PROJECT ABSTRACT

Campus FAQ Chatbot using Retrieval-Augmented Generation

A RAG chatbot answering college queries from an official FAQ corpus: all-MiniLM-L6-v2 embeddings, FAISS top-3 retrieval and a grounded generator that cites sources and abstains when unsure. Delivered built-to-order with corpus builder, indexing pipeline, backend, evaluation harness, web demo, PPT and viva kit.

01. Title

Campus FAQ Chatbot using Retrieval-Augmented Generation.

02. Introduction / Background

Large language models hallucinate — confidently inventing fee figures and exam dates — which makes them unsafe for official campus information. Retrieval-Augmented Generation (RAG) fixes this architecturally: the model may only answer from retrieved passages of an approved corpus, and every answer cites its sources. When nothing relevant is retrieved, the system abstains instead of guessing. For a college, where fees, dates and rules change yearly, RAG has a decisive operational advantage: updating a text file updates the chatbot, with zero retraining.

03. Problem Statement

College helpdesks answer the same questions — hostel fees, exam dates, bonafide certificates, library timings — hundreds of times each admission season, and static FAQ pages go unread. A plain LLM chatbot risks inventing official figures, which is worse than no chatbot. This project builds a grounded campus assistant: questions are embedded, the top-3 FAQ passages are retrieved from the college's own documents, and the generator answers strictly from those passages with citations — evaluated for faithfulness, retrieval hit-rate and abstention behavior.

04. Objectives

- Build a corpus pipeline: PDF/notice ingestion, 512-token chunking with 64-token overlap.
- Index chunks with all-MiniLM-L6-v2 embeddings in FAISS (cosine, top-3 retrieval).
- Implement grounded generation: answer only from retrieved passages, with source citations.
- Implement abstention: low retrieval scores produce an explicit ‘I don't have verified info’ response.
- Evaluate faithfulness, hit@3 and abstention rate on a 50-question test set; ship the chat web demo.

05. Proposed Methodology

(1) Corpus builder: college PDFs and notices are parsed, cleaned and chunked; a starter FAQ template ships for colleges starting from zero. (2) Each chunk is embedded with all-MiniLM-L6-v2 (384-dim) and indexed in FAISS. (3) At query time the question is embedded, top-3 passages retrieved by cosine similarity, and a grounded prompt (passages + question + ‘answer only from passages’ instruction) goes to the generator LLM. (4) A similarity threshold gates abstention. (5) Evaluation: 50 in-corpus questions scored for faithfulness (claims $\subseteq$ passages), 20 out-of-corpus questions scored for correct abstention.

06. System Architecture / Working

The pipeline is corpus → embedder → FAISS index → generator → cited answer. The demo chat UI shows retrieved passages with similarity scores alongside the generated answer, making the grounding auditable — a student can click through to the exact notice section behind any claim. Because knowledge lives in documents rather than weights, the college office updates answers by replacing a PDF; the project documents this workflow as the primary maintenance story, which examiners consistently probe.

07. Hardware / Software Requirements

- Python 3.10+, sentence-transformers, FAISS, Flask
- Generator LLM: API-based or local instruction-tuned model (configurable)
- College FAQ documents (PDF/text) — buyer compiles; starter template ships
- Runs on CPU; 4 GB RAM minimum

08. Expected Outcomes

A working RAG chatbot over the buyer's corpus, measured faithfulness/hit@3/abstention on the shipped test protocol, the corpus-builder and reindexing workflow, and the chat web demo. The report is explicit that answer quality follows corpus coverage — no generic accuracy is claimed.

09. Applications

- College admission and student helpdesks
- Department FAQ assistants (exam cell, hostel office)
- Any organization with document-grounded Q&A; needs

10. Future Scope

- Multilingual answers (Hindi/regional languages)
- Voice input via speech-to-text
- Ticket escalation to human staff on abstention