VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI – 590018, Karnataka

INTERNSHIP REPORT

ON

“LLM Wrapper”

Submitted in partial fulfilment for the award of degree(21)

BACHELOR OF ENGINEERING IN
Artificial Intelligence & Machine Learning

Submitted by:

HAREKA G D 1BI21AI021

Conducted at
COMPSOFT TECHONOGIES

BANGALORE INSTITUTE OF TECHNOLOGY

Department of Artificial Intelligence & Machine Learning

Accredited by NBA, New Delhi

K.R. Road, V.V. Pura, Bengaluru-560 004

BANGALORE INSTITUTE OF TECHNOLOGY

Department of Artificial Intelligence & Machine Learning

Accredited by NBA, New Delhi

K.R. Road, V.V. Pura, Bengaluru-560 004

CERTIFICATE

This is to certify that the Internship titled **"LLM Wrapper"** carried out by **Ms. Hareka G.D.,** a bonafide student of Bangalore Institute of Technology, in partial fulfillment for the award of **Bachelor of Engineering**, in **BRANCH** under Visvesvaraya Technological University, Belagavi, during the year 2022-2023. It is certified that all corrections/suggestions indicated have been incorporated in the report.

The project report has been approved as it satisfies the academic requirements in respect of Internship prescribed for the course Internship / Professional Practice (21CSI85)

Signature of Guide **Signature of HOD** **Signature of Principal**

External Viva:

Name of the Examiner Signature with Date

1)__________________
____

2)__________________
____

DECLARATION

I, **Hareka G D**, final year student of Branch, College Name - 560 082, declare that the Internship has been successfully completed, in **Compsoft Technologies**. This report is submitted in partial fulfillment of the requirements for award of Bachelor Degree in Branch name, during the academic year 2022-2023.

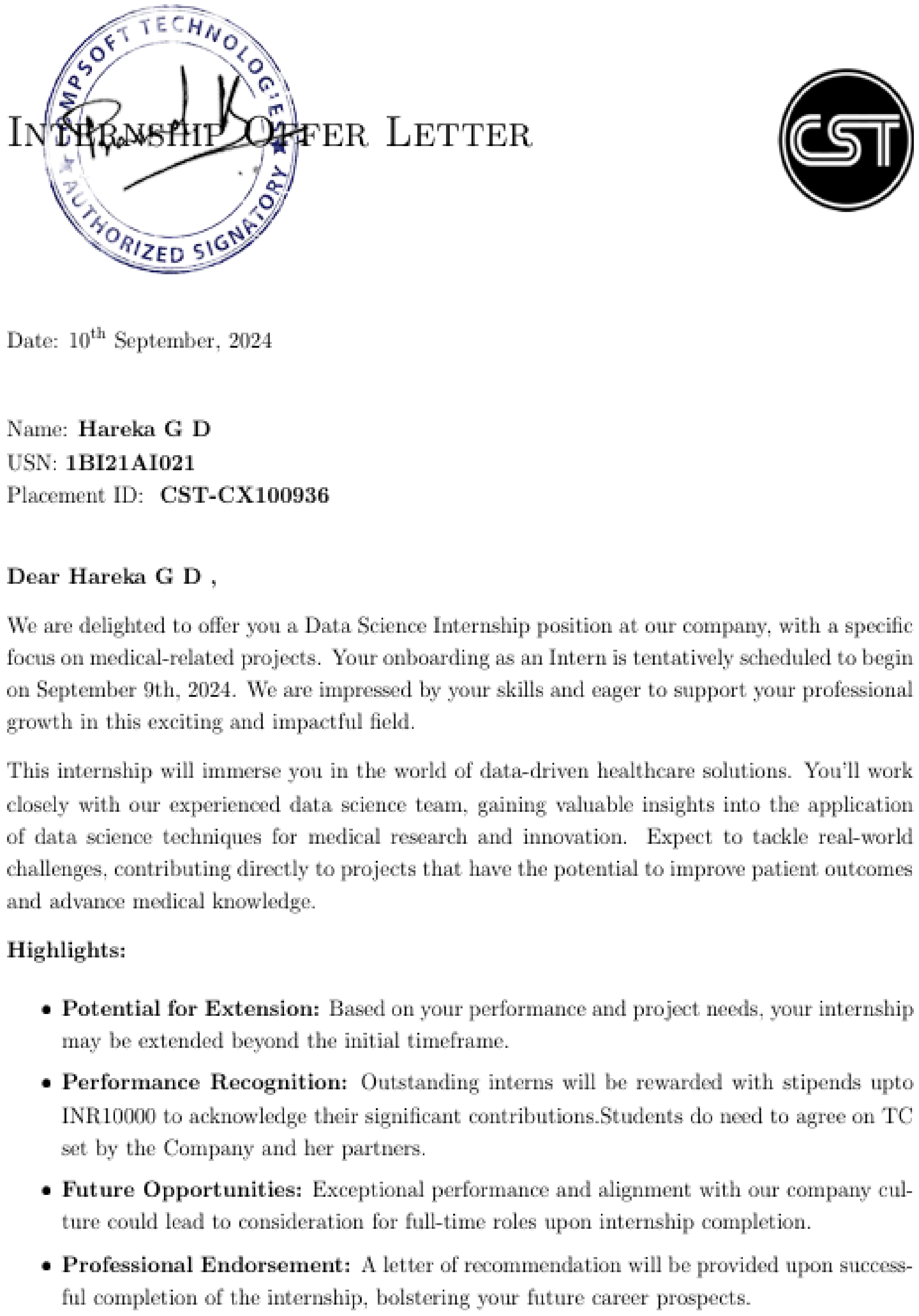
Date : 29-01-2025 :

Place : Bangalore

USN : 1BI21AI021

NAME : HAREKA G D

OFFER LETTER



Internship Offer Letter

MPSOFT TECHNOLOGIE
AUTHORIZED SIGNATORY

Date: 10th September, 2024

Name: **Hareka G D**
USN: **1BI21AI021**
Placement ID: **CST-CX100936**

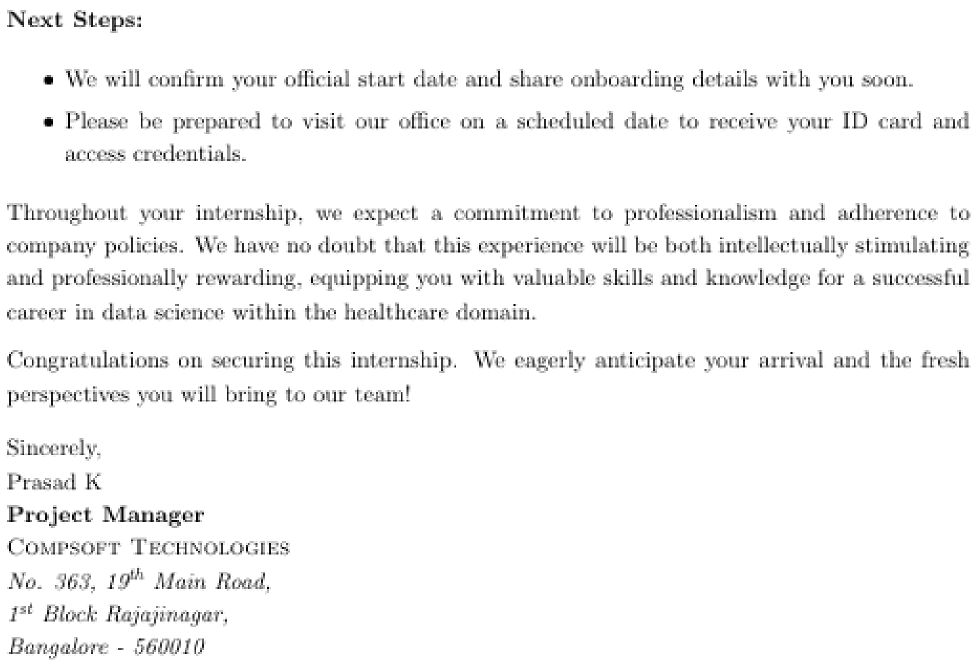
Dear Hareka G D ,

We are delighted to offer you a Data Science Internship position at our company, with a specific focus on medical-related projects. Your onboarding as an Intern is tentatively scheduled to begin on September 9th, 2024. We are impressed by your skills and eager to support your professional growth in this exciting and impactful field.

This internship will immerse you in the world of data-driven healthcare solutions. You'll work closely with our experienced data science team, gaining valuable insights into the application of data science techniques for medical research and innovation. Expect to tackle real-world challenges, contributing directly to projects that have the potential to improve patient outcomes and advance medical knowledge.

Highlights:

- **Potential for Extension:** Based on your performance and project needs, your internship may be extended beyond the initial timeframe.
- **Performance Recognition:** Outstanding interns will be rewarded with stipends upto INR10000 to acknowledge their significant contributions.Students do need to agree on TC set by the Company and her partners.
- **Future Opportunities:** Exceptional performance and alignment with our company culture could lead to consideration for full-time roles upon internship completion.
- **Professional Endorsement:** A letter of recommendation will be provided upon successful completion of the internship, bolstering your future career prospects.



Next Steps:

- We will confirm your official start date and share onboarding details with you soon.
- Please be prepared to visit our office on a scheduled date to receive your ID card and access credentials.

Throughout your internship, we expect a commitment to professionalism and adherence to company policies. We have no doubt that this experience will be both intellectually stimulating and professionally rewarding, equipping you with valuable skills and knowledge for a successful career in data science within the healthcare domain.

Congratulations on securing this internship. We eagerly anticipate your arrival and the fresh perspectives you will bring to our team!

Sincerely,
Prasad K
Project Manager
Compsoft Technologies
No. 363, 19th Main Road,
1st Block Rajajinagar,
Bangalore - 560010

ACKNOWLEDGEMENT

This Internship is a result of accumulated guidance, direction and support of several important persons. We take this opportunity to express our gratitude to all who have helped us to complete the Internship.

We express our sincere thanks to our Principal, **Dr. Aswath M.U.,** for providing usadequate facilities to undertake this Internship.

We would like to thank our Head of Dept, **Dr Kempanna M**, Dept. of Artificial Intelligence and Machine Learning, for providing us an opportunity to carry out Internship and for his valuable guidance and support.

We would like to thank **Mr**. **Suhas**, Compsoft Technologies for guiding us during the period of internship.

We express our deep and profound gratitude to our guide, **Dr. Kempanna M**, Professor and Head, for his keen interest and encouragement at every step in completing the Internship.

We would like to thank all the faculty members of our department for the support extended during the course of Internship.

We would like to thank the non-teaching members of our dept, for helping us during the Internship.

Last but not the least, we would like to thank our parents and friends without whose constant help, the completion of Internship would have not been possible.

HAREKA G D
1BI21AI021

ABSTRACT

This project presents a multi-LLM chatbot designed using **Streamlit** and **Replicate**, enabling users to interact with different large language models (LLMs) for diverse conversational experiences. The chatbot integrates **Meta Llama 3 (70B Instruct)**, **Mistral 7B Instruct**, and **Gemma 2B (Instruction-Tuned)**, randomly selecting a model to generate responses. Users can input queries, adjust parameters such as **temperature** and **top-p**, and receive AI-generated responses in real time. The system includes **session management**, **token limit enforcement**, and an interactive **chat history** feature. Future improvements include allowing manual model selection, better error handling, and UI enhancements for an optimized user experience. This chatbot demonstrates the potential of multi-model AI-driven interactions, offering flexibility in response quality and creativity

Table of Contents

CHAPTER 1

COMPANY PROFILE

1. COMPANY PROFILE

A Brief History of Compsoft Technologies

Compsoft Technologies, was incorporated with a goal ”To provide high quality and optimal Technological Solutions to business requirements of our clients”. Every business is a different and has a unique business model and so are the technological requirements. They understand this and hence the solutions provided to these requirements are different as well. They focus on clients requirements and provide them with tailor made technological solutions. They also understand that Reach of their Product to its targeted market or the automation of the existing process into e-client and simple process are the key features that our clients desire from Technological Solution they are looking for and these are the features that we focus on while designing the solutions for their clients.

Compsoft Technologies, strive to be the front runner in creativity and innovation in software development through their well-researched expertise and establish it as an out of the box software development company in Bangalore, India. As a software development company, they translate this software development expertise into value for their customers through their professional solutions.

They understand that the best desired output can be achieved only by understanding the clients demand better. Compsoft Technologies work with their clients and help them todefine their exact solution requirement. Sometimes even they wonder that they have completely redefined their solution or new application requirement during the brainstormingsession, and here they position themselves as an IT solutions consulting group comprising of high caliber consultants.

They believe that Technology when used properly can help any business to scale and achieve new heights of success. It helps Improve its efficiency, profitability, reliability; to put itin one sentence ” Technology helps you to Delight your Customers” and that is what we wantto achieve.

CHAPTER 2

ABOUT THE COMPANY

2. ABOUT THE COMPANY

In an era defined by the relentless race for digital transformation, Compsoft Technologies stands at the forefront of empowering businesses to thrive in a globally connected, on-demand world. As industries—from healthcare, finance, and retail to energy, gaming, and transportation—scramble to build event-driven, real-time applications that capture market opportunities, Compsoft bridges the gap between cutting-edge innovation and practical implementation. Specializing in research-based, actionable data and machine learning solutions, the company pioneers unexplored domains, delivering data-supported insights that guide startups and investment firms toward strategic decision-making.

Compsoft's expertise extends beyond theoretical exploration. With a proven track record in building scalable, real-time systems and world-class consulting processes, the company equips clients with tools to operationalize innovation efficiently. By integrating advanced machine learning models with event-driven architectures, Compsoft enables businesses to unlock transformative capabilities—whether enhancing financial reporting, optimizing operational performance, or deploying AI-driven medical research platforms. Their end-to-end approach combines rigorous Theoretical Analysis, bespoke Model Development, and granular Market Research to ensure solutions are not only technologically robust but also aligned with global trends and user demands.

Driven by a mission to simplify the development and deployment of real-time applications, Compsoft delivers unparalleled cost-benefits, consistent project execution, and measurable business outcomes. Their work spans strategic goal-setting, performance management, and financial optimization, all while adhering to strict confidentiality protocols through NDAs and proprietary safeguards. For clients and interns alike, Compsoft offers a dynamic platform to shape the future of industries, merging exploratory research with scalable, real-world impact in the fast-evolving landscape of data science and digital innovation.

CHAPTER 3

INTRODUCTION

3. INTRODUCTION

Introduction to Data Science

Data science is the study of data that helps us derive useful insight for business decision making. Data Science is all about using tools, techniques, and creativity to uncover insights hidden within data. It combines math, computer science, and domain expertise to tackle real-world challenges in a variety of fields.

Data Science processes the raw data and solve business problems and even make prediction about the future trend or requirement. For example, from the huge raw data of a company, data science can help answer following question:

- What do customer want?
- How can we improve our services?
- What will the upcoming trend in sales?
- How much stock they need for upcoming festival.

Data science involves these key steps:

- **Data Collection**: Gathering raw data from various sources, such as databases, sensors, or user interactions.
- **Data Cleaning**: Ensuring the data is accurate, complete, and ready for analysis.
- **Data Analysis**: Applying statistical and computational methods to identify patterns, trends, or relationships.
- **Data Visualization:** Creating charts, graphs, and dashboards to present findings clearly.
- **Decision-Making**: Using insights to inform strategies, create solutions, or predict outcomes.

Data Science Prerequisites

- **Machine Learning**: Machine Learning is the backbone of data science. Data Scientists need to have a solid grasp of ML in addition to basic knowledge of statistics.
- **Modeling**: Mathematical models enable you to make quick calculations and predictions based on what you already know about the data. Modeling is also a part of Machine Learning and involves identifying which algorithm is the most suitable to solve a given problem and how to train these models.
- **Statistics**: Statistics are at the core of data science. A sturdy handle on statistics can help you extract more intelligence and obtain more meaningful results.
- **Programming:** Some level of programming is required to execute a successful data science project. The most common programming languages are Python, and R. Python is especially

popular because it’s easy to learn, and it supports multiple libraries for data science and ML.

- **Database:** A capable data scientist needs to understand how databases work, how to manage them, and how to extract data from them.

Problem Statement

Design a chatbot that integrates multiple large language models (LLMs) using Streamlit and Replicate, allowing users to interact with AI models dynamically. The chatbot should support automated model selection, customizable response parameters, real-time conversation handling, and session management, ensuring an efficient and user-friendly experience

<u>CHAPTER 4</u>

SYSTEM ANALYSIS

4. SYSTEM ANALYSIS

1. Existing System

Currently, most chatbot implementations rely on a **single LLM** such as OpenAI's GPT, Google's Bard, or Meta's Llama, which limits flexibility in **model selection** and **response customization**. These systems have several drawbacks:

- **Lack of Model Diversity** – Users are restricted to a single model, which may not always provide the best response quality for different queries.
- **Limited Customization** – Many chatbots do not offer fine-tuned control over parameters like **temperature** and **top-p**, affecting creativity and response consistency.
- **Scalability Issues** – Running **large-scale models** like Llama 3 (70B) requires significant computational power, making them **slow** or **expensive** for regular users.
- **No Session Management** – Some chatbots do not effectively manage **chat history**, leading to repeated context loss.
- **Token Limit Constraints** – Long conversations can exceed model token limits, causing truncation or abrupt conversation resets.

2. Proposed System

With the growing advancements in **large language models (LLMs)**, users often face challenges in selecting the best model for their specific needs. Some models provide **highly detailed responses** but are resource-intensive, while others offer **faster responses** with lower computational costs. Additionally, users require an **interactive and user-friendly** interface to engage with these models effectively.

The goal of this project is to develop a **multi-LLM chatbot** that dynamically selects from multiple **state-of-the-art models**—**Meta Llama 3 (70B Instruct), Mistral 7B Instruct, and Gemma 2B IT**—to generate intelligent responses. The chatbot should offer:

- **Seamless integration** with **Replicate API** for accessing different LLMs
- **A simple and interactive UI** using **Streamlit**
- **Customizable parameters** (temperature, top-p) for fine-tuning response creativity
- **Efficient session management**, including chat history and token limit enforcement
- **Real-time response generation** with AI models

This chatbot serves as a **versatile AI assistant**, catering to various conversational needs while balancing computational efficiency and response quality.

3. Objective of the System

The primary objective of this chatbot is to overcome these drawbacks by integrating multiple **LLMs** and providing an **interactive and efficient** user experience. The key objectives include:

- **Multi-Model Integration** – Enable dynamic selection between **Llama 3, Mistral 7B, and Gemma 2B** for optimized performance.
- **User-Controlled Response Generation** – Allow customization of temperature and top-p for balancing randomness and coherence.
- **Efficient Session Management** – Maintain chat histor**y** and enable users to clear conversations when needed.
- **Token Limit Enforcement** – Prevent token overflow errors by monitoring input length and alerting users when exceeding the limit.
- **Streamlined UI & Accessibility** – Use **Streamlit** for a simple, interactive chatbot interface with real-time response streaming.

This implementation ensures a **versatile, user-friendly, and adaptive chatbot** capable of handling diverse queries effectively

<u>CHAPTER 5</u>

REQUIREMENT ANALYSIS

5. REQUIREMENT ANALYSIS

Hardware Requirement Specification

- **Processor:** Intel Core i5 (8th Gen) / AMD Ryzen 5 or higher
- **RAM:** 8 GB (16 GB recommended for smoother performance)
- **Storage:** At least 10 GB of free disk space
- **GPU (Optional):** NVIDIA GPU with CUDA support (recommended for local model inference)
- **Internet Connection:** Required for accessing Replicate's API and model hosting

Software Requirement Specification

Operating System:

- Windows 10/11, macOS, or Linux (Ubuntu 20.04+ recommended)

Programming Language & Frameworks:

- **Python 3.8+** (for backend development)
- **Streamlit** (for UI)

Dependencies & Libraries:

- `streamlit` – For creating the chatbot interface
- `replicate` – To interact with the Replicate API for LLM inference
- `transformers` – For tokenization (Hugging Face's `AutoTokenizer`)
- `os` – For managing environment variables
- `random` – To randomly select an LLM

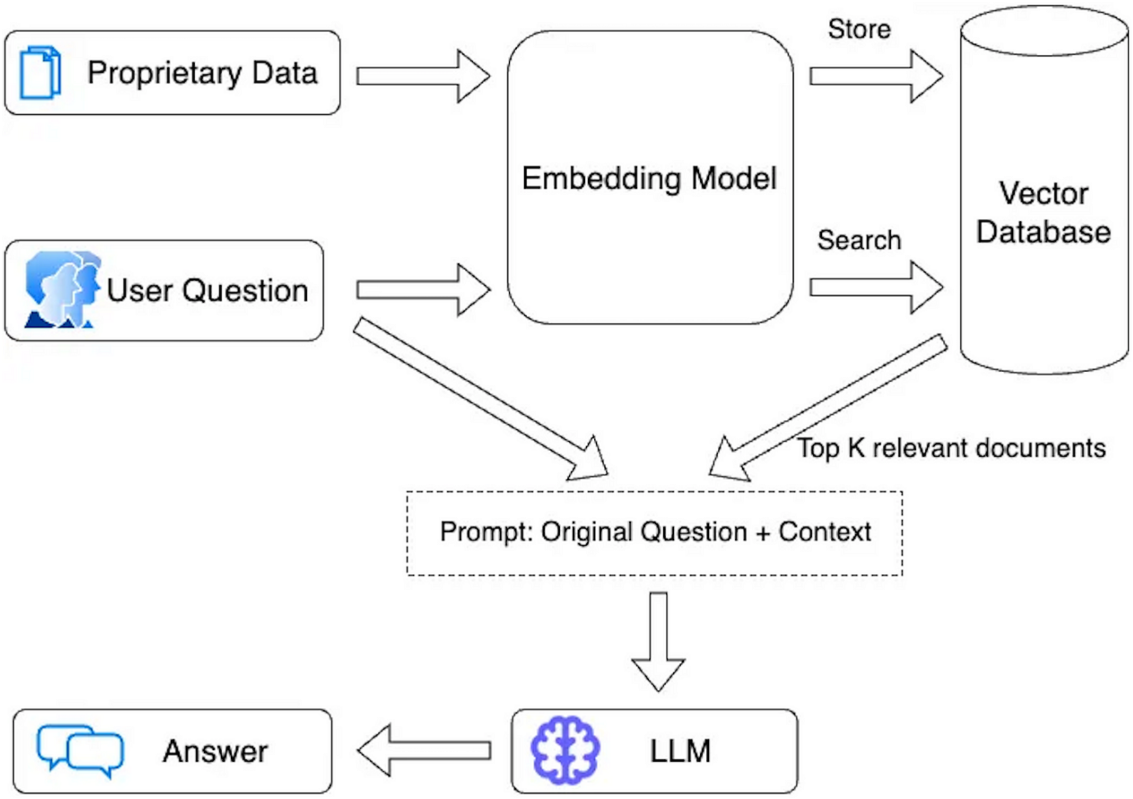
APIs & External Services:

- **Replicate API** – To access Llama 3, Mistral 7B, and Gemma 2B models
- **Hugging Face Models** – For tokenization (`huggyllama/llama-7b`

CHAPTER 6

DESIGN ANALYSIS

6. DESIGN & ANALYSIS



6.1 System Design

The chatbot follows a **client-server architecture** where:

1. The **frontend (client)** is built using **Streamlit**, which provides an interactive user interface.
2. The **backend (server)** interacts with **Replicate API** to fetch responses from various LLMs.

High-Level System Flow

- **User Input:** Users enter a query via the chat input field.
- **Session Handling:** The query is stored in **session state** for maintaining chat history.
- **Model Selection:** One of the predefined **LLMs** (**Meta Llama 3, Mistral 7B, or Gemma 2B**) is randomly selected.
- **Tokenization Check:** The input is **tokenized** to ensure it doesn’t exceed the token limit (3072 tokens).
- **API Call to Replicate:** The formatted prompt is sent to Replicate API to generate a response.
- **Streaming Response:** The chatbot **streams** the response in real-time for a smooth conversation experience.
- **Chat History Update:** The response is stored in session history for context retention.
- **Clear Chat Feature:** Users can **reset** the chat history if needed.

Component-Level Design

Component	Description
User Interface (UI)	Built using **Streamlit**, providing a clean and interactive chatbot interface.
Session Management	Uses **Streamlit's session state** to store chat history and user preferences.
Model Selection	Randomly selects from **Meta Llama 3, Mistral 7B, and Gemma 2B** models hosted on Replicate.
Parameter Control	Users can adjust **temperature** (creativity) and **top-p** (token filtering).
Token Management	Uses **Hugging Face's AutoTokenizer** to monitor input token limits.
Response Generation	Calls **Replicate API** for model inference and streams the response in real-time.
Error Handling	Detects API failures, incorrect token inputs, and excessive token lengths.

6.2Analysis of the System

Performance Analysis

- **Response Time:** Since the chatbot depends on Replicate API, latency depends on **model size** and **server response time**.
- **Efficiency:** Using **tokenization** ensures that responses do not exceed model limitations.
- **Scalability:** The system can scale efficiently by deploying on **cloud platforms** (AWS, Google Cloud, or Azure).

CHAPTER 7

IMPLEMENTATION

7. IMPLEMENTATION

Implementation is the stage where the theoretical design is turned into a working system. The most crucial stage in achieving a new successful system and in giving confidence on the new system for the users that it will work efficiently and effectively.

The system can be implemented only after thorough testing is done and if it is found to work according to the specification. It involves careful planning, investigation of the current system and it constraints on implementation, design of methods to achieve the change over and an evaluation of change over methods a part from planning.

Two major tasks of preparing the implementation are education and training of the users and testing of the system. The more complex the system being implemented, the more involved will be the system analysis and design effort required just for implementation.

The implementation phase comprises of several activities. The required hardware and software acquisition are carried out. The system may require some software to be developed. For this, programs are written and tested. The user then changes over to his new fully tested system and the old system is discontinued.

TESTING

The testing phase is an important part of software development. It is the Information zed system will help in automate process of finding errors and missing operations and also a complete verification to determine whether the objectives are met and the user requirements are satisfied. Software testing is carried out in three steps:

1. The first includes unit testing, where in each module is tested to provide its correctness, validity and also determine any missing operations and to verify whether theobjectives have been met. Errors are noted down and corrected immediately.
2. Unit testing is the important and major part of the project. So, errors are rectified easily in particular module and program clarity is increased. In this project entire system is divided into several modules and is developed individually. So, unit testing is conducted to individual modules.
3. The second step includes Integration testing. It need not be the case, the software whose modules when run individually and showing perfect results, will also show perfect results when run as a whole.

Unit Testing

Unit testing ensures that individual components of the chatbot function correctly.

Test Case ID	Component	Test Scenario	Expected Output	Status
UT-01	API Authentication	Validate Replicate API token input	Accept valid token, reject invalid ones	Pass
UT-02	Model Selection	Randomly choose an LLM from the list	One of the three models is selected	Pass
UT-03	Temperature Control	Ensure slider sets temperature correctly	Temperature value is within the range (0.01 - 5.0)	Pass
UT-04	Token Count Function	Verify that token count is accurate	Returns correct token count for a given prompt	Pass
UT-05	Chat History Management	Ensure chat history is stored and retrieved correctly	Messages persist in session state	Pass

Integration Testing

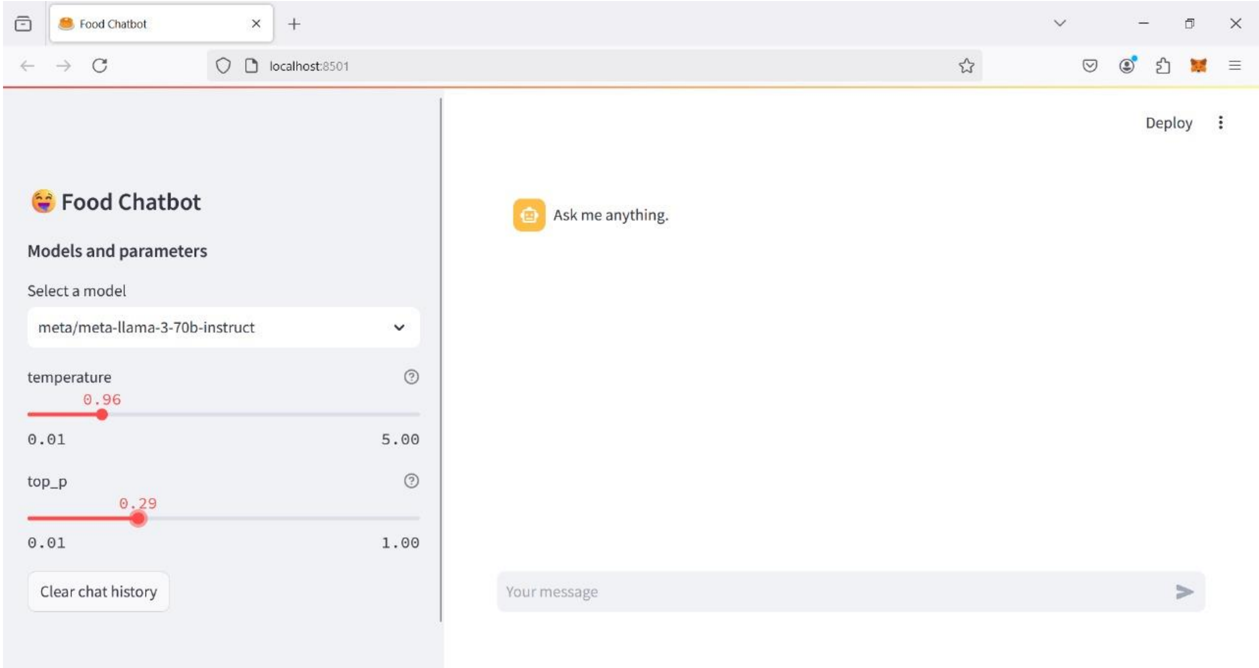
Integration testing ensures that different components work together seamlessly.

Test Case ID	Modules Involved	Test Scenario	Expected Outcome	Status
IT-01	API Authentication + Model Selection	Ensure API token validation and model selection occur correctly	Chatbot initializes with a selected model	Pass
IT-02	UI + Tokenization	Verify UI input passes through tokenizer correctly	Token count reflects actual input size	Pass
IT-03	Chat Input + Response Generation	Ensure user input triggers model response correctly	AI response is generated and displayed	Pass
IT-04	Chat History + UI	Verify chat history updates and displays properly	Previous messages remain visible	Pass
IT-05	Clear Chat + UI	Ensure clearing chat history resets both UI and session	Chat window resets completely	Pass

CHAPTER 8

SNAPSHOTS

8. SNAPSHOTS



Food Chatbot

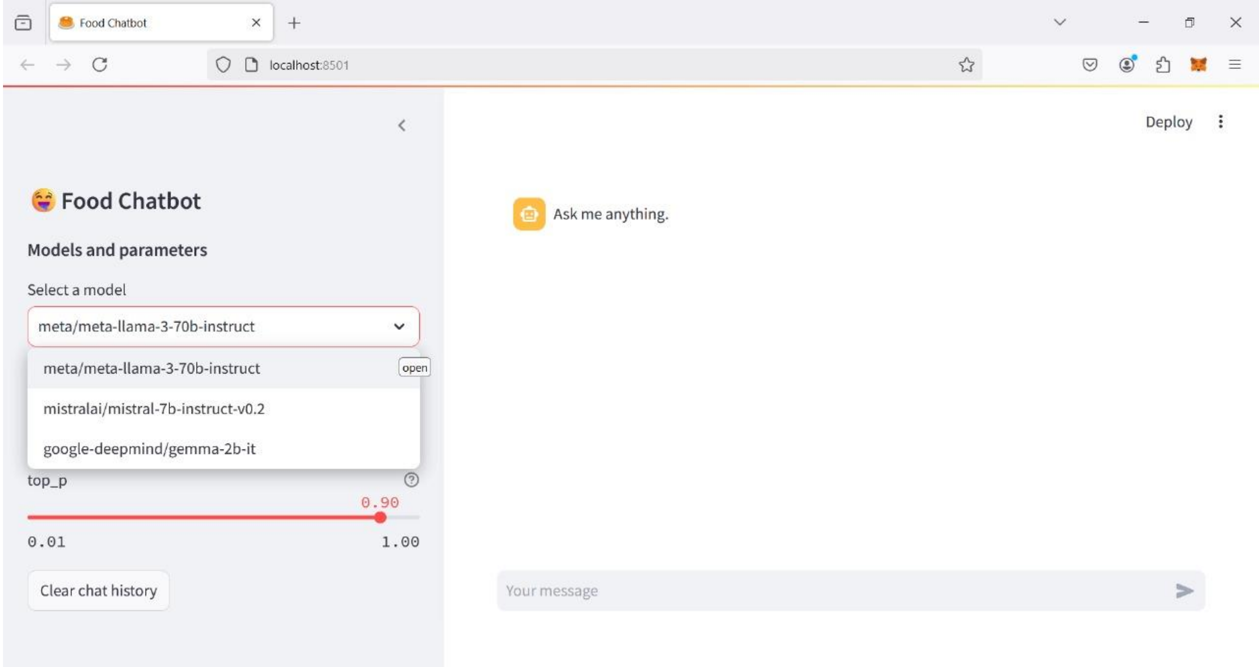
Models and parameters

Select a model: meta/meta-llama-3-70b-instruct

temperature: 0.96 (0.01 – 5.00)

top_p: 0.29 (0.01 – 1.00)

Fig 8.1: The parameters available for fine tuning the prompts



Food Chatbot
localhost:8501
Deploy
meta/meta-llama-3-70b-instruct
open
mistralai/mistral-7b-instruct-v0.2
google-deepmind/gemma-2b-it
Clear chat history
Ask me anything.
Your message

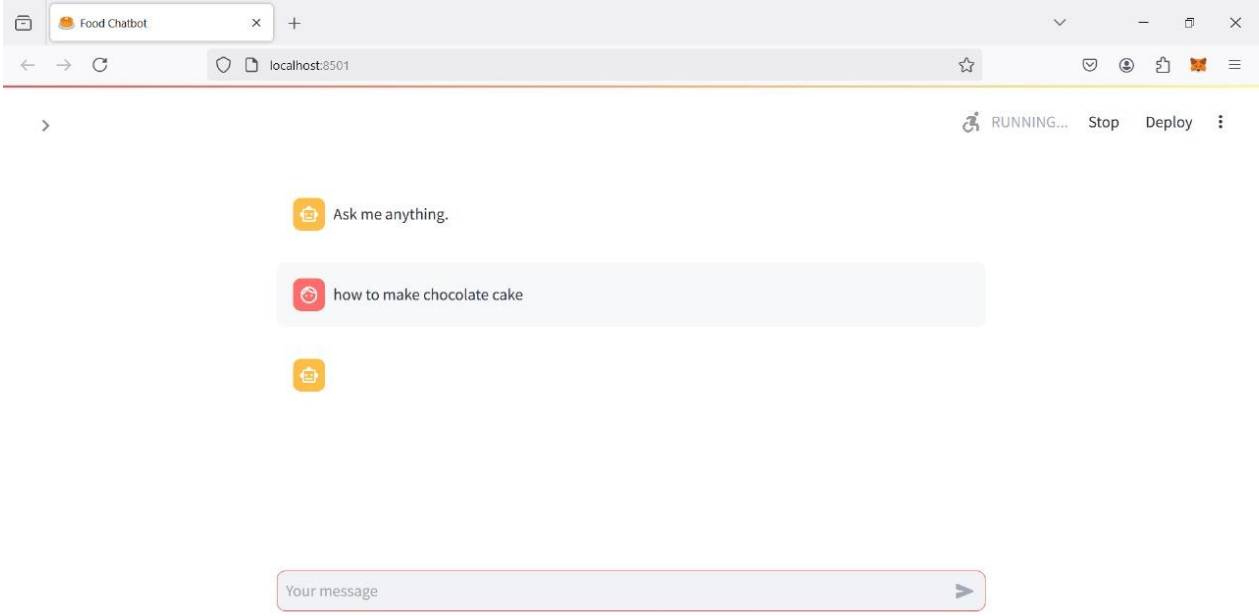
Food Chatbot

Models and parameters

Select a model: meta/meta-llama-3-70b-instruct

top_p: 0.90 (0.01 – 1.00)

Fig 8.2: The model selection choices available

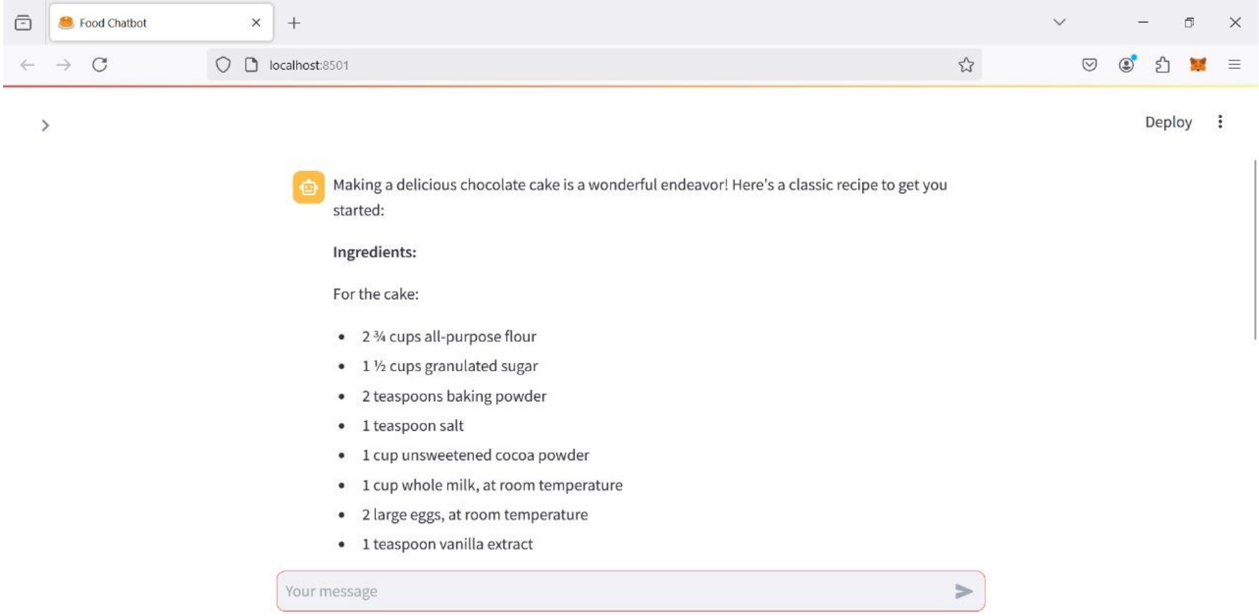


Ask me anything.

how to make chocolate cake

Your message

Fig 8.3: An example prompt



Making a delicious chocolate cake is a wonderful endeavor! Here's a classic recipe to get you started:

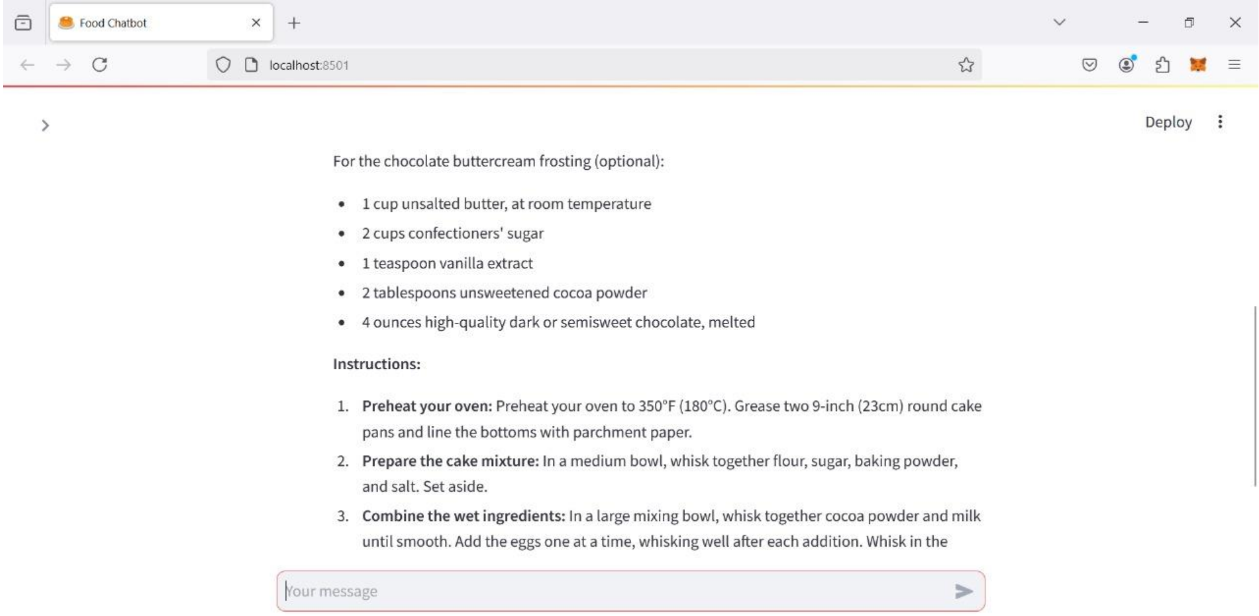
Ingredients:

For the cake:

- 2 ¾ cups all-purpose flour
- 1 ½ cups granulated sugar
- 2 teaspoons baking powder
- 1 teaspoon salt
- 1 cup unsweetened cocoa powder
- 1 cup whole milk, at room temperature
- 2 large eggs, at room temperature
- 1 teaspoon vanilla extract

Your message

Fig 8.4: The results for the prompt



For the chocolate buttercream frosting (optional):

- 1 cup unsalted butter, at room temperature
- 2 cups confectioners' sugar
- 1 teaspoon vanilla extract
- 2 tablespoons unsweetened cocoa powder
- 4 ounces high-quality dark or semisweet chocolate, melted

Instructions:

1. **Preheat your oven:** Preheat your oven to 350°F (180°C). Grease two 9-inch (23cm) round cake pans and line the bottoms with parchment paper.
2. **Prepare the cake mixture:** In a medium bowl, whisk together flour, sugar, baking powder, and salt. Set aside.
3. **Combine the wet ingredients:** In a large mixing bowl, whisk together cocoa powder and milk until smooth. Add the eggs one at a time, whisking well after each addition. Whisk in the

Fig 8.5: The results for the prompt

CHAPTER 9

CONCLUSION

9. CONCLUSION

The Multi-LLM Chatbot provides an interactive and scalable AI-powered conversational experience by integrating multiple large language models (LLMs) through the Replicate API. Designed using Streamlit, the chatbot ensures a user-friendly interface, supports session persistence, and allows dynamic model selection while maintaining token constraints to prevent errors.

By addressing the limitations of existing chatbots, such as single-model dependency and lack of customization, this system introduces multi-model adaptability, real-time response streaming, and adjustable response creativity. The chatbot is equipped with robust session management, error handling, and an intuitive UI, making it an ideal tool for various applications.

Through unit and integration testing, the chatbot's performance, reliability, and functionality have been validated, ensuring smooth operation across different use cases. The design and analysis highlight the chatbot's efficiency, scalability, and potential improvements, such as voice integration, database storage, and custom model selection.

In conclusion, this chatbot represents a significant advancement in AI-driven conversation systems, offering a versatile, efficient, and engaging user experience while paving the way for future enhancements in multi-model AI interactions.

10. REFERENCES

[1] www.geeksforgeeks.com

[2]www.w3schools.com

[3]www.simplilearn.com

[4]https://medium.com/cyberark-engineering/a-developer-guide-for-creating-a-multi-modal-chatbot-using-langchain-agents-9003ba0ffb4d