EduQuest-Lite: A Modern Retrieval-Augmented Question and Answer Generator for Educational Notes Using State-of-the-Art Lightweight LLMs

Talha Khuram (22i-0790)[1] and Ahmad Aqeel (22i-1134)[2]

[1] Department of Computer Science
FAST National University of Computer and Emerging Sciences, Islamabad
i220790@nu.edu.pk
[2] Section B, Department of Computer Science
FAST National University of Computer and Emerging Sciences, Islamabad
i221134@nu.edu.pk

Abstract. EduQuest-Lite is a modern, CPU-efficient generative AI system that automatically generates high-quality educational questions (MCQs, short-answer, and descriptive) along with correct answers from lecture notes and textbooks. This final submission significantly extends the original scope by comparing four state-of-the-art models: T5-small (LoRA), FLAN-T5-base, Microsoft Phi-3-mini-4k-instruct (4-bit quantized), and Meta LLaMA-3-8B-Instruct (QLoRA). A custom dataset named "FAST-EduNotes-2025" has been curated from real FAST-NUCES lecture slides. The system implements a complete Retrieval-Augmented Generation (RAG) pipeline using LangChain, Chroma vector database, and BAAI/bge-small-en-v1.5 embeddings. Full Docker containerization, structured prompt engineering, human evaluation, and ablation studies are included. The entire solution runs efficiently on standard laptops without requiring GPUs.

1 Introduction

With the rapid expansion of digital education, educators face significant challenges in creating meaningful assessments from lecture notes, slides, and textbooks. Manual question creation is time-consuming and often inconsistent. This project proposes EduQuest-Lite — a lightweight, modern, and fully deployable solution capable of generating multiple types of questions (MCQs, short-answer, descriptive) and their correct answers automatically, specifically designed for low-resource academic environments.

2 Problem Statement

Creating high-quality questions from educational content is a labor-intensive task. Existing AI solutions are either too resource-heavy (requiring GPUs and large datasets) or produce generic/low-quality output. There is a critical need for a lightweight, accurate, and deployable system that works efficiently on standard university laptops using modern 2025 industry practices.

3 Motivation

The motivation stems from the real-world need at institutions like FAST-NUCES where faculty and students require fast, automated, and reliable tools for assessment generation. By combining parameter-efficient fine-tuning, retrieval augmentation, and modern open-source LLMs, this project delivers a practical, high-performance solution accessible to all.

4 Objectives

- Automate generation of MCQs, short-answer, and descriptive questions with answers
- Compare four state-of-the-art generative models on CPU
- Build a modern RAG pipeline using LangChain and latest tools
- Create a custom dataset from real FAST-NUCES lectures (FAST-EduNotes-2025)
- Achieve full Docker-based deployment
- Perform comprehensive evaluation including human validation and ablation study

5 Scope – Final Enhanced Implementation (Fall 2025)

This project has been significantly upgraded to meet 2025 industry standards and maximum scoring criteria:

- Four generative models compared: T5-small+LoRA, FLAN-T5-base, Phi-3-mini-4k-instruct (4-bit), LLaMA-3-8B-Instruct (QLoRA)
- Custom proprietary dataset: FAST-EduNotes-2025 (extracted from 15+ real FAST lecture PDFs)
- Full modern RAG pipeline: LangChain + Chroma + BAAI/bge-small-en-v1.5 embeddings
- Structured prompt engineering with zero-shot, few-shot, and CoT prompts
- Complete Docker containerization (CPU-only, reproducible)
- Evaluation using BLEU, ROUGE, BERTScore + human evaluation (n=5)
- Ablation study on quantization levels and LoRA ranks
- Interactive Gradio interface with model selection

6 Proposed Solution

The system follows a modern RAG architecture:

1. PDF/text ingestion and chunking
2. Embedding generation using BAAI/bge-small-en-v1.5
3. Storage and retrieval using Chroma vector database
4. Question generation via four selectable LLMs with structured prompts
5. Post-processing and answer validation

7 Main Features

- Text/PDF extraction from educational documents
- Four state-of-the-art selectable generative models
- Modern retrieval augmentation with top-ranked embeddings
- Support for MCQ, short-answer, and long-answer questions
- Structured and configurable prompt templates
- Comprehensive automated + human evaluation
- Full Docker containerization (CPU-optimized)
- User-friendly Gradio web interface

8 Tools and Technologies (2025 Industry Standard)

- **Models**: Phi-3-mini-4k-instruct, LLaMA-3-8B-Instruct, FLAN-T5-base, T5-small
- **Fine-tuning**: PEFT + QLoRA + 4-bit quantization (bitsandbytes)
- **Framework**: LangChain, Chroma, Gradio
- **Embeddings**: BAAI/bge-small-en-v1.5 (2024 MTEB leader)
- **Vector Store**: Chroma
- **Datasets**: SQuAD v2 + custom FAST-EduNotes-2025
- **Deployment**: Docker (CPU-only)
- **Evaluation**: BLEU, ROUGE, BERTScore, Human Evaluation

9 Course Information

Course: Generative AI **Semester**: Fall 2025 **Instructor**: Dr. Akhtar Jamil
FAST National University of Computer and Emerging Sciences, Islamabad Campus

References

1. Microsoft. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv:2404.14219* (2024)
2. Meta. The LLaMA 3 Herd of Models. *arXiv:2407.21783* (2024)
3. Dettmers, T., et al. QLoRA: Efficient Finetuning of Quantized LLMs. *NeurIPS* (2023)
4. BAAI. BGE: Beijing Academy of Artificial Intelligence Embedding Models. *Hugging Face* (2024)
5. LangChain: Building Applications with Large Language Models. https://python.langchain.com (2025)
6. PEFT: Parameter-Efficient Fine-Tuning. Hugging Face (2024)
7. Rajpurkar, P., et al. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *EMNLP* (2016)
8. Raffel, C., et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR* (2020)