, which fail to address the complexities of efficient information retrieval. This results in higher cognitive demands and significant amounts of time spent searching through large datasets. With recent

advancements in artificial intelligence, new methods are emerging to enhance the way students interact with educational resources. Retrieval-Augmented Generation (RAG) is one such innovation, blending semantic search techniques with large language models to produce contextually relevant answers. Unlike systems that rely solely on pre-trained data, RAG frameworks pull pertinent details from external sources to create accurate and well-founded responses. This greatly increases the relevance and trustworthiness of AI-generated content. Within education, RAG technology empowers intelligent systems to deliver targeted explanations from transcripts, textbooks, and notes in response to natural language queries. Nevertheless, most current educational tools do not yet offer seamless integration of diverse content types, such as video and audio lectures, into a single cohesive knowledge base. While video platforms typically allow only basic caption searches and document readers support simple text lookup, they lack the sophistication for deeper, concept-driven retrieval.

2. PROBLEM STATEMENT:

In modern educational environments, students often struggle to find relevant explanations quickly from large volumes of learning materials such as lecture videos, audio recordings, and academic documents. Existing learning platforms primarily offer basic keyword search and manual navigation, making the learning process time-consuming and inefficient. Students are often required to search through lengthy videos or documents to locate specific concepts, leading to increased effort and reduced learning efficiency. Another issue is the lack of intelligent semantic understanding in traditional educational systems. Most platforms cannot understand the actual meaning or context of student queries, resulting in inaccurate or irrelevant search results. In addition, existing systems generally do not provide context-aware explanations or direct references such as lecture

timestamps or document sections, making it difficult for students to verify and understand the retrieved information effectively. Therefore, there is a need for an intelligent AI-based learning assistant that can process multimedia educational content, retrieve the most relevant information using semantic search, and generate accurate context-aware explanations through Retrieval-Augmented Generation (RAG). The system should support multiple learning formats, provide source references such as timestamps or document locations, and improve learning efficiency using modern AI technologies.

3. OBJECTIVE

To develop an intelligent AI-based learning assistant for educational content: The main objective of this research is to design an intelligent learning framework that helps students understand educational material more efficiently by using Retrieval-Augmented Generation (RAG). The system processes study materials such as lecture videos, audio recordings, and academic documents, enabling students to retrieve concept-specific explanations directly from their own learning resources.

To integrate Retrieval-Augmented Generation with Large Language Models: This research aims to combine semantic retrieval techniques with large language models to generate context-aware explanations. By retrieving relevant content segments from transcripts and documents and supplying them to the language model, the system can produce accurate and grounded answers related to the student's study material.

To enhance learning efficiency and knowledge accessibility for students: The objective is to improve how students locate and understand information within long lectures and documents by enabling semantic search across multiple learning materials. The system helps students quickly find the exact segment of a video or section of a document where a concept is explained.

To eliminate manual searching across multiple learning platforms: Another objective is to reduce the need for students to manually navigate through different platforms, such as video players, PDF readers, and note-taking applications. The proposed system provides a unified interface where students can upload study materials and directly ask questions in natural language.

To provide context-aware explanations with source references: The research aims to generate explanations that are grounded in the retrieved educational content while providing clear references, such as lecture timestamps or document page numbers. This approach ensures transparency and allows students to verify explanations within the original study material. To support multimodal educational content processing: The objective is to design a system capable of processing multiple learning formats, including MP4 lecture videos and MP3 audio recordings. By converting these materials into structured text through transcription and extraction techniques, the system builds a unified, searchable knowledge base.

To improve the effectiveness of AI-driven academic assistance: The proposed system aims to provide an intelligent academic assistant by improving learning outcomes and reducing the time required for revision.

4. LITERATURE SURVEY

An AI teaching assistant system that enables instructors to upload course PDFs to generate customized Q&A platforms. The system is based on a Retrieval Augmented Generation (RAG). The implementation integrates RAG with responsive web technologies and is evaluated using a standardized test question bank. Experimental results demonstrate that the system achieves an average answer accuracy of up to 86%, indicating a strong performance in an educational context. The results also confirm that the integration of RAG technology significantly improves the accuracy of the response and semantic relevance, with an average accuracy increase of 9.85% compared to the native LLM alone. [1]

Fang et al. (2025) introduce a Retrieval-Augmented Generation (RAG) framework that uses open-source tools and local LLMs to automate the evaluation of simulated teaching audio from teacher trainees. The RAG framework and local LLM system, using the superior InternLM2-Chat-7B model, successfully automated teaching evaluation. Expert scoring confirmed the system's high mean scores for Logic and organisation (4.50) and Content accuracy (4.13). However, its effectiveness is limited by a small, audio-only sample and its inability to capture key teaching elements in multimodal subjects. The lowest score of 3.30 for Practicality and innovation

highlights its limitations in generating in-depth, innovative content. [2]

The Retrieval-Augmented Generation (RAG) approach is a general-purpose fine-tuning method that combines a pre-trained sequence-to-sequence model (parametric memory, BART) with a neural retriever (DPR) over a non-parametric Wikipedia knowledge index. RAG achieved state-of-the-art Exact Match (EM) scores on open-domain Question Answering (QA) tasks, including Natural Questions and Web Questions, outperforming purely parametric and extractive models. For generation

tasks like MS-MARCO and Jeopardy Question Generation, RAG produced responses that were significantly more factual and specific than the BART baseline, while also demonstrating higher diversity. A core advantage is the ability to easily update the model's world knowledge by simply replacing the non-parametric index at test time, without requiring costly retraining of the parametric component. RAG successfully unifies retrieval into a single, end-to-end learned architecture, making it highly effective for a wide range of knowledge-intensive NLP tasks. [3]

The research paper presents VITA (Virtual Teaching Assistants), an adaptive distributed learning (ADL) platform embedding an LLM-powered chatbot to provide dialogic support and integrity-aware, formative assessment for data science education. The introduction highlights that the system's motivation is to address the urgent challenge of scaling data science programs while maintaining quality, using OpenAI's models to create personalised and scalable educational experiences that enhance faculty productivity. The system automates formative assessments and tracks specific metrics, such as the collective counts of Passed or Failed for each quiz and over 206,000 statements in deployed courses, which are used to drive personalised adaptive learning pathways. [4]

The research paper explains that Retrieval-Augmented Generation (RAG) improves large language models (LLMs) by letting them use external data for more accurate and reliable answers. It shows that RAG reduces hallucinations, increases domain-specific accuracy, and makes responses more interpretable and controllable compared to normal LLMs. The paper concludes

that RAG-based systems are essential for building trustworthy, domain-adapted AI applications across fields like healthcare, law, and education [5]

The study presents Retrieval-Augmented Generation (RAG) as a powerful approach that enhances large language models by linking them to external and domain-specific data sources for more accurate and trustworthy responses. It integrates vector databases and embedding models to reduce hallucinations and strengthen factual grounding, ensuring that the AI's outputs are both precise and verifiable. The findings demonstrate that RAG significantly improves response accuracy, reliability, and explainability, making it highly effective for real-world applications such as education and enterprise AI assistants. [6]

Retrieval-Augmented Generation (RAG) is a critical framework designed to mitigate LLM hallucinations and enhance factual accuracy for knowledge-intensive tasks. The RAG architecture integrates the LLM's parametric memory with an external non-parametric memory a vector database containing organisational or domain-specific data. The pipeline operates by retrieving contextually relevant data, augmenting the user query, and subsequently generating a grounded response. This process significantly improves contextual understanding and allows for grounding, providing verifiable references to source materials. While RAG offers a cost-effective alternative to model fine-tuning, its successful implementation necessitates robust data management. [7]

Retrieval-Augmented Generation (RAG) was primarily introduced to overcome the limitations of large language models (LLMs) in precisely handling knowledge and preventing hallucinations, by combining a parametric seq2seq model (e.g., BART-large, 400M parameters) with an explicit non-parametric memory (a 21M document Wikipedia index). This hybrid architecture achieves State-of-the-Art performance on knowledge-intensive tasks, such as 44.5 Exact Match (EM) on Natural Questions, requiring fewer trainable parameters than purely parametric models like. RAG's strength in ensuring grounded responses makes it an essential foundation for specialized applications like educational chatbots, where its simple implementation is used predominantly for factual access to source knowledge and personalized learning support. [8]

The paper, "Leveraging In-Context Learning and Retrieval-Augmented Generation for Automatic Question Generation in Educational Domains," addresses the need for contextually relevant and pedagogically sound questions, as current automated methods often generate irrelevant questions. The authors explore In-Context Learning (ICL) using GPT-4, Retrieval-Augmented Generation (RAG) using BART and a retrieval module, and a novel Hybrid Model combining both, to improve question generation. In evaluation, ICL with k=7 demonstrated superior performance in automated metrics like ROUGE-L and BERT Score. However, the Hybrid Model (combining retrieval and ICL) consistently achieved the highest scores in human evaluation metrics, including grammaticality, appropriateness, and especially complexity. Overall, both the ICL and Hybrid models significantly outperformed traditional fine-tuned baselines like T5-large and BART-large, underscoring the benefits of these advanced techniques for creating high-quality educational questions. [9]

The paper, "Investigating the Intersection of LLMs, RAG, and Learning Analytics in Higher Education," is a mixed-methods study that analyses the integration of Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and Learning Analytics in the educational sector. The findings indicate that these tools are effective in significantly boosting student engagement and facilitating personalized learning, with faculty also reporting increased time efficiency for administrative tasks. However, the study also highlights significant challenges that must be addressed, including critical concerns over data privacy, the potential for algorithmic bias, and the urgent need for adequate faculty training to ensure effective and equitable deployment. In terms of performance parameters, the study gauged the tools' perceived effectiveness on a 5-point scale, resulting in an average rating of 3.35 by students and 3.75 by faculty, reinforcing the positive impact while underscoring the need for further refinement. [10]

5. METHODOLOGIE:

The proposed system integrates Retrieval Augmented Generation (RAG), semantic embedding's, and large language models to provide an intelligent academic assistant capable of

retrieving and explaining educational content from multimedia learning materials. The methodology consists of the following major components:

Learning Material Ingestion Module: Educational resources such as lecture videos, audio recordings, textbooks, and digital documents are collected and uploaded into the system. The supported formats include MP4 lecture videos and MP3 audio recordings. Video files are converted into audio format using media processing tools, while textual data is extracted from audio. This module ensures that learning materials from different sources are gathered and converted into a unified format suitable for further processing.

Content Processing and Transcription Module: The uploaded learning materials are processed to extract meaningful textual information. Audio and video lectures are transcribed using automated speech recognition models to generate time stamped text transcripts (JSON format). The extracted data is then cleaned and normalised by removing noise, unnecessary symbols, and formatting artefacts. This module converts unstructured multimedia learning materials into structured textual data that can be used for semantic analysis.

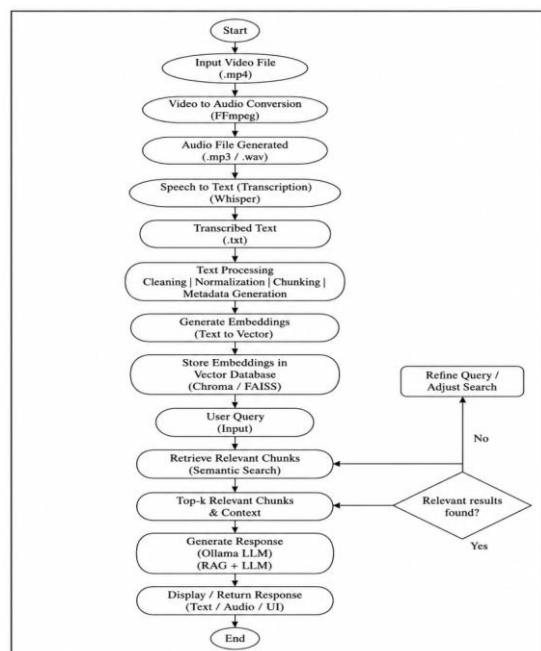


Fig. 2: System Flowchart

Fig. 1 Flowchart

Semantic Chunking and Embedding Module:

The processed textual content is divided into smaller semantic segments to improve retrieval accuracy. Each segment is transformed into a vector representation using embedding models capable of capturing semantic relationships between words and sentences. These embeddings represent the contextual meaning of the content and are stored along with metadata such as video timestamps. The resulting vector representations enable efficient similarity search and semantic matching between user queries and relevant educational content.

Retrieval-Augmented Generation Module: When a student submits a question, the system converts the query into a vector embedding using the same embedding model used during indexing. A similarity search is then performed against the stored embeddings to identify the most relevant content segments. The retrieved segments are supplied as contextual information to a large language model, which generates a clear and context-aware explanation for the user’s query. By grounding responses in retrieved educational content, this module improves the accuracy of generated explanations while providing references such as lecture timestamps.

6. SYSTEM ARCHITECTURE:

The architecture illustrates the integration of Retrieval-Augmented Generation (RAG), semantic embeddings, and large language models to enable intelligent retrieval and explanation of educational content. The system processes multimedia learning resources such as lecture videos, audio recordings transforming them into a unified semantic knowledge base. The architecture consists of multiple interconnected components responsible for data processing, semantic retrieval, and AI-based response generation.

1. Learning Content Source

The system accepts educational materials from multiple sources, such as lecture videos, recorded tutorials, audio lectures, and academic documents. These sources include,

Video Lectures: Recorded classroom lectures or online learning videos that contain conceptual explanations. **Audio Recordings:** Educational

podcasts, lecture recordings, or audio-based learning resources are uploaded, and they are processed to extract structured textual information. The processing layer performs the following tasks:

1. Video and Audio Transcription:

- Speech recognition models convert lecture audio into time stamped transcripts.
- Document Text Extraction: Audio is parsed to extract textual content.

2. Embedding and Vector Storage Layer :

The processed text is segmented into smaller semantic chunks to improve retrieval accuracy. Each chunk is then converted into vector embeddings that represent the contextual meaning of the content:

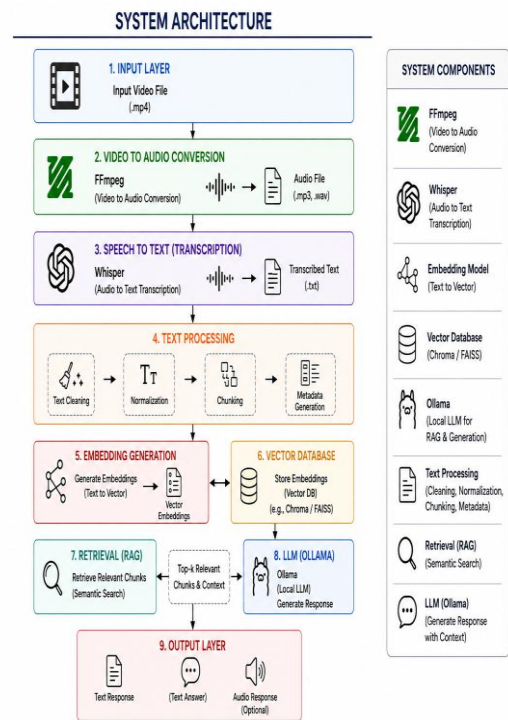


Fig. 2: System Architecture

These embeddings are stored in a database, which enables efficient similarity-based search operations. These embeddings are stored in a database, which enables efficient similarity-based search operations. The stored data includes:

- Semantic embeddings of text segments.
- Meta data such as video time stamps.
- References to original learning materials. This layer enables fast retrieval of

contextually relevant information based on semantic similarity.

3. Retrieval and Query Processing Layer

When a user submits a query, the system processes the request through the retrieval engine. The query is converted into a vector embedding using the same embedding model used during indexing:

- The retrieval engine performs a similarity search against the stored embeddings and identifies the most relevant content segments from the knowledge base.

4. Response Generation Layer: The response generation layer uses a large language model to produce explanations based on the retrieved context. The language model combines the user query with the retrieved content segments to generate accurate and context-aware answers.

5. User Interface Layer The user interface provides a simple and interactive platform for students to interact with the system. Through the interface, users can:

- Ask questions related to the uploaded content
- View explanations generated by the AI assistant.
- Navigate to relevant lecture timestamps or document sections.
- The interface ensures a seamless interaction between the user and the learning assistant.

7. RESULTS AND ANALYSIS

Evaluation Metrics : To assess the effectiveness of the proposed RAG system, the following performance metrics were considered:

Retrieval Accuracy: Measures the correctness of retrieved content segments relevant to user queries.

Response Latency: Time taken to generate a response after receiving a query, contextual information. Relevance and clarity of responses.

Experimental Setup The system was evaluated using a simulated academic environment consisting of diverse educational resources.

Dataset: lecture videos (MP4), audio lectures (MP3)

Pipeline: Speech-to-text transcription for audio/video, Semantic chunking of extracted text, Vector embeddings using transformer-based models Similarity search using vector database, Baseline for

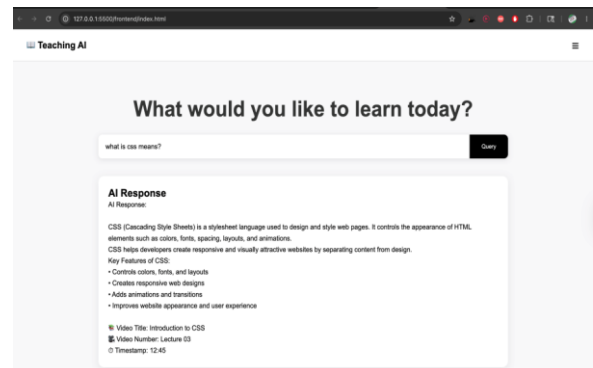


Fig.3: User Interface

8. CONCLUSION

The proposed RAG-Based AI Teaching Assistant provides an effective solution for retrieving useful information from educational video content. The system combines speech transcription; semantic search, vector embeddings, and Retrieval-Augmented Generation (RAG) to help users quickly access important learning materials from large video collections. By displaying accurate answers along with related video titles and timestamps, the system reduces the time required for manual searching and improves the overall learning experience.

Unlike traditional keyword-based search systems, the developed assistant understands the meaning and context of user queries using Natural Language Processing techniques. This enables the system to deliver more relevant and context-aware responses. The integration of vector embeddings and semantic similarity search also improves retrieval accuracy and supports scalability for handling multiple educational resources efficiently.

The implementation results demonstrate that the system can successfully convert video lectures into searchable knowledge, generate meaningful responses, and assist learners in understanding concepts more effectively. The user-friendly interface further enhances accessibility and interaction for students.

Overall, the project highlights the practical use of Artificial Intelligence in the education domain. The proposed system improves the accessibility.

9. REFERENCES

<https://ssrn.com/abstract=5199042>

- 1) Developing a Local Generative AI Teaching Assistant System: Utilizing Retrieval-Augmented Generation Technology to Enhance the Campus Learning Environment <https://www.mdpi.com/2079-9292/14/17/3402>
- 2) Title: Evaluating simulated teaching audio for teacher trainees using RAG and local LLMs <https://doi.org/10.1038/s41598-025-87898-5>
- 3) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks <https://arxiv.org/abs/2005.11401>
- 4) Anaroua, F. I., Li, Q., Tang, Y., & Liu, H. P. (2025). AI-driven formative assessment and adaptive learning in data-science education: Evaluating an LLM-powered virtual teaching assistant. <https://arxiv.org/abs/2509.20369>
- 5) Zhao, S., Yang, Y., Wang, Z., He, Z., Qiu, L. K., & Qiu, L. (2024). Retrieval-Augmented Generation (RAG) and beyond: A comprehensive survey on how to make your LLMs use external data more wisely. <https://arxiv.org/abs/2409.14924>
- 6) Merit, R. (2025, January 31). What is Retrieval-Augmented Generation (RAG)? NVIDIA Blog <https://blogs.nvidia.com/blog/what-is-retrievalaugmented-generation/>
- 7) Klesel, M., & Wittmann, H. F. (2025). Retrieval-Augmented Generation (RAG). *Business & Information Systems Engineering*, 67(4), 551–561. <https://doi.org/10.1007/s12599-025-00945-3>
- 8) Swacha, J. Gracel, M. Retrieval-Augmented Generation (RAG) Chatbots for Education: A Survey of Applications. *Appl. Sci.* 2025, 15, 4234. <https://doi.org/10.3390/app15084234>
- 9) Leveraging In-Context Learning and Retrieval-Augmented Generation for Automatic Question Generation in Educational Domains <https://arxiv.org/abs/2501.17397>
- 10) Investigating the Intersection of LLMs, RAG, and Learning Analytics in Higher Education