A Retrieval-Augmented Generation (RAG) Framework for Generating AN Accurate and Real Time Educational Content

Rashmi P
Faculty, ICIS, Srinivas University
doi.org/10.64643/IJIRTV12I12-206817-459

***Abstract*—The growing demand for intelligent learning systems has brought to the fore the significance of context-aware, real-time content generation technologies. The present paper presents a comprehensive Retrieval-Augmented Generation (RAG) system with Google Gemini models for facilitating accurate question-answering from user-uploaded learning documents. The proposed system offers multimodal input — text, voice and picture-based search queries — enabling users to interact intuitively and naturally. Leaning on FAISS for optimized vector indexing and LangChain for runtime chaining of dynamic retrievals, the platform delivers high-precision matching of documents. Enriched with user authentication, session logging, inappropriate query detection and feedback logging, the system also has an admin dashboard to monitor user behavior and control content too. Offline text-to-speech synthesis and auto-generated query resolution reminders are some features that enhance the user experience quite significantly too. The solution suggested offers a flexible, secure and user-centered approach to the presentation of educational content with potentially helpful implications for customized learning environments.**

***Index Terms*—FAISS Vector store, Google Gemini, LLM, LangChain, Retrieval-Augmented Generation (RAG), Streamlit**

I. INTRODUCTION

In the context of modern education, the proposed multimodal Retrieval-Augmented Generation (RAG) system addresses critical challenges of scalability, accuracy and accessibility inherent in current digital learning platforms. With the proliferation of diverse input modalities and the demand for context-sensitive learning support, the system offers a transformative approach to knowledge retrieval and generation for educational institutions, teachers and learners.The integration of Large Language Models (LLMs) with retrieval mechanisms ensures that learners receive grounded and evidence-based answers rather than hallucinated outputs, which are common when relying solely on generative models [7], [26]. By incorporating vector similarity search through FAISS, the system is capable of rapidly retrieving semantically relevant documents from massive datasets, making it suitable for deployment in digital universities, national knowledge portals and e-learning management systems where scalability is critical [9], [27].

Furthermore, the inclusion of Optical Character Recognition (OCR) and speech-to-text functionalities enhances accessibility, enabling users to submit queries via handwritten notes or spoken questions, thus bridging literacy and disability gaps [2], [11]. The feedback mechanism embedded within the system supports continuous model refinement based on user responses, fostering adaptive learning pathways tailored to users' needs [16], [22].

For teachers and administrators, the system serves as an intelligent assistant capable of generating summaries, quizzes and structured lesson plans, thereby reducing instructional workload while maintaining high content quality [8], [14]. In addition, features such as authentication, chat moderation and query filtering ensure secure and responsible usage, which is essential for institutional compliance and the prevention of misuse [4], [21].

The admin dashboard provides actionable analytics, including login trends, frequently asked questions and session durations, empowering administrators to make

data-driven decisions on content curation, resource allocation and student engagement strategies [18], [23]. Moreover, the automatic monitoring of unanswered queries ensures that the knowledge base remains dynamic, continuously updated with newly uploaded documents to address evolving informational requirements [12], [25].

Importantly, the multimodal capabilities of the interface enhance inclusivity by supporting text, voice and image inputs, thus catering to learners with diverse preferences and abilities. This is particularly significant in multilingual and resource-constrained educational contexts, where language barriers and digital literacy challenges often limit equitable learning opportunities [3], [20].

Overall, the proposed multimodal RAG system exemplifies how machine learning, retrieval-augmented generation and human-centred design can converge to create intelligent, adaptive educational support systems. By bridging the gap between static content and dynamic user queries, it contributes to building responsive, culturally sensitive and scalable learning environments for the digital age [1], [6], [28].

II. OBJECTIVE

- To create a multimodal Retrieval-Augmented Generation (RAG) platform that seamlessly blends text, voice and image-based inputs to generate accurate, context-aware answers for educational questions in real-time.
- To enable a secure, user-focused authentication system that facilitates account management, query moderation, user blocking and session tracking for responsible use in institutional deployments.
- In order to incorporate Optical Character Recognition (OCR) and speech-to-text capabilities for accessibility improvement, allowing students to enter their queries with handwritten notes or verbal questions without any interruption.
- In order to create a smart admin dashboard that offers actionable insights on user behaviour, query patterns, session length and feedback to aid data-driven decision-making for content creation and platform improvement.
- To create an ongoing pipeline for improving the knowledge base by tracking open questions and dynamically refreshing document indexes to keep the system responsive to changing learning needs.

III. METHODOLOGY

This study presents a structured approach for delivering accurate, context-aware educational content using Retrieval-Augmented Generation (RAG) integrated with multimodal input processing. The system is designed to handle diverse user queries submitted through text, voice and image inputs to ensure continuity, accessibility and relevance throughout the knowledge retrieval and generation process.

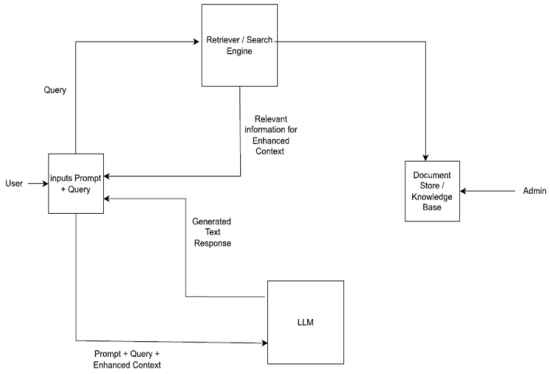

Fig. 1. RAG System diagram

The proposed system is a multimodal Retrieval-Augmented Generation (RAG) system designed to facilitate accurate educational content generation. It is based on a modular, scalable system with multiple dependent modules. The underlying approach involves document ingestion, semantic indexing, multimodal input processing and context-sensitive response generation using Google Gemini language models in LangChain. The following subsections outline the system's key components.

User Management

The system implements a safe and convenient user authentication process to offer responsible access. Users can register by creating a username, password and selection of security question to restore an account. Credentials are verified against saved hashed values using SHA-256 encryption when a user logs in. The system also offers password recovery via the

user's selected security question to allow accessibility without compromising security. User sessions are tracked to monitor activity and develop analytics and an admin can manage user logs as well as watch for patterns of behavior.

Document Parsing and Ingestion
To generate responses based on domain-specific content, users or admins can upload a file in PDF, DOCX, TXT format or even an image containing text. For text files, the system utilizes LangChain's loaders to extract raw content.

For images, the system uses EasyOCR to extract textual content via Optical Character Recognition (OCR). The retrieved content is chunked through recursive splitting algorithms to maintain context coherence before vector embedding and the parsing pipeline effectively enables efficient representation and retrieval of knowledge in the querying process.

Vector store with FAISS
All document pieces are then loaded with Google Generative AI Embeddings to transform them into semantic vectors of high dimension. These embeddings are subsequently indexed within a FAISS (Facebook AI Similarity Search) index, enabling efficient similarity search.

The system keeps scanning for updated, deleted or new documents and auto-updates the FAISS index to ensure that queries are responded against the most appropriate content at all times.

Query flow using LangChain
When a user submits a query text, voice, or image-based the system converts it into a text format, passes it through the FAISS retriever and returns top-matching documents.

These documents are passed to a LangChain create_retrieval_chain that passes them to a Google Gemini LLM with a prompt template structure.

The model generates a response only from the context retrieved, keeping hallucinations low and accuracy high. The system supports model-switching between different Gemini models, allowing flexibility in cost and performance.

Inappropriate Query Handling
For ethical and responsible usage, the system maintains a database of offensive or banned words. All user searches are filtered for these words and repeat offenders get temporarily or permanently blocked. Blocking results are tracked by user offenses and punishments are assigned accordingly. This behavior is logged as well and gets stored in the user's session history that can be viewed by the admin for moderation.

Admin Interface: Feedback and Analytics
Secured admin dashboard is accomplished by the implementation of Streamlit for system activity control.

The key features include:

- Document Upload and Management – Admins can view and delete uploaded documents or convert image documents to DOCX documents after OCR.
- User Activity Logs – Admins can monitor login/logout timestamps, session duration and usage patterns.
- Query Trends – A user query log tracks frequency of user queries, enabling visualization of top topics through bar charts and word clouds.
- Feedback Monitoring – User feedback presented by users is collected and analyzed to derive improvement suggestions. Admins can also opt to clear logs and feedback at regular intervals. This admin layer adds transparency, administrability and long-term flexibility to the system.

Technologies and Tools Used in the System
User Authentication:
User authentication is provided using Streamlit, JSON and SHA-256 for password encryption.

Security Questions System:
A Python logic provides functionality for recovering passwords by predefined security questions.

Document Parsing:
Parsing of the documents is done using LangChain loaders and EasyOCR with support for PDF, DOCX, TXT and image formats.

Text Vector Embedding & Indexing: Semantic vectorizing is done using Google Generative AI Embedding and indexing is done using FAISS library.
Integration With Language Model: Response generation is performed by Google Gemini API using LangChain framework.
Retrieval Chain: LangChain's retrieval chain combines language model and document retrieval processes.
Input Modes: The input can be text, voice and image types based on Streamlit, EasyOCR and Speech Recognition packages.
Inappropriate Query Filter: Custom JSON-based filter for inappropriate query blocking.
Session Tracking & Chat History Logging: Activity tracking and chat history logging using JSON, Pickle and Streamlit.
Feedback Collection & Admin Dashboard: Collect feedback using Streamlit, monitoring activities in the dashboard using Plotly and Pandas.
Query Analysis: Analysis of user queries using Plotly and WordCloud.
OCR, Voice & Speech: Image recognition using EasyOCR; Voice-to-text conversion using Speech Recognition and vice versa using pyttsx3/gTTS.

IV. ANALYSIS AND OBSERVATION

The deployment of the multimodal RAG learning platform identified a few important learnings during testing and assessment. First, combining text, voice and image-based query entries greatly enhanced system accessibility, enabling users with diverse preferences and capabilities to use the system seamlessly. Speech-to-text output precisely captured spoken queries in quiet settings, while the OCR module successfully gleaned text from legible printed pictures, though minor errors were noticed with handwritten or smudged inputs.
The search performance with FAISS-based vector similarity search was efficient even for large sets of documents and responded with accurate results in seconds. It was noticed that the response accuracy relied greatly on the quality and coverage of the documents to be uploaded, emphasizing the need for extensive and well-curated knowledge base.
User moderation and authentication functionalities, such as query filtering and user blocking options, functioned smoothly, maintaining responsible use and reinforcing platform security. User feedback analysis showed that users enjoyed the real-time response production and support across multiple modes, although some recommended refinement in the clarity of the interface for new users.
The admin dashboard successfully summarized user activity trends, query analysis and session lengths, which allowed administrators to make data-driven decisions regarding content refreshes and user engagement. A significant finding was the fact that unanswerable questions, when automatically re-submitted on the arrival of new document uploads, enhanced the adaptability of the system and minimized unresolved queries in the long run.
In total, the analysis proves that the system based on the proposal effectively closes the gap between static learning materials and dynamic learning questions and thus offers a scalable, responsive and user-driven solution for contemporary educational settings.

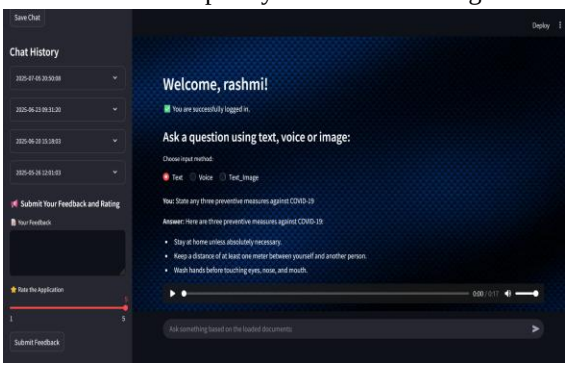

Fig. 2. Sample Output

V. CONCLUSION

This project proposes a robust and interactive Retrieval-Augmented Generation (RAG) system for content presentation in education. Through the integration of document retrieval and Google Gemini large language models in LangChain, the system provides context-related, precise output to user queries. Its support of multimodal inputs like text, voice and image augments its ease of access and usability. The major features such as user authentication, stripping out objectionable queries, immediate response and monitoring chat history improve both usability and security.
The application of a FAISS-based vector store ensures performance-intensive document indexing and similarity search, enabling the system to scale with

huge datasets without performance degradation. Furthermore, the admin dashboard provides the administrator with the authority to monitor user activity, manage document uploads and derive actionable insights from query and feedback trends.

Overall, the framework bridges the gap between static learning materials and dynamic with smart support. It provides the foundation for scalable, ethical and adaptive learning tools that will be capable of adapting with evolving students' demands. The future can reach as far as multilingual support, integration with Whisper for improved voice transcription and deployment on cloud platforms for wider use.

REFERENCES

[1] X. Wang et al., "Searching for best practices in retrieval-augmented generation," arXiv preprint arXiv:2407.01219, 2024.

[2] T. B. Brown et al., "Language models are few-shot learners," arXiv preprint arXiv:2005.14165, 2020.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," Communications of the ACM, vol. 64, no. 10, pp. 610–623, 2021.

[5] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," arXiv preprint arXiv:2109.07958, 2021.

[6] K. Zhou et al., "Knowledge augmented language models: A survey of methods and challenges," arXiv preprint arXiv:2202.05026, 2022.

[7] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," arXiv preprint arXiv:2005.11401, 2020.

[8] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-augmented language model pre-training," arXiv preprint arXiv:2002.08909, 2020.

[9] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," arXiv preprint arXiv:2004.04906, 2020.

[10] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.

[11] M. Wang et al., "Knowledge mechanisms in large language models: A survey and perspective," arXiv preprint arXiv:2407.15017, 2024.

[12] A. Vaswani et al., "Attention is all you need," arXiv preprint arXiv:1706.03762, 2017.

[13] N. Goyal et al., "Deep learning for natural language processing: Creating neural networks with Python," Apress, New York, 2018.

[14] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open-domain question answering," in Proc. EACL, pp. 874–880, 2021.

[15] M. S. Parvez et al., "Retrieval-augmented code generation and summarization," arXiv preprint arXiv:2108.11601, 2021.

[16] Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," arXiv preprint arXiv:2312.10997, 2024.

[17] H. Li et al., "A survey on retrieval-augmented text generation," arXiv preprint arXiv:2202.01110, 2022.

[18] D. Cai et al., "Recent advances in retrieval-augmented text generation," in Proc. SIGIR, pp. 3417–3419, 2022.

[19] P. Zhang et al., "Retrieve anything to augment large language models," arXiv preprint arXiv:2310.07554, 2023.

[20] K. Lee et al., "Latent retrieval for weakly supervised open-domain question answering," arXiv preprint arXiv:1906.00300, 2019.

[21] Anthropic, "Contextual retrieval," Available: https://www.anthropic.com/news/contextual-retrieval, accessed Jan. 18, 2025.

[22] Y. Jiang et al., "LongRAG: Enhancing retrieval-augmented generation with long-context LLMs," arXiv preprint arXiv:2406.15319, 2024.

[23] L. Zhao et al., "LongRAG: A dual-perspective retrieval-augmented generation paradigm for long-context question answering," in Proc. EMNLP, pp. 1570–1581, 2024.

[24] A. Bechard and C. Ayala, "Reducing hallucination in structured outputs via retrieval-augmented generation," arXiv preprint arXiv:2404.08189, 2024.

[25] G. Izacard et al., “Atlas: Few-shot learning with retrieval augmented language models,” arXiv preprint arXiv:2208.03299, 2022.
[26] A. Bansal and S. Suddala, “Enhancing generative AI capabilities through retrieval-augmented generation systems and LLMs,” Library Progress International, vol. 44, no. 3, pp. 17765–17775, 2024.
[27] S. Borgeaud et al., “Improving language models by retrieving from trillions of tokens,” arXiv preprint arXiv:2112.04426, 2022.
[28] R. Jozefowicz et al., “Exploring the limits of language modeling,” arXiv preprint arXiv:1602.02410, 2016.
[29] A. Vaswani et al., “Attention is all you need,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 5998–6008, 2017.
[30] T. B. Brown et al., “Language models are few-shot learners,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.