South East Technological University
Carlow

Design and Evaluation of a Lightweight Hallucination-Aware Retrieval-Augmented Generation Framework

Ranganayakulu Chowdary Daggubati | Email: C00317521@setu.ie | Supervisor: Dr. Milu Philip
MSc in Data Science | Department of Computing, South East Technological University, Carlow

Introduction

- Large Language Models (LLMs) generate fluent but sometimes incorrect answers (hallucination).
- In domain-specific environments such as university websites, reliable and accurate responses are critical.
- Traditional chatbot systems lack adaptability, while standalone LLMs may generate misleading answers.
- Retrieval-Augmented Generation (RAG) combines retrieval with generation to improve response grounding by using external data sources.

Dataset

- The dataset consists of publicly available university website content.
- Collected from multiple university web pages (official sources).
- Includes admissions, courses, policies, and general information.
- Preprocessed into structured text format (cleaning, filtering, formatting).
- Split into chunks for efficient retrieval.

Methodology

- Collect and preprocess university website data
- Perform text chunking (300 / 600 tokens)
- Generate embeddings using **MiniLM / BGE**
- Store embeddings in FAISS / ChromaDB
- Retrieve top-k relevant chunks (k = 3 / 5 / 8)
- Generate responses using lightweight LLM (Gemma / Mistral)
- Apply confidence-based mechanism to reduce hallucination

Research Questions

- How effectively can a lightweight RAG system reduce hallucination in domain-specific question answering?
- How do retrieval parameters influence answer relevance and groundedness?

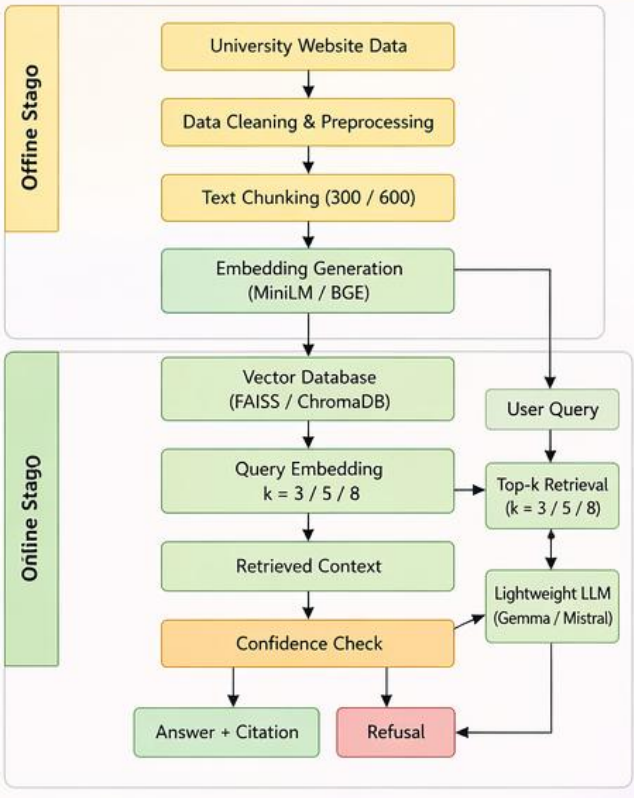
Literature Review

- RAG-based systems have been shown to improve factual accuracy and grounding in LLMs.
- Existing chatbot systems lack systematic evaluation of retrieval parameters and their impact on responses.
- Limited research focuses on hallucination measurement in lightweight (resource-efficient) models.
- Retrieval optimisation remains an underexplored area in domain-specific systems, especially for university or organisational data.

Next Steps

- Implement full RAG pipeline
- Conduct retrieval optimisation experiments
- Evaluate hallucination and groundedness
- Analyse trade-offs between accuracy and efficiency

Design Strategy / System Workflow



Technologies

Python | LangChain | FAISS | ChromaDB | MiniLM | BGE | Gemma | Mistral | FastAPI