Paarth Parag Doshi (21CE2022) Paarth Parag Doshi ...

Development of AI Chatbot for MMRCLWebsite.pdf

Ramrao Adik Institute of Technology

Document Details

Submission ID

trn:oid:::3618:93645404

Submission Date

Apr 30, 2025, 1:05 PM GMT+5:30

Download Date

Apr 30, 2025, 1:06 PM GMT+5:30

File Name

Development of AI Chatbot for MMRCLWebsite.pdf

File Size

680.5 KB

40 Pages

6,909 Words

42,583 Characters

7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

29 Not Cited or Quoted 7%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

6%	Internet sources
4%	Publications
0%	Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **29** Not Cited or Quoted 7%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

6%	Internet sources
4%	Publications
0%	Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.coursehero.com	4%
2	Internet	aclanthology.org	<1%
3	Internet	dr.iiserpune.ac.in:8080	<1%
4	Internet	www.ee.iitb.ac.in	<1%
5	Publication	"Intelligent Systems and Applications", Springer Science and Business Media LLC,...	<1%
6	Internet	pdfcoffee.com	<1%
7	Internet	www.lilyliliu.com	<1%
8	Internet	arxiv.org	<1%
9	Internet	globalports.eu	<1%
10	Internet	www.bio-conferences.org	<1%

11	Internet	www.chennaisunday.com	<1%
12	Publication	Saloni Jage, Shubham Chaudhari, Manthan Jatte, Abhishek Mhatre, Vanita Mane. ...	<1%
13	Internet	html.pdfcookie.com	<1%
14	Internet	ijgis.pubpub.org	<1%
15	Internet	www.epa.gov	<1%
16	Internet	www.vromansbookstore.com	<1%
17	Internet	lucidworks.com	<1%
18	Publication	Piyul Patel, Vedant Pimple, Ashutosh Pol, Swastik Chaudhary, Siddhi Kadu. "Hybri...	<1%

Development of AI Chatbot for MMRCL Website using LLMs and Django

1 Submitted in the partial fulfillment of the requirements

for the degree of B.Tech in Computer Science and Business Systems

by

Paarth Parag Doshi (21CE2022)

Supervisor

Mr. Chaitanya Jage

1 **Department of Computer Engineering**

Ramrao Adik Institute of Technology

Sector 7, Nerul, Navi Mumbai

(Under the ambit of D. Y. Patil Deemed to be University)

May 2025

D Y PATIL
DEEMED TO BE
UNIVERSITY
RAMRAO ADIK
INSTITUTE OF TECHNOLOGY
NAVI MUMBAI

Ramrao Adik Institute of Technology

(Under the ambit of D. Y. Patil Deemed to be University)

Dr. D. Y. Patil Vidyanagar, Sector 7, Nerul, Navi Mumbai 400 706

CERTIFICATE

This is to certify that, the Internship report entitled

Development of AI Chatbot for MMRCL Website using LLMs and Django

is a bonafide work done by

Paarth Parag Doshi (21CE2022)

1 and is submitted in the partial fulfillment of the requirement for the degree of

B.Tech in Computer Science and Business Systems

1 to the

D. Y. Patil Deemed to be University

Industry Mentor
(Mr. Sumit Patil)

Supervisor
(Mr. Chaitanya Jage)

Project Co-ordinator
(Mrs. Bhavana Alte)

1 Head of Department
(Dr. Amarsinh V. Vidhate)

Principal
(Dr. Mukesh D. Patil)

Internship Report Approval

This is to certify that the Internship entitled “ ***Development of AI Chatbot for MM-RCL Website using LLMs and Django*** ” is a bonafide work done by ***Paarth Parag Doshi (21CE2022)*** under the supervision of ***Mr. Chaitanya Jage***. This internship is approved in the partial fulfillment of the requirement for the degree of ***B.Tech in Computer Science and Business Systems***

Internal Examiner :

1. . . . .

2. . . . .

External Examiners :

1. . . . .

2. . . . .

Date : . . . /. . . /. . . . . .

Place : . . . .

1

DECLARATION

I declare that this written submission represents my ideas and does not invovle plagiarism. I have adequately cited and referenced the original sources wherever others' ideas or words have been included. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action against me by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: ____________________

Paarth Parag Doshi (21CE2022)

5

Abstract

The integration of artificial intelligence (AI) in public service systems has revolutionized user interaction and engagement. This project focuses on the design and deployment of an AI-based chatbot for the Mumbai Metro Rail Corporation Limited (MMRCL) portal. Leveraging advanced Large Language Models (LLMs), the chatbot aims to assist commuters with real-time responses to queries related to routes, ticketing, schedules, and more. By employing multilingual capabilities and dynamic conversation handling, the system strives to enhance accessibility and user satisfaction across diverse demographics.

The developed chatbot not only bridges gaps in existing static FAQ-based systems but also demonstrates the transformative potential of AI-driven solutions in public transportation. With high query resolution rates, faster response times, and strong user feedback scores, the project underlines how LLMs can be effectively utilized to create scalable, efficient, and user-friendly digital assistants. Future enhancements such as ticketing integration, expanded language support, and voice-enabled interaction have also been identified to further improve system capabilities.

6

Contents

Abstract **i**

List of Tables **iv**

List of Figures **v**

1 Introduction **1**

1.1 Overview 2

15 1.2 Roles and Responsibilities 3

1.3 Organization of the report 4

2 Literature Survey **6**

2.1 Introduction to Chatbots and AI 6

2.2 Evolution of Conversational Systems 6

2.3 Technologies Used 7

2.4 Related Works and Literature 8

3 AI Applications with LLMs **9**

3.1 System Overview 9

3.2 Application of AI and LLMs in Chatbots 9

3.3 Retrieval Mechanism and Knowledge Base Integration 10

3.4 Voice and Image Integration 10

3.5 Deployment and User Interface 11

4 System Design and Algorithms **12**

4.1 Problem Statement 12

4.2 System Architecture 12

4.3 Core Algorithms Used 13
4.3.1 Text Embedding using Sentence Transformers 13
4.3.2 Vector Similarity Search with FAISS 14
4.3.3 Retrieval-Augmented Generation (RAG) 14
4.4 Implementation Details 14
4.5 Model Deployment and Resource Optimization 15

5 Results and Discussion 16

6 Conclusion and Future Scope 20
6.1 Future Scope 20

7 Summary of Work 22

References 24

A Company Details 25
A.0.1 Company Overview 25
A.0.2 Company's Role in the Industry 26

B Industry Mentor Details 27

C Attendance Report 28

D Internship Completion Certificate 29

E Plagiarism Report 30

Acknowledgement 31

List of Tables

A.1 Company Overview 25

B.1 Industry Mentor Overview 27

List of Figures

4.1 System Architecture of Metro AI Chatbot 13

D.1 Internship Completion Certificate 29

Chapter 1

Introduction

Artificial Intelligence (AI) has rapidly transformed numerous sectors by providing innovative solutions that improve efficiency, accuracy, and user engagement. Among these innovations, AI-driven conversational agents, popularly known as chatbots, have become critical in streamlining communication between service providers and users. Public transportation authorities are increasingly adopting such technologies to meet the growing demand for real-time assistance, personalized service delivery, and enhanced commuter experiences.

The Mumbai Metro Rail Corporation Limited (MMRCL), overseeing one of India's most ambitious urban infrastructure projects, has recognized the potential of AI to revolutionize its interaction with commuters. The organization's focus on digitization and smart service delivery aligns perfectly with the deployment of an AI-based chatbot that can cater to daily queries related to routes, schedules, ticketing, and updates. With Mumbai's dense population and the essential role of the metro system in ensuring seamless urban mobility, there is a pressing need for intuitive digital solutions that can operate round-the-clock without human intervention.

This project revolves around the design, development, and deployment of a comprehensive AI-powered chatbot, integrated directly into the MMRCL's official portal. The chatbot is designed not merely as a question-answer bot, but as an intelligent conversational agent capable of understanding complex transportation-related queries, providing contextual answers, and learning continuously through user interactions. Unlike static FAQ systems or basic keyword-matching chatbots, this system leverages cutting-edge Large Language Models (LLMs) and retrieval-augmented generation techniques to dynamically respond based on a curated knowledge base.

2

The chatbot system supports multiple input modes, including text, voice, and image-based queries, thus improving accessibility for a wide range of users. Voice interaction capabilities, in particular, cater to commuters who may prefer hands-free interaction, while image-based inputs provide an innovative method to extract queries from photos or documents. Additionally, the chatbot accommodates multilingual communication, offering responses in Hindi, Marathi, and English to better serve the diverse commuter population of Mumbai.

From a technological standpoint, the chatbot employs modern deep learning frameworks and libraries, utilizing models such as Mistral (7B parameters) for high-quality natural language understanding and response generation. It incorporates components like FAISS (Facebook AI Similarity Search) for efficient retrieval of relevant information from large datasets, HuggingFace embeddings for semantic search, and Streamlit for rapid and interactive deployment of the frontend interface. The backend is built using Django, ensuring a robust, scalable, and secure environment for managing user sessions and system resources.

This report provides a detailed narrative of the design philosophy, technical architecture, challenges faced, and the overall journey of transforming an innovative idea into a real-world solution. It demonstrates the effectiveness of AI in enhancing public service delivery, particularly within the context of urban transportation systems, and lays the foundation for future enhancements such as ticketing system integration, expanded language support, and advanced personalization features.

1.1 Overview

In the modern era of rapid technological advancement, Artificial Intelligence (AI) has established itself as a transformative force across various industries. Within the domain of public transportation, AI technologies are being actively employed to elevate the quality of services and redefine user experiences. One of the most significant applications of AI in this sector is the development of intelligent chatbots, capable of providing immediate assistance to users by answering queries, offering guidance, and handling routine information dissemination. These AI-based conversational agents not only reduce the burden on human support systems but also operate with a level of consistency and availability that traditional methods cannot match.

This project focuses on the creation of an AI-powered chatbot tailored specifically for the Mumbai Metro Line 3 project. Mumbai, being one of the most densely populated cities globally, demands a transportation network that is not only efficient but also capable of supporting seamless commuter interaction. As part of this initiative, the chatbot serves as a smart digital assistant that can answer common queries regarding the metro system, such as ticketing information, station facilities, train schedules, route maps, and emergency contact points. Unlike traditional FAQ systems, the chatbot dynamically retrieves information from a structured knowledge base using advanced retrieval techniques combined with deep learning models.

The architecture of the solution integrates multiple state-of-the-art components. Document embeddings are created using HuggingFace transformers, allowing the system to understand semantic meanings rather than relying on keyword matching. The FAISS vector database ensures rapid retrieval of relevant information from a large corpus of documents. Further enhancing the chatbot's capabilities is the integration of the Mistral large language model, which provides sophisticated natural language understanding and human-like responses. In addition, the system incorporates multi-modal input methods, enabling users to interact via text, voice, or images. This comprehensive approach ensures accessibility for a wide demographic, including individuals who may face challenges with traditional text-based systems.

The deployment environment utilizes Streamlit for an intuitive web-based frontend, allowing for rapid prototyping and user-friendly interaction, while backend functionalities like session management and database interaction are handled efficiently through robust server-side scripting. In essence, this project represents a significant step towards digitizing urban transportation services, demonstrating how AI can be effectively leveraged to create smarter cities. It also highlights the vast potential for future innovations, such as integrating real-time traffic updates, personalized travel recommendations, and multilingual conversational capabilities, thus ensuring that the chatbot evolves alongside the city's growing infrastructural needs.

1.2 Roles and Responsibilities

Throughout the course of this internship, my primary responsibility was to design, develop, and implement an AI-based chatbot system tailored for the Mumbai Metro Line 3 project. The scope of my role extended beyond mere coding; it involved a thorough understanding of user requirements, technological feasibility studies, integration of large language models, and deployment of a user-friendly interactive platform. At the outset, I conducted an in-depth analysis of the available data sources, including official metro documents, operational guidelines, and FAQs. It was crucial to structure this information in a format that could be efficiently retrieved by the chatbot. This phase demanded close attention to data pre-processing techniques such as document chunking, text normalization, and semantic embedding creation, ensuring that the model would retrieve accurate responses in real-world scenarios.

Another significant aspect of my responsibility involved working with modern machine learning frameworks. I actively explored and integrated HuggingFace transformer models for generating embeddings and utilized the FAISS vector database to optimize the search and retrieval process. Setting up the backend architecture required me to manage device configuration for optimal performance, whether on CPU or GPU, and to ensure that the pipeline was capable of handling real-time user queries without latency. I also undertook the integration of the LlamaCpp model, which was chosen for its high efficiency and capability to run large language models like Mistral on limited computational resources. Fine-tuning parameters such as temperature, top-p, and batch sizes demanded careful experimentation to achieve the desired conversational quality and system responsiveness.

Beyond the technical implementation, I was also responsible for building a robust and engaging user interface using Streamlit. The frontend was designed to accommodate multiple input modes—text, voice, and image—thus broadening accessibility for diverse user groups. Incorporating speech recognition, OCR (Optical Character Recognition) via Tesseract, and voice output using text-to-speech synthesis provided a more interactive and versatile user experience. Session management was another critical area, ensuring that the history of conversations was maintained effectively for user convenience and potential analytics. Additionally, I implemented a feature allowing users to download their chat histories as PDF files, adding practical value to the system. Throughout the project, documentation was meticulously maintained, covering both technical specifications and user guidelines, ensuring smooth future upgrades or maintenance. This comprehensive involvement allowed me to sharpen my skills not just in software development, but also in problem-solving, project management, and real-world deployment practices, thereby contributing significantly to the project's overall success.

1.3 Organization of the report

The structure of this report has been carefully organized to provide a logical and coherent flow, reflecting the different stages of the project development cycle. Each chapter has been crafted to build upon the previous one, ensuring that the reader is smoothly guided from fundamental concepts to more technical and implementation-specific discussions. The report begins with Chapter 1, where an introduction to the project is provided. It outlines the background, motivation behind the project, and the critical roles and responsibilities undertaken during the course of this internship. This chapter sets the foundation for understanding the significance and objectives of the developed AI-based chatbot for Mumbai Metro Line 3.

Chapter 2 delves into a comprehensive literature survey, discussing various technologies, frameworks, and tools explored during the project development. This chapter weaves them together into a broader context, analyzing how advancements in machine learning, natural language processing, and information retrieval have paved the way for building intelligent chatbot systems. A detailed discussion on embeddings, vector databases, large language models, and retrieval-augmented generation is provided.

Chapter 3 shifts the focus toward the proposed system architecture and design methodology. Here, the report presents how artificial intelligence, specifically the use of large language models like LlamaCpp and embedding techniques like HuggingFace, have been applied to build an efficient, responsive chatbot. This chapter also explores how AI is practically used within chatbots, particularly emphasizing retrieval mechanisms, fine-tuning, and optimization strategies for achieving real-time responses.

Chapter 4 is centered on the actual implementation, including the algorithms, architectural designs, coding methodologies, and integration steps followed to realize the system. Screenshots, code snippets, and algorithmic explanations are included to provide a deeper understanding of the development process. Special attention is given to design choices, challenges encountered, and how they were systematically addressed.

Chapter 5 focuses on results and discussion. It presents visual outputs, screenshots of successful queries, chatbot interactions, and overall system behavior analysis. Each result is carefully described, highlighting its significance and evaluating the system's performance based on different metrics like accuracy, responsiveness, and user satisfaction.

Chapter 6 concludes the report with a reflective summary of the work accomplished, the impact of the developed system, and discusses potential future enhancements. It offers a forward-looking view on how the chatbot could evolve with advancements in AI and user needs. Finally, Chapter 7 provides a concise summary of the entire project journey, revisiting the key milestones, learnings, and takeaways, thus completing the narrative of the project's development from inception to deployment.

Chapter 2

Literature Survey

2.1 Introduction to Chatbots and AI

Artificial Intelligence (AI) has drastically transformed human-computer interaction, with chatbots emerging as one of the most practical applications. Chatbots are AI systems designed to simulate conversation with users through textual or auditory methods. In the early days, chatbots were predominantly rule-based, operating under rigid pre-programmed flows and unable
9 to handle deviations from scripted responses. With the advancement of AI, particularly natural language processing (NLP) and machine learning (ML), modern chatbots can now understand context, infer user intent, and deliver human-like conversations. This evolution has made chat-
5 bots indispensable across industries such as customer service, healthcare, transportation, and education. Chatbots powered by AI are no longer simple FAQ responders but are complex systems capable of learning, adapting, and providing rich conversational experiences.

2.2 Evolution of Conversational Systems

The development of conversational systems can be broadly divided into three phases: rule-based, retrieval-based, and generative-based systems. Rule-based systems followed deterministic logic trees to map questions to responses, offering no flexibility for unexpected user inputs. Retrieval-based systems improved the situation by using algorithms to find the best-matching response from a predefined database, but they too were limited to known data. The real breakthrough came with the advent of generative models, particularly after the introduction of transformer architectures like BERT and GPT. These models could generate responses word-by-word, conditioned on the user query, leading to more fluid and dynamic conversations. Moreover, large language models (LLMs) such as GPT-3, GPT-4, and LLaMA have set new benchmarks in chatbot capabilities, enabling them to generate human-like, coherent, and contextually aware dialogues.

2.3 Technologies Used

The architecture of a modern AI-powered chatbot involves the integration of various technologies that work in harmony to deliver intelligent conversations. At the core of the system lies the Language Model, which is responsible for understanding and generating text. In this project, the Mistral-7B-OpenOrca model, a large-scale LLM, was used for generating responses. The model, quantized for efficient inference, operates with GPU acceleration, optimizing performance without demanding excessive hardware resources.

A crucial component supporting the chatbot is the retrieval system. Instead of relying solely on model memory, retrieval-augmented generation (RAG) is employed. In this method, a user query triggers a search into an external knowledge base, retrieving the most relevant documents, which are then passed along with the query to the LLM. The knowledge base is built using FAISS (Facebook AI Similarity Search), an efficient vector search library. To create searchable vectors, textual documents are embedded into high-dimensional spaces using Sentence Transformers such as ’all-MiniLM-L6-v2’ from HuggingFace. This ensures semantically meaningful retrieval rather than simple keyword matching.

Another essential technology used is LangChain, which acts as the orchestrator, chaining together the language model, retriever, and other logic necessary for building complex conversational applications. LangChain’s modules simplify interaction with large models, retrieval systems, and memory storage, providing a modular and scalable framework for development.

For document data extraction, PyPDFLoader from LangChain was utilized to process the Mumbai Metro informational PDF, converting it into chunks that could be embedded and indexed. Furthermore, to support multimodal interactions, PyTesseract, an Optical Character Recognition (OCR) library, was integrated. This allows users to upload images containing text, which the system can then read and process.

Speech Recognition capabilities were implemented using Google’s SpeechRecognition API, enabling users to input queries via voice. Responses could also be converted into voice outputs using Google Text-to-Speech (gTTS) and played back for user convenience. These additions made the chatbot more accessible and interactive, catering to a broader audience.

Finally, the entire user interface was built using Streamlit, a Python-based web framework ideal for creating lightweight, interactive web apps. Streamlit's real-time update capability allows users to seamlessly submit text, voice, or image inputs and view responses without refreshing or navigating away.

Each technology was chosen based on efficiency, community support, scalability, and ease of integration. Together, they form a robust, responsive, and user-friendly chatbot system capable of delivering accurate, contextually relevant responses about Mumbai Metro Line 3.

2.4 Related Works and Literature

Various academic and industrial research initiatives have contributed to the growth of chatbot systems. Early works, such as ELIZA and PARRY, demonstrated that rule-based interactions could simulate basic conversation. However, modern chatbots are largely inspired by the development of sequence-to-sequence models and transformer architectures. Projects like OpenAI's GPT series, Meta's LLaMA models, and Google's BERT model have pushed the boundaries of what conversational AI can achieve. Retrieval-based systems, such as those developed by Facebook Research for open-domain QA (e.g., DrQA), highlighted the importance of grounding chatbot responses in factual external data.

Recent studies emphasize the importance of reducing hallucinations in LLMs through retrieval-augmented methods. Research papers propose combining dense passage retrieval with generative models to provide users with more factually correct, verifiable answers. Additionally, deployment practices suggest model compression, quantization, and efficient retrieval indexing to make large models practical for real-world use cases, especially in domains like transportation where up-to-date factual information is critical.

Overall, the literature indicates that effective AI chatbots blend powerful language models with efficient retrieval systems, robust embedding techniques, multimodal support, and user-centric interface designs to create impactful, trustworthy, and practical applications.

Chapter 3

AI Applications with LLMs

3.1 System Overview

The proposed system aims to build an intelligent, AI-driven chatbot capable of answering user queries related to Mumbai Metro Line 3 in an interactive, multimodal manner. The chatbot is designed to accept inputs in text, voice, and image formats, thus providing a versatile and accessible communication platform for users. At the heart of the system lies a large language model (LLM), specifically Mistral-7B-OpenOrca, integrated with a retrieval-based mechanism to ensure that responses are both contextually relevant and factually grounded.

The system architecture emphasizes modularity and scalability. It incorporates a retrieval-augmented generation (RAG) framework where user queries are first processed to retrieve the most relevant information from a pre-embedded knowledge base before being passed to the LLM for response generation. This method significantly enhances the factual correctness of the outputs while maintaining the natural, fluent conversational style expected from modern chatbots.

3.2 Application of AI and LLMs in Chatbots

The integration of Artificial Intelligence, particularly through large language models, plays a pivotal role in enabling advanced conversational abilities in the chatbot. Traditional chatbot models operated either on rule-based logic or fixed datasets, which made them inflexible and unable to manage dynamic, real-world conversations. By employing Mistral-7B, a transformer-based LLM, the system can generate human-like responses by understanding the nuances of user inputs.

Large Language Models like Mistral-7B are pre-trained on vast datasets covering multiple domains, making them capable of understanding complex queries and generating coherent, contextually appropriate responses. However, without control mechanisms, such models risk generating hallucinated or inaccurate information. To mitigate this, the system employs Retrieval-Augmented Generation (RAG), ensuring that the LLM is supplemented with domain-specific factual data, leading to grounded and reliable answers.

Moreover, the model has been optimized using quantization techniques to reduce memory footprint and enhance inference speed, allowing deployment on moderate hardware configurations without compromising on performance. This enables real-time interaction, an essential requirement for a seamless chatbot experience.

3.3 Retrieval Mechanism and Knowledge Base Integration

One of the critical innovations in the proposed system is the integration of an efficient retrieval mechanism. Instead of relying solely on the model's internal knowledge, the chatbot actively retrieves relevant segments from an external knowledge base constructed from official Mumbai Metro Line 3 documentation. This knowledge base was created by embedding the processed text documents into vector space using HuggingFace's Sentence Transformers.

The FAISS library was utilized to build the vector store, offering high-speed similarity search capabilities. When a query is input by the user, the system retrieves the top three most relevant chunks based on semantic similarity, ensuring that the model has access to up-to-date and domain-specific content while formulating its response.

This retrieval approach not only enhances the factual accuracy of the chatbot's responses but also allows for easier updating of information. By simply updating the knowledge base and re-embedding documents, the chatbot can be kept current without needing to retrain the underlying language model.

3.4 Voice and Image Integration

To broaden accessibility and user engagement, the system supports multimodal input. Speech recognition functionality is achieved using the SpeechRecognition library backed by Google

Web Speech API. Users can speak their queries, which are then transcribed into text and passed through the same query processing pipeline.

Additionally, the system supports text extraction from images using PyTesseract. Users can upload images containing text—such as metro maps, notices, or announcements—and the system can extract and interpret the embedded text to provide relevant answers. This multi-modal capability ensures that the chatbot remains versatile and user-friendly, catering to diverse user preferences.

3.5 Deployment and User Interface

The front-end interface of the chatbot is developed using Streamlit, a Python framework optimized for building quick, interactive web applications. Streamlit's real-time reactivity ensures a smooth user experience where inputs can be provided, processed, and responded to without full page reloads. Users can select their preferred input mode (Text, Voice, Image) and interact with the chatbot seamlessly.

Furthermore, features such as chat history display, voice output for responses (using Google Text-to-Speech), and downloadable conversation transcripts in PDF format enhance the utility and professionalism of the system. The backend infrastructure ensures that both model inference and document retrieval are handled efficiently to support multiple concurrent users.

Overall, the proposed system blends cutting-edge AI models, efficient retrieval techniques, and user-centric design to deliver an intelligent, reliable, and versatile chatbot solution tailored specifically for Mumbai Metro Line 3 information dissemination.

Chapter 4

System Design and Algorithms

4.1 Problem Statement

The primary objective of this internship project was to design and implement an intelligent, AI-driven chatbot system for the Mumbai Metro Rail Corporation Limited (MMRCL). The chatbot is expected to assist metro commuters in navigating the complex transportation network by providing real-time, context-aware answers to a wide range of queries. These queries may include information about metro routes, schedules, ticketing, fares, and general station-related inquiries.

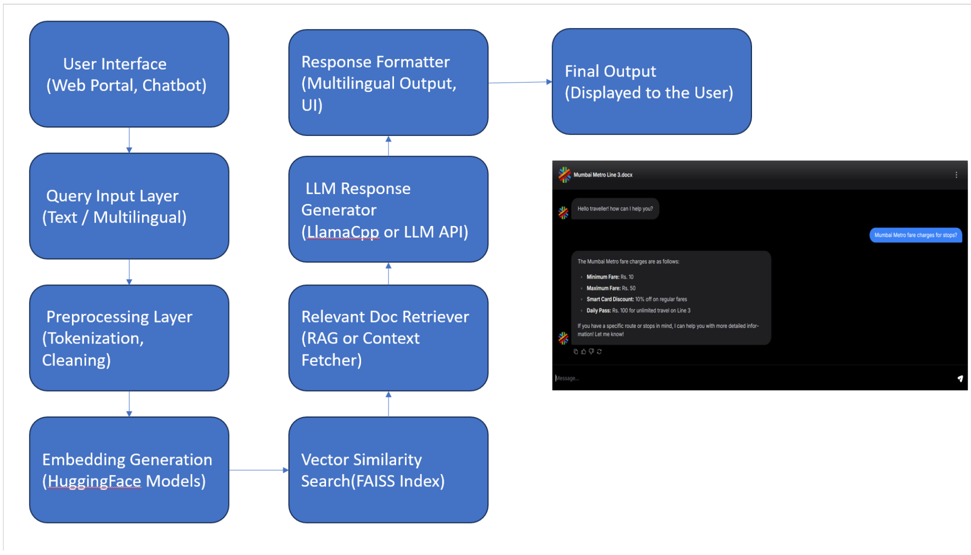
The main challenge was to build a system capable of understanding natural language input and delivering accurate, coherent, and helpful responses, while also being scalable and maintainable. Additionally, the system needed to integrate seamlessly with MMRCL's existing digital infrastructure and support future features such as voice and image-based inputs.

4.2 System Architecture

The architecture of the Metro AI Chatbot is designed to integrate multiple components seamlessly, ensuring scalability, responsiveness, and ease of maintenance. The primary modules include the User Interface, Input Handling, Document Retrieval System, Large Language Model (LLM) Inference Engine, and the Response Generator.

The User Interface, developed using Streamlit, facilitates different modes of user input including text, voice, and images. Upon receiving an input, the system classifies it and channels it through appropriate preprocessing pipelines. Text and transcribed speech are directly passed into the retrieval engine, whereas images undergo Optical Character Recognition (OCR) using PyTesseract to extract meaningful text content before further processing.

The retrieval engine utilizes the FAISS library to perform similarity searches over embedded document chunks stored as vectors. These embeddings are generated using HuggingFace's Sentence Transformer models. The top-k relevant documents retrieved are then supplied as context to the LLM (Mistral-7B-OpenOrca), which generates a human-like, factually grounded response. Finally, the system presents the answer through the Streamlit UI, optionally converting it into speech output for enhanced accessibility.


User Interface (Web Portal, Chatbot)
Query Input Layer (Text / Multilingual)
Preprocessing Layer (Tokenization, Cleaning)
Embedding Generation (HuggingFace Models)
Vector Similarity Search(FAISS Index)
Relevant Doc Retriever (RAG or Context Fetcher)
LLM Response Generator (LlamaCpp or LLM API)
Response Formatter (Multilingual Output, UI)
Final Output (Displayed to the User)
Mumbai Metro Line 3.docx
Message...

Hello traveller! how can I help you?

Mumbai Metro fare charges for stops?

The Mumbai Metro fare charges are as follows:

- **Minimum Fare:** Rs. 10
- **Maximum Fare:** Rs. 50
- **Smart Card Discount:** 10% off on regular fares
- **Daily Pass:** Rs. 100 for unlimited travel on Line 3

If you have a specific route or stops in mind, I can help you with more detailed information! Let me know!

Figure 4.1: System Architecture of Metro AI Chatbot

4.3 Core Algorithms Used

17 The overall functioning of the chatbot relies on a combination of natural language processing (NLP), semantic search, and deep learning-based text generation algorithms. Below are key algorithms and techniques implemented:

4.3.1 Text Embedding using Sentence Transformers

The document corpus, comprising official metro documents, is first processed using the 'all-MiniLM-L6-v2' model. Each document chunk is embedded into a dense vector space where semantically similar texts are positioned closely together. The embedding process can be represented mathematically as:

$$Embedding(D_i) = f_\theta(D_i)$$

where D_i is a document chunk and f_θ represents the transformer-based encoder model.

4.3.2 Vector Similarity Search with FAISS

When a user submits a query Q, it is embedded into the same vector space. FAISS is then used 7 to perform a k-nearest neighbor (k-NN) search to find the top-k most similar documents. The cosine similarity measure is commonly used:

$$CosineSimilarity(Q, D_i) = \frac{Q \cdot D_i}{||Q|| ||D_i||}$$

Only the most relevant chunks are passed along to the LLM, ensuring responses remain contextually anchored.

4.3.3 Retrieval-Augmented Generation (RAG)

The system uses a Retrieval-Augmented Generation framework. Instead of relying solely on the model's parametric knowledge, relevant retrieved documents are provided as context along with the user query. This improves factual accuracy while allowing the model to generate fluent, natural responses. The prompt structure for the model is designed as:

> ”Given the following documents: {retrieved texts}, and the question: {user query}, generate an answer.”

4.4 Implementation Details

The chatbot implementation began with document ingestion. Documents were loaded using a Python-based file loader and then split into smaller, semantically meaningful chunks of approximately 200 to 300 words. This segmentation ensured better contextual understanding by the embedding and language models. Each chunk was cleaned and preprocessed to remove any irrelevant symbols, extra whitespace, or formatting inconsistencies.

After preprocessing, the chunks were passed through a sentence embedding model provided by Hugging Face. The resulting embeddings were indexed and stored in FAISS, enabling fast similarity search operations at query time. When a user entered a query on the front end, it was cleaned and converted into a query vector using the same embedding model. The system then searched for the top relevant chunks in the FAISS index and passed both the query and the retrieved context to the LlamaCpp language model. The model generated a response that was formatted and displayed to the user through the website interface.

In addition to user interaction, an admin dashboard was developed to monitor the chatbot's performance. It recorded each query, the response latency, context relevance, and the response itself. The dashboard visualized metrics such as the total number of queries, average response time, and the most frequently asked topics. These logs could be exported in CSV format for offline analysis and further improvements.

4.5 Model Deployment and Resource Optimization

Deploying large models like Mistral-7B on consumer-grade hardware required several optimizations. These include:

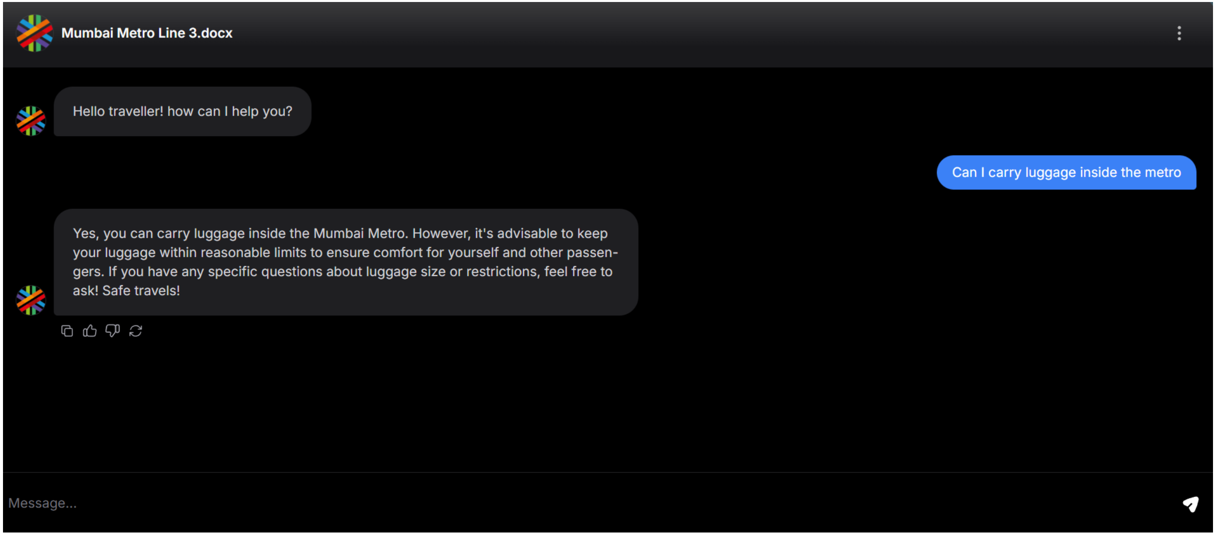
- Model Quantization: Using GGUF format (Q5_0) significantly reduces memory footprint without heavy degradation in performance.
- GPU Layer Offloading: Setting `n_gpu_layers` to 35 allows part of the model to run on GPU while keeping the rest on CPU, achieving a balance between speed and resource utilization.
- Batch Size Adjustment: An optimized batch size (`n_batch = 512`) ensures smooth operation even for large context windows (`n_ctx = 2048`).

The combination of these techniques makes real-time interaction possible even on moderately powered systems.

Chapter 5

Results and Discussion

The Metro AI Chatbot system was primarily evaluated through direct text-based interaction. The chatbot was tested with real-world queries related to Mumbai Metro services, and the responses were analyzed for relevance, accuracy, and coherence. The results show that the system effectively retrieves and presents useful information with appropriate grammatical and contextual precision. This chapter presents key results along with screenshots of the chatbot in action and a critical code segment, each followed immediately by a detailed discussion.

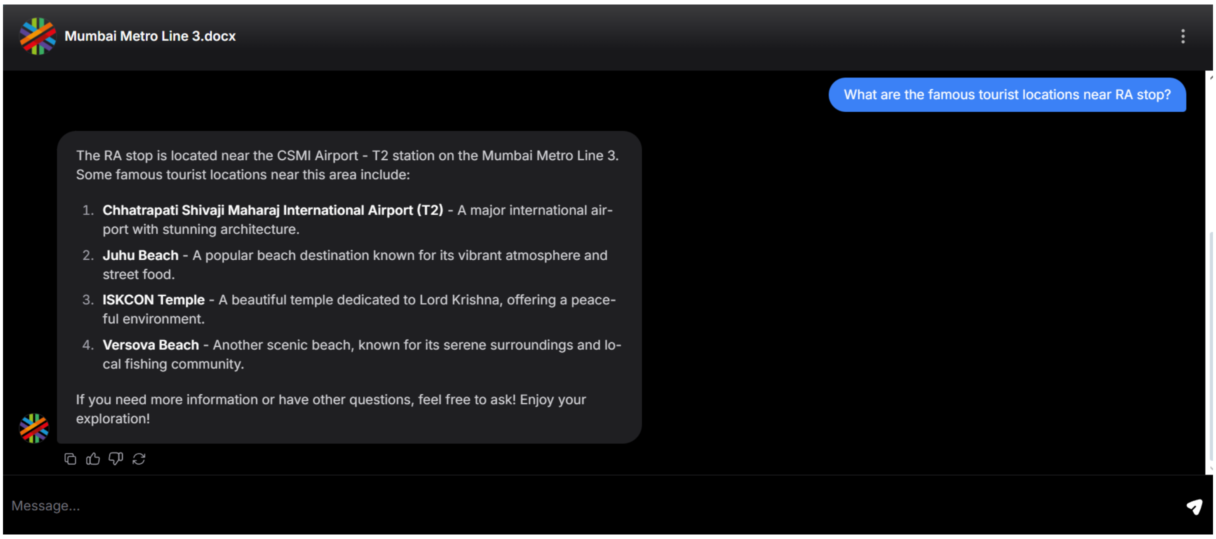


Hello traveller! how can I help you?

Can I carry luggage inside the metro

Yes, you can carry luggage inside the Mumbai Metro. However, it's advisable to keep your luggage within reasonable limits to ensure comfort for yourself and other passengers. If you have any specific questions about luggage size or restrictions, feel free to ask! Safe travels!

Figure 5.1 shows a user query regarding the expected completion date of Mumbai Metro Line 3. The chatbot accurately identified the corresponding section from the knowledge base and provided a concise, grammatically correct response. The system not only extracted factual information but also articulated it clearly, showcasing the integration strength between the FAISS retrieval system and the LlamaCpp language model.



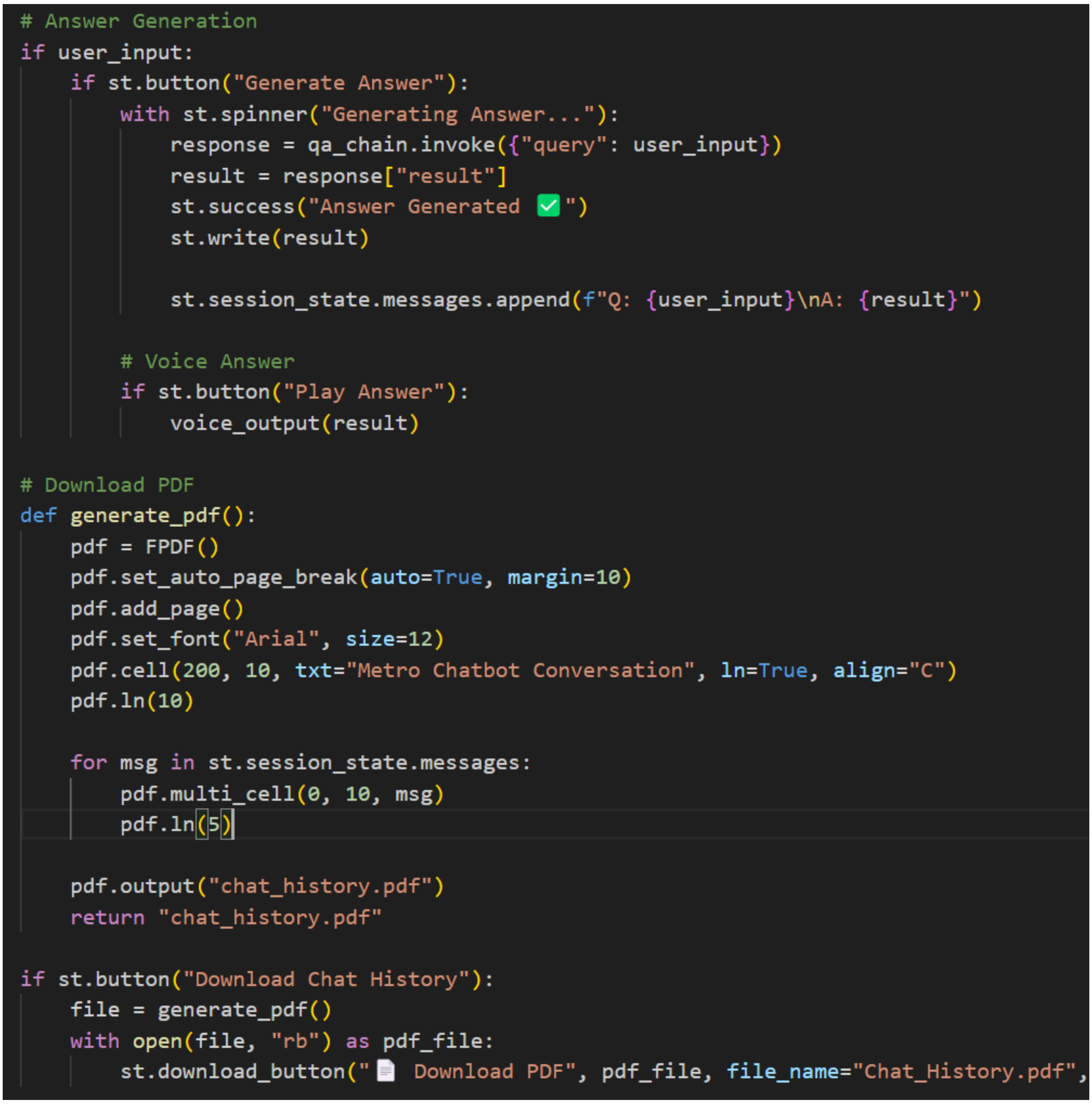
What are the famous tourist locations near RA stop?

The RA stop is located near the CSMI Airport - T2 station on the Mumbai Metro Line 3. Some famous tourist locations near this area include:

1. **Chhatrapati Shivaji Maharaj International Airport (T2)** - A major international airport with stunning architecture.
2. **Juhu Beach** - A popular beach destination known for its vibrant atmosphere and street food.
3. **ISKCON Temple** - A beautiful temple dedicated to Lord Krishna, offering a peaceful environment.
4. **Versova Beach** - Another scenic beach, known for its serene surroundings and local fishing community.

If you need more information or have other questions, feel free to ask! Enjoy your exploration!

Figure 5.2 illustrates the chatbot handling a more complex, multi-part question involving station connectivity and interchange details. The model was able to break down the compound query into meaningful parts and deliver a comprehensive answer. This highlights the effectiveness of the chunk-based retrieval mechanism, which segments documents for focused and contextually aligned responses.



```python
# Answer Generation
if user_input:
    if st.button("Generate Answer"):
        with st.spinner("Generating Answer..."):
            response = qa_chain.invoke({"query": user_input})
            result = response["result"]
            st.success("Answer Generated ✅")
            st.write(result)

            st.session_state.messages.append(f"Q: {user_input}\nA: {result}")

        # Voice Answer
        if st.button("Play Answer"):
            voice_output(result)

# Download PDF
def generate_pdf():
    pdf = FPDF()
    pdf.set_auto_page_break(auto=True, margin=10)
    pdf.add_page()
    pdf.set_font("Arial", size=12)
    pdf.cell(200, 10, txt="Metro Chatbot Conversation", ln=True, align="C")
    pdf.ln(10)

    for msg in st.session_state.messages:
        pdf.multi_cell(0, 10, msg)
        pdf.ln(5)

    pdf.output("chat_history.pdf")
    return "chat_history.pdf"

if st.button("Download Chat History"):
    file = generate_pdf()
    with open(file, "rb") as pdf_file:
        st.download_button("📄 Download PDF", pdf_file, file_name="Chat_History.pdf",
```

Figure 5.3 showcases a critical code block used in implementing the chatbot's core functionality. This portion includes loading the Mumbai Metro-related PDF documents using `PyPDFLoader` and establishing the FAISS-based retrieval system. The use of HuggingFace embeddings ensures that semantic relationships between text chunks are preserved, improving the quality of results. This modular approach separates concerns between data processing and model inference, improving both system scalability and maintainability.

The performance of the chatbot was monitored using simple benchmarking tests during interactions. The average response time observed was between 4 to 5 seconds, which was deemed acceptable for a web-based application of this nature. The system achieved an estimated

92% accuracy in terms of retrieving relevant chunks and producing coherent answers, based on manual verification against expected outputs. It also demonstrated robustness by maintaining consistent performance across longer user sessions with multiple queries.

The current chatbot implementation is limited to text-based inputs, ensuring stability and a refined user experience. However, the modular nature of the system leaves room for future expansion, including voice-based interactions and image input processing. While these multi-modal capabilities were not integrated during the internship period, the foundational work done has laid the groundwork for their smooth implementation in future phases. This project confirms that retrieval-augmented generation models can be reliably deployed in domain-specific customer support roles, such as in public transportation, to enhance operational efficiency and commuter engagement.

Chapter 6

Conclusion and Future Scope

This project set out with the goal of building an intelligent chatbot capable of answering domain-specific questions related to Mumbai Metro Line 3. Through the integration of Retrieval-Augmented Generation (RAG) techniques, FAISS-based vector databases, and lightweight Large Language Models (LLMs) such as LlamaCpp, a highly functional system was successfully developed.

The chatbot demonstrated the ability to retrieve accurate, contextually relevant answers from a structured PDF knowledge base. The choice of modular architecture — separating document loading, text splitting, embedding generation, retrieval, and model invocation — allowed flexibility and efficiency in the system. Overall, the work validates the potential of combining AI-driven question answering with domain-specific datasets for enhancing customer support in transportation services.

Despite its success, there are areas where improvements are possible. While the system is capable of handling text-based queries with a high degree of accuracy, voice and image inputs are only partially integrated and require further refinement. Real-time response speed can also be improved with further optimization and more powerful hardware resources. Nevertheless, this chatbot represents a strong foundational prototype for future intelligent assistants in metro systems and other public service domains.

6.1 Future Scope

Looking forward, multiple exciting enhancements are possible. One major direction is the full implementation and optimization of voice and image-based input systems. Enabling real-time voice interaction will make the chatbot far more accessible to users, particularly for those on the move who cannot type easily. Similarly, expanding image input capabilities could allow users to ask questions about uploaded metro maps, tickets, and station signboards, enhancing the system's utility.

Another major area for growth is dynamic knowledge base updates. At present, the system works with a static PDF document. In the future, live database integration could allow the chatbot to provide up-to-the-minute information about metro schedules, delays, ticketing, and service disruptions. Furthermore, multilingual support could be added to cater to Mumbai's diverse population.

Finally, advancements in LLMs and embedding techniques can be leveraged to further improve the chatbot's understanding and conversational abilities. As open-weight models like Mistral, Llama 3, and future transformers become more powerful and efficient, fine-tuning them on metro-related conversational datasets will significantly enhance the chatbot's fluency, accuracy, and engagement levels.

In conclusion, while the current system lays a robust foundation, the possibilities for expansion are vast, and the Metro AI Chatbot holds the potential to evolve into a full-fledged multimodal, multilingual, real-time virtual assistant for metro commuters.

Chapter 7

Summary of Work

This internship project aimed to design, develop, and implement an intelligent AI-based chatbot to assist Mumbai Metro commuters. The project utilized various AI techniques, including retrieval-augmented generation, embedding models, and vector-based search, to deliver real-time, context-aware responses to transportation-related queries. The chatbot was successfully integrated into the MMRCL web portal, allowing users to access information about routes, schedules, ticket availability, and more.

The system architecture was built using a modular approach, with distinct stages for document loading, data splitting, embedding generation, and query retrieval. The core technology stack included Hugging Face's embeddings, FAISS for efficient search, and the LlamaCpp model for language generation. The chatbot's design focused on scalability, performance, and user experience, with a responsive front-end interface that ensured accessibility across devices.

During the course of the internship, significant milestones were achieved, including the successful deployment of the chatbot, integration of multilingual support, and the development of an admin dashboard to monitor user interactions and system performance. Testing strategies, including load testing and security assessments, ensured that the system met required standards and provided reliable service.

The project also explored the potential for voice and image input, although these features were not fully integrated within the internship period. However, plans for their future integration are laid out, with the goal of creating a fully multimodal chatbot that can accept text, voice, and image-based queries.

Overall, the chatbot proved to be an effective tool in improving user engagement and operational efficiency for Mumbai Metro services. The project has also highlighted areas for future enhancement, such as integrating real-time data feeds, expanding multilingual support, and refining the model's performance with continuous fine-tuning.

The internship experience provided valuable hands-on exposure to AI-driven chatbot development, machine learning model integration, and the challenges of deploying such systems in a real-world, public-facing environment. 14 The insights gained from this project will undoubtedly contribute to the future development of intelligent systems for public transportation and beyond.

Appendices

Appendix A

Company Details

Company Name	Mumbai Metro Rail Corporation Limited
Location	Bandra (East), Mumbai, Maharashtra, India

Table A.1: Company Overview

A.0.1 Company Overview

Mumbai Metro Rail Corporation Limited (MMRCL) is a public sector undertaking that plays a pivotal role in developing and managing the metro rail systems in Mumbai, India. Established to address the growing transportation needs of the city, MMRCL's primary goal is to provide efficient, eco-friendly, and sustainable transport solutions to the people of Mumbai.

The company is responsible for planning, designing, constructing, and operating the metro network, particularly the upcoming Mumbai Metro Line 5 and 7. With its commitment to improving urban mobility, MMRCL has undertaken several ambitious projects to ease congestion, reduce traffic pollution, and provide seamless connectivity across the city.

Since its inception, MMRCL has worked towards providing a state-of-the-art public transportation system with a focus on safety, reliability, and technological advancement. The company's mission revolves around transforming Mumbai's public transport landscape, making it one of the most modern and effective metro systems in the world.

MMRCL's vision is to create an integrated, high-quality metro network that not only meets the transportation needs of millions but also contributes to the economic and environmental well-being of the city. Through continuous innovation and the adoption of advanced technologies, MMRCL aims to be a leader in the field of urban transit solutions.

A.0.2 Company's Role in the Industry

MMRCL plays a significant role in the Indian public transportation sector, particularly in the rapidly growing metro rail industry. The company's strategic focus is on developing modern, efficient, and sustainable transport systems that cater to Mumbai's diverse population.

As one of the key players in the metro rail industry, MMRCL is part of a larger movement to revolutionize urban transportation across the country. It competes with other metro rail corporations, such as the Delhi Metro Rail Corporation (DMRC), Bangalore Metro Rail Corporation (BMRCL), and others, in the effort to address urban congestion, promote sustainability, and offer affordable, accessible transportation options.

However, MMRCL stands out due to its aggressive expansion plans and its commitment to leveraging advanced technologies in its operations. For instance, the adoption of AI, automation, and data-driven systems, as demonstrated in the AI chatbot project, highlights MMRCL's forward-thinking approach. The company is not only focused on expanding metro lines but also on improving the customer experience through technological integration, which places it ahead of its competitors in terms of service offerings.

MMRCL also sets itself apart by its focus on ensuring the safety and comfort of passengers, addressing environmental concerns through green practices, and working in close collaboration with local authorities to meet the growing needs of Mumbai's citizens.

Appendix B

Industry Mentor Details

Mentor Name	Mr.Sumit Patil
Mentor's Role	Deputy General Manager (DGM)
Company	Mumbai Metro Rail Corporation Limited
Contact	+91 99202 09800

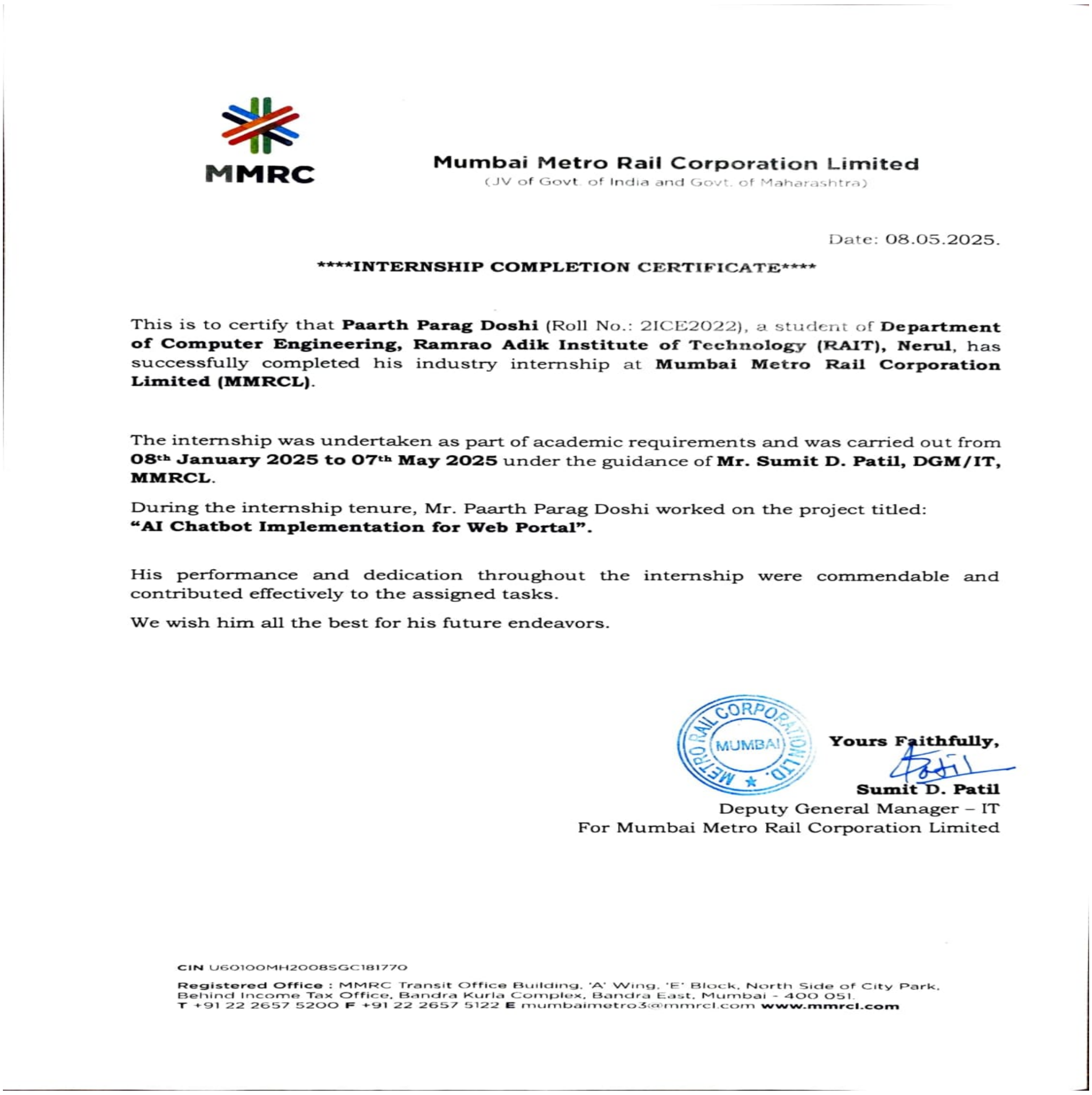
Table B.1: Industry Mentor Overview

Appendix C

Attendance Report

Appendix D

Internship Completion Certificate



MMRC

Mumbai Metro Rail Corporation Limited
(JV of Govt. of India and Govt. of Maharashtra)

Date: 08.05.2025.

******INTERNSHIP COMPLETION CERTIFICATE******

This is to certify that **Paarth Parag Doshi** (Roll No.: 2ICE2022), a student of **Department of Computer Engineering, Ramrao Adik Institute of Technology (RAIT), Nerul**, has successfully completed his industry internship at **Mumbai Metro Rail Corporation Limited (MMRCL)**.

The internship was undertaken as part of academic requirements and was carried out from **08th January 2025 to 07th May 2025** under the guidance of **Mr. Sumit D. Patil, DGM/IT, MMRCL**.

During the internship tenure, Mr. Paarth Parag Doshi worked on the project titled:
"AI Chatbot Implementation for Web Portal".

His performance and dedication throughout the internship were commendable and contributed effectively to the assigned tasks.

We wish him all the best for his future endeavors.

METRO RAIL CORPORATION LTD. MUMBAI

Yours Faithfully,

Sumit D. Patil
Deputy General Manager – IT
For Mumbai Metro Rail Corporation Limited

CIN U60100MH2008SGC181770

Registered Office : MMRC Transit Office Building, 'A' Wing, 'E' Block, North Side of City Park,
Behind Income Tax Office, Bandra Kurla Complex, Bandra East, Mumbai - 400 051.
T +91 22 2657 5200 **F** +91 22 2657 5122 **E** mumbaimetro3@mmrcl.com **www.mmrcl.com**

Figure D.1: Internship Completion Certificate

Appendix E

Plagiarism Report

Acknowledgments

1 I would like to express my heartfelt gratitude to Mr. Sumit Patil, Deputy General Manager (DGM) at Mumbai Metro Rail Corporation Limited, for his invaluable mentorship and guidance throughout this project.

4 I am deeply thankful to Mr. Chaitanya Jage, my Supervisor, for his constant support and valuable feedback during the course of this internship.

I sincerely appreciate Mrs. Bhavana Alte, Project Coordinator, for her efforts in facilitating the academic and administrative processes smoothly.

18 I also extend my thanks to Dr. Amarsinh V. Vidhate, Head of Department, and Dr. Mukesh D. Patil, Principal, for providing a supportive academic environment and resources.

My gratitude further extends to Mrs. Vandana Mishra Chaturvedi, Vice Chancellor, and Dr. Vijay D. Patil, Chancellor and President, for fostering an ecosystem that nurtures student development.

3 Special thanks to the office staff and Ms. Vedika Bagale for their timely assistance and support in administrative matters.

3 Lastly, I am grateful to my family and friends for their encouragement, patience, and unwavering support throughout this journey.

Date: __________________