OFFLINE AI CHATBOT FOR VARIOUS PURPOSES

Submitted in partial fulfillment of the requirements of the degree

BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

By

Krish Bhatia 07

Vikalp Bora 09

Priyanka Jotwani 36

Prof. Prerna Solenke

Vivekanand Education Society's Institute of Technology,

An Autonomous Institute affiliated to University of Mumbai

HAMC, Collector's Colony, Chembur,

Mumbai-400074

University of Mumbai (AY 2024-25)

CERTIFICATE

This is to certify that the Mini Project entitled **" Offline Chat Bot"** is a bonafide work of **Krish Bhatia (07) , Vikalp Bora (09) & Priyanka Jotwani (36)** submitted to the University of Mumbai in partial fulfillment of the requirement for the award of the degree of **"Bachelor of Engineering"** in **"Computer Engineering" .**

(Prof. Prerna Solenke)
Mentor

(Prof._____)
Head of Department

(Prof._____)
Principal

Mini Project Approval

This Mini Project entitled "Reality Check" by **Krish Bhatia (07), Vikalp Bora (09) & Priyanka Jotwani (36)** is approved for the degree of **Bachelor of Engineering** in **Computer Engineering.**

Examiners

1……
(Internal Examiner Name & Sign)

2……
(External Examiner name & Sign)

Date: Place:

Contents

Abstract:

The proliferation of misinformation poses significant challenges in educational and military environments, where reliable information is crucial. This highlights the need for effective offline solutions that can deliver accurate responses without internet dependency. While existing chatbots often rely on online databases, they fail to serve areas with limited connectivity. Our project presents an innovative offline chatbot that operates in two distinct modes: utilizing a large language model (LLM) for generating answers to random questions and implementing a Retrieval-Augmented Generation (RAG) model for responding based on a curated dataset. This chatbot is designed to be a valuable resource in schools, enhancing learning by providing instant information, and supporting military personnel by offering quick access to essential knowledge. The offline functionality, powered by a continuously updated dataset, ensures that the chatbot remains relevant and accurate, making it a vital tool in environments where internet access is restricted. We will assess the performance of both models, aiming to refine the chatbot's accuracy and user experience in diverse contexts.

Acknowledgement:

We would like to express our sincere gratitude to Vivekanand Education Society's Institute of Technology for their support and resources throughout our project. Their assistance in providing project-related data was invaluable to the success of our chatbot system.

We are especially thankful to Ms.Prerna Solanke, our TE Mini-Project Mentor, for his continuous support and expert advice in guiding the development of the offline chatbot. His insights on integrating LLaMA2 and the RAG model were pivotal in shaping the project's two-way functionality.

Our deepest appreciation goes to Dr. (Mrs.) Nupur Giri, the Head of the Computer Department, and Dr. (Mrs.) J.M. Nair, the Principal, for granting us the opportunity to work on this project. Their encouragement and resources enabled us to create a solution beneficial for environments without internet access, such as military applications and educational settings.

We would also like to extend our heartfelt thanks to all the teaching and non-teaching staff for their unwavering encouragement, which played a crucial role in the successful completion of this project.

List of Abbreviations:

LLM : Large Language Model

RAG: Retrieval Augmented Generation

List of Figures:

List of Tables:

1. Introduction

1.1 Introduction:

In today's information-driven world, reliable access to accurate information is crucial. However, in certain environments where internet connectivity is limited or unavailable—such as remote military operations or educational settings—access to real-time information becomes a significant challenge. To address this, we have developed an offline chatbot system that operates in two distinct ways: first, by utilizing the LLaMA2 model to answer a wide range of general questions, and second, through a Retrieval-Augmented Generation (RAG) model that sources answers directly from pre-fed data in the form of PDFs.

The primary goal of our project is to create an efficient, self-contained system that ensures users in low or no-internet zones can still access accurate and relevant information. By leveraging the power of LLMs like LLaMA2 for general inquiries and RAG for domain-specific queries, we aim to provide a versatile solution. This dual-model approach not only supports the retrieval of general knowledge but also offers specialized information based on pre-uploaded documents.

One of the key advantages of this system is its adaptability. While many chatbots rely heavily on internet access to retrieve and validate information, our solution works entirely offline, making it invaluable for applications in the military, where secure communication is essential, or for students in remote areas without access to the internet. By feeding relevant data into the RAG model, the chatbot can provide specific, accurate responses to user queries, significantly reducing the need for internet-based resources.

Our project’s emphasis on offline functionality opens new avenues for deploying AI-driven tools in environments where traditional internet-based models may not be feasible. The chatbot ensures information remains both accessible and reliable, even in the most remote or restricted settings.

This version aligns closely with your project's purpose, highlighting the significance of offline access, the use of LLaMA2 and RAG models, and the potential applications for your system in remote or secure environments. Let me know if you’d like any further refinements!

1.2 Motivation

The rise of misinformation has become a global issue, infiltrating various fields such as sports, science, and history. While general misinformation can often be recognized and debunked, misinformation in these specialized domains is more subtle and complex. Inaccurate information can mislead people, distort public understanding, and influence opinions based on false or misleading claims, especially when real-time fact-checking is unavailable in offline environments.

In sports, misinformation may manifest as fabricated statistics, rumors about athletes, or misrepresented historical outcomes of matches or achievements. Such inaccuracies can confuse fans, damage reputations, and even impact athletes' careers. In the scientific domain, the consequences are even more severe. Misinformation around discoveries, health guidelines, or emerging technologies—such as unverified claims about vaccines, climate change, or medical treatments—can lead to widespread public distrust and dangerous decision-making. The increasing spread of pseudoscience and manipulated studies makes it even harder for individuals to discern credible information. Similarly, in history, misinformation often takes the form of revisionism, where facts are distorted, or events are reinterpreted to serve specific agendas, affecting public understanding and collective memory.To combat these issues, a robust and accurate fact-checking mechanism is essential. However, general-purpose fact-checking systems lack the depth and domain-specific knowledge required to handle the complexity of misinformation in these areas. This gap highlights the need for a system that can perform offline and deliver accurate, domain-specific responses, especially in environments where internet access is limited or unavailable.

Our project aims to bridge this gap by integrating advanced models like LLaMA2 and RAG. By training the system on authoritative data from sources such as government websites, encyclopedias, and peer-reviewed journals, our chatbot will be equipped to accurately distinguish between fact, opinion, and misinformation in specialized fields like sports, science, and history. The combination of fine-tuning language models and leveraging pre-fed, trusted data through RAG ensures a reliable offline fact-checking solution. This capability is especially critical for military and educational environments, where users must rely on verified information without real-time access to the internet.

1.3 Problem Statement & Objectives

In environments where internet connectivity is limited or unavailable, accessing accurate and up-to-date information becomes a significant challenge. Existing general-purpose tools for information retrieval and fact-checking often depend on internet access and are not optimized for specialized knowledge in areas like sports, science, and history. Additionally, the complexity of domain-specific information requires deeper understanding and context that general-purpose chatbots struggle to provide.

For instance, answering a question about an athlete's performance or an event's outcome requires detailed knowledge of the sport's history and statistics. In science, responding to inquiries about discoveries or health guidelines involves rigorous knowledge of established scientific principles and references. Similarly, understanding historical claims demands familiarity with verified records and sources. Without the capability to accurately retrieve domain-specific information offline, users in these environments—such as the military, students in remote locations, or professionals working in secure areas—are left without reliable resources.

Thus, the need arises for a self-contained, offline chatbot system that can provide accurate, domain-specific information even in environments without internet access. This chatbot system should be capable of leveraging advanced language models like LLaMA2 for general knowledge inquiries and a Retrieval-Augmented Generation (RAG) model for fact-based responses sourced from pre-loaded PDFs.

Objectives:

- **Offline Domain-Specific Information Retrieval**: Build an offline chatbot system that provides accurate, domain-specific answers in areas such as sports, science, and history by utilizing pre-fed data from trusted sources such as government publications, academic papers, and credible reference materials.
- **Ensuring Accuracy and Reliability**: Train the LLaMA2 and RAG models on carefully curated, high-quality datasets to ensure the chatbot provides reliable and precise answers, reducing the risk of misinformation or errors.
- **Enhancing Usability in No-Internet Environments**: Address the gap in traditional chatbots that rely on internet access by ensuring the system functions entirely offline, offering efficient and accurate responses without requiring real-time data retrieval from online sources.
- **Optimizing Response Time**: Streamline the data retrieval and processing capabilities of the chatbot to ensure fast and accurate fact-checking, even when working with large, complex datasets stored locally.
- **User-Friendly Interface**: Develop an intuitive and accessible user interface, allowing users to interact with the chatbot seamlessly through text, voice, or search queries, catering to various environments like military bases or educational institutions.
- **Transparency and Trust**: Ensure that all responses provided by the chatbot are backed by verifiable sources, with clear references to the original PDFs or documents, fostering trust and discouraging the spread of misinformation.

1.4 Organization of the Report

Chapter 1: Introduction

This chapter introduces the core concept behind the offline chatbot project, which aims to develop a domain-specific information retrieval system for sports, science, and history. It discusses the motivation for creating such a system, highlights the problem statement related to misinformation, and outlines the key objectives that guide the project's direction.

Chapter 2: Review of Related Work

This chapter provides an in-depth review of existing chatbot systems, fact-checking tools, and relevant literature. It examines previous research, studies, and technologies related to information retrieval and fact-checking, emphasizing the limitations of current systems in verifying domain-specific information. Additionally, it identifies gaps and challenges in the fields of sports, science, and history, setting the stage for the unique contributions of the offline chatbot.

Chapter 3: Proposed System

The third chapter describes the architecture and functionality of the offline chatbot system in detail. It covers the integration of LLaMA2 and the Retrieval-Augmented Generation (RAG) approach, as well as the hardware and software components utilized in the system's development. This chapter includes a presentation of experimental results and analysis, summarizing the performance and validation of the chatbot. It concludes with a discussion of potential future enhancements, providing a comprehensive overview of the system's development and its future directions.

Literature Survey

2.1 Survey of Existing System

1. Meta AI Researchers (Hugo Touvron et al.) – February 26, 2023

- **Key Takeaways:**
  - Efficient Performance with Smaller Models: LLaMA models (7B–65B parameters) perform better than larger models like GPT-3 (175B), Chinchilla-70B, and PaLM-540B.
  - Public Data for Training: LLaMA leverages publicly available datasets, such as CommonCrawl, Wikipedia, and GitHub, instead of proprietary data.
- **Limitations:**
  - Resource-Intensive: LLaMA-65B required 2048 A100 GPUs and 21 days for training, making it resource-heavy.
  - Performance Gaps: It falls short on tasks like massive multitask language understanding (MMLU) compared to Chinchilla-70B and PaLM-540B.

2. Fei Liu, Zejun Kang, Xing Han – January 2018

- **Key Takeaways:**
  - RAG Overview: Combines traditional information retrieval with generative models by retrieving PDFs or relevant documents and generating contextual answers.
  - Locally Deployed Ollama Models: Ensures data privacy and compliance by working without internet access, suitable for environments with strict data policies.
- **Limitations:**
  - Hardware Resource Constraints: Running large models locally demands high GPU/TPU capacity.
  - Scalability and Maintenance: Updating models is labor-intensive, especially with frequent changes to automotive manuals or documents.

3. O. Hourrane, H. Ouchra, A. Hafsa, El. Eddaoui, H. Benlahmar, O. Zahour – April 2020

- **Key Takeaways:**
  - NLP-based Chatbot: Uses NLP to answer student queries about educational programs and career options.

    - Personalized Guidance Support: Helps students access accurate guidance despite the challenges they face.
- **Limitations:**
    - Cultural and Linguistic Specificity: Tailored to the Moroccan context, limiting its usability outside the region.
    - Limited Personalization: Offers general advice but struggles with personalized recommendations based on individual profiles.

4. R. Patel, N. Bhagora, P. Singh, K. Namdev – April 2020

- **Key Takeaways:**
    - Cloud-based Chatbot: Manages student information efficiently using cloud computing for scalability and accessibility.
- **Limitations:**
    - Limited Interaction Scope: May struggle with complex queries beyond predefined responses.
    - Dependency on Cloud Services: Risks related to data privacy and security are not thoroughly addressed.

2.2 Mini Project Contribution

Sustainable Development Goal: Peace, Justice, and Strong Institutions

The primary objective of our project is to combat misinformation by enabling rapid and accurate fact-checking across various domains, with a particular emphasis on scientific and historical data. The proposed offline chatbot system leverages advanced methodologies for real-time fact verification using Large Language Models (LLMs) like LLaMA2, Retrieval-Augmented Generation (RAG), and pre-fed datasets, ensuring precise and timely validation of information.

By utilizing pre-uploaded data from authoritative sources, the system delivers reliable responses crucial for fact-checking in rapidly changing domains. The incorporation of Knowledge Graphs (KGs) enhances the accuracy of information verification by providing a structured, contextual understanding of interconnected facts. This allows for a deeper analysis of complex relationships, further supporting accurate verification of claims.

Our approach aims to mitigate the spread of misinformation by offering users reliable, up-to-date facts, thereby reducing the social and economic impacts associated with false information. Additionally, this project seeks to raise public awareness about the significance of fact-checking, promoting digital literacy, and fostering a more informed society. The offline nature of the chatbot ensures that even in environments with limited internet access—such as military operations or remote educational settings—users can still rely on verified information, contributing to overall peace, justice, and strong institutions.

Proposed System

3.1 Introduction

The proposed offline chatbot system aims to build an advanced fact-checking framework by leveraging cutting-edge technologies such as Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG). In an age where misinformation spreads rapidly, particularly in domains like sports, science, and history, real-time verification of information has become crucial. This system addresses these challenges by ensuring the accuracy and reliability of facts, even in environments with limited or no internet access.

Large Language Models (LLMs), like LLaMA2, are powerful AI models trained on vast amounts of text data. They excel in natural language understanding and generation, making them ideal for tasks like fact-checking, where precision in interpreting and responding to queries is critical. However, while LLMs are highly capable, they rely on static, pre-existing knowledge, which can pose challenges in keeping up with rapidly evolving information, especially in specialized domains.

To overcome this limitation, the system incorporates a Retrieval-Augmented Generation (RAG) approach. RAG enhances the chatbot's capabilities by integrating LLMs with a retrieval mechanism that utilizes pre-fed data from authoritative sources, such as PDFs of scientific papers and historical documents. This setup allows the **system to pull relevant,** accurate information for fact-checking queries, ensuring that users receive timely and precise responses.

Furthermore, to deepen the system’s ability to understand and analyze complex relationships between entities, the incorporation of Knowledge Graphs (KGs) is planned. Knowledge graphs represent information through entities (nodes) and their relationships (edges), providing a structured, interconnected way of reasoning about facts. By integrating KGs, the chatbot can ensure not only accuracy but also a profound understanding of the context surrounding the facts being verified.

This hybrid approach, combining LLMs for processing, RAG for data retrieval, and KGs for structured reasoning, forms a highly adaptable and modular system. Each component works together to enhance fact-checking accuracy while also allowing for future extensions and improvements as the information landscape continues to evolve, particularly in environments where users lack real-time internet access.

3.2 Architectural Framework / Conceptual Design

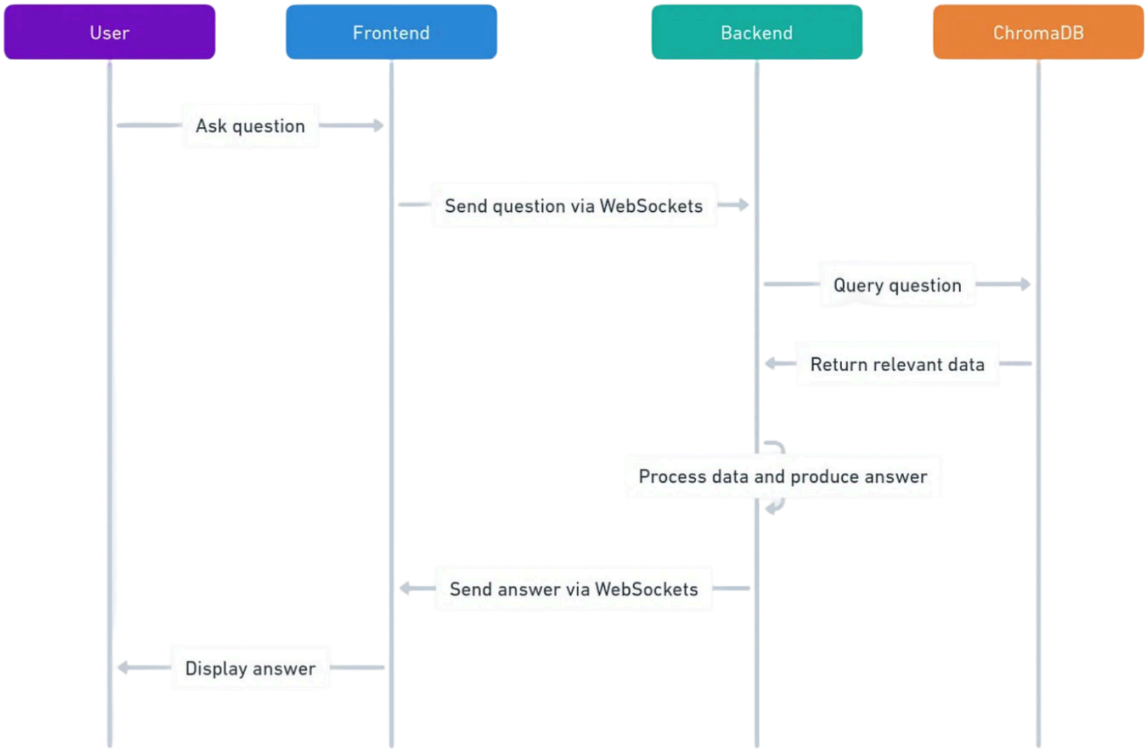


Figure 1: Architectural Framework

Figure 2: Architecture for model for various purposes

3.3 Algorithm and Process Design

This offline chatbot operates through two distinct modes:

1. **General Query Handling using the LLaMA 2 Model**
2. **Domain-Specific Query Handling using a RAG Model with Pre-loaded PDFs**

The design leverages fine-tuning and retrieval-augmented generation techniques to achieve robust performance in internet-restricted environments. Below is a step-by-step breakdown of the algorithm and process flow.

1. General Query Handling with LLaMA 2 Model

- **Input**: User asks a general question (e.g., "What is the capital of France?").
- **Process**:
    - The LLaMA 2 model, fine-tuned on a diverse QA dataset, processes the question.
    - It searches its internal knowledge learned during fine-tuning to generate a relevant response.
- **Output**: A general answer is returned, e.g., *"The capital of France is Paris."*

Domain-Specific Query Handling with RAG Model

- **Input**: User asks a specific question (e.g., "What are the safety protocols for electric vehicles?").
- **Process**:
    - **Query Encoding**: The user query is converted into a vector using an embedding model.
    - **Document Search**: The query vector is matched against vectorized chunks of pre-loaded PDFs using a similarity search algorithm (e.g., FAISS).
    - **Relevant Data Retrieval**: The top-matching chunks from the PDFs are retrieved.
    - **Response Generation**: The RAG model uses both the retrieved chunks and its own knowledge to generate a precise response.
- **Output**: A specific and accurate answer based on the PDF content is returned, e.g., *"Ensure all personnel wear insulated gloves and goggles when handling battery components."*

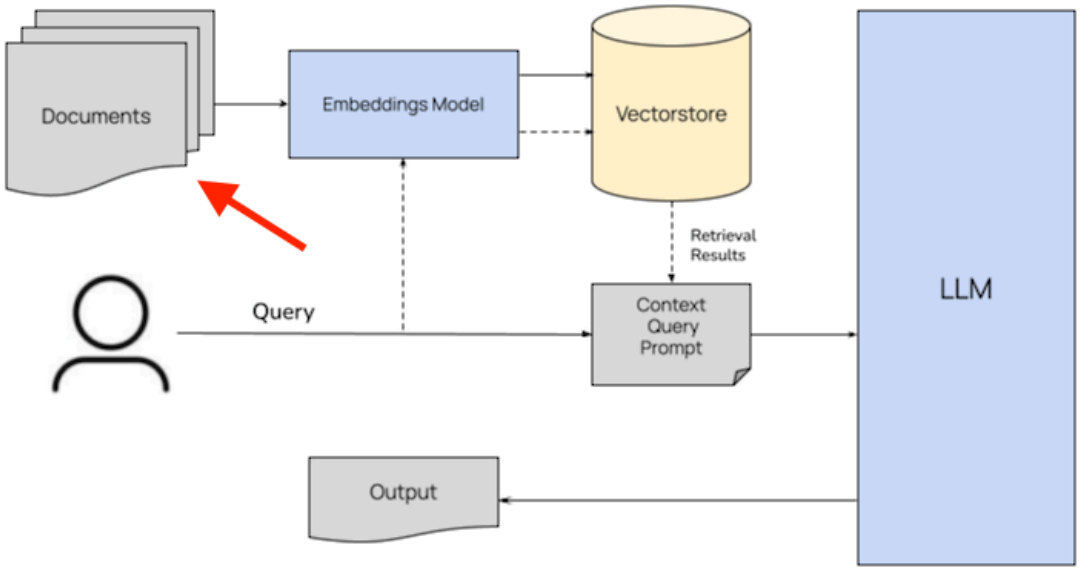


Figure 3 RAG Implementation

3.4 Methodology Applied

Methodology Employed- Components Working

1. **Desired LLM**: Houses the pre-trained large language model, serving as the backbone for generating responses to user queries.
2. **Ollama Integration**: Orchestrates the interaction between user inputs, the LLM, with web UI ensuring seamless communication and processing.
3. **Python**: We used langchain to give structure to chatbot, and streamlit for temporary GUI. We have also used ChatPromptTemplate which is a library in python to provide the basic chat template.

Methodology Employed- Actual Flow

- Upon receiving user inputs through the web UI's chat input box, the text data is transmitted to the Ollama module for interpretation.
- Ollama plays a crucial role in parsing and analyzing the user queries, extracting relevant information, and preparing the input for the LLM.
- Once processed, the interpreted data is forwarded to the LLM, which utilizes its pre-trained knowledge to generate appropriate responses.
- These responses are then transmitted back to the python, where they are presented to the user for further interaction.

3.5 Hardware & Software Specifications

HARDWARE:

- Processor: A modern multi-core processor (e.g., Intel i5/i7 or AMD Ryzen)
- RAM: At least 8GB, but 16GB or more
- Storage: SSD with at least 256GB of space.
- Graphics Card (GPU): A dedicated GPU (e.g., NVIDIA GTX 1660 or better)
-

SOFTWARE:

- OLLAMA: Orchestrates the interaction between user inputs and the LLM.
- Python: Providing a user-friendly interface whilst making connectivity with the LLM.

Backend:

The backend is built using Python and leverages several key libraries:

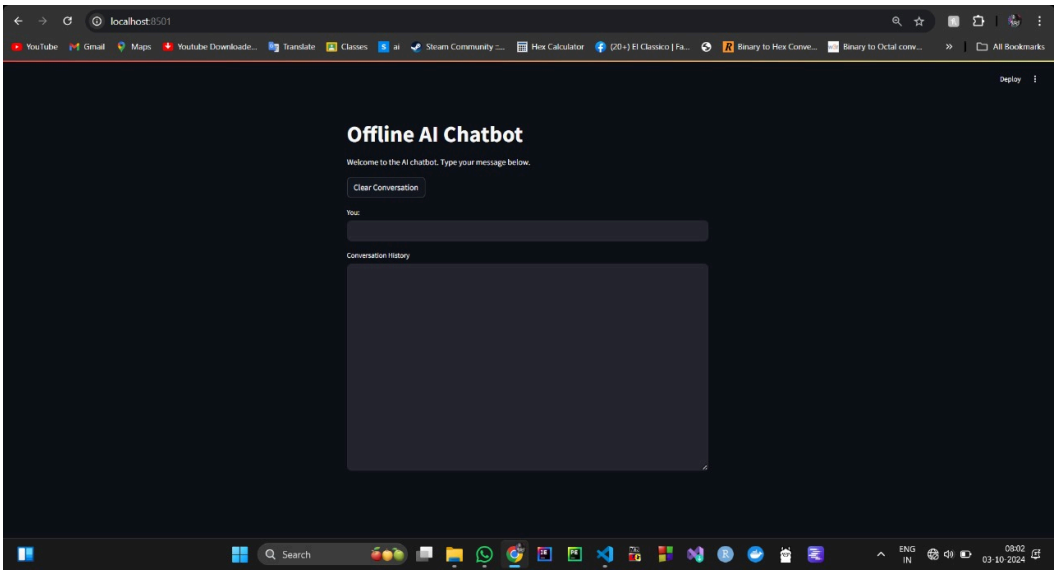
- **Langchain:** For building the RAG model.
- **Websockets:** For real-time communication.
- **ChromaDB:** For vectorstore management.
- **TOML:** For configuration management.

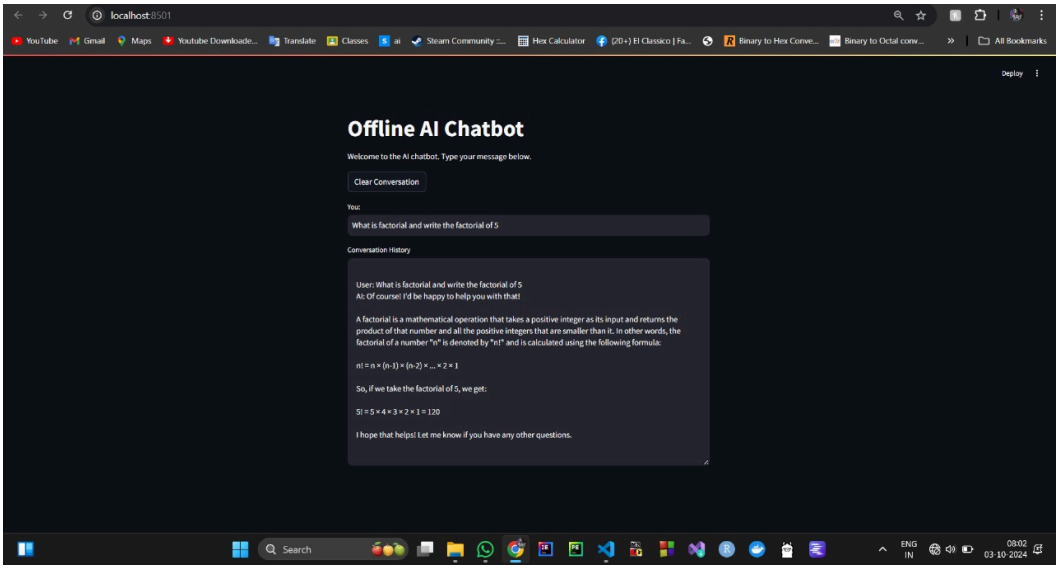
The core component is the RAG model, which will integrate the language model with the vectorstore to provide context-aware answers.The model will maintain a session history and utilizes a Retrieval-Augmented Generation chain to process user queries and return relevant responses.

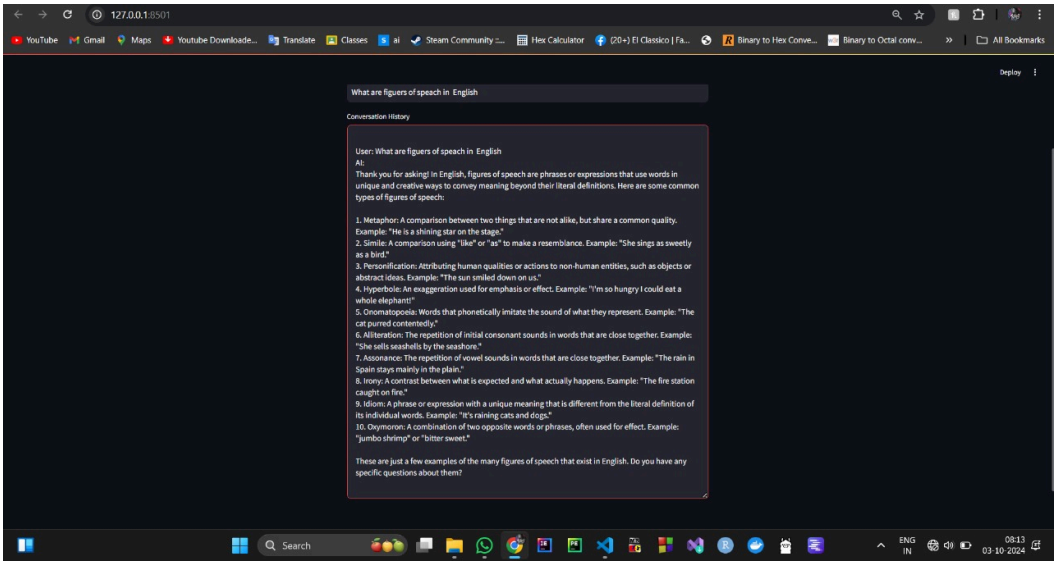
Frontend:

The frontend is a React.js application that will provides a simple interface for users to interact with the chatbot. It establishes a WebSocket connection with the backend to send user queries and display responses in real-time. We can also give the ability to the user to alter the UI as per needs.

3.6 Experiment and Results for Validation and Verification







3.7 Result Analysis and Discussion

The primary objective of this project was to develop an offline chatbot capable of providing accurate responses without internet access, making it suitable for use in settings such as the military and remote educational environments. The chatbot operates in two modes:

1. **General Question Answering**: It utilizes a fine-tuned LLaMA 2 model to answer a broad range of random questions.
2. **Domain-Specific Fact Retrieval**: It leverages a Retrieval-Augmented Generation (RAG) model, which references specific information from pre-fed PDFs to generate precise responses for fact-based queries.

Initially, the LLaMA 2 model was trained on a semi-structured dataset composed of scraped data from fields like sports, science, and history. However, the results were subpar—responses were often vague or irrelevant, lacking the precision necessary for reliable fact verification. This raised challenges, particularly for offline use cases where the chatbot needs to function autonomously and deliver high-quality answers.

To improve the performance, the dataset was restructured into a **QA format** with clear input-output pairs, allowing the model to better interpret user queries and generate more relevant responses. This change significantly enhanced the chatbot's ability to respond with increased accuracy. Figures (4) and (5) demonstrate how the shift to structured data improved the LLaMA 2 model's reasoning capabilities and reduced redundant or incorrect outputs.

In addition to fine-tuning the LLaMA 2 model, the integration of the **RAG model** brought further improvements. The RAG model could retrieve specific, domain-relevant information from pre-loaded PDFs, ensuring more precise answers for targeted queries. This hybrid approach allowed the chatbot to effectively combine general knowledge with domain-specific data, thereby improving its utility in scenarios with no internet connectivity.

In conclusion, the combination of **QA-format fine-tuning** and **RAG-based retrieval** greatly enhanced the chatbot's performance. The project highlights the importance of proper data structuring and retrieval mechanisms in offline environments, ensuring that users—such as military personnel or students in remote areas—can access reliable information without the need for internet connectivity.

3.8 Conclusion and Future Work

The offline chatbot developed in this project successfully demonstrated the capability to provide accurate responses without internet connectivity, addressing the specific needs of environments like military operations and remote education. By leveraging two operational modes—a fine-tuned **LLaMA 2 model** for general inquiries and a **RAG model** for fact-based queries using pre-fed PDFs—the system offers both versatility and precision. The transition to a QA-format dataset played a key role in enhancing the performance of the LLaMA 2 model, while the RAG integration ensured reliable information retrieval from domain-specific documents.

This hybrid approach ensures that users can access both general knowledge and detailed domain-specific content without external dependencies. The project highlights the critical importance of data structuring and retrieval mechanisms in offline systems, making it an effective solution for internet-restricted environments.

Future Work

While the chatbot has shown promising results, several areas remain open for future improvements:

1. **Expansion of Domain-Specific Content**: Adding more PDFs across various fields to further enhance the RAG model's knowledge base.
2. **Model Optimization**: Exploring quantization and pruning techniques to reduce the computational load, ensuring smoother performance on low-resource devices.
3. **Multilingual Capabilities**: Extending the chatbot’s abilities to support multiple languages, which would make it more useful for diverse users.
4. **User Feedback Loop**: Implementing feedback mechanisms to fine-tune the chatbot continuously based on user interactions and errors.
5. **Enhanced Security Measures**: Incorporating encryption and access control for sensitive data, particularly for use in military settings.

These future directions will further enhance the chatbot’s adaptability, usability, and performance, making it an even more valuable tool for offline environments.

Reference

1. https://medium.com/@mutazyounes/bring-your-data-to-life-creating-a-chatbot-with-llm-langchain-vector-db-locally-on-docker-ed647e546f85
2. B. A. Shawar and E. Atwell, "Chatbots: Are they Really Useful?", LDV Forum, vol. 22, no. 1, pp. 29-49, January 2007, [online] Available: https://www.researchgate.net/publication/220046725_Chatbots_Are_they_Really_Useful.
3. C. Lebeuf, M. Storey and A. Zagalsky, "Software Bots", IEEE Software, vol. 35, no. 1, pp. 18-23, January/February 2018.
4. O. Hourrane, H. Ouchra, A. Hafsa, EL. Eddaoui, H. Benlahmar and O. Zahour, "Towards a Chatbot for educational and vocational guidance in Morocco: Chatbot E-Orientation", International Journal of Advanced Trends in Computer Science and Engineering, vol. 9, no. 2, pp. 2479-2487, April 2020.
5. https://www.semanticscholar.org/paper/NLP-for-Chatbot-Application-Nithyanandam-Kasinathan/796e648e8190c483cc2df9370e4e4e4ba8b0f4ad