“TITLE: Running GenAI on Intel AI Laptops and Simple LLM Inference on CPU and fine-tuning of LLM Models using Intel® OpenVINO™”

“SUBTITLE: Medical And Drug Assistant Chatbot”

A PROJECT REPORT

Submitted to

Intel® Unnati Industrial Training Program 2024

Submitted By

1. Mr. Atharva Prakash Pawar.
2. Miss. Ankita Kiran Phalke.
3. Miss. Sonali Madhukar Lokare.

Under the Guidance of Mentor

Prof. K. P. Jagtap

Computer Science and Engineering

YSPM’s Yashoda Technical Campus

Faculty of Engineering, Wadhe, Satara-415011 2024- 25

ABSTRACT

With the use of cutting-edge AI technologies, the Medical and pharmacological Assistant Chatbot can offer trustworthy, up-to-date medical advice and pharmacological information. The chatbot offers high accuracy and quick reaction times by utilizing the optimized TinyLlama model and Intel® ® OpenVINO™ for efficient CPU-based inference. The Flask backend manages smooth communication between the user interface and the AI model, while the Streamlit frontend guarantees an interface that is easy to use. The solution is accessible and reasonably priced, as it is made to operate on Intel AI laptops and does not require specialized hardware. The chatbot is a useful tool for both personal and professional use because it can instantly respond to user inquiries regarding symptoms, diseases, drugs, and general health issues. The system is flexible and scalable, and it receives frequent upgrades to keep it up to date with the most recent medical data. The chatbot runs around-the-clock and provides constant access to medical assistance while putting a priority on user data protection and privacy. By providing consumers with timely, accurate, and personalized medical information, this project sets a new benchmark for digital health solutions and illustrates the transformational potential of AI in healthcare.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Our innovative Medical and pharmaceutical Assistant Chatbot offers reliable, current medical and pharmaceutical assistance in real time. Our chatbot provides clients with accurate and personalized advice, allowing them to receive vital medical assistance from the comfort of their homes, by utilizing the enhanced TinyLlama model and the powerful capabilities of Intel® OpenVINO™. Streamlit was used in the design of the chatbot's user-friendly interface, enabling simple interaction and query submission. Users can enter inquiries about drugs or medical problems, and the system will process them quickly. The Flask-powered backend manages communication between the model and the user interface, making sure that requests are handled quickly and answers are sent out on time.

Our solution is centred around the Model Inference component, which utilises the TinyLlama model optimised with Intel® OpenVINO™. This optimisation enables efficient CPU-based inference on Intel AI laptops, resulting in a system that is both highly accessible and cost-effective. Additionally, because the solution runs on standard CPUs instead of expensive GPU resources, a wider range of users can benefit from advanced AI capabilities. Our chatbot, which combines cutting-edge AI technology with optimised hardware, sets a new standard for digital health solutions by guaranteeing prompt and accurate responses, improving user experience, and fostering system trust. The chatbot's efficient processing and personalised recommendations demonstrate the revolutionary potential of Intel's AI technologies.

In conclusion, the Medical and Drug Assistant Chatbot is a prime example of the digital healthcare of the future, where artificial intelligence (AI) and technology are essential to providing prompt, tailored medical support. This study demonstrates the substantial advantages of fusing AI with hardware optimization to enhance user experience and healthcare delivery, increasing the effectiveness and accessibility of advanced medical support.

1.1 MOTIVATION

The motivation for developing the Medical and Drug Assistant Chatbot is to meet the growing need for accessible, real-time medical advice and drug information. Traditional healthcare services can be costly and time-consuming, creating barriers for timely medical assistance. By leveraging Intel® OpenVINO™ and the TinyLlama model, our chatbot provides accurate, immediate, and personalized recommendations. This project aims to enhance healthcare accessibility and efficiency, offering a cost-effective solution that reduces dependency on conventional healthcare systems, ensuring users receive essential medical information whenever they need it.

1.2 PROJECT OVERVIEW

Streamlit was used to create the chatbot's user-friendly interface, which makes it simple to interact with and submit queries. Users can enter inquiries about drugs or medical problems, and the system will process them quickly. The Flask-powered backend manages communication between the model and the user interface, making sure that requests are handled quickly and answers are sent out on time.

The Model Inference component, which makes use of the TinyLlama model optimised with Intel® OpenVINO™, is the central element of our approach. This optimization makes the system both affordable and very accessible by enabling effective CPU-based inference on Intel AI laptops. Because the system runs on regular CPUs, a larger variety of users can access advanced AI capabilities as it does not require costly GPU resources.

1.3 NEED OF PROJECT

The increasing need for quick access to trustworthy medical information makes the Medical and Drug Assistant Chatbot indispensable. Many people lack prompt access to expert medical advice because traditional healthcare services might be costly and unavailable when needed. By offering a system that instantly provides precise medication information and medical advice, our project seeks to address these problems and improve the efficiency and accessibility of healthcare. The TinyLlama model and Intel® OpenVINO™ allow the chatbot to run smoothly on regular CPUs without the need for expensive GPU resources. This ensures that more people can benefit from individualized medical help by making powerful AI capabilities accessible to a wider audience. The chatbot is a useful tool in the

current healthcare environment because of its integration with user-friendly interfaces and reliable backend systems, which further increase its usefulness and efficacy.

CHAPTER 2

PROBLEM DEFINITION AND SCOPE

3.1 PROBLEM STATEMENT

The Medical and Drug Assistant Chatbot aims to address the growing demand for accessible and reliable medical advice and drug information. Traditional healthcare services can be expensive and time-consuming, creating barriers for timely medical assistance. The chatbot seeks to provide an efficient solution by leveraging AI technology to deliver immediate, personalized recommendations.

3.2 SCOPE

The scope of the project includes:

1. Developing a user-friendly interface using Streamlit for easy interaction and query submission.
2. Implementing a robust backend with Flask to manage communication and process user queries efficiently.
3. Integrating the TinyLlama model optimized with Intel® OpenVINO™ for CPU-based inference on Intel AI laptops.
4. Ensuring scalability to handle a diverse range of medical queries and provide accurate responses.
5. Enhancing accessibility by operating on standard CPUs, making advanced medical support cost-effective and widely available.

3.3 GOALS & OBJECTIVE

Goals:

1. To provide real-time medical advice and drug information through an intuitive chatbot interface.
2. To optimize AI model inference using Intel® OpenVINO™ for efficient CPU performance.
3. To enhance accessibility to medical assistance, particularly for users in remote or underserved areas.

Objectives:

1. Develop a functional Streamlit interface for user interaction.

2. Implement Flask backend to handle query processing and response delivery.

3. Integrate and optimize the TinyLlama model with Intel® OpenVINO™ for CPU-based inference.

4. Ensure accuracy and reliability of medical recommendations provided by the chatbot.

5. Evaluate the system's performance and user satisfaction through testing and feedback.

6. Implementing the Medical and Drug Assistant Chatbot involves integrating advanced technologies such as Intel® OpenVINO™, LLM (Language Model), and fine-tuning methodologies to ensure accurate and efficient delivery of medical advice and drug information.

1. Instantaneous medical advice

Based on the questions that users ask, the chatbot offers precise and immediate medical advice. Users can inquire about ailments, symptoms, and general health issues, and they will promptly obtain pertinent information.

2. Information on Drugs:

Comprehensive information regarding different medications, such as how to take them, how much of them to take, any possible side effects, how they interact with other medications, and when they shouldn't be used, is available to users. With the aid of this tool, people can choose their prescriptions with knowledge.

3. Interface Friendly:

The user experience is made easy and intuitive with the Streamlit frontend. A straightforward, interactive design makes it simple for users to submit their questions and get prompt, understandable answers.

4. Effective Query Handling:

By utilizing Intel® OpenVINO™ for CPU-based inference, the chatbot effectively responds to user inquiries. This guarantees quick reaction times and nearly instantaneous information delivery to users.

5. Excellent Relevance and Accuracy:

The chatbot is powered by the refined TinyLlama model, which provides incredibly precise and pertinent responses. Because the model is trained on a large body of medical knowledge, the accuracy of the information is guaranteed.

6. Dynamism and Flexibility:

The system is made to be both flexible and scalable. It can be adjusted and updated on a regular basis to take into account user comments and the most recent medical research, making sure it stays current and useful over time.

7. Economical Implementation:

The solution is affordable and easily accessible because it is designed to function on Intel AI laptops. This increases the possible user base by doing away with the necessity for specialised, pricey gear.

8. Security and Privacy:

User data security and privacy are guaranteed by the chatbot. Because all transactions are conducted securely and in accordance with accepted data protection guidelines, users are more confident and regulatory standards are met.

9. Constant Accessibility:

The chatbot is always open, giving consumers constant access to medical knowledge and assistance. This increases the system's overall usefulness by guaranteeing that users can always access the assistance they require.

10. Ability to Integrate:

Other healthcare services and apps can be incorporated with the system. As a result, it can function as a comprehensive tool within a larger healthcare ecosystem, improving its utility and reach.

CHAPTER 3

SYSTEM SPECIFICATION

3.1 HARDWARE REQUIREMENTS :

Processor: Intel Core i5 or higher

Memory: Minimum 8 GB RAM (16 GB recommended)

Storage: Minimum 256 GB SSD (512 GB recommended)

Graphics: Integrated graphics

Networking: Reliable internet connection

Intel AI Laptop: Optimized for AI workloads with Intel hardware

3.2 SOFTWARE REQUIREMENTS:

Operating System: Windows 10/11, macOS, or Linux

Programming Languages: Python 3.8 or higher

Frameworks and Libraries:

Streamlit (frontend)

Flask (backend)

Intel® OpenVINO™ (model optimization and inference)

TinyLlama model

Additional Python libraries: numpy, pandas, requests, etc.

Development Tools:

IDE: PyCharm, Visual Studio Code, or Jupyter Notebook

Version Control: Git

CHAPTER 4

SOFTWARE DESIGN

4.1 Process Flow:

A process flow diagram (PFD) visually represents the sequence of steps or activities involved in a process. It typically uses standardized symbols to depict different types of actions, inputs, outputs, and decision points, making it easier to understand the flow and interactions within a process.

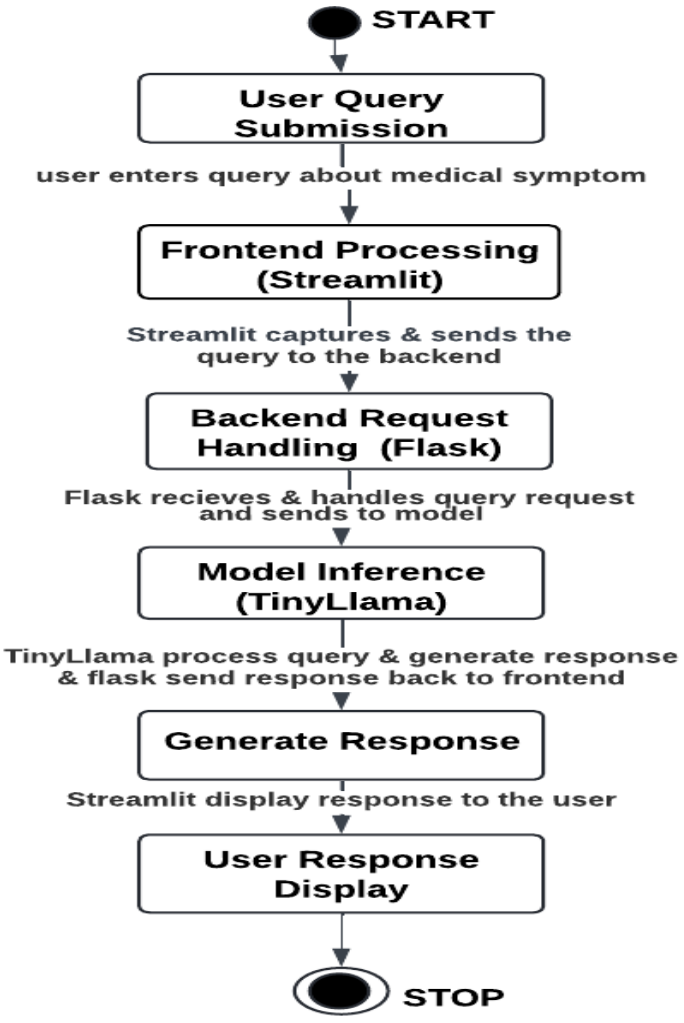


Fig No 5.1

4.2 System Architecture

System architecture diagrams provide a visual illustration of a system's various components and show how they communicate and interact with each other. These diagrams document a system's structure and architecture.

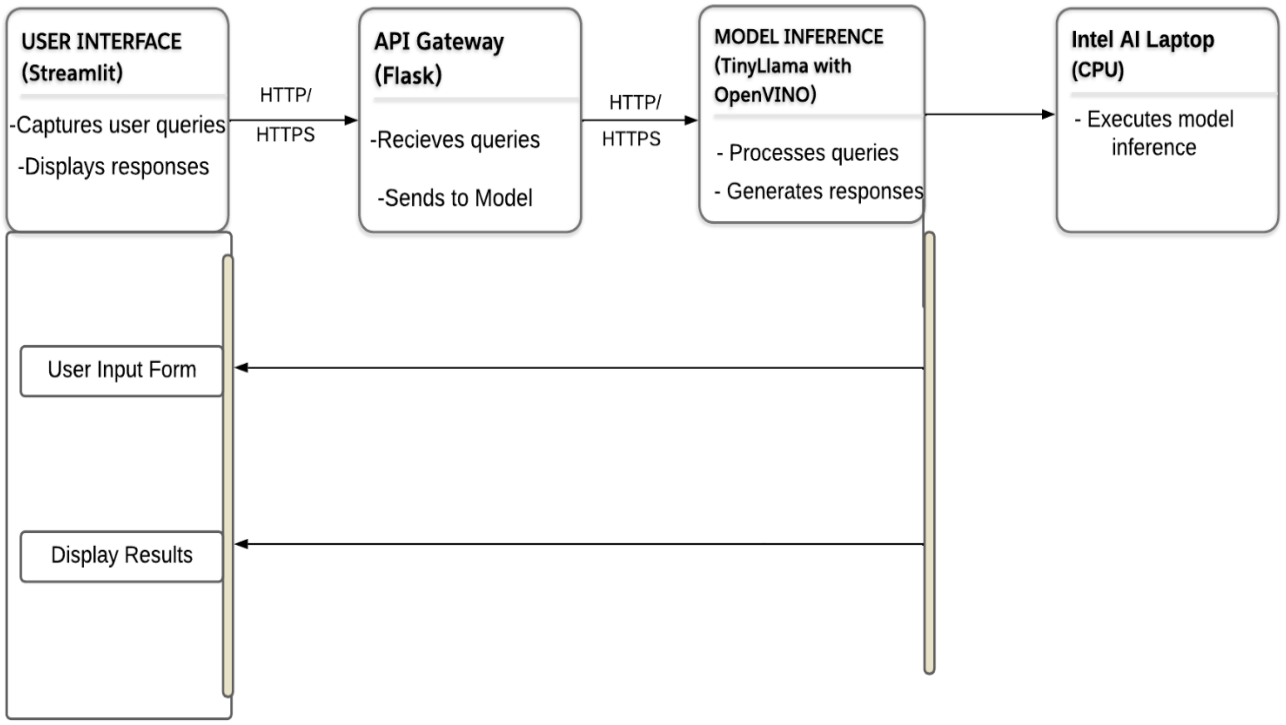


Fig No 5.3

4.3 UML Diagrams

4.4.1 Use Case Diagram

A Use Case Diagram in Unified Modeling Language (UML) is a visual representation of the functional requirements of a system from the user's perspective. It describes the various ways users interact with a system and the different functionalities the system provides in response. Use Case Diagrams are widely used in software development to capture and communicate the system's intended behavior.

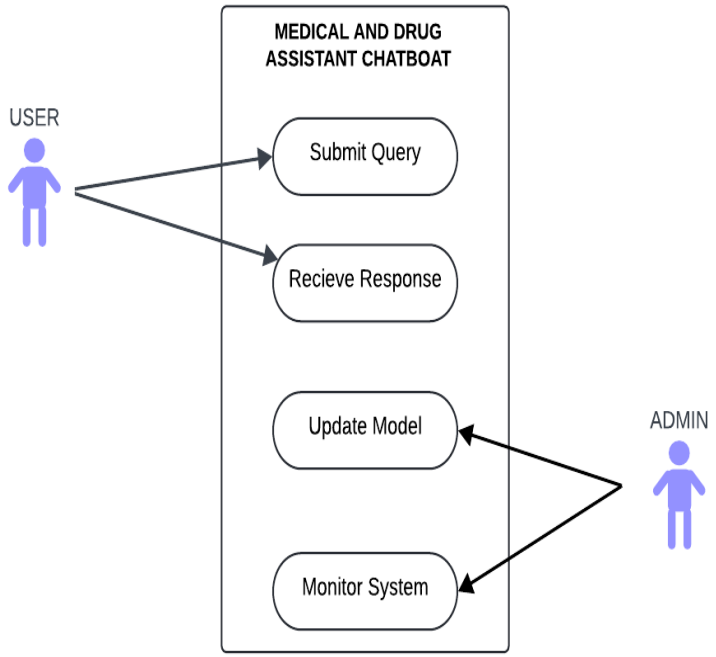


Fig 5.4.1

4.4.2 Sequence Diagram

A Sequence Diagram in Unified Modeling Language (UML) is a dynamic modeling diagram that represents the interactions between objects or components over time. It illustrates the sequence of messages exchanged among objects or components within a particular scenario of a use case. Sequence diagrams are particularly useful for visualizing the flow of events in a system and understanding how different elements collaborate.

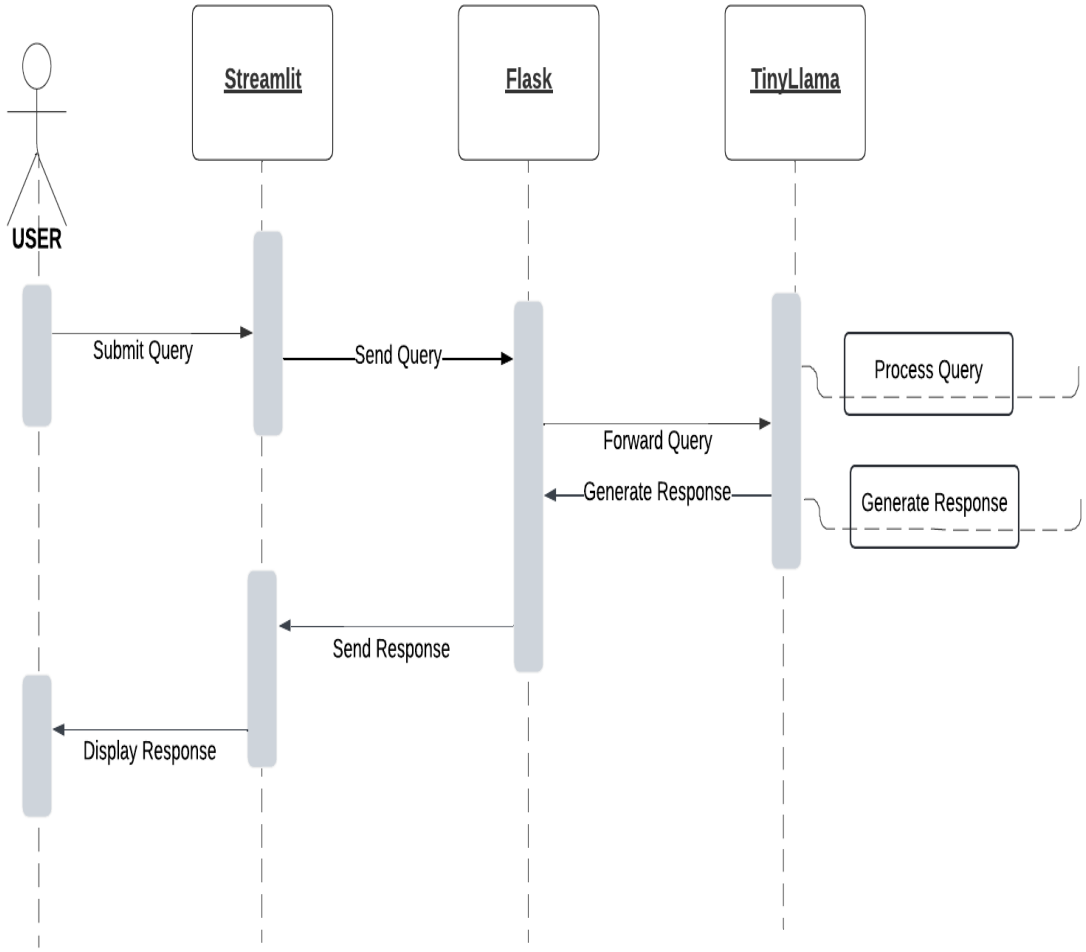


Fig 4.4.2

CHAPTER 5

IMPLEMENTATION DETAILS

5.1 MODULES AND THEIR FUNCTIONALITY

1. Intel® OpenVINO™:

An array of tools called Intel® OpenVINO™ (Open Visual Inference and Neural Network Optimisation) is intended to help deep learning models perform best with Intel technology. It permits faster inferencing on FPGAs, GPUs, and Intel CPUs, enabling high-performance and efficient operation of AI systems. OpenVINO™ optimises models for deployment across edge devices and servers by supporting many frameworks including as TensorFlow, PyTorch, and ONNX. OpenVINO™ accelerates AI inference through hardware acceleration, increasing its speed and effectiveness. This makes it perfect for real-time processing applications like image recognition and medical diagnostics.

2. LLM (Language Model):

A LLM is a language model that is intended to comprehend and produce writing that is similar to that of a human, such as GPT (Generative Pre-trained Transformer) models. Large volumes of text data are used to train these models, which can then be adjusted to specialise in a variety of industries or jobs, such as healthcare. Because of their superior ability to comprehend and produce natural language, LLMs are well-suited for applications such as information retrieval, content creation, and conversational AI. LLMs in the healthcare industry are capable of deciphering medical queries, giving precise answers, and making tailored recommendations based on the given data. To appropriately grasp medical language and context in the context of the chatbot, LLMs have undergone specific training using medical literature and datasets. By fine-tuning, the model becomes more capable of responding to user inquiries about medication information and medical advice in a precise and context-aware manner, resulting in dependable and educational encounters.

3. Fine-Tuning:

Fine-tuning is a process where a pre-trained LLM is further trained on domain-specific data to adapt it to a particular task or application. In the context of the Medical and Drug Assistant Chatbot, fine-tuning involves training the LLM on medical literature, patient records, and drug databases to enhance its understanding of medical terminology and context. This process improves the model's accuracy and relevance in generating medical advice and drug information, ensuring that the chatbot provides reliable and context-aware recommendations to users.

4. Streamlit:

The frontend interface, Streamlit, provides customers with an easy-to-use platform to enter medical inquiries and receive replies in a style that is easy to navigate. Its framework, which is based on Python, makes development easier and allows interactive applications to be quickly prototyped and deployed. Streamlit's smooth integration with Python modules makes data visualisation easier and boosts user engagement by giving prompt, understandable reply on queries.

5. **Flask**:

The backend API, Flask, controls the flow of data between the AI model and the user interface. It manages front-end HTTP requests, effectively responds to user inquiries, and coordinates data flow to and from the AI model. Strong backend functionality is ensured while retaining responsiveness and scalability thanks to Flask's lightweight and modular architecture, which enables flexible integration with a variety of Python modules and frameworks.

6. **TinyLlama with OpenVINO™**:

The chatbot leverages TinyLlama, a specialized model optimized with Intel® OpenVINO™ for CPU-based inference on Intel AI laptops. This optimization enhances performance and efficiency in processing medical queries, enabling rapid response times without relying on high-end GPU resources. OpenVINO™ accelerates AI workloads by utilizing hardware acceleration capabilities, ensuring that the chatbot operates seamlessly even under heavy computational loads, making it suitable for real-time medical advice delivery.

7. **LLM (Language Model)**:

LLMs are integrated and fine-tuned to interpret and generate human-like text responses based on extensive training data. In the context of the chatbot, LLMs are specifically trained on medical

literature and datasets to understand medical terminology and context accurately. Fine-tuning enhances the model's ability to provide precise and context-aware responses to user queries regarding medical advice and drug information, ensuring reliable and informative interactions.

8. **Testing and Deployment**:

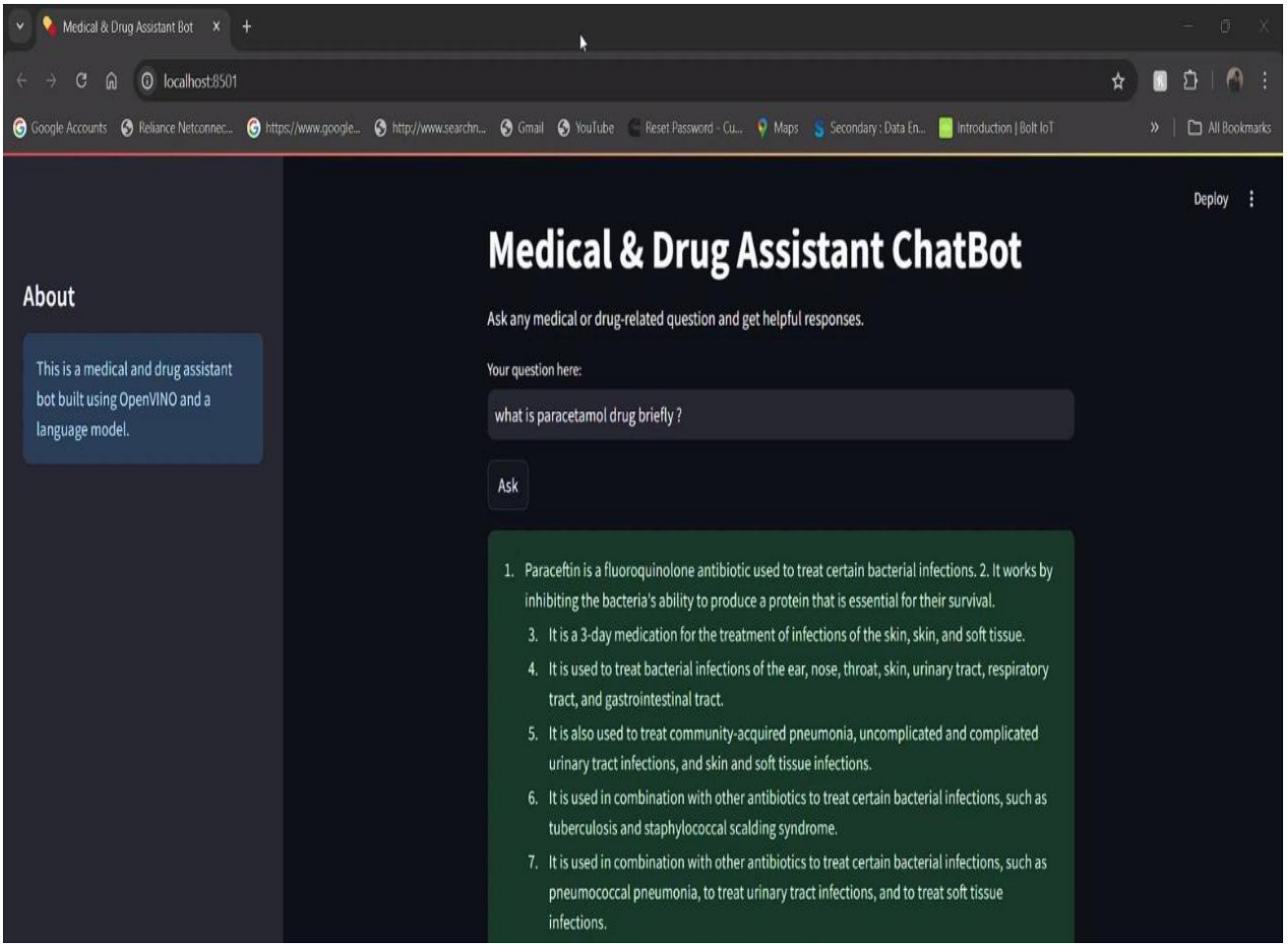
The chatbot undergoes rigorous testing to validate functionality, accuracy, and reliability across various scenarios and user inputs. Testing ensures that the chatbot delivers accurate medical recommendations and responds effectively to user queries. Once tested, the chatbot is deployed on Intel AI hardware to leverage optimized performance and scalability. Deployment on Intel AI laptops ensures that the chatbot meets performance requirements while maintaining accessibility and affordability, fulfilling its role in enhancing healthcare accessibility through advanced technology integration.

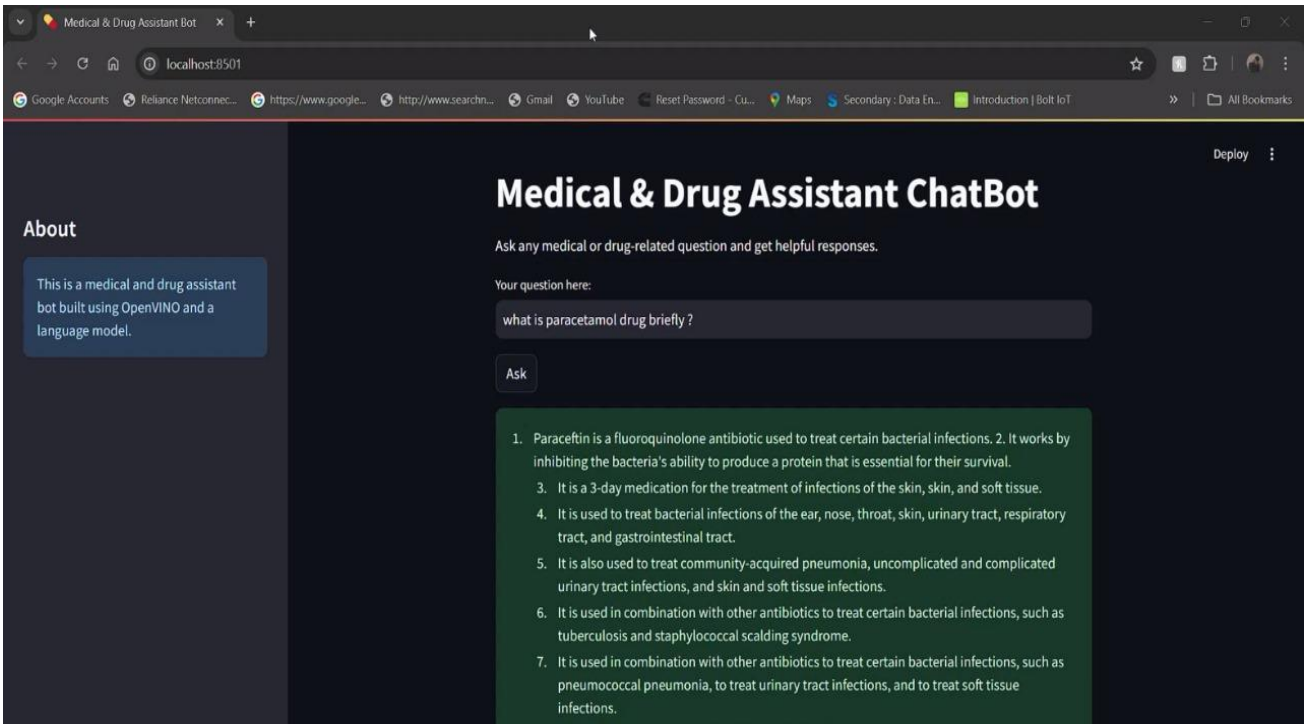
9. **Integration and Optimization**:

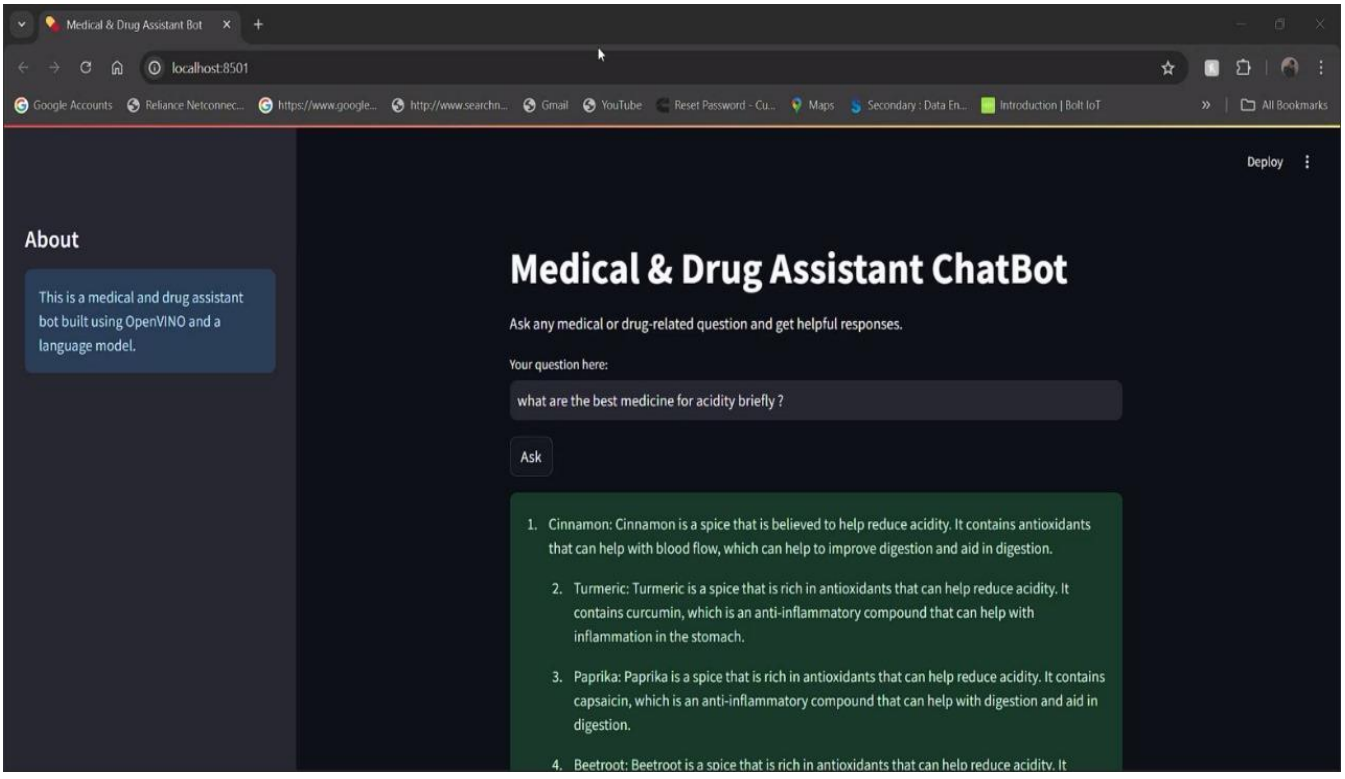
Integration is the process of integrating the TinyLlama model with the Flask backend to guarantee smooth data transfer and peak performance. The model's features may be easily incorporated thanks to Flask's modular architecture, which makes handling user inquiries and responses more effective. The optimizations of OpenVINO™ improve inference speed and accuracy even more, optimising resource usage and guaranteeing prompt delivery of medical advice. By meeting user expectations for rapid and accurate assistance, this integrated strategy helps the chatbot achieve its goal of promptly and effectively providing trustworthy healthcare information.

CHAPTER 6
RESULT AND ANALYSIS

6.1 SNAPSHOT AND GUI:




Medical & Drug Assistant Bot
localhost:8501
Deploy
About
This is a medical and drug assistant bot built using OpenVINO and a language model.
Medical & Drug Assistant ChatBot
Ask any medical or drug-related question and get helpful responses.
Your question here:
what is paracetamol drug briefly ?
Ask
1. Paraceftin is a fluoroquinolone antibiotic used to treat certain bacterial infections. 2. It works by inhibiting the bacteria's ability to produce a protein that is essential for their survival.
3. It is a 3-day medication for the treatment of infections of the skin, skin, and soft tissue.
4. It is used to treat bacterial infections of the ear, nose, throat, skin, urinary tract, respiratory tract, and gastrointestinal tract.
5. It is also used to treat community-acquired pneumonia, uncomplicated and complicated urinary tract infections, and skin and soft tissue infections.
6. It is used in combination with other antibiotics to treat certain bacterial infections, such as tuberculosis and staphylococcal scalding syndrome.
7. It is used in combination with other antibiotics to treat certain bacterial infections, such as pneumococcal pneumonia, to treat urinary tract infections, and to treat soft tissue infections.


Medical & Drug Assistant Bot
localhost:8501
Deploy
About
This is a medical and drug assistant bot built using OpenVINO and a language model.
Medical & Drug Assistant ChatBot
Ask any medical or drug-related question and get helpful responses.
Your question here:
what are the best medicine for acidity briefly ?
Ask
1. Cinnamon: Cinnamon is a spice that is believed to help reduce acidity. It contains antioxidants that can help with blood flow, which can help to improve digestion and aid in digestion.
2. Turmeric: Turmeric is a spice that is rich in antioxidants that can help reduce acidity. It contains curcumin, which is an anti-inflammatory compound that can help with inflammation in the stomach.
3. Paprika: Paprika is a spice that is rich in antioxidants that can help reduce acidity. It contains capsaicin, which is an anti-inflammatory compound that can help with digestion and aid in digestion.
4. Beetroot: Beetroot is a spice that is rich in antioxidants that can help reduce acidity. It

CHAPTER 7
CONCLUSION

The Medical and pharmacological Assistant Chatbot offers trustworthy, real-time medical advice and pharmacological information by integrating state-of-the-art AI technologies. With the help of the optimized TinyLlama model and Intel® OpenVINO™ for effective CPU-based inference, our chatbot achieves high performance and accuracy at a reasonable cost. A user-friendly interface is ensured by using Streamlit for the frontend and Flask for the backend, which effectively manages communication between the AI model and the user interface.

The initiative shows how AI has the power to completely transform healthcare by giving users quick, personalized answers to their questions. This service is very helpful for both individual users and medical professionals since it makes it simple for users to get information about symptoms, treatments, and general health advice. Because the system is designed to function well on Intel AI laptops, it may be deployed in a variety of situations without the requirement for specialized hardware.

Our chatbot also demonstrates how important it is for AI systems to be adaptable and constantly developing. The model may be updated and modified frequently to incorporate user feedback and the most recent medical data, keeping it useful and up to date over time.

The Medical and Drug Assistant Chatbot, in conclusion, is a prime example of the revolutionary effects of AI in healthcare. It showcases the potential of Intel's AI technology by providing a reliable, effective, and user-friendly medical advice solution. With substantial advantages for both patients and healthcare providers, this project raises the bar for digital health so