Lluminating Language

Summer Project 2024
Brain and Cognitive Society (BCS), IIT Kanpur

Abstract

ChatIITK, developed by BCS for IIT Kanpur's Summer Project - Illuminating Language, is an advanced Retrieval-Augmented Generation (RAG) chatbot. It combines information retrieval from IIT Kanpur websites with natural language generation, using ChromaDB for efficient data access. Supporting various models and offering terminal and Streamlit UI interfaces, ChatIITK enhances information access and user interaction for the IIT Kanpur community.

Objective

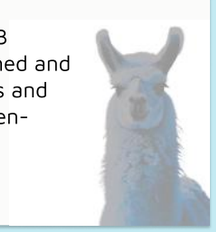
Our project aims to develop an advanced chatbot capable of implementing Retrieval-Augmented Generation (RAG) techniques to effectively utilize a diverse array of IITK data sources. By using web scraping tools like Selenium, we have collected extensive data from various college-related websites. The chatbot is built upon the powerful LLaMA base model, sourced from the Hugging Face library, to ensure high performance in natural language understanding and generation. Equipped with the conversational abilities of an LLM, enhancing it with comprehensive knowledge of IITK data, our chatbot will be able to provide accurate and contextually relevant responses, making it an invaluable tool for accessing information from multiple data repositories for the IITK community.

Implementation

LLMs

Large Language Models (LLMs), built on Transformer architectures, excel in natural language processing by pre-training on vast textual datasets to learn language patterns. During inference, LLMs utilize self-attention mechanisms to capture long-range dependencies, ensuring they generate contextually relevant and coherent responses. This capability enables LLMs to simulate human-like conversational flow, and hence integrating LLM's are a vital part of our project.

We selected LLaMA 3-instruct for its specialized training in multi-agent communication scenarios, as Llama 3 instruction-tuned models are fine-tuned and optimized for dialogue/chat use cases and outperform many of the available open-source chat models on common benchmarks.



Database

We utilized Selenium to web scrape data from IITK related websites, by automating the extraction of web page content.

Data we scrapped

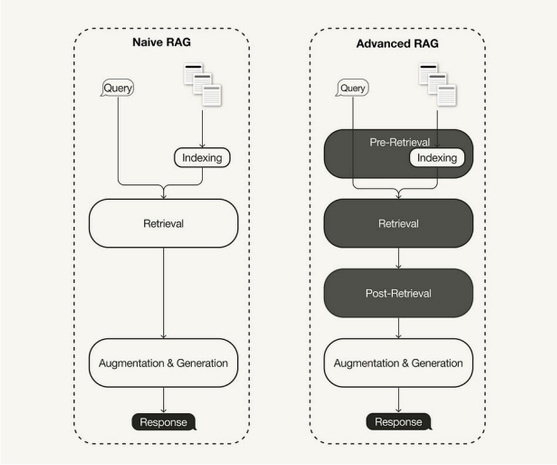
1. Alumni Data and IITK Startups Data
2. Relevant articles of Vox Populi
3. SPO, E Cell
4. Gymkhana website
5. The Constitution IITK
6. Brain & Cognitive Society (BCS), ICG, Electronics Club, Programming Club etc
7. PoR Handbook's data and UG manual

And more...

RAG

Retrieval-Augmented Generation (RAG) is simply retrieving relevant information from external sources to provide further context to our LLM.

RAG can be broken down to a few steps:



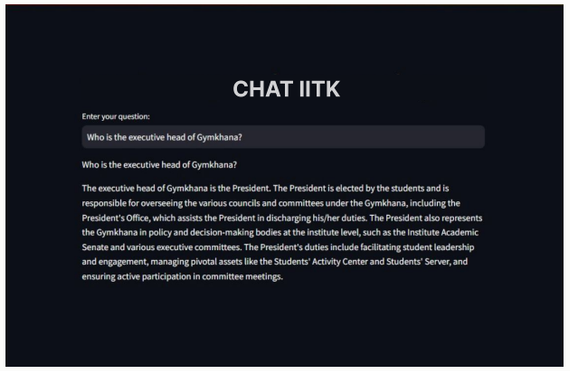
The LLM model generates a response by synthesizing information using context from both the retrieved documents and its pre-existing knowledge.

Naive RAG approach improves contextual relevance and factual correctness, although it struggles with retrieval quality.

Advanced RAG optimizes pre-retrieval, retrieval, and post-retrieval processes, enhancing query formulation, data modification, search ranking, and prompt compression for higher-quality responses.

Interface

Streamlit is used for the development of an interactive and user-friendly interface for our chatbot. This allows for easy deployment and interaction with the chatbot by the IITK community.



Conclusion

By integrating Advanced RAG techniques on the corpus of IITK data, our chatbot has access to up-to-date and comprehensive information. This AI tool will significantly enhance user experience by providing accurate, relevant responses to queries related to IITK resources, admissions, courses, and more. By doing so, it serves as an essential resource for faculty, students, and anyone seeking quick assistance without navigating multiple sources themselves.

Contributions

Mentees: Ameer Zaman, Aritra Ambudh Dutta, Abhinav Pandey, Siddhant Shekhar, Vaibhav Itauriya, Minshul Agrawal, Karan Kostha, Rounak Mishra, Harsh Sarawagi, Sumit Vishwkarma, Harsh Wardhan
Mentors: Udbhav Agarwal, Himanshu Shekhar, Arin Dhariwal, Shreya Gupta

Documentation

Git Hub Repository
https://github.com/udbhav-44/BCS-Lluminating-Language

Document
https://www.overleaf.com/project/668145c580154952b369fa52

Results