IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, June 2026
Impact Factor: 8.2

Retrieval-Augmented Generation (RAG) Based AI Teaching Assistant for Personalized Learning

Sachin[1], Abhishek[2], Ms. Vandana Swami[3], Dr. Rajendra Singh[4]
Student[1,2], Assistant Professor[3], Dean[4]
Department of Computer Science and Engineering[1,2,3,4]
Raffles University, Neemrana, Rajasthan, India
Sachindiwa2003@gmail.com[1], abhiydv8955@gmail.com[2],
vandana.swami@rafflesuniversity.edu.in[3], rajendra.singh@rafflesuniversity.edu.in[4]

Abstract: *Retrieval-Augmented Generation (RAG) combines information retrieval and large language models to provide accurate, context-aware answers. This research develops an AI Teaching Assistant that retrieves educational content from a knowledge base before generating responses. The system improves learning outcomes, reduces hallucinations, and supports personalized education. The architecture integrates document ingestion, embeddings, vector databases, semantic search, and response generation.*
Retrieval-Augmented Generation (RAG) combines information retrieval and large language models to provide accurate, context-aware answers. This research develops an AI Teaching Assistant that retrieves educational content from a knowledge base before generating responses. The system improves learning outcomes, reduces hallucinations, and supports personalized education. The architecture integrates document ingestion, embeddings, vector databases, semantic search, and response generation.
Retrieval-Augmented Generation (RAG) combines information retrieval and large language models to provide accurate, context-aware answers. This research develops an AI Teaching Assistant that retrieves educational content from a knowledge base before generating responses. The system improves learning outcomes, reduces hallucinations, and supports personalized education. The architecture integrates document ingestion, embeddings, vector databases, semantic search, and response generation.

Keywords: Retrieval-Augmented Generation (RAG) Based AI Teaching Assistant for Personalized Learning

I. INTRODUCTION

Artificial Intelligence is transforming modern education. Traditional learning methods often fail to provide personalized support to every learner. AI-powered educational assistants can help students access information quickly. However, conventional chatbots rely only on pre-trained knowledge and may generate inaccurate responses. Retrieval-Augmented Generation addresses this limitation by retrieving relevant information before generating answers. This project focuses on building a RAG-based teaching assistant capable of providing reliable educational guidance.

The demand for personalized learning is increasing due to the growth of online education platforms. Students require instant access to subject-specific knowledge and adaptive learning resources. The proposed system bridges this gap by combining semantic retrieval and generative AI.

Artificial Intelligence is transforming modern education. Traditional learning methods often fail to provide personalized support to every learner. AI-powered educational assistants can help students access information quickly. However, conventional chatbots rely only on pre-trained knowledge and may generate inaccurate responses. Retrieval-Augmented Generation addresses this limitation by retrieving relevant information before generating answers. This project focuses on building a RAG-based teaching assistant capable of providing reliable educational guidance.

The demand for personalized learning is increasing due to the growth of online education platforms. Students require instant access to subject-specific knowledge and adaptive learning resources. The proposed system bridges this gap by combining semantic retrieval and generative AI.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, June 2026
Impact Factor: 8.2

Artificial Intelligence is transforming modern education. Traditional learning methods often fail to provide personalized support to every learner. AI-powered educational assistants can help students access information quickly. However, conventional chatbots rely only on pre-trained knowledge and may generate inaccurate responses. Retrieval-Augmented Generation addresses this limitation by retrieving relevant information before generating answers. This project focuses on building a RAG-based teaching assistant capable of providing reliable educational guidance.

The demand for personalized learning is increasing due to the growth of online education platforms. Students require instant access to subject-specific knowledge and adaptive learning resources. The proposed system bridges this gap by combining semantic retrieval and generative AI.

II. PROBLEM STATEMENT

Students frequently struggle to find accurate educational resources and personalized guidance. Existing chatbots often produce hallucinated answers and lack contextual awareness. Educational institutions require systems capable of delivering trustworthy and up-to-date information. Therefore, a RAG-based solution is proposed to improve answer quality and learning effectiveness.

Students frequently struggle to find accurate educational resources and personalized guidance. Existing chatbots often produce hallucinated answers and lack contextual awareness. Educational institutions require systems capable of delivering trustworthy and up-to-date information. Therefore, a RAG-based solution is proposed to improve answer quality and learning effectiveness.

Students frequently struggle to find accurate educational resources and personalized guidance. Existing chatbots often produce hallucinated answers and lack contextual awareness. Educational institutions require systems capable of delivering trustworthy and up-to-date information. Therefore, a RAG-based solution is proposed to improve answer quality and learning effectiveness.

III. OBJECTIVES

1. Develop a personalized AI teaching assistant.
2. Improve answer accuracy through retrieval mechanisms.
3. Reduce hallucination in generated responses.
4. Support adaptive and self-paced learning.
5. Enhance student engagement and academic performance.
6. Provide scalable educational support.

IV LITERATURE REVIEW

Recent studies show that Retrieval-Augmented Generation significantly improves factual accuracy in AI systems. Lewis et al. introduced the RAG framework for integrating retrieval and generation. Vaswani et al. proposed the Transformer architecture that powers modern language models. Devlin et al. developed BERT, which improved contextual understanding in NLP. Research indicates that combining retrieval mechanisms with generative models reduces misinformation and improves reliability.

Several educational platforms have experimented with AI tutors. However, many systems still suffer from outdated knowledge and limited personalization. The RAG approach provides a practical solution by grounding responses in external knowledge sources.

Recent studies show that Retrieval-Augmented Generation significantly improves factual accuracy in AI systems. Lewis et al. introduced the RAG framework for integrating retrieval and generation. Vaswani et al. proposed the Transformer architecture that powers modern language models. Devlin et al. developed BERT, which improved contextual understanding in NLP. Research indicates that combining retrieval mechanisms with generative models reduces misinformation and improves reliability.

Several educational platforms have experimented with AI tutors. However, many systems still suffer from outdated knowledge and limited personalization. The RAG approach provides a practical solution by grounding responses in external knowledge sources.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, June 2026
Impact Factor: 8.2

Recent studies show that Retrieval-Augmented Generation significantly improves factual accuracy in AI systems. Lewis et al. introduced the RAG framework for integrating retrieval and generation. Vaswani et al. proposed the Transformer architecture that powers modern language models. Devlin et al. developed BERT, which improved contextual understanding in NLP. Research indicates that combining retrieval mechanisms with generative models reduces misinformation and improves reliability.

Several educational platforms have experimented with AI tutors. However, many systems still suffer from outdated knowledge and limited personalization. The RAG approach provides a practical solution by grounding responses in external knowledge sources.

V. SYSTEM ARCHITECTURE

The proposed architecture consists of six major modules: document ingestion, preprocessing, embedding generation, vector database storage, retrieval engine, and response generation. Educational documents are uploaded and converted into chunks. Embeddings are generated using transformer-based models and stored in a vector database. User queries are embedded and matched with relevant chunks through similarity search. Retrieved content is passed to a Large Language Model to generate accurate answers.

The proposed architecture consists of six major modules: document ingestion, preprocessing, embedding generation, vector database storage, retrieval engine, and response generation. Educational documents are uploaded and converted into chunks. Embeddings are generated using transformer-based models and stored in a vector database. User queries are embedded and matched with relevant chunks through similarity search. Retrieved content is passed to a Large Language Model to generate accurate answers.

The proposed architecture consists of six major modules: document ingestion, preprocessing, embedding generation, vector database storage, retrieval engine, and response generation. Educational documents are uploaded and converted into chunks. Embeddings are generated using transformer-based models and stored in a vector database. User queries are embedded and matched with relevant chunks through similarity search. Retrieved content is passed to a Large Language Model to generate accurate answers.

V. METHODOLOGY

The methodology begins with data collection from educational resources. Documents are cleaned and divided into smaller chunks. Embeddings are generated using Hugging Face models and stored in FAISS. When a user submits a query, the system converts it into an embedding and retrieves relevant information using semantic similarity. Retrieved content is combined with the user query and passed to the language model. The generated answer is then displayed through a web interface.

Evaluation metrics include response accuracy, retrieval relevance, latency, and user satisfaction.

The methodology begins with data collection from educational resources. Documents are cleaned and divided into smaller chunks. Embeddings are generated using Hugging Face models and stored in FAISS. When a user submits a query, the system converts it into an embedding and retrieves relevant information using semantic similarity. Retrieved content is combined with the user query and passed to the language model. The generated answer is then displayed through a web interface.

Evaluation metrics include response accuracy, retrieval relevance, latency, and user satisfaction.

The methodology begins with data collection from educational resources. Documents are cleaned and divided into smaller chunks. Embeddings are generated using Hugging Face models and stored in FAISS. When a user submits a query, the system converts it into an embedding and retrieves relevant information using semantic similarity. Retrieved content is combined with the user query and passed to the language model. The generated answer is then displayed through a web interface.

Evaluation metrics include response accuracy, retrieval relevance, latency, and user satisfaction.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, June 2026
Impact Factor: 8.2

VI. TECHNOLOGIES USED

Python is used as the primary programming language. LangChain orchestrates retrieval and generation workflows. FAISS provides efficient vector similarity search. Hugging Face Embeddings convert text into dense vectors. Gemini/OpenAI APIs generate responses. Streamlit is used to create an interactive user interface. These technologies collectively enable efficient implementation of the RAG pipeline.

Python is used as the primary programming language. LangChain orchestrates retrieval and generation workflows. FAISS provides efficient vector similarity search. Hugging Face Embeddings convert text into dense vectors. Gemini/OpenAI APIs generate responses. Streamlit is used to create an interactive user interface. These technologies collectively enable efficient implementation of the RAG pipeline.
Python is used as the primary programming language. LangChain orchestrates retrieval and generation workflows. FAISS provides efficient vector similarity search. Hugging Face Embeddings convert text into dense vectors. Gemini/OpenAI APIs generate responses. Streamlit is used to create an interactive user interface. These technologies collectively enable efficient implementation of the RAG pipeline.

VII. IMPLEMENTATION

The implementation phase includes dataset preparation, chunk creation, embedding generation, vector indexing, prompt engineering, and frontend integration. Educational documents are stored in PDF format and processed automatically. The retrieval module identifies relevant content, while the language model generates coherent responses. Streamlit enables users to interact with the system through a simple interface.

Testing was conducted using educational queries from computer science and engineering subjects.

The implementation phase includes dataset preparation, chunk creation, embedding generation, vector indexing, prompt engineering, and frontend integration. Educational documents are stored in PDF format and processed automatically. The retrieval module identifies relevant content, while the language model generates coherent responses. Streamlit enables users to interact with the system through a simple interface.
Testing was conducted using educational queries from computer science and engineering subjects.
The implementation phase includes dataset preparation, chunk creation, embedding generation, vector indexing, prompt engineering, and frontend integration. Educational documents are stored in PDF format and processed automatically. The retrieval module identifies relevant content, while the language model generates coherent responses. Streamlit enables users to interact with the system through a simple interface.
Testing was conducted using educational queries from computer science and engineering subjects.

VIII. RESULTS AND DISCUSSION

Experimental results indicate that the RAG-based teaching assistant performs significantly better than traditional chatbots. The system achieved higher factual accuracy and lower hallucination rates. Users reported improved satisfaction due to context-aware responses. Retrieval grounding ensured that generated answers remained relevant and trustworthy. The architecture also demonstrated scalability for large educational datasets.

Experimental results indicate that the RAG-based teaching assistant performs significantly better than traditional chatbots. The system achieved higher factual accuracy and lower hallucination rates. Users reported improved satisfaction due to context-aware responses. Retrieval grounding ensured that generated answers remained relevant and trustworthy. The architecture also demonstrated scalability for large educational datasets.
Experimental results indicate that the RAG-based teaching assistant performs significantly better than traditional chatbots. The system achieved higher factual accuracy and lower hallucination rates. Users reported improved satisfaction due to context-aware responses. Retrieval grounding ensured that generated answers remained relevant and trustworthy. The architecture also demonstrated scalability for large educational datasets.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, June 2026
Impact Factor: 8.2

IX. ADVANTAGES

The proposed system offers several benefits, including improved accuracy, personalized learning experiences, scalability, reduced misinformation, efficient knowledge retrieval, and support for continuous learning. It can assist students across multiple domains and educational levels.

The proposed system offers several benefits, including improved accuracy, personalized learning experiences, scalability, reduced misinformation, efficient knowledge retrieval, and support for continuous learning. It can assist students across multiple domains and educational levels.
The proposed system offers several benefits, including improved accuracy, personalized learning experiences, scalability, reduced misinformation, efficient knowledge retrieval, and support for continuous learning. It can assist students across multiple domains and educational levels.

X. LIMITATIONS

The effectiveness of the system depends on the quality of the knowledge base. Poorly structured data may reduce retrieval accuracy. Large-scale datasets may require significant storage and computational resources. Response latency may increase when handling extensive document collections.

The effectiveness of the system depends on the quality of the knowledge base. Poorly structured data may reduce retrieval accuracy. Large-scale datasets may require significant storage and computational resources. Response latency may increase when handling extensive document collections.
The effectiveness of the system depends on the quality of the knowledge base. Poorly structured data may reduce retrieval accuracy. Large-scale datasets may require significant storage and computational resources. Response latency may increase when handling extensive document collections.

XI. FUTURE SCOPE

Future enhancements include multilingual support, voice-based interaction, adaptive learning analytics, recommendation systems, and integration with Learning Management Systems. Advanced vector databases and fine-tuned language models can further improve performance.

Future enhancements include multilingual support, voice-based interaction, adaptive learning analytics, recommendation systems, and integration with Learning Management Systems. Advanced vector databases and fine-tuned language models can further improve performance.
Future enhancements include multilingual support, voice-based interaction, adaptive learning analytics, recommendation systems, and integration with Learning Management Systems. Advanced vector databases and fine-tuned language models can further improve performance.

XII. CONCLUSION

The RAG-based AI Teaching Assistant demonstrates the potential of combining retrieval systems with generative AI for education. By grounding responses in retrieved knowledge, the system improves accuracy, reliability, and personalization. The project contributes to the development of intelligent educational tools capable of supporting students in modern learning environments.

The RAG-based AI Teaching Assistant demonstrates the potential of combining retrieval systems with generative AI for education. By grounding responses in retrieved knowledge, the system improves accuracy, reliability, and personalization. The project contributes to the development of intelligent educational tools capable of supporting students in modern learning environments.
The RAG-based AI Teaching Assistant demonstrates the potential of combining retrieval systems with generative AI for education. By grounding responses in retrieved knowledge, the system improves accuracy, reliability, and personalization. The project contributes to the development of intelligent educational tools capable of supporting students in modern learning environments.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, June 2026
Impact Factor: 8.2

REFERENCES

[1] Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2020.
[2] Vaswani et al., Attention Is All You Need, 2017.
[3] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers, 2018.
[4] LangChain Documentation.
[5] FAISS Documentation.
[6] Hugging Face Documentation