ADScAI
Applied Data Science & Artificial Intelligence Symposium

Self-Optimizing RAG System With SLM for Domain Specific Learning

E.A Akindu Himan
Department of Computer Science and Engineering
University of Moratuwa
akindu.22@cse.mrt.ac.lk

***Keywords*—Retrieval-Augmented Generation, Reinforcement Learning, AI in Education, Vector Databases, Domain-Specific Learning, Small Language Models**

I. Introduction

This research presents a Self-Optimizing Retrieval-Augmented Generation (RAG) system integrated with Small Language Models (SLMs) to support domain-specific learning. The system is designed to transform structured and unstructured content, such as domain-related documents, lecture recordings, and scanned notes, into a searchable, intelligent knowledge base. By leveraging lightweight and efficient SLMs, the solution offers cost-effective and scalable AI assistance tailored to specific subject areas.

The system enhances traditional RAG architectures by incorporating reinforcement learning from human feedback (RLHF) to enable self-optimization. User feedback dynamically improves both the retrieval quality from vector databases and the response generation from the language model. In addition to interactive question-answering, the system provides personalized learning paths, curated reference materials, and progress insights.

II. Literature Review

Modern learners face challenges managing information from different sources, such as lectures, slides, and reference materials. Some key issues are: [1]

1) It's hard to remember important concepts from long lectures or large documents.
2) There's no easy integration between spoken content and written materials, making it hard to fully understand.
3) Existing tools don't adapt to individual needs or offer relevant answers.
4) Students waste time searching for specific content or clarifying doubts.
5) Students must search for answers to each question separately.

III. Materials and Methods

A. Materials

The proposed Self-Optimizing Retrieval-Augmented Generation (RAG) System requires the following key components:

1) Data Sources:

- Live lecture videos (YouTube, Zoom, MS Teams)
- Lecture slides (PDFs, PowerPoint presentations)
- Academic papers and textbooks (Open-access research databases)
- Web documents (trusted educational sources)

B. Methods

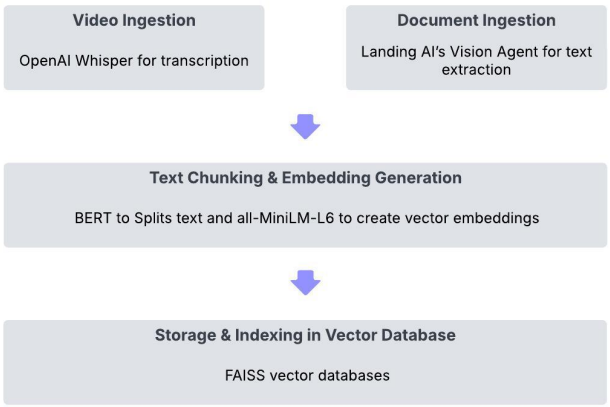

Fig. 1. Ingestion Algorithm

1) Preprocessing Tools:

- Speech-to-Text: OpenAI Whisper [2] for real-time transcription
- Optical Character Recognition (OCR): Landing AI's Vision Agent [3] for better extraction

2) Vector Databases:

- BERT – Splits text into meaningful sections and creates embeddings.
- FAISS and SBERT– Efficiently searches vectorized lecture and document embeddings.

3) Small Language Model (SLM):

- Open-source GPT-based model – Generates answers by retrieving relevant data from vector databases. (e.g., Mistral, LLAMA)

https://doi.org/10.31705/ADScAI.2025.25

- LangChain – Connects AI models with tools to enhances retrieval.

4) Reinforcement Learning from Human Feedback:

- Retrieves external content for knowledge expansion using web search.
- Proximal Policy Optimization (PPO) to improve accuracy based on user feedback.

C. Abstract System Diagram

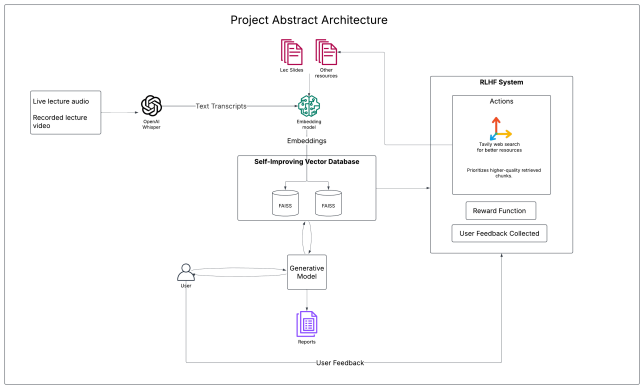

Fig. 2. Project Abstract Architecture

IV. Comparison with Existing Solutions

A. Google NotebookLM

Google NotebookLM [4] is a retrieval-augmented tool that allows users to upload documents and interact with them via LLM-powered summarization and question answering.

B. YouTube Note-Taking Applications

Several YouTube-based note-taking tools (e.g., Eightify, VidSummize) convert lecture videos into summaries or timestamped notes.

C. Performance and Feature Comparison

TABLE I
Feature Comparison of Existing Tools vs. Proposed System

Feature	NotebookLM	YouTube Apps	Proposed System
RAG-based Q&A	✓	✗	✓
Supports Multi-modal Input	Partial	Partial	✓
Feedback Optimization	✗	✗	✓
Personalized Study Plans	✓	✗	✓
Lightweight SLM Integration	✗	✗	✓
Long-term Adaptability	✗	✗	✓

V. Results

To evaluate the effectiveness of the self-optimization mechanism, we performed a comparative analysis of the system's answer quality before and after optimization. Two key metrics were used for this assessment:

- **BERT Score:** A semantic similarity metric evaluating the overlap between generated answers and ground-truth responses.
- **LLM-as-a-Judge Score:** A qualitative score given by a Large Language Model acting as an evaluator, rating response relevance and accuracy.

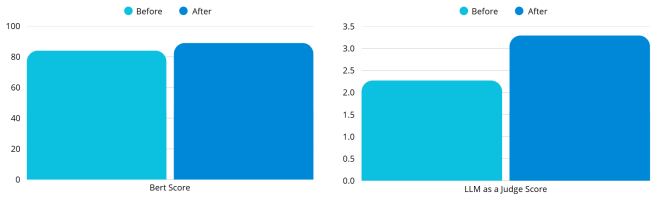

Fig. 3. Results

VI. Future Work

- **User Feedback:** Gather user feedback through quantitative evaluations and refine the performance of the system.
- **Adaptive Indexing:**Introduce an adaptive indexing method to improve efficiency.
- **Visual Understanding:** Integration with image captioning to understand visual stills from videos.

VII. Conclusion

This work presents a novel Self-Optimizing Retrieval-Augmented Generation (RAG) system integrated with Small Language Models (SLMs), specifically designed to support domain-specific learning. The proposed system addresses key limitations in current AI learning tools by introducing mechanisms for continuous improvement through Reinforcement Learning from Human Feedback (RLHF), enabling it to adapt to user behavior and evolving knowledge needs.

The self-optimization component allows the system to refine its retrieval accuracy, semantic indexing, and answer generation over time, enhancing the relevance and reliability of its outputs. Moreover, by leveraging lightweight, domain-adaptable language models, the system achieves a balance between performance and efficiency, making it suitable for deployment in resource-constrained environments.

References

[1] D. E. Whitehorse, "Three common student learning challenges with lectures," https://www.dremilywhitehorse.com/blog/three-common-student-learning-challenges-with-lectures, 2023, accessed: 2025-03-12.

[2] OpenAI, "Whisper: A robust speech recognition system," https://openai.com/index/whisper/, 2023, accessed: 2025-03-12.

[3] L. AI, "Visionagent: Ai for visual intelligence in business," https://landing.ai/visionagent, 2023, accessed: 2025-03-12.

[4] Google, "Notebooklm: An ai-powered research assistant," https://notebooklm.google/, 2023, accessed: 2025-03-12.