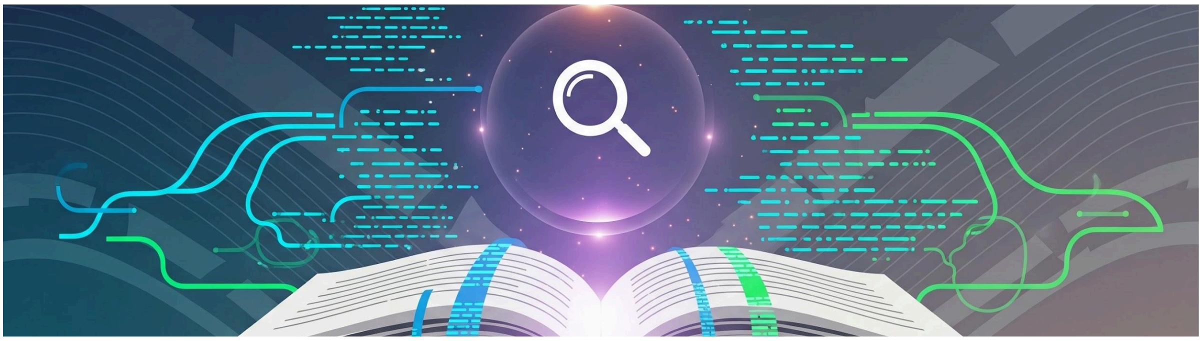

Curriculum-Based AI Tutor using Retrieval-Augmented Generation (RAG)

NCERT Class 8 Science

1. Abstract

This project presents a curriculum-based AI tutor designed to answer questions strictly from the NCERT Class 8 Science textbook. The system uses a Retrieval-Augmented Generation (RAG) approach, combining semantic search with a generative language model to ensure factual, syllabus-aligned responses. Textbook content is embedded using sentence-transformer models and indexed with FAISS for efficient retrieval. A local transformer-based language model generates answers based only on retrieved textbook passages, reducing hallucinations. The system also provides source citations for transparency and evaluates answer quality using BLEU and ROUGE-L metrics. A Streamlit-based chat interface demonstrates real-time interaction with the AI tutor.

2. Introduction

Generative AI systems often produce fluent but ungrounded answers, which can be problematic in educational settings. To address this issue, this project focuses on building an AI tutor constrained to a specific curriculum—NCERT Class 8 Science. By restricting the model's knowledge source to textbook content, the system ensures reliability, grade-appropriate language, and reduced misinformation. The tutor is intended as a learning aid rather than a replacement for teachers, helping students clarify textbook concepts interactively.

3. Dataset & Preprocessing

The dataset consists of the NCERT Class 8 Science textbook, covering all chapters in the syllabus. The textbook was extracted from PDF files, cleaned to remove non-textual artifacts, and stored in a structured JSONL format. Each chapter was split into overlapping text chunks to preserve context while enabling efficient retrieval. These chunks form the knowledge base used by the retrieval system.

4. Methodology

4.1 Retrieval-Augmented Generation (RAG)

The system follows a Retrieval-Augmented Generation architecture. When a user submits a question, it is first converted into a semantic embedding. This embedding is used to retrieve the most relevant textbook passages from a FAISS vector index. The retrieved passages are then provided as context to a language model, which generates an answer strictly based on this content.

4.2 Embeddings & Vector Search

Text chunks are embedded using the all-MiniLM-L6-v2 sentence-transformer model, which captures semantic meaning efficiently. FAISS is used as the vector database to enable fast nearest-neighbor search over the embeddings.

4.3 Answer Generation

To avoid dependency on paid APIs, a local transformer-based model (FLAN-T5) is used for answer generation. Prompt engineering ensures that the model answers only using retrieved textbook content and returns a fallback response for out-of-syllabus questions.

4.4 Transparency & Logging

Each response includes citations indicating the chapter and chunk from which the information was retrieved. All interactions are logged for future error analysis and evaluation.

5. Evaluation

The system was evaluated using 10 representative textbook questions. Generated answers were compared with reference answers using BLEU and ROUGE-L metrics. BLEU scores were generally low, which is expected for open-ended question-answering tasks due to paraphrasing. ROUGE-L scores provided a better indication of content overlap and showed reasonable alignment with textbook references. These results confirm that the system retrieves relevant content and generates grounded responses.

Metric	Average Score	Interpretation
BLEU	[A low average BLEU score]	Expected low score due to high paraphrase rate in answers.
ROUGE-L	[A moderate average ROUGE-L score]	Indicates reasonable overlap in content between generated and reference answers.

6. Results & Discussion

The AI tutor successfully answered curriculum-based questions with relevant textbook grounding and clear citations. Out-of-syllabus queries were handled gracefully using a predefined fallback response. While answer fluency is limited by the lightweight local model, retrieval accuracy and factual correctness were consistently maintained. The Streamlit-based interface enabled smooth real-time interaction.

7. Limitations & Future Work

Current limitations include verbose or loosely structured answers due to the use of a small local language model. Future improvements include upgrading to larger open-source models, applying fine-tuning using curated question–answer pairs, and enhancing retrieval granularity. Additional evaluation using human judgment and expansion to other NCERT subjects are also potential future directions.

8. Conclusion

This project demonstrates that a reliable and transparent AI tutor can be built using a Retrieval-Augmented Generation approach with open-source tools. By grounding responses in textbook content and providing citations, the system offers a safe and effective educational assistant aligned with curriculum requirements.