FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

STUDENT
SCIENTIFIC CONFERENCE 2024

Retrieval-Augmented Generation (RAG) using Large Language Models (LLMs)

Adam Frémund[1], Martin Bulín[2]

1 Introduction

The convergence of Retrieval-augmented generation (RAG) methodologies with the robust computational prowess of Large Language Models (LLMs) heralds a new era in natural language processing, promising unprecedented levels of accuracy and contextual relevance in text generation tasks.

Pre-trained large language models, also referred to as foundation models, typically lack the ability to learn incrementally, may exhibit hallucinations, and can inadvertently expose private data from their training corpus. Addressing these shortcomings has sparked increasing interest in retrieval-augmented generation methods. RAG enhances the predictive capabilities of large language models by integrating an external datastore during inference. This approach enriches prompts with a blend of context, historical data, and pertinent knowledge, resulting in RAG LLMs.

Compared to LLMs operating without retrieval mechanisms, RAG LLMs demonstrate significantly superior prediction accuracy despite having fewer parameters. Furthermore, they possess the flexibility to update their knowledge base by substituting retrieval corpora, and they furnish users with citations for straightforward verification and evaluation of predictions.

Figure 1 illustrates the ecosystem of Retrieval-augmented generation (RAG), showcasing the interconnected blocks of External Data Stores, QA system, USER, and LLM. In this framework, External Data Stores serve as repositories of additional knowledge, enriching the prompts provided to Large Language Models (LLMs). The QA system facilitates the interaction between the user and the RAG system, allowing for queries and responses to be processed seamlessly. Through this collaborative setup, RAG optimally leverages external data to enhance the quality and relevance of generated outputs.

2 Motivation

Implementing a chatbot application tailored to interact with study regulations presents a compelling opportunity to revolutionize student support services at our university. By introducing this innovative tool, we aim to address the pressing need for a more accessible, efficient, and user-friendly platform for students to navigate the complexities of university regulations. Currently, students often face challenges in understanding and adhering to study regulations due to the cumbersome nature of traditional administrative processes. With the integration of a chatbot application, however, we envision a transformative solution that empowers students to effortlessly access crucial information about course requirements, credit distributions, academic deadlines, and more, all through intuitive natural language interactions. This initiative not only

[1] Ph.D. student of Cybernetics, focused on Artificial Intelligence, e-mail: afremund@kky.zcu.cz
[2] Ph.D. student of Cybernetics, focused on Artificial Intelligence, e-mail: bulinm@kky.zcu.cz

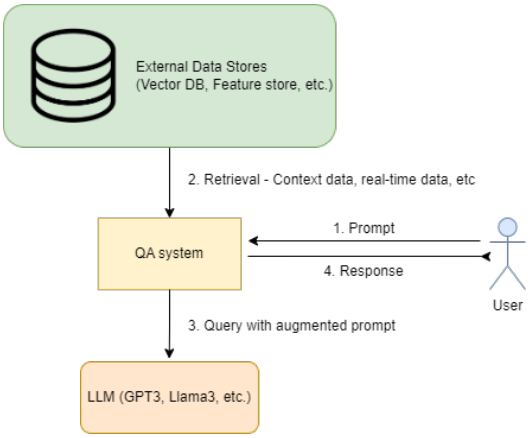

Figure 1: RAG architecture

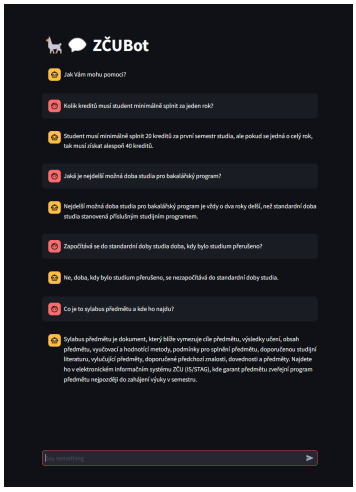

Figure 2: GUI for ZČUBot web application

promises to enhance the overall student experience but also demonstrates our commitment to leveraging cutting-edge technology to streamline administrative processes and foster a culture of transparency and accessibility within our academic community. By embarking on this journey to implement a chatbot for study regulations, we are poised to revolutionize student support services and propel our university into the forefront of innovation in higher education.

3 Implementation

Our Retrieval-augmented Generation (RAG) system, utilizing the Llama3 70b instruct model and the Langchain library, is designed to assist students at the University of West Bohemia with navigating study regulations. The system architecture includes data extraction from PDFs, integration with Llama3, and user interaction through a FastAPI-based API and a Streamlit GUI. We extract information from university documents. This data is structured and indexed for efficient retrieval.

The Llama3 70b instruct model is fine-tuned with domain-specific data to generate relevant responses. The Langchain library enhances the model by appending retrieved context from the PDFs to the input prompts, improving accuracy and relevance.

The FastAPI-based API handles user queries, invoking the retrieval mechanism and Llama3 model to generate responses. The Streamlit GUI provides an intuitive interface for students to interact with the chatbot, ensuring clear and accessible communication.

In summary, our RAG system integrates advanced language modeling with effective data retrieval to provide a robust tool for student support, enhancing accessibility and efficiency in navigating university regulations. This application is currently a proof of concept (POC).

Acknowledgement

This work was supported by the University of West Bohemia grant, project No. SGS-2022-017

References

Patrick Lewis and Ethan Perez et al. (2021) *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.