



Project Report

Project Title: USTC AI Assistance

Course Title: Artificial Intelligence & Expert System Lab

Course Code: CSE-324

Submitted To:

Debabrata Mallick

Lecturer, CSE, FSET, USTC.

Submitted By:

Team: SHUNNO

Team Members

Name: Md. Absar

Id: 21070104

Name: Shanto Chandra Das

Id: 221010133

Name: Md. Arafat-ul Alam

Id: 221010132

Name: Golam Ali

Id: 221010133



Report: USTC AI Assistance

University of Science and Technology Chittagong (USTC)

Abstract

This report presents the development of an AI-powered chatbot for **USTC** that automates responses to student queries related to admissions, courses, faculty, and academic life. The chatbot integrates **LangChain** for document processing, **Pinecone** for semantic vector-based search, and **NVIDIA models** for natural language processing tasks. The system's architecture utilizes **meta/llama3-70b-instruct** for generating contextual responses and **nvidia/llama-3.2-nv-embedqa-1b-v1** for embedding queries and retrieving relevant information. The UI is developed using **Gradio** and **Flask**, ensuring a seamless and user-friendly experience. The chatbot offers an optimal solution for improving student engagement and administrative efficiency. **Future work** includes expanding multilingual support, context-aware memory, and mobile platform integration.

Keywords

AI Chatbot, USTC, Pinecone, LangChain, meta/llama3-70b-instruct, nvidia/llama-3.2-nv-embedqa-1b-v1, Semantic Search, Gradio, Flask

I. Introduction

Universities, especially large ones like **USTC**, are faced with the challenge of managing a large volume of student queries. The manual handling of these queries is inefficient and resource-draining. To

overcome this, we have developed an **AI-powered chatbot** designed to respond to frequently asked questions (FAQs) about admissions, courses, faculty, and university resources.

The **USTC AI Chatbot** uses **advanced NLP models** like **meta/llama3-70b-instruct** and **nvidia/llama-3.2-nv-embedqa-1b-v1** to generate accurate responses and retrieve relevant information from a large dataset stored in **Pinecone**. **LangChain** is used to combine documents retrieved from Pinecone to provide more comprehensive responses. The user interface is built with **Gradio** for a responsive web-based experience.

II. Methodology

A. Data Collection and Preprocessing

The data for the chatbot was collected from **USTC's official website** [9], which includes **FAQs**, **course catalogs**, **faculty bios**, and other administrative details. This data was cleaned, tokenized, and preprocessed using **NLTK** [3] for basic text processing and **Hugging Face Transformers** [2] for more advanced NLP tasks such as entity recognition.

B. Model Integration

Two pre-trained models were integrated into the chatbot:

1. **meta/llama3-70b-instruct** [8]: Used for generating contextually relevant responses based on the input query.
2. **nvidia/llama-3.2-nv-embedqa-1b-v1** [7]: Used to generate query embeddings and perform semantic search by retrieving related documents from the **Pinecone** vector database [6].

LangChain components are utilized for chaining and combining retrieved documents, allowing the chatbot to provide more coherent and detailed answers.

C. User Interface Development

The user interface was developed using **Gradio** [5], a Python library for creating interactive UIs. **Flask** [4] was used to handle the backend, process the user queries, and fetch responses from the AI models.

D. LangChain Integration for Document Handling

LangChain provides seamless integration of **Pinecone** for storing and retrieving semantically similar queries. The key functions used include:

- **create_retrieval_chain:** Retrieves documents from **Pinecone** based on the semantic similarity of the query.
- **create_stuff_documents_chain:** Combines retrieved documents into a coherent response tailored to the user's query.

III. System Architecture and Implementation

A. System Flowchart

The system operates through a series of steps, from user query input to final response generation:

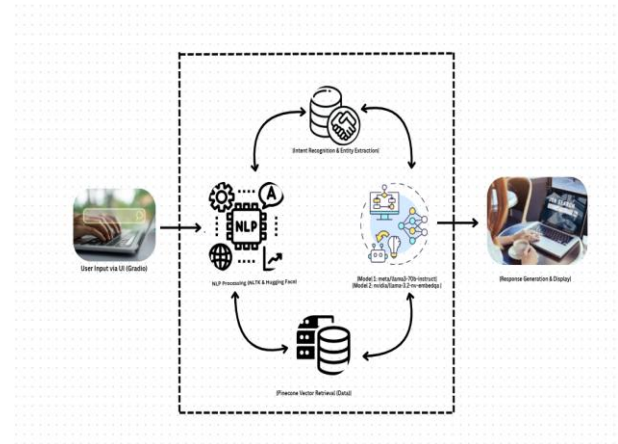


Fig 01:- Architecture

B. LangChain Components

The system uses the following LangChain components:

1. **ChatNVIDIA:** Used to handle conversational queries and generate responses using the **meta/llama3-70b-instruct** model.
2. **Embedding:** Uses **nvidia/llama-3.2-nv-embedqa-1b-v1** to generate embeddings of queries and store them in **Pinecone** for efficient search.
3. **create_retrieval_chain:** This LangChain function retrieves relevant documents from the **Pinecone** database based on query similarity.
4. **create_stuff_documents_chain:** Combines multiple documents retrieved to form a coherent and contextually relevant response.

IV. Dataset and Database Overview

A. Dataset Details

The chatbot uses a diverse dataset sourced from **USTC's website** [9] and includes FAQs, course catalogs, faculty bios, and fee structures.

Dataset Type	Data Source	Example Data
FAQ	University FAQs	Admission requirements, exam dates
Course Catalogs	CSE Department	Course descriptions, prerequisites
Faculty Bios	Faculty Database	Faculty research interests, contact info
Fee Structure	Finance Office	Tuition fees, scholarship availability

B. Pinecone Database Structure

The **Pinecone** vector database stores embeddings of queries for **semantic retrieval**. Each query is represented as a high-dimensional vector for efficient matching.

Field	Description
Query ID	Unique identifier for each query
Embedding Vector	High-dimensional vector representing query meaning
Metadata	Additional data (e.g., source, timestamp)

V. Results and Output


A. Output Example

Here is an example of how the chatbot responds to a student query:


Input Query:


What are the admission requirements for the CSE department?

Output Response:

 **USTC AI-Powered Chatbot**

Chat

What are the admission requirements for the CSE department? 

 To be eligible for admission into the 1st year 1st semester of the CSE department, you must have secured at least GPA-2.5 separately in both the S.S.C. and H.S.C. examinations or GPA-2.0 in either S.S.C or H.S.C and GPA-6.0 in aggregate or GPA-2.0 in either S.S.C or H.S.C and GPA-5.0 in aggregate (if either or his/her parents are a freedom fighter).

Ask something...

Send ▶

B. System Output Screenshot

Below is a screenshot of the chatbot interface showing the user query and response:

Explanation:

- **User Query:** "What are the admission requirements for the CSE department?"
- **Response Generation:** The chatbot uses **meta/llama3-70b-instruct** to generate a contextual answer and retrieves related documents from **Pinecone** using the **nvidia/llama-3.2-nv-embedqa-1b-v1** model.

VI. Conclusion

The **USTC AI Chatbot** integrates advanced **NLP models**, **semantic vector search**, and modern **document-handling frameworks** to provide an efficient, scalable solution for responding to student queries. The system leverages **LangChain**, **Pinecone**, and **NVIDIA models** to generate accurate, context-aware responses. By automating the query-answering process, the chatbot reduces administrative workload and enhances student engagement. This solution is optimal for the current needs of **USTC** and provides a robust foundation for future enhancements.

VII. Future Work

1. **Multilingual Support:** Expanding language capabilities to support multiple languages, enabling access to a broader audience.
2. **Context-Aware Memory:** Implementing memory features to retain context over multiple interactions, improving the chatbot's ability to provide personalized responses.
3. **Mobile Integration:** Developing mobile applications to provide

students with access to the chatbot on-the-go.

4. **Adaptive Learning:** Using student feedback and interactions to adapt the system's responses, improving accuracy over time.

VIII. References

- [1] RasaHQ, "Rasa Open-Source Chatbot Framework," GitHub Repository, [Online]. Available: <https://github.com/RasaHQ/rasa>. [Accessed: 08-Nov-2024].
- [2] Hugging Face, "Hugging Face Transformers Documentation," [Online]. Available: <https://huggingface.co/transformers/>. [Accessed: 08-Nov-2024].
- [3] NLTK, "NLTK Documentation," [Online]. Available: <https://www.nltk.org/>. [Accessed: 08-Nov-2024].
- [4] Flask, "Flask Documentation," [Online]. Available: <https://flask.palletsprojects.com/>. [Accessed: 08-Nov-2024].
- [5] Gradio, "Gradio Documentation," [Online]. Available: <https://gradio.app/>. [Accessed: 08-Nov-2024].
- [6] Pinecone, "Pinecone Vector Database," [Online]. Available: <https://www.pinecone.io/>. [Accessed: 08-Nov-2024].
- [7] NVIDIA AI, "NVIDIA LLaMA 3.2 NV EmbedQA Model Documentation," [Online]. Available: <https://catalog.ngc.nvidia.com/orgs/nvidia/models/llama-3.2-nv-embedqa-1b-v1>. [Accessed: 08-Nov-2024].
- [8] Meta AI, "LLaMA 3 Meta AI Model Documentation," [Online]. Available: <https://ai.meta.com/research/publications/llama>. [Accessed: 08-Nov-2024].

[9] USTC, "University of Science and Technology Chittagong (USTC) Official Website," [Online]. Available: <https://www.ustc.ac.bd/>. [Accessed: 08-Nov-2024].

[10] LangChain, "LangChain Chains and Document Combination Documentation," [Online]. Available: <https://python.langchain.com/en/latest/modules/chains.html>. [Accessed: 08-Nov-2024].

[11] IEEE Xplore, "The Role of Chatbots in Academic Institutions," [Online]. Available: <https://ieeexplore.ieee.org/>. [Accessed: 08-Nov-2024].

[12] USTC RAG-2048 "USTC AI Assistance ChatBot," GitHub Repository, [Online]. Available: https://github.com/absar-mahmud13/USTC-RAG-2048-_AI_Assistance.git
