**AI Powered Health Assistant**

A Project Report

submitted in partial fulfillment of the requirements

of

AICTE Internship on AI: Transformative Learning

with

TechSaksham – A joint CSR initiative of Microsoft & SAP

by

**Sayandeep Dutta, 22052754@kiit.ac.in**

Under the Guidance of

**Jay Rathod**

**ACKNOWLEDGEMENT**

I would like to take this opportunity to express our deep sense of gratitude to all individuals who helped us directly or indirectly during this thesis work.

Firstly, I would like to thank my supervisor, Jay Rathod, for being a great mentor and the best adviser I could ever have. His advice, encouragement and the critics are a source of innovative ideas, inspiration and causes behind the successful completion of this project. The confidence shown in me by him was the biggest source of inspiration for me. It has been a privilege working with him during this one month internship program. He always helped me during my project and many other aspects related to the program. His talks and lessons not only help in project work and other activities of the program but also make me a good and responsible professional.

#### **ABSTRACT**

With the increasing demand for accessible healthcare information, AI-powered medical assistants have become crucial in providing preliminary medical insights. This project presents an AI-powered healthcare system utilizing a fine-tuned version of the LLaMA 3 8B Instruct model, designed specifically for answering medical queries. The model, trained on the AI Medical Dataset, enables informative responses to diverse health-related inquiries.

The primary objective of this project is to develop a system that assists users by generating contextually accurate medical information. The methodology involves integrating the pre-trained **ruslanmv/ai-medical-model-32bit** from Hugging Face with Google Colab, leveraging **transformers, BitsAndBytes quantization, and CUDA acceleration** for efficient inference. The system processes user queries by tokenizing the input, generating responses using the model, and refining outputs to enhance readability.

Key results demonstrate the model's ability to provide detailed explanations regarding medical conditions, symptoms, and physiological processes. While the AI assistant serves as a valuable informational tool, it is not a replacement for professional medical advice. The findings highlight the potential of AI-driven healthcare solutions in assisting users with preliminary knowledge while emphasizing the necessity of expert consultation.

This project underscores the effectiveness of deep learning in medical NLP applications and provides a foundation for future enhancements, such as integrating real-time databases and improving response accuracy.

**TABLE OF CONTENT**

**Abstract I**

**Chapter 1.**  **Introduction 1**

1.1 Problem Statement 1

1.2 Motivation 1

1.3 Objectives 2

1.4. Scope of the Project 2

**Chapter 2.**  **Literature Survey 3**

**Chapter 3.**  **Proposed Methodology**

**Chapter 4.**  **Implementation and Results**

**Chapter 5. Discussion and Conclusion**

**References**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Caption** | **Page No.** |
|  | Overview of the Architecture | **11** |
|  | Sample query output 1 | **13** |
|  | Sample query output 2 | **13** |
|  | Sample query output 3 | **14** |
|  | Error encountered while deploying in Streamlit | **14** |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**CHAPTER 1**

**Introduction**

* 1. **Problem Statement:**

The availability of accurate and timely healthcare information is a major concern in today’s world. Many people turn to the internet for medical advice, often encountering unreliable or misleading information. Additionally, healthcare professionals face challenges in managing a high volume of patient inquiries, leading to increased workloads and potential delays in providing care.

This project aims to address these challenges by developing an **AI-powered healthcare system** that provides **accurate and informative responses to medical queries** using a fine-tuned deep learning model. The system assists users in understanding medical conditions, symptoms, and potential causes, thus improving accessibility to medical knowledge. While it does not replace professional medical consultation, it serves as a **preliminary information resource**, reducing misinformation and enhancing healthcare accessibility.

* 1. **Motivation:**

The motivation for this project stems from the increasing reliance on **artificial intelligence in healthcare** and the need for a **trustworthy AI-based medical assistant**. The integration of AI in the healthcare domain can **enhance knowledge dissemination, assist medical practitioners, and empower individuals with reliable health-related insights**.

### ****Potential Applications and Impact:****

1. **Medical Information Retrieval** – Helps individuals understand medical conditions, symptoms, and treatments.
2. **Assisting Healthcare Professionals** – Reduces the burden of answering repetitive queries.
3. **Reducing Misinformation** – Provides AI-generated responses based on a medically curated dataset.
4. **Educational Tool** – Useful for medical students and researchers in obtaining detailed explanations.
   1. **Objective:**

The primary objectives of this project are:

1. To **develop an AI-powered medical assistant** capable of answering user queries on healthcare-related topics.
2. To **fine-tune and deploy a pre-trained LLaMA 3 8B Instruct model** optimized for medical question-answering.
3. To **integrate the system with a user-friendly interface** for seamless interaction.
4. To **provide accurate, informative, and contextually relevant responses** based on a curated medical dataset.
5. To **enable real-time inference using Google Colab** due to GPU constraints on local machines.

.

* 1. **Scope of the Project:**

### ****In-Scope Features:****

* Implementing **LLM-based medical Q&A system** using the **ruslanmv/ai-medical-model-32bit** from Hugging Face.
* **Processing natural language queries** and generating relevant medical insights.
* Using **transformers and quantization techniques** for efficient computation.
* Deploying the system on **Google Colab** for execution due to hardware constraints.

### ****Limitations:****

* The system **does not replace professional medical consultation** and should only be used for informational purposes.
* The model’s **accuracy depends on the quality of training data**, and it may not always provide definitive answers.
* The AI may **struggle with complex diagnoses** that require human expertise.
* **No real-time patient data integration**—the system is limited to answering general medical queries.

**CHAPTER 2**

**Literature Survey**

## ****2.1 Review of Relevant Literature****

The use of **AI in medical diagnosis and healthcare support** has been an area of extensive research. Several studies have explored the application of **natural language processing (NLP) and deep learning** for providing medical insights.

### ****Related Works:****

* **Chatbots for Healthcare** – AI-powered chatbots like **MedPaLM 2 and BioGPT** have been developed to address medical queries.
* **LLMs in Medicine** – Research on **BERT, BioBERT, and Med-BERT** has shown that NLP models can effectively process medical literature.
* **Medical Question Answering Systems** – Various models have been proposed for extracting and summarizing medical knowledge from datasets.

## ****2.2 Existing Models, Techniques, and Methodologies****

### ****Pre-existing Models in Medical NLP:****

1. **BioBERT** – A transformer-based model trained on biomedical text.
2. **MedPaLM 2** – A Google AI model designed for medical question-answering.
3. **BioGPT** – A generative pre-trained model for biomedical literature analysis.
4. **LLM-based Medical Assistants** – OpenAI’s GPT models and Meta’s LLaMA models have been used in healthcare applications.

### ****Techniques Used:****

* **Transformer-based NLP models** for medical text processing.
* **Fine-tuning LLMs** on medical datasets.
* **Quantization techniques (BitsAndBytes)** to optimize large models for inference.

## ****2.3 Gaps in Existing Solutions & Project Contribution****

### ****Identified Gaps in Existing Models:****

* Many models are **too large to run on local machines**, requiring high-end hardware.
* Some AI-powered chatbots **lack domain-specific fine-tuning**, leading to generic answers.
* There are **limited open-source models** optimized for answering healthcare queries.
* Some models **do not process user input efficiently**, leading to long response times.

### ****How This Project Addresses the Gaps:****

* Uses a **fine-tuned version of LLaMA 3** trained on an extensive **medical dataset**.
* Implements **quantization techniques (BitsAndBytes)** for efficient execution on limited hardware.
* Provides a **streamlined pipeline for medical Q&A**, enhancing response relevance.
* Uses **Google Colab for deployment**, enabling accessibility despite hardware limitations.

**CHAPTER 3**

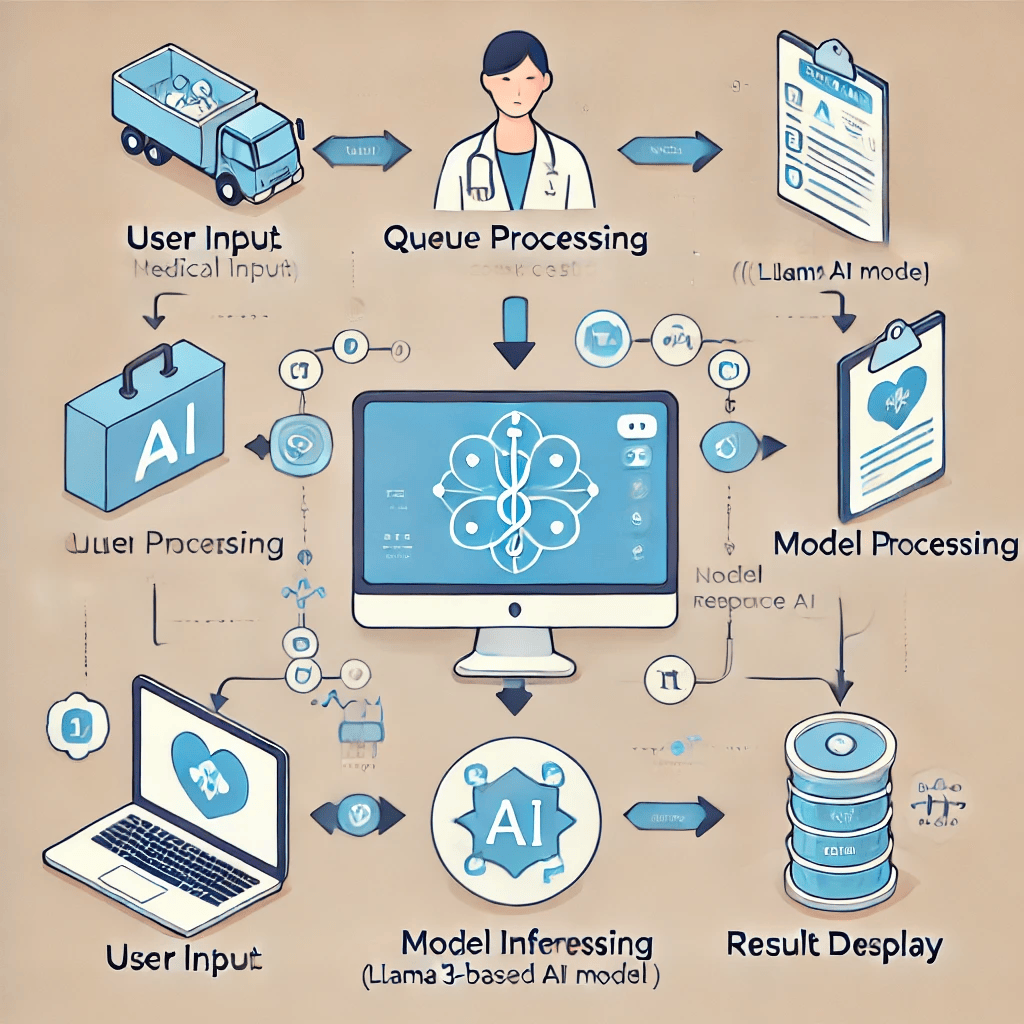
**Proposed Methodology**

* 1. **System Design**

### ****Proposed System Architecture****

The AI-powered healthcare system follows this architecture:

1. **User Input:** The user enters a medical question.
2. **Query Processing:** The input is tokenized and passed to the AI model.
3. **Model Inference:** The **LLaMA 3-based medical AI model** generates an answer.
4. **Response Processing:** The output is formatted for clarity.
5. **Result Display:** The system returns the response to the user.



**Fig 1: Overview of the Architecture**

* 1. **Requirement Specification**

Mention the tools and technologies required to implement the solution.

**Hardware Requirements:**

· **Google Colab** (Cloud-based GPU support)

· **Minimum 8GB RAM (Colab Pro recommended for better performance)**

· **CUDA-enabled GPU (Required for fast model inference)**

**Software Requirements:**

· **Python 3.x** – Programming language for model integration.

· **Hugging Face Transformers** – To load and use the pre-trained model.

· **Torch 2.2.1** – Deep learning framework for model execution.

· **BitsAndBytes** – Library for model quantization and optimization.

· **Accelerate** – For efficient model execution on limited hardware.

· **Google Colab** – Platform for running the model.

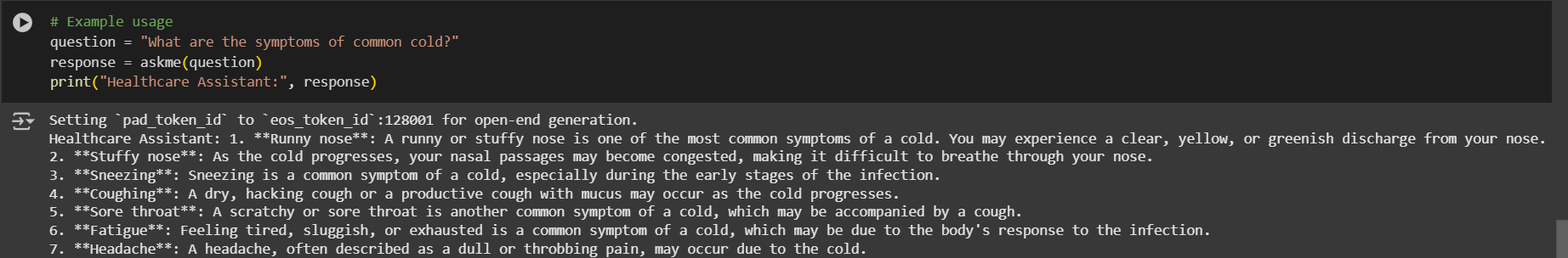
**CHAPTER 4**

**Implementation and Result**

* 1. **Snap Shots of Result:**

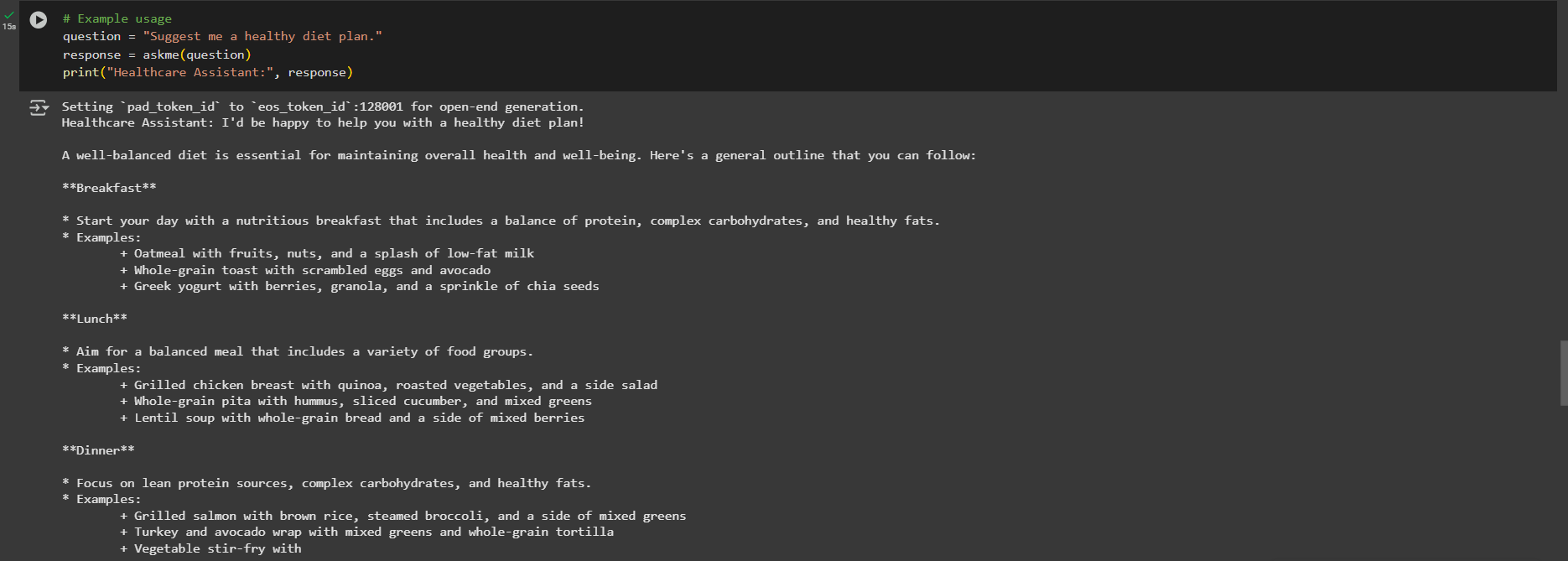
Kindly provide 2-3 Snapshots which showcase the results and output of your project and after keeping each snap explain the snapshot that what it is representing.

**Snapshot 1:**

****

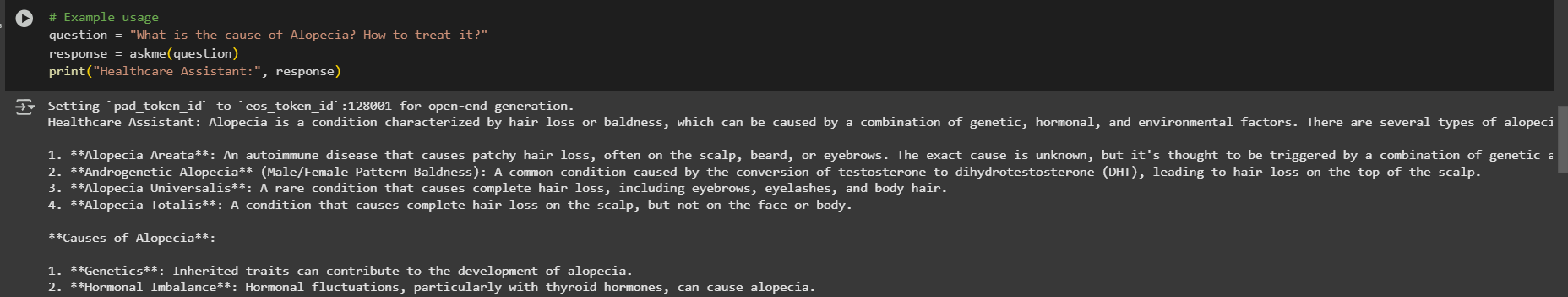
The user asked question,”What are the symptoms of common cold?”  
The model answered properly as shown in the output.

**Snapshot 2:**

****

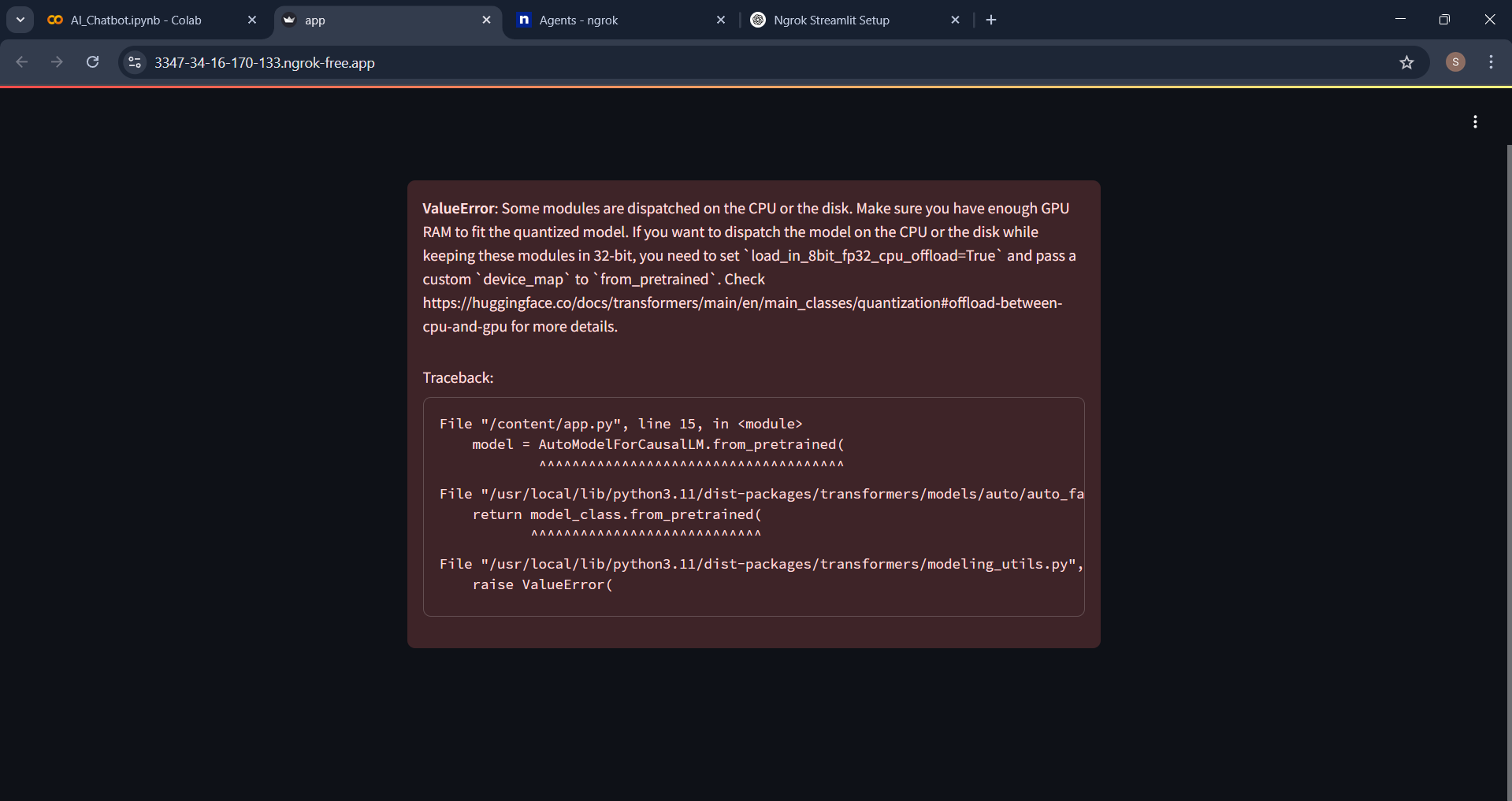
The user asked question,”Suggest me a healthy diet plan.”  
The model answered properly as shown in the output.

**Snapshot 3:**

****

The user asked question,”What is the cause of Alopecia?”  
The model answered properly as shown in the output.

**Error encountered while deploying in Streamlit:**

****

Due to hardware limitations the model could not be deployed to streamlit.

* 1. **GitHub Link for Code:**

**CHAPTER 5**

**Discussion and Conclusion**

* 1. **Future Work:**

There are several areas where the AI-powered healthcare system can be improved and expanded in the future:

**Enhanced Accuracy and Robustness:**

* 1. Fine-tune the model further using more diverse and real-world medical datasets.
  2. Integrate reinforcement learning with human feedback (RLHF) to improve response reliability.

**Multimodal Integration:**

* 1. Extend the model to analyze **medical images, lab reports, and patient histories** in addition to text-based queries.
  2. Incorporate **natural language processing (NLP) and computer vision** for a comprehensive AI healthcare assistant.

**Real-Time Consultation and Chatbot Capabilities:**

* 1. Implement a **conversational AI assistant** that interacts dynamically with patients and healthcare professionals.
  2. Enable the system to **handle follow-up questions** for a more personalized healthcare experience.

**Compliance with Medical Regulations:**

* 1. Improve the system to adhere to **HIPAA (Health Insurance Portability and Accountability Act) and GDPR** for secure handling of patient data.
  2. Develop a **medical expert verification system** to cross-check AI-generated responses with licensed professionals.

**Deployment as a Web and Mobile Application:**

* 1. Deploy the system on **cloud-based platforms** for real-time accessibility.
  2. Develop a **mobile-friendly version** for ease of use by patients and doctors.

**Multilingual and Regional Adaptation:**

* 1. Expand language support to cater to non-English speaking users.
  2. Adapt responses based on region-specific medical guidelines and drug availability.
  3. **Conclusion:**

This project successfully developed an **AI-powered healthcare system** that utilizes the **LLaMA 3-based fine-tuned model** to provide **accurate and informative medical responses**. By leveraging a **pretrained medical dataset**, the system is capable of assisting users with medical inquiries in a structured and reliable manner.

The project demonstrates the potential of **artificial intelligence in the healthcare industry**, particularly in **providing quick, preliminary insights** for medical conditions. While the model is informative, it is not a replacement for **professional medical advice** and should be used as a supplementary tool for patients and healthcare providers.

Key contributions of this project include:

* Successful integration of a **state-of-the-art AI model** for medical question-answering.
* Effective use of **Hugging Face Transformers and Google Colab** for cloud-based execution.
* Development of a structured pipeline involving **query processing, model inference, and result formatting** for optimal output.

Future enhancements, such as **improving model accuracy, real-time chatbot integration, and expanding multimodal capabilities**, can further improve the system's utility. The project serves as a **foundation for AI-driven medical assistance**, contributing to the broader goal of **accessible and AI-enhanced healthcare solutions**.

**REFERENCES**

1. Ruslanmv, "AI Medical Model 32-bit," Hugging Face, Available at: <https://huggingface.co/ruslanmv/ai-medical-model-32bit>
2. Meta, "Meta-Llama-3-8B-Instruct," Hugging Face, Available at: <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>
3. AI Medical Dataset, Available at: <https://huggingface.co/datasets/ruslanmv/ai-medical-dataset>
4. Brown et al., "Language Models are Few-Shot Learners," NeurIPS, 2020.
5. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.  
     
     
   .