CSE 366 Course Project Report

EWU RAG Chatbot

Md. Iftekhar Hossain Khan (2020-3-60-073)
Md. Iftakher Alam (2020-2-60-003)
Hasnain Ahmed(2020-1-60-092)
Nourin Nahar Hridy (2021-1-60-102)

May 20, 2024

Abstract

This project presents the development and implementation of the EWU RAG Chatbot, a Retrieval-Augmented Generation (RAG) application designed to answer questions based on a custom dataset. The primary objective is to leverage advanced language models and embedding techniques to create an efficient and accurate chatbot. The dataset, originally in a text format, was preprocessed by segmenting it into manageable chunks. These chunks were then embedded using Ollama's nomic-embed-text to generate dense vector representations. For the question-answering tasks, we utilized the GROQ API LPU Inference Engine, integrating state-of-the-art language models such as Llama3 and Llama3 80B. This integration enables the chatbot to effectively comprehend and generate relevant responses based on the embedded dataset. Our implementation demonstrates the feasibility and effectiveness of combining RAG with advanced LLMs and embedding strategies to enhance information retrieval and user interaction in chatbot applications. The results indicate significant improvements in response accuracy and relevance, highlighting the potential of this approach for various practical applications.

Contents

1 Introduction

Since the first invention of chatbot in 1966 by psychotherapist Joseph Weizenbaum, the practice of developing innovative artificial intelligence has never stopped. The integration of AI in numerous sectors has notably transformed how information is dispersed and obtained. Chatbots are interactive tools to provide users with instant generic responses, allowing humans to communicate with digital devices. This project is about the creation of a chatbot, named EWU RAG Chatbot, to provide users info on East West University (EWU). Located in Dhaka, Bangladesh, East West University is a prominent institution that offers a wide range of academic programs and services.

The EWU RAG Chatbot intends to amplify user experience in gathering information about EWU by offering attainable, accurate, and timely responses about the university's admissions process, faculty info, undergraduate programs for different departments, faculty members, campus facilities, and other relevant aspects. The principal of this project is to create an efficient and easily-accessible chatbot that can assist prospective students, current students, faculties, and other possible users in obtaining their queries on EWU.

The EWU RAG Chatbot is able to comprehend and reply to a wide range of questions by utilising natural language processing and machine learning technology. This enhances the user experience for users who are looking for information about East West University. This paper describes the steps involved in developing the chatbot, including the design considerations, implementation specifics, and difficulties faced. It also talks about how the chatbot might affect user satisfaction and the university's communication plans.

1.1 Background Information

Chatbots are being used widely in numerous aspects; such as customer service, health care, education etc. As they are continuously being trained to be able to perform any queries from the users, these interactive tools are also teaching themselves by improving the efficiency of information dissemination.

World is now developing AI tools to make life easier. Data accessing and collecting have become easier due to the rapid development of AI. East West University (EWU) is one of the prominent private universities in Bangladesh. As this institute has diversed graduate and undergraduate programs, info is also diverse and numerous. This diverse info sometimes makes students and other potential info gatherers confused and lost. To save their time and make the access of info easier, the idea for the EWU RAG Chatbot was conceived.

The EWU RAG Chatbot is an innovative project intended to address the growing need for efficient and easily accessible information services within the university. It seeks to respond to a wide range of queries, from admissions and academic programs to faculty details and campus facilities, ensuring that users have a seamless and informative experience. With the addition of advanced AI technologies, the EWU RAG Chatbot aims to set a new standard for how educational institutions may utilise digital tools to enhance their services.

1.2 Literature Review

Although information from many sources were implemented to build this project, ideas from 12 videos from YouTube were basically utilized in this project. The 1st video focused on how to build a RAG application using open-source models, where it showed how the chatbot can be able to read PDFs using Llama2 and the 2nd one showed Llama's usage in building the RAG.

The 3rd and 4th video from the playlist helped us with learning how to apply tools to the chabot for it to read the PDFs using Llama and it used the concept of Embeddings in here.

The 5th video introduced us to the Langchain, which is a framework, specifically designed to streamline AI application development. By the 6th video, it introduced us with the 'Nomic-Embedded-Text'.

Concepts of Natural Language Processing and Machine Learning are taken from the 7th and 8th video. These videos explain how NLP enables chatbots to understand and respond to human language effectively, and how ML allows them to learn and improve over time.

From the 7th to 12th video helped us understand how Llama, Groq, and LangChain worked together in a chatbot, consequently assisting us with building one. [2]

1.3 Problem Statement

East West University (EWU) is a leading private institution offering a wide range of undergraduate and graduate programs. As the university continues to expand its academic offerings and student services, the volume of information that students, prospective students, faculty, and other possible bodies need to access has also grown significantly. This abundance of information often leads to confusion and difficulty in quickly finding accurate and relevant details.

Currently, the process of obtaining information about admissions, academic programs, faculty details, and campus facilities is time-consuming and inefficient. Students and other users must navigate through multiple webpages, contact various administrative offices, or wait for email responses to get the information they need. This not only affects user satisfaction but also places a significant burden on the university's administrative staff.

The lack of a centralized, easily accessible source for information poses several problems:

1. Inefficiency in Information Access: Users face challenges in quickly finding accurate and up-to-date information due to the scattered nature of resources and the need for direct human assistance.

2. User Frustration: The complexity and time required to obtain necessary information lead to frustration among students, prospective students, and other stakeholders, potentially affecting their perception of the university.

3. Administrative Burden: The university's administrative staff is overwhelmed with repetitive inquiries, which could be better managed through an automated system, allowing them to focus on more critical tasks.

4. Inconsistent Information: Multiple sources of information can lead to inconsistencies, causing confusion and potential misinformation among users.

To address these issues, there is a clear need for an efficient, user-friendly solution that can provide immediate information to all potential users. The EWU RAG Chatbot project aims to fulfill this need by developing an advanced chatbot that leverages natural language processing (NLP) and machine learning (ML) technologies. This chatbot will serve as a centralized information hub, capable of answering a wide range of queries related to the university,

1.4 Research Objectives

The primary goal of the EWU RAG Chatbot project is to develop an efficient and user-friendly chatbot that can serve as a comprehensive information resource for East West University (EWU). Goals are:

1. To Develop a Centralized Information Hub
2. To Enhance User Experience
3. To Reduce Administrative Workload
4. To Ensure Consistent and Reliable Information Dissemination
5. To Improve Access to Information
6. To Evaluate the Impact on User Satisfaction and Communication
7. To Provide Personalized Assistance

2 Methodology

2.1 Data Preparation

The dataset was prepared by collecting relevant information from the East West University Website. There were 6 sections in the website and each of them have several subsections. We decided to distribute these 6 sections among the members and started gathering information on which the RAG application will be built upon.

'About' , 'Authority' section was handled by **Hasnain Ahmed**. 'Admissions', 'News & Events' section was handled by **Md. Iftakher Alam**. 'Faculties' & 'Departments' section was handled by **Nourin Nahar Hridy**. From the 'Departments' Section, we decided to collect only the Science Faculty's five departments. Those are:

- Department of Computer Science and Engineering (CSE)
- Department of Electrical and Electronic Engineering (EEE)
- Department of Genetic Engineering and Biotechnology (GEB)
- Department of Pharmacy (Pharma.)
- Deparmento of Civil Engineering (CE)

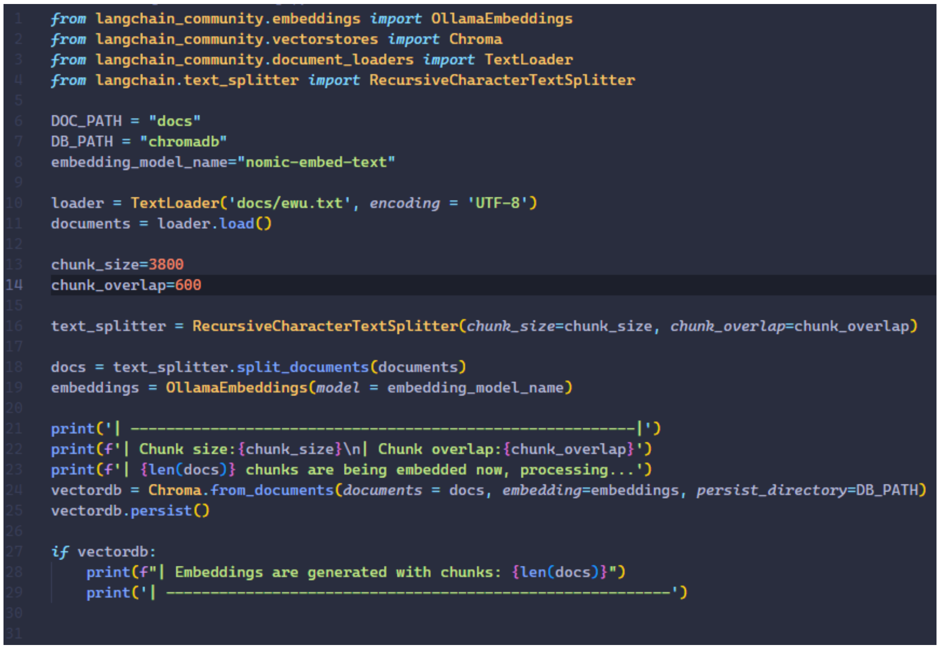
All these information gathered into a single text format document (txt) and fine tuned it sequentially in such a way the RAG application gets the relevant contexts when the questions are asked.

Afterwards we chunked these documents into smaller segments as it's an essential technique that helps optimize the relevance of the content we get back from a vector database once we use the LLM to embed content.

2.2 Embedding Process

'RecursiveCharacterTextSplitter'[5] is the library responsible for chunking the documents into smaller segments. It has two important parameters. One is 'chunk_size' and another is 'chunk_overlap'. Our document has total of 73 thousands characters. We defined the 'chunk_size' to '3800' and 'chunk_overlap' to '600'. These two parameters tries to split the document in order until the chunks are small enough. This has the effect of trying to keep all paragraphs (and then sentences, and then words) together as long as possible, as those would generically seem to be the strongest semantically related pieces of text.

This chunking process is the crucial step for our application as some LLM models has some limitations related to tokens. '**llama-3-8b**' LLM model has the context window size about 8 thousand tokens. So we defined the chunk_size to '3900' and chunk_overlap = '500' in the end. Ollama's 'nomic-embed-text'[4] Embeddings was used to embed the chunked documents and store it in a vector database called 'ChromaDB'[1]. This section was handled by **Md. Iftekhar Hossain Khan**. Here is the code snippet below:



```python
from langchain_community.embeddings import OllamaEmbeddings
from langchain_community.vectorstores import Chroma
from langchain_community.document_loaders import TextLoader
from langchain.text_splitter import RecursiveCharacterTextSplitter

DOC_PATH = "docs"
DB_PATH = "chromadb"
embedding_model_name="nomic-embed-text"

loader = TextLoader('docs/ewu.txt', encoding = 'UTF-8')
documents = loader.load()

chunk_size=3800
chunk_overlap=600

text_splitter = RecursiveCharacterTextSplitter(chunk_size=chunk_size, chunk_overlap=chunk_overlap)

docs = text_splitter.split_documents(documents)
embeddings = OllamaEmbeddings(model = embedding_model_name)

print('| ---------------------------------------------------------|')
print(f'| Chunk size:{chunk_size}\n| Chunk overlap:{chunk_overlap}')
print(f'| {len(docs)} chunks are being embedded now, processing...')
vectordb = Chroma.from_documents(documents = docs, embedding=embeddings, persist_directory=DB_PATH)
vectordb.persist()

if vectordb:
    print(f"| Embeddings are generated with chunks: {len(docs)}")
    print('| ----------------------------------------------------------')
```

Figure 1: Code snippet of Generating Embeddings

2.3 Model Integration

In the development of the EWU RAG Chatbot, the integration of advanced language models and retrieval mechanisms is a critical step. This section gives the details of the process of integrating the GROQ API LPU Inference Engine with high-performance language models like Llama3 and Llama3 80B, ensuring efficient and accurate responses from the chatbot.

What is GROQ API? Groq is an AI solutions company delivering ultra-low latency inference with the first ever LPU™ Inference Engine. Groq API enables developers to integrate state-of-the-art LLMs such as Llama-3, Mixtral and so on into low latency applications. [3]

2.4 Response Retrieval Process

We need to choose the specific LLM Model when submitting query and getting response. GROQ API has those LLM models on their interface. Our application has the ability to choose any of the LLM models GROQ API has provided on their interface. Which makes the model more unique and more capaple of getting variation of response from those LLM models. For our testing purpose, we used their '**llama-3-8b-8192**' and '**llama3-70b-8192**'. LLM models are listed below which GROQ API has provided:

1. **Google's**
   - gemma-7b-it
2. **Meta's**
   - llama3-70b-8192
   - llama3-8b-8192
3. **Mistral AI's**
   - mixtral-8x7b-32768

As the embeddings are already generated in the ChromaDB vectorstore, these vectorstore then work as a retriever for the LLMs to retrieve the response based on the query. Below is the process flowchart of the RAG application.

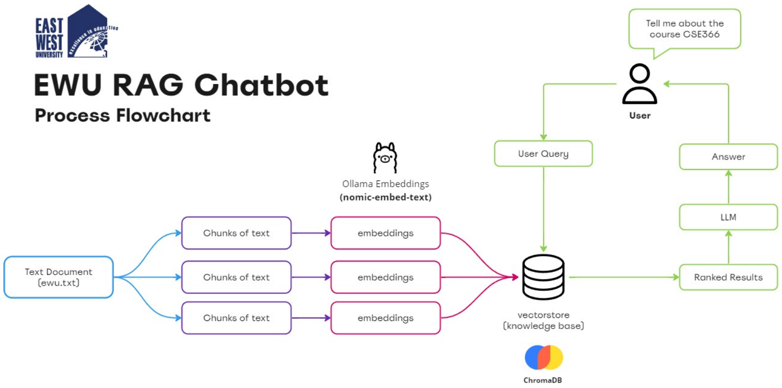


Figure 2: EWU RAG Chatbot process flowchart

2.5 Application's framework design

Our application is built with the framework name '**Streamlit**' [6]. With this framework, we managed to build our app in a very user friendly way which users may find it easily operable. Glimpse of our app can be viewed from our Github repository.

3 Results

3.1 Testing accuracy based on questions

Our application were tested on different queries based on the context of East West University. It responded most of the queries accurately but some problems were still observable when asking particular questions. Below is a list of queries which were asked and their accuracy of the responses.

Questions	Response Accuracy
how many faculties are there in ewu?	correct response
what is ewu?	correct response
what is the mission of ewu?	correct response
what is the vision of ewu?	correct response
where is ewu located?	correct response
what are the degrees do they offer in ewu for undergrads?	correct response
what are the degrees do they offer in ewu for grads?	**incorrect response**
Who is the Vice Chancellor of East West University?	correct response
Who is the founder of East West University?	correct response
List of the Syndicate Members	correct response
what is the class System of EWU	**incorrect response**
provide semester system of ewu	**incorrect response**
provide semester system of ewu	**incorrect response**

Table 1: Queries and Response Accuracy

Questions	Response Accuracy
What are the Admission Eligibilites for B.Pharm	correct response
What are the Admission Eligibilites except B.Pharm?	correct response
What are the Admission Eligibilites except B.Pharm?	correct response
Table of the tuition fee per credit of all programs of Undergraduates	correct response
give me the course summary of cse	**incorrect response**
how many majors are there in cse department for undergrads?	correct response
what are the course of data science major for undergrads?	**incorrect response**
what are the courses of software major for undergrads?	correct response
provide the chairperson names of all departments of science faculty	correct response
give me the list of faculty members of cse	correct response
is cse366 a core course of cse?	correct response
is cse303 a core course of cse?	correct response
what is the name of the course cse366?	correct response
what is the prerequisite for the course cse366?	correct response
what is the prerequisite for the course cse303?	correct response
what is the prerequisite for the course cse405?	correct response
what is the name of the course cse303?	correct response
what are the core course of cse for undergrads?	correct response
is there any course name cse487?	correct response

Table 2: Queries and Response Accuracy

3.2 Testing accuracy results on the LLMs

Questions	Response Accuracy
llama3-8b-8192	89%
llama3-70b-8192	92%

Table 3: LLM models with their response accuracy score

The response accuracy is calculated based on dividing the number of correct response the LLMs have provided by the number of questions we tested with.

$$Accuracy = No.of correctanswers/No.of totalanswers$$

Other LLM models like **mixtral-8x7b-32768** and **gemma-7b-it** were also tested but results were very poor. So we decided not to apply test queries any more on them.

4 Discussion

4.1 Implications of Findings

The landscape of chatbots in higher education will be influenced by this research. We demonstrated the possibility of using artificial intelligence (AI) to enhance user experience and information sharing in universities by developing the EWU RAG Chatbot. Other higher education establishments looking to widen access to information, increase user satisfaction, and speed up administrative processes can learn from the successful implementation of the chatbot.

Besides solving the specific challenges that East West University is dealing with, the EWU RAG Chatbot contributes to the ongoing discussion about the role of chatbots in higher education. The chatbot has set a new level of how academic entities should employ digital technologies to respond to the evolving requirements of their students, staff, and other users by providing instant, accurate, and personalized responses to user queries. The chatbot provides a sustainable solution for long-term information management because its scalability ensures that it can handle an increasing number of queries and adapt to any future changes.

4.2 Comparison With Prior Work

Our research's contrast to other research on the subject illustrates the important advances and developments made possible by the creation of the EWU RAG Chatbot. Few studies have explicitly examined private colleges in Bangladesh, even though previous study has examined the usage of chatbots in educational contexts. Our initiative closes this gap by offering a specially designed solution that takes into account the particular difficulties that East West University has, like the variety of academic offerings.

Furthermore, what distinguishes our chatbot from conventional rule-based systems is its utilization of modern AI technologies like machine learning and natural language processing. These technologies improve the chatbot's capacity to handle complex queries and improve user experience by enabling more natural and intuitive conversations. This contrast shows how crucial it is to apply modern methods when developing chatbots in order to maximize their efficacy and satisfy users.

4.3 Limitations

The EWU RAG Chatbot has some limitations even though it functions correctly. First of all, not all of the data from the EWU website could be included in the text file, which is how our bot absorbed the data that consumers would eventually see. Our chatbot will not be aware of any modifications to the information on the website and responses will become inaccurate or inconsistent as a result. Users' familiarity and willingness to interact with the chatbot affect its success.

Furthermore, non-English speakers or users with disabilities may find using this chatbot challenging. Subsequent research and development will be required to increase the chatbot's functionalities and suit the diverse needs of its users to overcome these limitations.

5 Conclusion

In a nutshell, the creation of the EWU RAG Chatbot is a significant step for the educational aspect as it accelerates distribution of information, enhances user experience by providing info, and optimizes administrative operations in a faster and better way. This project illustrates deeper implications for institutions as it shows how chatbots can make communication for higher education easier and straightforward, with least complications.

Further studies are needed on augmenting the chatbot's functionalities and limitations. This requires prospects for cooperation and fusion with different university systems and platforms. EWU RAG Chatbot can be made a valuable source for information by keeping potential users engaged with the tool to keep it updated and evolved.

6 References

References

[1] ChromDB Open Source embeddings database. https://www.trychroma.com/.

[2] Collection of Youtube videos on RAG. https://www.youtube.com/playlist?list=PLtJqR8jwVvHwT_NAGwBE7R5Y1YtppV8X2/.

[3] GROQ API. https://docs.api.groq.com/index.html.

[4] Ollama's nomic-embed-text. https://ollama.com/library/nomic-embed-text.

[5] RecursiveCharacterTextSplitter. https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/#:~:text=This%20text%20splitter%20is%20the,%22%20%22%2C%20%22%22%5D%20.

[6] Streamlit. https://streamlit.io/.

7 Github Link

Here is the Github repository link of the RAG application. There is detailed description given on how to run the application and what are the requirement to run the app. https://github.com/ihkcreations/EWU_RAG_Chatbot