

Offline Retrieval-Augmented Academic Assistant using Phi-2 for Educational Question Answering

Sangini Shee

KLE Technological University, Hubballi, Karnataka, India

K. Sindhu

KLE Technological University, Hubballi, Karnataka, India

Kshitij Heblikar

KLE Technological University, Hubballi, Karnataka, India

Chirag S. Hanamaratti

KLE Technological University, Hubballi, Karnataka, India

Uday Kulkarni

KLE Technological University, Hubballi, Karnataka, India

Aishwarya Netagal

KLE Technological University, Hubballi, Karnataka, India

Abstract

The emergence of Retrieval-Augmented Generation (RAG) has greatly enhanced the robustness of Large Language Models (LLMs) by embedding the generated response in an external source of knowledge. Most educational assistants, however, leverage cloud-based LLMs, leading to reliance on the Internet, issues of privacy and high computational costs. In this paper, we present an Offline Retrieval-Augmented Academic Assistant that leverages academic textbooks for educational Q&A through MiniLM semantic embeddings, FAISS vector indexing, and Microsoft's Phi-2 Small Language Model. Academic PDF textbooks are preprocessed with PyMuPDF, segmented into overlapping chunks, semantically embedded, and indexed locally with FAISS. At inference time, relevant textbook chunks are retrieved and provided to Phi-2 with the help of a Retrieval-Augmented Generation pipeline for generating response. To ensure the feasibility of running the system on consumer-grade hardware, Phi-2 is optimized using 4-bit NF4 quantization, drastically reducing the memory consumption. Experimental results show efficient semantic retrieval, lower hallucination rate, better contextual accuracy, inference speed, and fully offline operation.

Keywords

Retrieval-Augmented Generation, Phi-2, FAISS, MiniLM Offline AI, Educational Chatbot, Semantic Retrieval, Quantization.