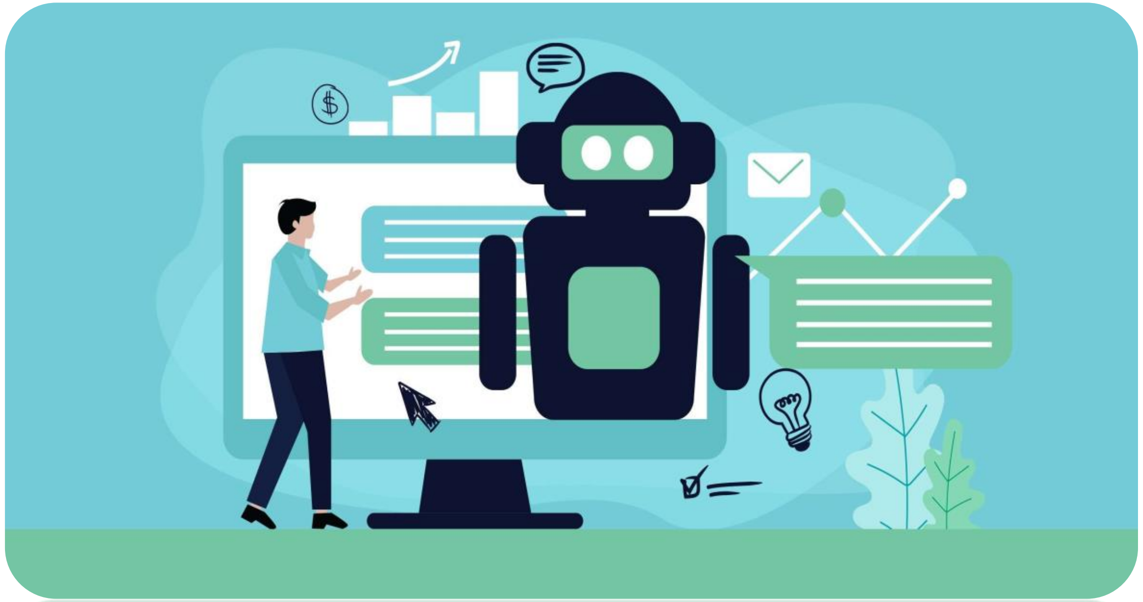
CHATBOT WITH WEB EXTRACTION



Project Leadership

Sharat Manikonda
Director at Innodatatics and Sponsor
linkedin.com/in/sharat-chandra

Team Members

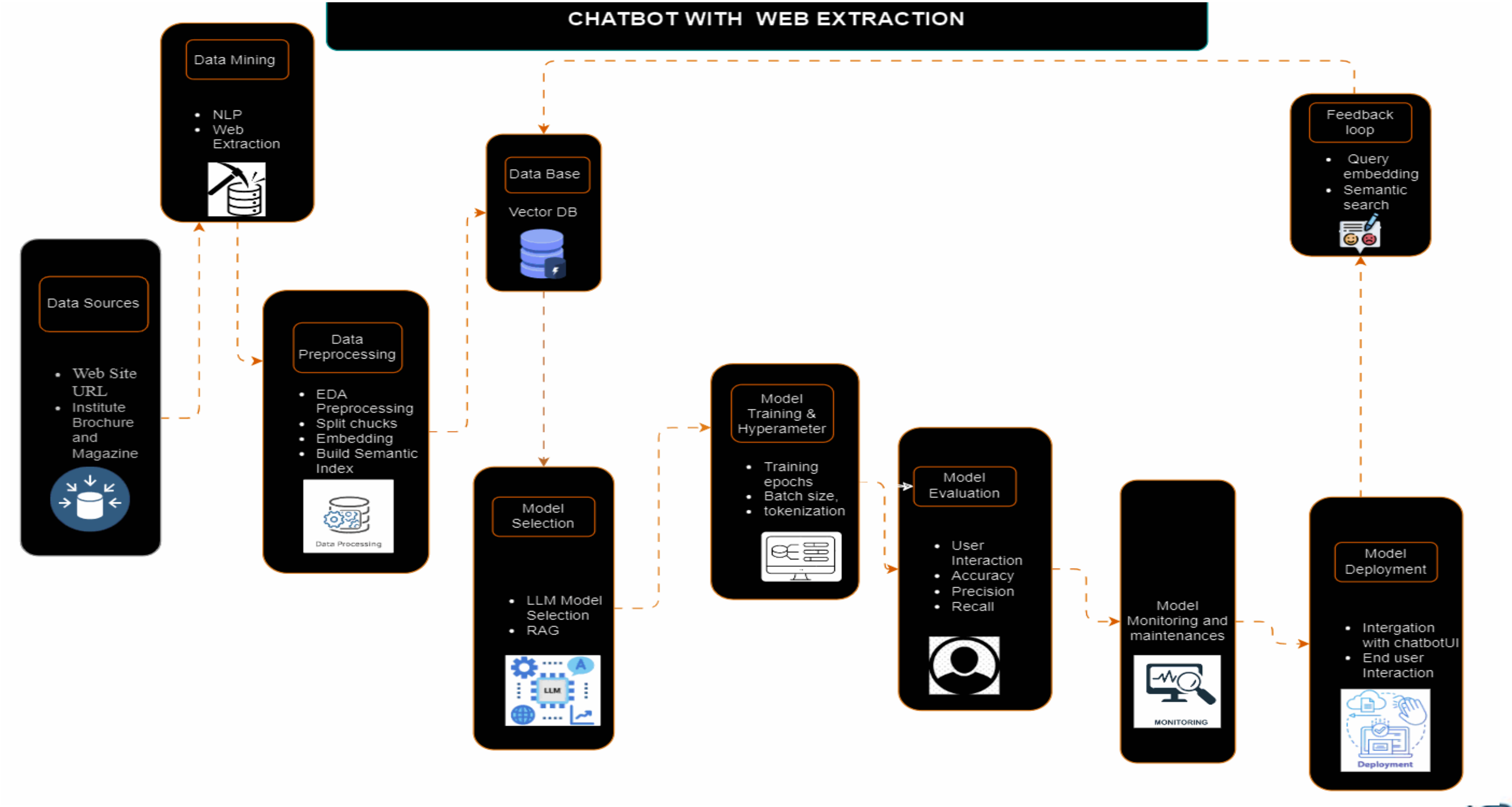
Name: Jeevan M

www.linkedin.com/in/m-j-ab2472196

Contents

- Business Objective
- Business Constraints
- Project Architecture
- Data collection and details
- Exploratory Data Analysis
- LLM model used
- Modeling
- Evaluation
- Deployment

Project Overview and Scope



CHATBOT WITH WEB EXTRACTION

Data Sources
- Web Site URL
- Institute Brochure and Magazine

Data Mining
- NLP
- Web Extraction

Data Preprocessing
- EDA Preprocessing
- Split chucks
- Embedding
- Build Semantic Index

Data Base

Vector DB

Model Selection
- LLM Model Selection
- RAG

Model Training & Hyperameter
- Training epochs
- Batch size,
- tokenization

Model Evaluation
- User Interaction
- Accuracy
- Precision
- Recall

Model Monitoring and maintenances

Model Deployment
- Intergation with chatbotUI
- End user Interaction

Feedback loop
- Query embedding
- Semantic search

INNODATATICS
Innovation • Data • Analytics

Business Problem

- **Business Problem: Student has to communicate with the salesperson over the call or any manager has to go to the website of JNTUH to search for any details regarding JNTUH.**

Business Objective

Objective

Business Objectives : Minimize the manual work

Constraints

- ***Maximize the data privacy***
- ***Business Success Criteria : Maximize the data privacy***
- ***Machine Learning success criteria: Accuracy of the output Should be 98%***
- ***Economic Success Criteria : Communicating with our application should increase 20% revenue***

Innovation • Data • Analytics

CRISP-ML(Q) Methodology

There are six stages of CRISP-ML(Q) Methodology

1. **Business and data understanding**
2. **Data preparation**
3. **Model building**
4. **Model evaluation**
5. **Model deployment**
6. **Monitoring and maintenance**

Technical Stacks

- **Web Extraction:** Develop reliable web scraping or extraction mechanisms to pull relevant data from the JNTUH website.
- **Machine Learning**: Employ ML algorithms to achieve a 98% accuracy rate in providing relevant information and responses.
- **Language Models (LLM):** Implement different language models, such as LLama2, for natural language processing and understanding.
- **Split Chunking:** Utilize techniques for splitting and chunking text data to efficiently process and analyze information
- **Vector-based Representations:** Leverage vector embeddings to encode and represent textual information in a numerical format for machine learning applications.
- **Embedding:** Implement embedding techniques to map words or phrases into numerical vectors, enhancing semantic understanding.
- **Semantic Streamlit:** Build a user-friendly interface using Streamlit that focuses on semantic and intuitive design for effective communication..
- **User Engagement Metrics:** Incorporate analytics tools to measure and track user engagement, aiming for an 80% increase in communication.
- **Revenue Tracking:** Set up systems to monitor and analyze revenue growth, attributing it to improved communication with the application.

Data Collection and Understanding

Data collection:

Web Scraping : Utilize web scraping tools such as Beautiful Soup or Selenium to extract information from the JNTUH website.

Downloading : Magazine and Brochures.

Data Understanding:

Structured and Unstructured Data:

- Identify and collect both structured data (e.g., tables, forms) and unstructured data (e.g., text from paragraphs, announcements).
- This diversity helps in training a more versatile language model.

Data Information

Student inquiries and requests related to JNTUH details, courses, and procedures.

- **Affiliations & Academic Audit**
- Vice Chancellor
- Directors
- Mandate of the cell
- Deputy director
- Assistant director
- Staff members
- Vision & mission
- Objectives
- Goals
- Timelines of affiliation activities
- Procedure for grant of affiliation
- News letters
- Status of affiliated colleges
- Affiliated colleges
- Download
- Contact us
- Student corner

- **UCEST-Facilities**
- Library
- Built up Area Entrepreneurship Development Cell(EDC)
- Physical Education
- National Service Scheme (NSS)

System Requirements

- Windows/mac/linux- 16 GB of VRAM. Graphics : NVIDIA RTX3090/4090 GPUs works.
- Google Colab –T4 GPU

Exploratory Data Analysis [EDA]

1. Data Collection:

- Gather relevant data from user interactions, website extractions, and sources.

2. Data Cleaning:

- Identify and handle missing values and inconsistencies for robust analysis.

Missing Values Observation

During web scraping extracted file text files, CSV files, and Pdf files, files or empty.

Data miss due to information sources is unstructured or structured data

Therefore extract data concerning the format.

- Tabular data in CSV format

- Text data in Txt format

- Image and image text data in PDF format

Data Preprocessing

Data Collection:

1. Gather user queries and interactions.
2. Extract information from the JNTUH website using web scraping tools.

1.Data Cleaning:

1. Handle missing values in user queries or extracted data.
2. Remove duplicates to ensure data integrity.
3. Correct inconsistencies in the data obtained from web extraction.

2.Text Cleaning:

1. Remove irrelevant characters, HTML tags, or special symbols.
2. Tokenize and normalize text to a consistent format.

LLM Models

Large language models (LLMs) are a type of artificial intelligence (AI) program that can recognize and generate text. They're trained on large amounts of data and use transformer models to recognize, translate, predict, or generate text or other content.

- zephyr-7b-alpha
- Mistral-7B-Instruct1
- zephyr-7b-beta
- Llama2

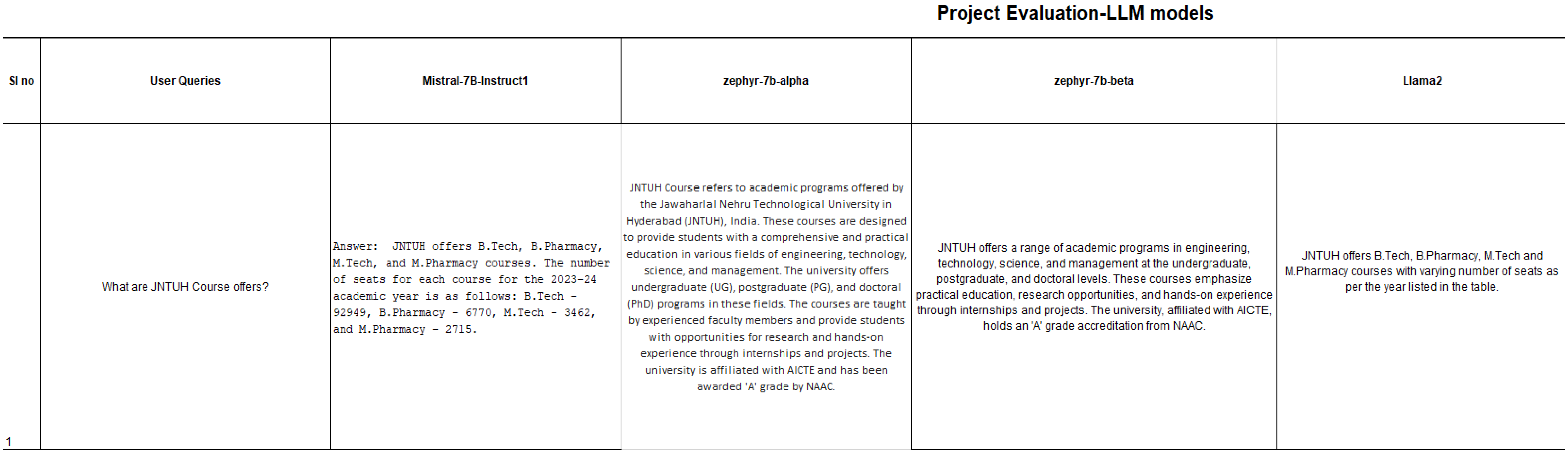
Model Building

1. Loads documents, splits text into chunks, creates embeddings, vector store, and an LLMS (Language Model) model for conversational retrieval.
2. Set up a conversation history, memory, and a retrieval chain for conversational responses.

Packages used in - “zephyr-7b-alpha” code

- langchain
- torch
- accelerate
- sentence_transformers
- streamlitstreamlit_chat
- faiss-cpu
- tiktoken
- huggingface-hub
- Pypdf
- ctransformers

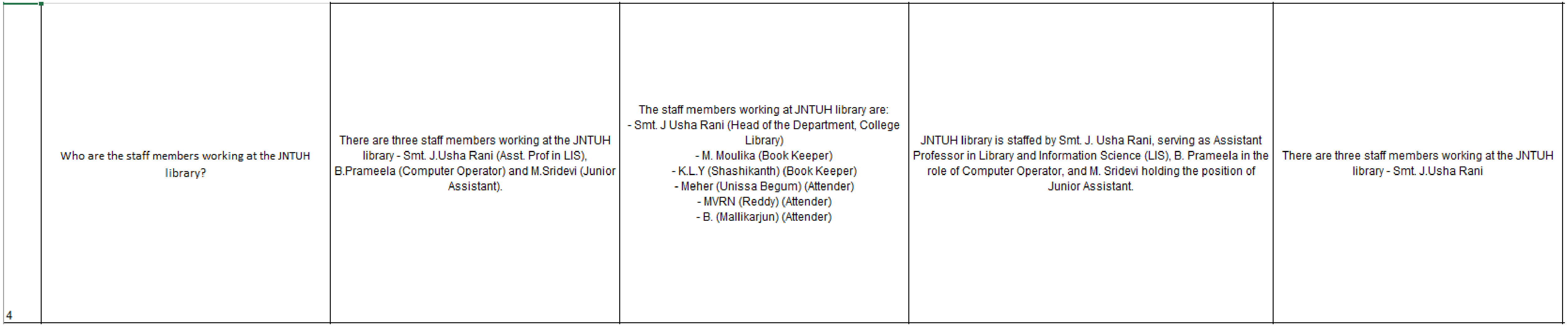
Model Accuracy Comparison



Project Evaluation-LLM models

Sl no	User Queries	Mistral-7B-Instruct1	zephyr-7b-alpha	zephyr-7b-beta	Llama2
1	What are JNTUH Course offers?	Answer: JNTUH offers B.Tech, B.Pharmacy, M.Tech, and M.Pharmacy courses. The number of seats for each course for the 2023-24 academic year is as follows: B.Tech - 92949, B.Pharmacy - 6770, M.Tech - 3462, and M.Pharmacy - 2715.	JNTUH Course refers to academic programs offered by the Jawaharlal Nehru Technological University in Hyderabad (JNTUH), India. These courses are designed to provide students with a comprehensive and practical education in various fields of engineering, technology, science, and management. The university offers undergraduate (UG), postgraduate (PG), and doctoral (PhD) programs in these fields. The courses are taught by experienced faculty members and provide students with opportunities for research and hands-on experience through internships and projects. The university is affiliated with AICTE and has been awarded 'A' grade by NAAC.	JNTUH offers a range of academic programs in engineering, technology, science, and management at the undergraduate, postgraduate, and doctoral levels. These courses emphasize practical education, research opportunities, and hands-on experience through internships and projects. The university, affiliated with AICTE, holds an 'A' grade accreditation from NAAC.	JNTUH offers B.Tech, B.Pharmacy, M.Tech and M.Pharmacy courses with varying number of seats as per the year listed in the table.

Model Accuracy Comparison



4	Who are the staff members working at the JNTUH library?	There are three staff members working at the JNTUH library - Smt. J.Usha Rani (Asst. Prof in LIS), B.Prameela (Computer Operator) and M.Sridevi (Junior Assistant).	The staff members working at JNTUH library are: - Smt. J Usha Rani (Head of the Department, College Library) - M. Moulika (Book Keeper) - K.L.Y (Shashikanth) (Book Keeper) - Meher (Unissa Begum) (Attender) - MVRN (Reddy) (Attender) - B. (Mallikarjun) (Attender)	JNTUH library is staffed by Smt. J. Usha Rani, serving as Assistant Professor in Library and Information Science (LIS), B. Prameela in the role of Computer Operator, and M. Sridevi holding the position of Junior Assistant.	There are three staff members working at the JNTUH library - Smt. J.Usha Rani

Best Model – Zephyr-7b-alpha

- Zephyr-7b-alpha is a specific language model used in the code (CTrasformers model).
- It's considered the best model in this context based on its performance in generating responses for the Enquiry ChatBot.
- The choice of model might be based on factors like training data, architecture, and fine-tuning for specific tasks.

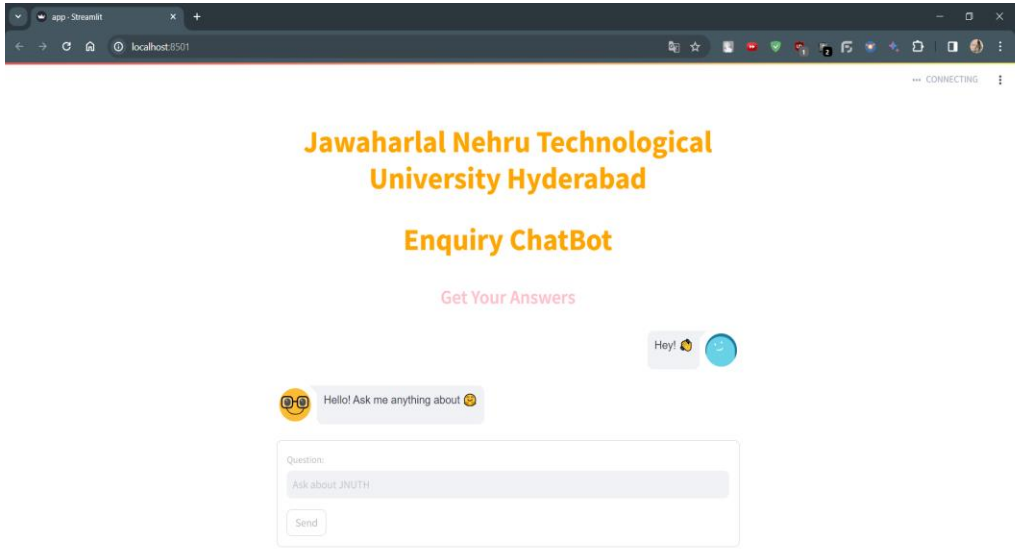
Model Deployment - Strategy

- The deployed model should be integrated into a web interface using Streamlit or similar frameworks.

- It requires hosting the model and associated resources on a server or cloud platform.

- Deployment should include monitoring for performance, scalability, and user feedback integration.

Screen shot of output



Jawaharlal Nehru Technological University Hyderabad

Enquiry ChatBot

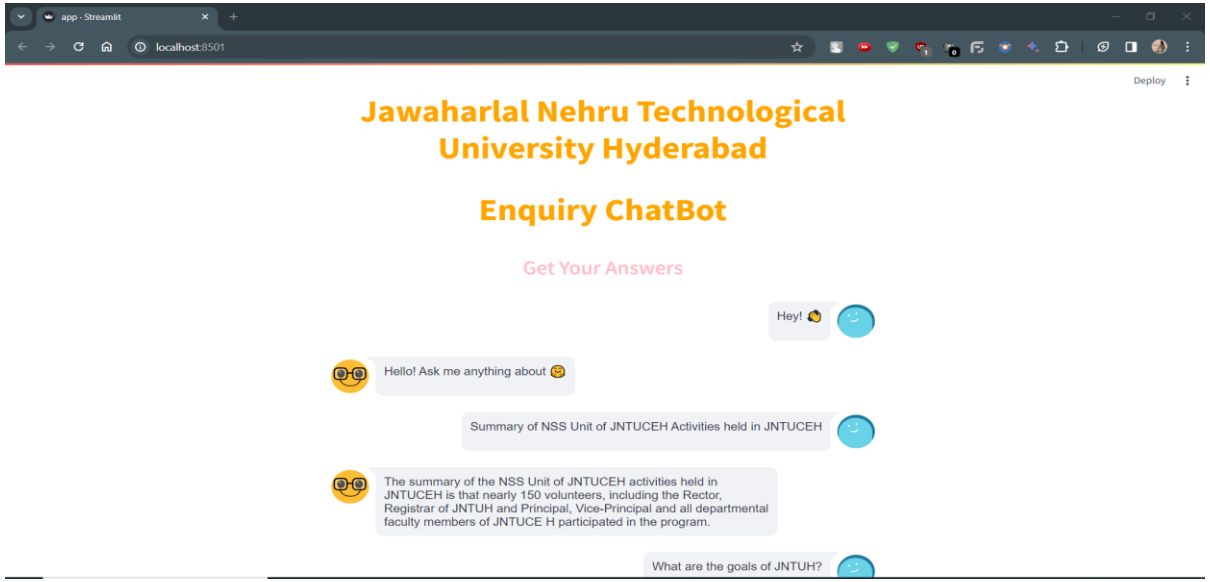
Get Your Answers

Hey! 👋

Hello! Ask me anything about 🤗

Question:

Ask about JNUTH



Jawaharlal Nehru Technological University Hyderabad

Enquiry ChatBot

Get Your Answers

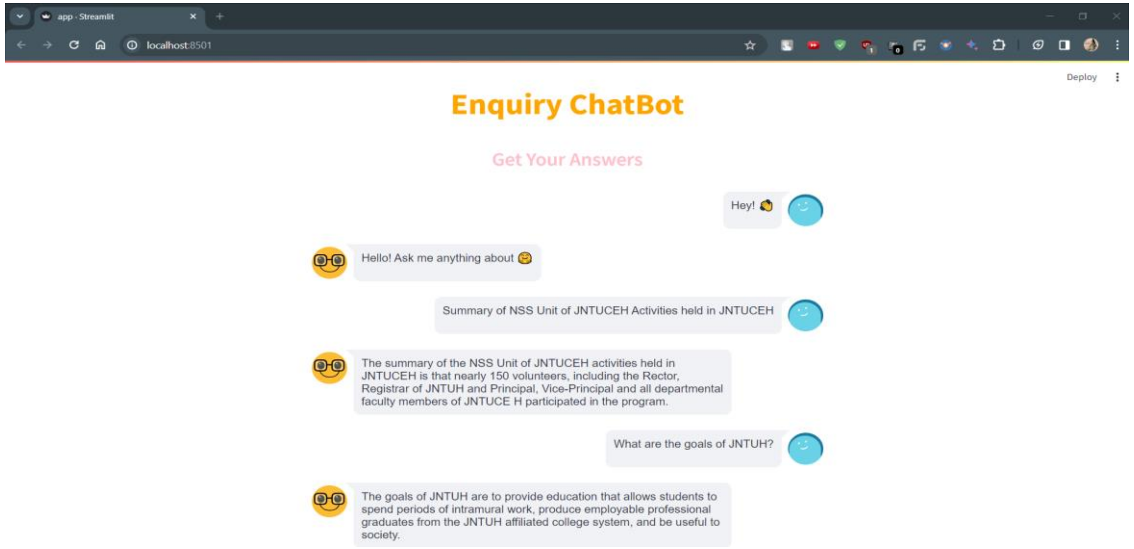
Hey! 👋

Hello! Ask me anything about 🤗

Summary of NSS Unit of JNTUCEH Activities held in JNTUCEH

The summary of the NSS Unit of JNTUCEH activities held in JNTUCEH is that nearly 150 volunteers, including the Rector, Registrar of JNTUH and Principal, Vice-Principal and all departmental faculty members of JNTUCE H participated in the program.

What are the goals of JNTUH?



Enquiry ChatBot

Get Your Answers

Hey! 👋

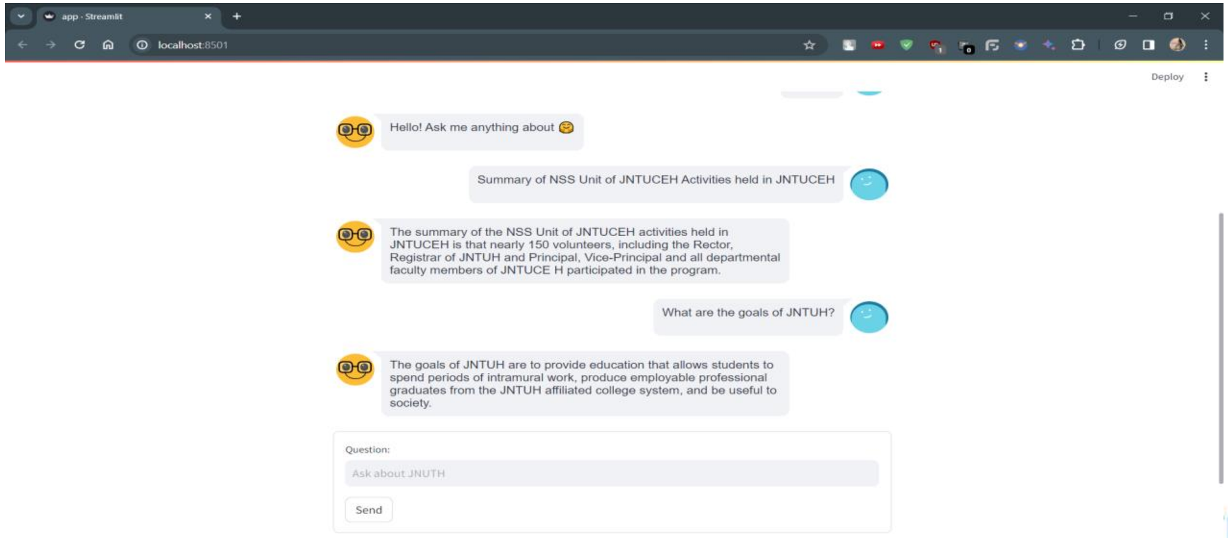
Hello! Ask me anything about 🤗

Summary of NSS Unit of JNTUCEH Activities held in JNTUCEH

The summary of the NSS Unit of JNTUCEH activities held in JNTUCEH is that nearly 150 volunteers, including the Rector, Registrar of JNTUH and Principal, Vice-Principal and all departmental faculty members of JNTUCE H participated in the program.

What are the goals of JNTUH?

The goals of JNTUH are to provide education that allows students to spend periods of intramural work, produce employable professional graduates from the JNTUH affiliated college system, and be useful to society.



Hello! Ask me anything about 🤗

Summary of NSS Unit of JNTUCEH Activities held in JNTUCEH

The summary of the NSS Unit of JNTUCEH activities held in JNTUCEH is that nearly 150 volunteers, including the Rector, Registrar of JNTUH and Principal, Vice-Principal and all departmental faculty members of JNTUCE H participated in the program.

What are the goals of JNTUH?

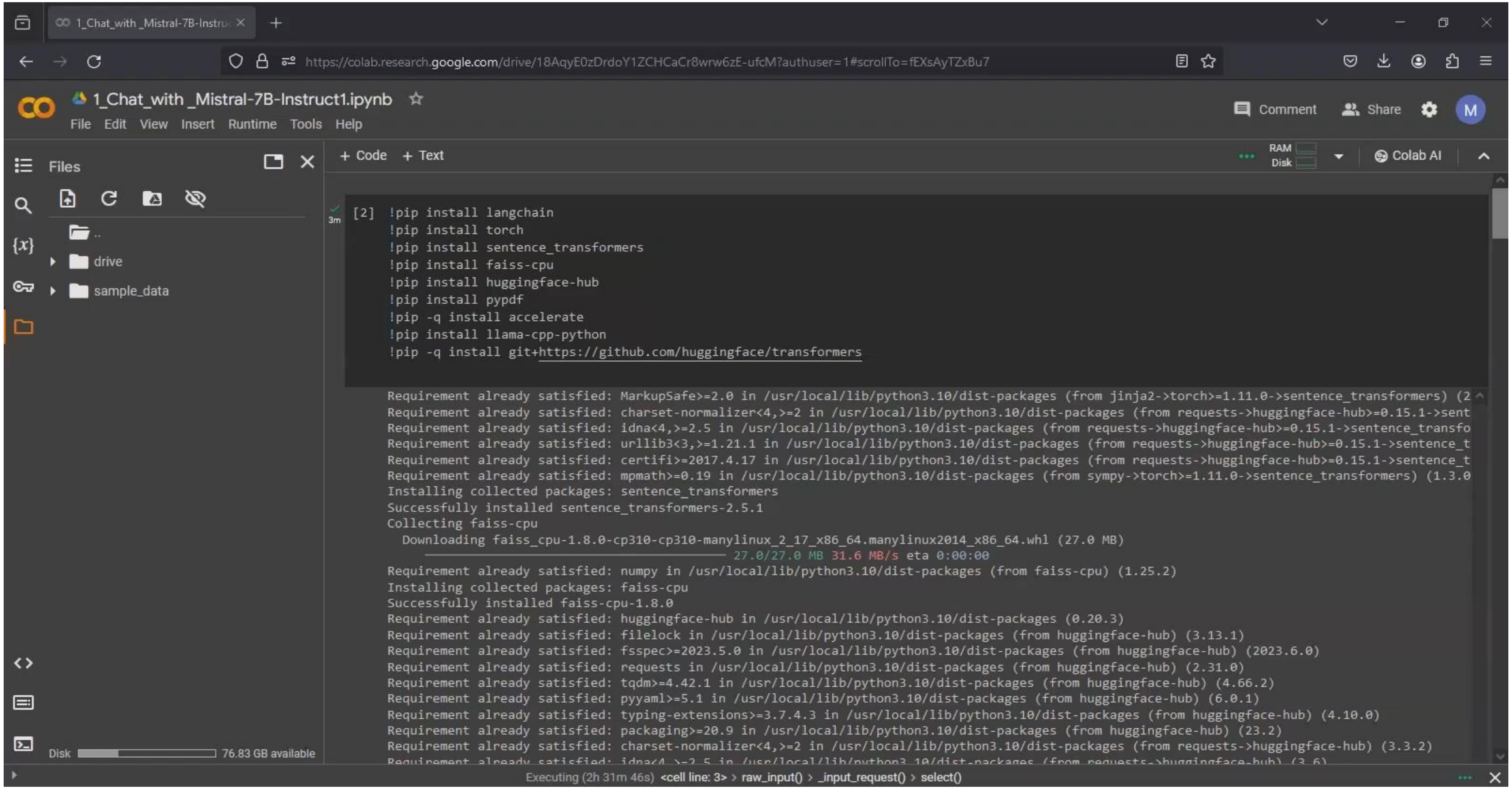
The goals of JNTUH are to provide education that allows students to spend periods of intramural work, produce employable professional graduates from the JNTUH affiliated college system, and be useful to society.

Question:

Ask about JNUTH

Innovation • Data • Analytics

Video of output- Google Colab

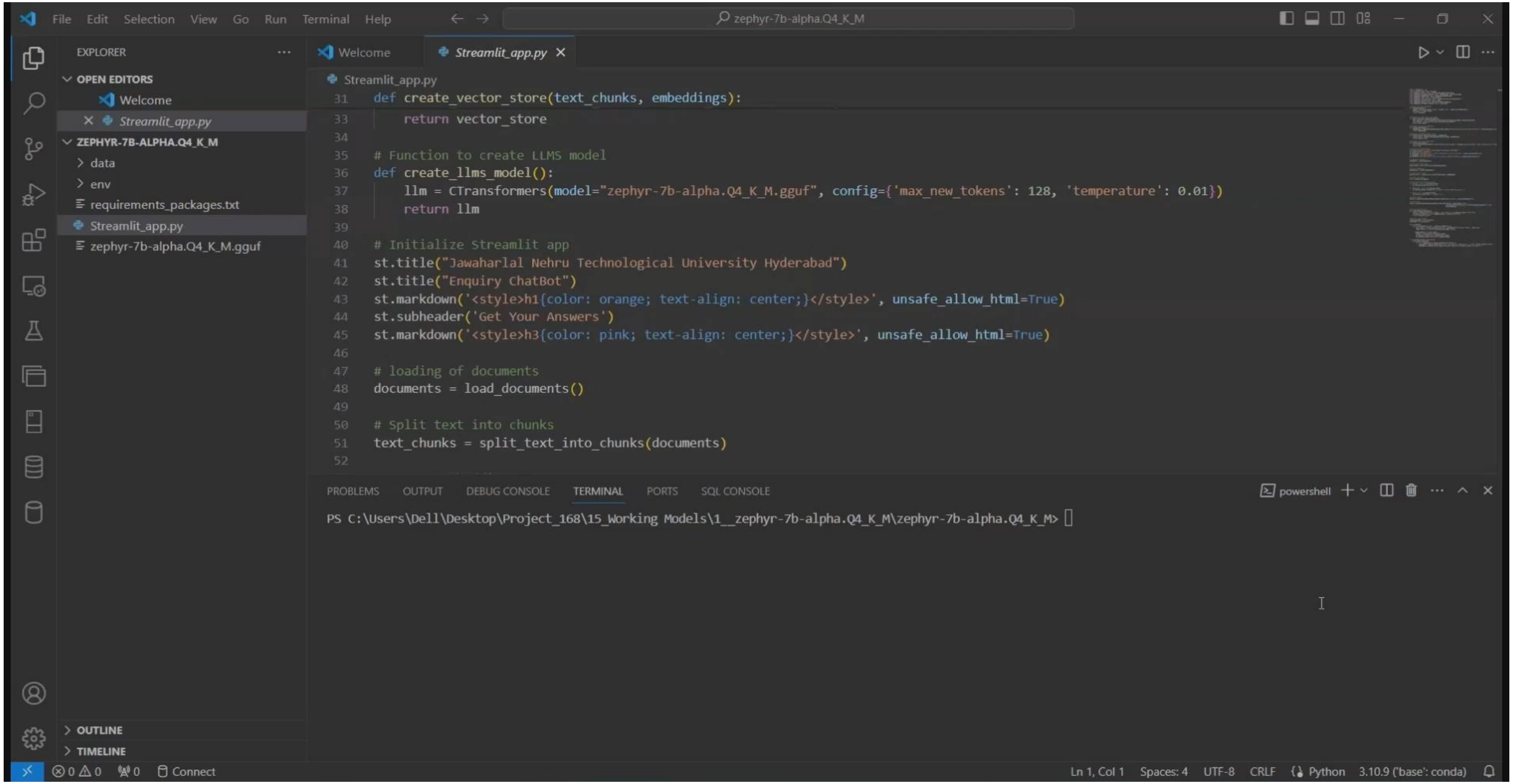

1_Chat_with _Mistral-7B-Instru
https://colab.research.google.com/drive/18AqyE0zDrdoY1ZCHCaCr8wrw6zE-ufcM?authuser=1#scrollTo=fEXsAyTZxBu7
1_Chat_with _Mistral-7B-Instruct1.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
M
Files
+ Code + Text
RAM
Disk
Colab AI
..
drive
sample_data
Disk 76.83 GB available
3m
Executing (2h 31m 46s) <cell line: 3> › raw_input() › _input_request() › select()

```
[2] !pip install langchain
    !pip install torch
    !pip install sentence_transformers
    !pip install faiss-cpu
    !pip install huggingface-hub
    !pip install pypdf
    !pip -q install accelerate
    !pip install llama-cpp-python
    !pip -q install git+https://github.com/huggingface/transformers
```

```
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.10/dist-packages (from jinja2->torch>=1.11.0->sentence_transformers) (2
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.15.1->sent
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.15.1->sentence_transfo
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.15.1->sentence_t
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub>=0.15.1->sentence_t
Requirement already satisfied: mpmath>=0.19 in /usr/local/lib/python3.10/dist-packages (from sympy->torch>=1.11.0->sentence_transformers) (1.3.0
Installing collected packages: sentence_transformers
Successfully installed sentence_transformers-2.5.1
Collecting faiss-cpu
  Downloading faiss_cpu-1.8.0-cp310-cp310-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (27.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 27.0/27.0 MB 31.6 MB/s eta 0:00:00
Requirement already satisfied: numpy in /usr/local/lib/python3.10/dist-packages (from faiss-cpu) (1.25.2)
Installing collected packages: faiss-cpu
Successfully installed faiss-cpu-1.8.0
Requirement already satisfied: huggingface-hub in /usr/local/lib/python3.10/dist-packages (0.20.3)
Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from huggingface-hub) (3.13.1)
Requirement already satisfied: fsspec>=2023.5.0 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub) (2023.6.0)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from huggingface-hub) (2.31.0)
Requirement already satisfied: tqdm>=4.42.1 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub) (4.66.2)
Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub) (6.0.1)
Requirement already satisfied: typing-extensions>=3.7.4.3 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub) (4.10.0)
Requirement already satisfied: packaging>=20.9 in /usr/local/lib/python3.10/dist-packages (from huggingface-hub) (23.2)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->huggingface-hub) (3.6)
```

INNODATATICS
Innovation • Data • Analytics

Video of output-VS Code -Streamlit



```python
def create_vector_store(text_chunks, embeddings):
    return vector_store

# Function to create LLMS model
def create_llms_model():
    llm = CTransformers(model="zephyr-7b-alpha.Q4_K_M.gguf", config={'max_new_tokens': 128, 'temperature': 0.01})
    return llm

# Initialize Streamlit app
st.title("Jawaharlal Nehru Technological University Hyderabad")
st.title("Enquiry ChatBot")
st.markdown('<style>h1{color: orange; text-align: center;}</style>', unsafe_allow_html=True)
st.subheader('Get Your Answers')
st.markdown('<style>h3{color: pink; text-align: center;}</style>', unsafe_allow_html=True)

# loading of documents
documents = load_documents()

# Split text into chunks
text_chunks = split_text_into_chunks(documents)
```

```
PS C:\Users\Dell\Desktop\Project_168\15_Working Models\1__zephyr-7b-alpha.Q4_K_M\zephyr-7b-alpha.Q4_K_M>
```

Challenges

- Data quality: Ensuring the quality and relevance of documents used for training and retrieval.

- Model performance: Tuning the models for optimal response generation and retrieval accuracy.

- Scalability: Handling large volumes of queries and maintaining low response latency.

- User experience: Designing an intuitive and responsive interface for seamless interaction.

- Integration: Integrating the model with existing systems and workflows.

Future Scopes

- Enhancing conversational capabilities: Improving the model's ability to handle diverse queries and provide accurate responses.
- Personalization: Customizing responses based on user preferences and past interactions.
- Multimodal capabilities: Integrating text, audio, and visual inputs for a richer user experience.
- Domain expansion: Extending the model's knowledge and adaptability to new domains beyond education.
- Collaboration: Collaborating with domain experts and stakeholders for continuous improvement and refinement of the system.

Queries ?

Thanks