**Project Documentation: Personalized Chatbot with Dynamic Responses**

Graduation Project of Generative AI Professional Track 🎓



**Team Members**

Esraa Samir Amin

Rahma Abdulla Ramadan Ahmed

Mohamed Hesham Ismail

Ahmed Abdelrahman Abdelfattah Azab

Mariam Essam Abdelaziz

Mostafa Ebrahim Abdelrahman

**Supervised by:**

Eng. Abdullah El-Shbasi

**Date:**

19/10/2024

Table of Contents

[**1.** **Project Idea** 3](#_Toc180085252)

[**1.1.** **Brief Description of the Problem Being Solved:** 3](#_Toc180085253)

[**1.2.** **Overview of the Proposed Solution (High-Level Idea):** 4](#_Toc180085254)

[**1.3.** **Unique Value Proposition:** 4](#_Toc180085255)

[**2.** **Motivation and Problem Statement** 5](#_Toc180085256)

[**2.1.** **Motivation** 5](#_Toc180085257)

[**2.2.** **Problem Statement** 5](#_Toc180085258)

[**3.** **Objectives** 6](#_Toc180085259)

[**3.1.** **Develop a Multilingual Medical Chatbot:** 6](#_Toc180085260)

[**3.2.** **Enable Medical Abbreviation Disambiguation:** 6](#_Toc180085261)

[**3.3.** **Provide Real-Time, Accurate Medical Information:** 6](#_Toc180085262)

[**3.4.** **Optimize Performance and Resource Usage:** 6](#_Toc180085263)

[**3.5.** **Create a User-Friendly Interface:** 6](#_Toc180085264)

[**3.6.** **Support Healthcare Providers with Quick, Accurate Data Access:** 6](#_Toc180085265)

[**4.** **End users and Features** 7](#_Toc180085266)

[**4.1.** **End User Personas:** 7](#_Toc180085267)

[**4.2.** **Key Features that Address User Needs** 7](#_Toc180085268)

[**4.3.** **How the Features Solve End-User Problems** 8](#_Toc180085269)

[**5.** **Data Flow** 9](#_Toc180085270)

[**5.1.** **Data Collection:** 9](#_Toc180085271)

[**5.2.** **Data Storage:** 9](#_Toc180085272)

[**5.3.** **Data Access:** 9](#_Toc180085273)

[**5.4.** **CSV Data Features:** 9](#_Toc180085274)

[**6.** **Programming Languages + Frameworks** 10](#_Toc180085275)

[**6.1.** **Programming Languages** 10](#_Toc180085276)

[**6.2.** **Frameworks and Tools** 10](#_Toc180085277)

[**6.3.** **Supporting Technologies** 11](#_Toc180085278)

[**7.** **Methodology** 12](#_Toc180085279)

[**7.1.** **Development Workflow:** 12](#_Toc180085280)

[**7.2.** **Data Preparation** 12](#_Toc180085281)

[**7.3.** **Model Fine-Tuning** 12](#_Toc180085282)

[**7.4.** **Models Used:** 13](#_Toc180085283)

[**7.5.** **Inference Optimization** 13](#_Toc180085284)

[**8.** **Challenges and Solutions** 13](#_Toc180085285)

[**9.** **Testing and Results** 14](#_Toc180085286)

[**10.** **Conclusion** 14](#_Toc180085287)

[**References** 15](#_Toc180085288)

# **Project Idea**

This project aims to build a Personalized Chatbot with Dynamic Responses using fine-tuning techniques on Hugging Face datasets. Specifically, it is a Multilingual Medical Assistant with LoRA-Optimized FastLanguageModel (Gemma-7B PEFT). The chatbot leverages efficient inference techniques like 4-bit quantization for performance optimization and RoPE scaling to handle extended conversations.

* 1. **Brief Description of the Problem Being Solved:**

In the medical field, patients, healthcare providers, and professionals frequently encounter complex medical terminology and abbreviations that are not easily understood without specific expertise. Moreover, accessing accurate and up-to-date medical information across different languages can be challenging, especially in multilingual environments. This can lead to misunderstandings, incorrect diagnoses, or a delay in obtaining proper medical advice. The issue is further compounded by the lack of user-friendly platforms capable of providing real-time, accurate medical information in a way that is accessible to both professionals and the public.

* 1. **Overview of the Proposed Solution (High-Level Idea):**

This project proposes a multilingual medical chatbot that leverages Generative AI to respond to medical inquiries in real-time, providing accurate, context-aware explanations of medical terminology, abbreviations, symptoms, diagnoses, and treatments. The system is built on top of the Gemma-7B model—a powerful language model fine-tuned using the MeDAL [1] dataset (Medical Dataset for Abbreviation Disambiguation)—and is optimized using LoRA (Low-Rank Adaptation) to reduce memory usage and speed up processing. The chatbot is integrated with a Streamlit-based user interface, which allows users to input queries in multiple languages and receive comprehensive medical information instantly.

* 1. **Unique Value Proposition:**

The project's unique value lies in its ability to:

* **Medical Expertise and Abbreviation Disambiguation:** Unlike general-purpose chatbots, this system is trained specifically on MeDAL, a specialized dataset focused on medical abbreviations and terminology, ensuring high accuracy when dealing with complex medical jargon.
* **Support Multiple Languages:** The chatbot is designed to handle inquiries in multiple languages, including English, Arabic, and Spanish, making it accessible to users from diverse linguistic backgrounds.
* **Efficient and Scalable Model**: By utilizing LoRA to optimize the model, the project achieves faster performance and lower resource consumption, making it scalable and cost-effective for widespread use, even on devices with limited computational power.
* **User-Friendly Interface:** Through the Streamlit interface, the chatbot provides an intuitive and seamless experience for users, whether they are medical professionals looking for detailed explanations or patients seeking easy-to-understand advice. This design ensures that both technical and non-technical users can interact with the chatbot effortlessly.

# **Motivation and Problem Statement**

* 1. **Motivation**

In today's rapidly evolving healthcare environment, both patients and professionals face challenges in accessing accurate and timely medical information. For patients, understanding complex medical terminology, diagnoses, or treatment plans can be overwhelming, particularly when unfamiliar medical abbreviations are used or when information is not available in their native language. Healthcare providers, on the other hand, need to stay up-to-date with the latest medical developments, often requiring quick access to detailed and specialized information during patient consultations. Moreover, in multilingual settings, the need for accurate medical communication across different languages becomes even more critical.

The motivation behind this project is to bridge these gaps by creating an accessible, multilingual chatbot capable of providing real-time, accurate responses to medical inquiries. By leveraging state-of-the-art machine learning models, this solution can help both patients and healthcare professionals’ access medical knowledge more easily and confidently, while also supporting diverse linguistic needs.

* 1. **Problem Statement**

The key problem this project addresses is the lack of accessible, accurate, and real-time medical information for both the general public and healthcare providers, particularly in multilingual environments. The challenges include:

* **Understanding complex medical terminology and abbreviations:** Medical jargon, especially abbreviations, is often difficult for patients to interpret, leading to confusion or misunderstandings about diagnoses and treatments.
* **Language barriers in accessing medical information**: Non-English speakers or those less proficient in medical terminology struggle to find reliable medical resources in their native languages.
* **Limited resources for healthcare professionals:** Healthcare providers often need fast, accurate access to specialized information during consultations, and traditional resources may not offer the immediacy required.

This project seeks to address these challenges by developing a chatbot that delivers precise, multilingual, and contextually relevant medical information to end users, enhancing both patient understanding and professional efficiency.

# **Objectives**

The primary objectives of the project are:

* 1. **Develop a Multilingual Medical Chatbot:**
* Create a chatbot capable of interpreting and responding to medical inquiries in multiple languages, including English, Arabic, and Spanish. The chatbot will leverage generative AI to provide real-time, accurate answers to medical questions, improving accessibility for users across different linguistic backgrounds.
  1. **Enable Medical Abbreviation Disambiguation:**
* Implement a robust system for disambiguating medical abbreviations. This feature will allow the chatbot to recognize and clarify complex medical abbreviations, ensuring users receive clear and understandable explanations of medical terms.
  1. **Provide Real-Time, Accurate Medical Information:**
* Ensure the chatbot delivers real-time responses that are based on verified medical data, primarily from the MeDAL dataset. The system will aim to provide contextually accurate and relevant answers, tailored to both professional medical practitioners and patients seeking healthcare advice.
  1. **Optimize Performance and Resource Usage:**
* Use Low-Rank Adaptation (LoRA) to optimize the chatbot model, ensuring efficient use of computational resources while maintaining high performance. This optimization will enable the chatbot to run effectively on devices with limited hardware capabilities, ensuring scalability and cost-effectiveness.
  1. **Create a User-Friendly Interface:**
* Develop an intuitive and easy-to-use interface using Streamlit, allowing users to seamlessly interact with the chatbot. The interface will cater to both technical and non-technical users, ensuring a smooth and accessible user experience for a diverse audience.
  1. **Support Healthcare Providers with Quick, Accurate Data Access:**
* Empower healthcare professionals by providing them with quick, on-demand access to detailed medical information, supporting them in making informed decisions during patient consultations.

By achieving these objectives, the project will deliver a powerful, scalable solution that improves the accessibility and accuracy of medical information for both healthcare professionals and patients.

# **End users and Features**

* 1. **End User Personas:**

1. **Patients and the General Public:**

* Individuals seeking reliable medical information and clarification on symptoms, treatments, or medical terms, without necessarily having a healthcare background.
* Users in non-English-speaking regions or those who prefer interacting with technology in their native language (e.g., Arabic, Spanish).

1. **Healthcare Providers:**

* Doctors, nurses, and other healthcare professionals who require quick, accurate interpretations of medical abbreviations or up-to-date medical information during patient interactions.
* Professionals are seeking to streamline their workflow by using a tool that efficiently clarifies complex medical terminology.

1. **Medical Students and Researchers:**

* Individuals involved in medical education or research need access to accurate explanations of terminology, abbreviations, and medical concepts.
* Users are looking to enhance their learning by interacting with an AI that provides detailed information based on comprehensive medical datasets.
  1. **Key Features that Address User Needs**

1. **Multilingual Medical Query Support:**

* Allows users to ask medical-related questions in multiple languages (including Arabic, English, and Spanish).
* Supports global access, providing a user-friendly experience for non-English-speaking individuals.

1. **Medical Abbreviation Disambiguation:**

* The chatbot can interpret and disambiguate complex medical abbreviations, offering expanded definitions and context-sensitive explanations.
* This feature helps professionals quickly understand medical jargon and assists patients in comprehending complex diagnoses.

1. **Real-Time Responses:**

* Provides instant, accurate answers to medical queries, including detailed explanations of symptoms, treatments, and conditions.
* Enhances the user experience by delivering immediate, trustworthy information at the point of need.

1. **Streamlit-Based User Interface:**

* Offers a simple, intuitive interface that allows users to input their queries and receive responses effortlessly.
* Includes additional options for displaying data insights (e.g., memory usage, dataset information) to enhance transparency and usability.

1. **Optimized for Resource Efficiency (LoRA Integration):**

* The chatbot is optimized to function with low resource consumption while maintaining high performance. This allows for scalability, making the solution suitable for devices with limited computing power or low-cost deployment environments.
* Key for institutions with budget constraints or users in regions with limited technology access.
  1. **How the Features Solve End-User Problems**

1. **For Patients and the General Public:**

* Problem: Difficulty in understanding complex medical terms and limited access to accurate medical information in their native language.
* Solution: The multilingual support ensures that users can communicate in their preferred language, while the real-time responses provide immediate, accurate, and easy-to-understand medical information. The medical abbreviation disambiguation clarifies technical terms, making healthcare knowledge accessible to non-experts.

1. **For Healthcare Providers:**

* Problem: Healthcare professionals often encounter unfamiliar abbreviations or need quick references to medical concepts in fast-paced environments.
* Solution: The medical abbreviation disambiguation feature allows providers to efficiently interpret abbreviations during patient consultations, improving decision-making speed. The real-time response capability ensures they get accurate information instantly, enhancing patient care quality.

1. **For Medical Students and Researchers:**

* Problem: Medical students and researchers often need detailed information about specific conditions or abbreviations that may not be readily available from a single source.
* Solution: The detailed explanations provided by the chatbot offer valuable insights into medical terms, supporting educational and research needs. The user interface makes the process of querying for information simple and fast, saving time during study sessions or research tasks.

# **Data Flow**

* 1. **Data Collection:**
* Data is collected from **MeDAL** in the form of **CSV files** that contain medical abbreviations and their corresponding explanations and contexts.
* The dataset is structured with fields such as text (the medical term/abbreviation), location (where in the document the abbreviation appears), and label (the expanded form or meaning of the abbreviation).
  1. **Data Storage:**
* Since the dataset is non-relational, the data is loaded dynamically using Python libraries such as datasets and pandas, which provide access to the CSV data in-memory for efficient retrieval and use by the model.
* The model itself doesn't require long-term storage of the data, as it processes input dynamically.
  1. **Data Access:**
* When a user queries the system, the chatbot processes the input, uses the MeDAL dataset to interpret abbreviations or medical terms, and generates a real-time response.
* Data access involves loading and parsing the CSV files, tokenizing the medical data, and then passing the parsed data into the AI model to interpret and generate accurate medical information.
  1. **CSV Data Features:**
* **Features**: The CSV files used in this project contain fields like text, label, and location, where text represents the medical term or abbreviation, label provides the expanded meaning, and location gives context for the term’s use.
* **Shape**: Typically, the dataset includes several thousand rows, each representing a unique medical abbreviation or term.
* **Balanced or Not**: The dataset is generally balanced as it covers a wide range of medical abbreviations and terms across various medical domains (respiratory, cardiovascular, etc.), ensuring fair representation.

# **Programming Languages + Frameworks**

* 1. **Programming Languages**
* **Python**: The primary programming language used for this project. Python was selected due to its versatility, rich ecosystem of machine learning libraries, and ease of integration with frameworks like **Hugging Face Transformers** and **Streamlit**.
  1. **Frameworks and Tools**

1. **Transformers (by Hugging Face):**

* **Purpose**: Used to load, fine-tune, and run the **Gemma-7B** model for generating medical text responses. The Transformers library provides efficient model loading, training, and deployment functionalities.
* **Why**: It is the go-to library for handling large language models like **Gemma-7B**, offering pre-trained models and tokenizers that are easy to use with state-of-the-art performance.

1. **Unsloth:**

* **Purpose**: Used to download and manage the **FastLanguageModel** versions like **Gemma-7B**. This tool is specifically designed for handling generative AI models that require efficient training and inference.
* **Why**: Provides a streamlined API for handling large-scale models, with built-in optimizations such as **LoRA** (Low-Rank Adaptation).

1. **Streamlit:**

* **Purpose**: A Python-based framework for building web applications, used here to create an interactive user interface for the chatbot.
* **Why**: Streamlit allows for fast, straightforward deployment of web apps without needing extensive front-end development experience. This makes it ideal for quickly building and deploying the chatbot interface.

1. **PyTorch:**

* **Purpose**: A deep learning framework used to implement and run the **Gemma-7B** model on GPUs. It is also used for handling tensor operations and model inference.
* **Why**: PyTorch is widely used for machine learning and is compatible with **Hugging Face Transformers**. It offers GPU acceleration for handling large models like **Gemma-7B** efficiently.

1. **Datasets (by Hugging Face):**

* **Purpose**: A library used to load and preprocess the **MeDAL dataset**. It handles the structured CSV data and prepares it for model input.
* **Why**: Datasets provides an efficient way to load, process, and manage large datasets, making it easy to integrate with machine learning workflows.
  1. **Supporting Technologies**

1. **APIs:**

* The **Streamlit** application can be enhanced by integrating with **external medical APIs** or knowledge bases if additional real-time medical data is required. APIs could provide up-to-date medical articles, symptom checkers, or drug information.

1. **Cloud Platforms:**

* **Hugging Face Model Hub**: If the fine-tuned model is hosted for public access, it can be deployed directly through the Hugging Face API, simplifying model updates and versioning.

# **Methodology**

* 1. **Development Workflow:**

(data collection → preprocessing → model fine-tuning → testing). This workflow ensures an organized approach to the chatbot's development, from gathering data to testing the final model.

1. **Data Collection**: The primary dataset used is the **MeDAL (Medical Dataset for Abbreviation Disambiguation)**, which contains well-organized medical terms, abbreviations, and their expanded definitions.
2. **Preprocessing**: Data is preprocessed to ensure compatibility with the model, including tokenization, removing special characters, and inserting EOS tokens.
3. **Model Fine-Tuning**: The pre-trained **Gemma-7B** model, from the **Unsloth** library, is fine-tuned using the **MeDAL dataset**, optimized through **LoRA** to minimize resource consumption.
4. **Testing**: After fine-tuning, the chatbot was tested for accuracy, responsiveness, and the ability to handle various medical queries in multiple languages.
   1. **Data Preparation**
      1. **Dataset Used:**

* **MeDAL Dataset:** The dataset used is the MeDAL (Medical Dataset for Abbreviation Disambiguation), which is a comprehensive source for medical abbreviations and terminology. This dataset is essential for training the chatbot to recognize and explain medical terms accurately, especially within multilingual contexts.
  + 1. **Preprocessing Steps:**
* **Tokenization:** Prepared data for the model using Hugging Face tokenizers to ensure the model interprets input properly.
* **EOS Token Insertion:** End-of-sequence (EOS) tokens were added to prevent infinite generation loops during response generation.
  1. **Model Fine-Tuning**
* **Parameter-Efficient Fine-Tuning:**
* **LoRA (Low-Rank Adaptation)**: The project leverages **LoRA** to fine-tune the pre-trained **Gemma-7B** model. This method minimizes computational overhead by allowing only specific layers of the model to be fine-tuned, which reduces memory consumption and speeds up the training process.
* **Gradient Checkpointing**: Gradient checkpointing is enabled during fine-tuning to further reduce memory usage, allowing the model to handle larger input sequences without requiring excessive computational resources.
  1. **Models Used:**
  + Gemma-7B: The core model used in the project, Gemma-7B, is a pre-trained model optimized for handling medical-related text generation tasks. It was fine-tuned using the MeDAL dataset and further optimized for efficient inference.
  1. **Inference Optimization**

**1. Quantization:**

* + Applied 4-bit quantization for faster inference and reduced memory usage.

**2. RoPE Scaling:**

* + Implemented Rotational Positional Embeddings (RoPE) to handle longer input sequences effectively, enabling the chatbot to maintain coherent, context-aware conversations.

# **Challenges and Solutions**

* **Dataset Size:** The MEDAL dataset was too large to process efficiently.
  + **Solution:** We used the Unsloth[2] library to manage loading and reduce memory constraints.
* **Memory Constraints:**
  + **Solution:** Implemented gradient checkpointing and 4-bit quantization to reduce memory usage during training and inference.
    - **Peak Reserved Memory**: 6.273 GB
    - **Memory Reserved During Training**: 1.773 GB (~12.02% of total system memory).
* **Long Sequence Handling:**
  + **Solution:** Used RoPE scaling to maintain conversation coherence across extended interactions.
* **Response Latency:**
  + **Solution:** Fine-tuned models with LoRA to improve inference speed for real-time responses.
* **Dynamic Responses with Constraints:** Ensuring the chatbot generates personalized responses while adhering to predefined constraints such as length limits and topic relevance.
  + **Solution:** Used prompt engineering and rule-based constraints to guide response generation, ensuring that responses are dynamically adjusted based on the context while adhering to the rules.

# **Testing and Results**

* **Training Time**:
  + **Total Time**: 634.1 seconds (~10.57 minutes)
* **Training Loss Reduction** (across 60 steps):
  + Initial Loss: **3.452600**
  + Final Loss: **2.207700**
* **Memory Usage**:
  + **Peak Reserved Memory**: 6.273 GB
  + **Memory Usage During Training**: 1.773 GB (~12% of system memory).

# **Conclusion**

In this project, we successfully developed a personalized multilingual medical chatbot using fine-tuning techniques on the Hugging Face platform, specifically leveraging the Gemma-7B model with LoRA-optimized adaptations. The chatbot efficiently addresses the challenges of providing real-time, accurate medical information across multiple languages while maintaining resource efficiency through optimizations like 4-bit quantization and RoPE scaling.

The core objectives of enabling multilingual support, medical abbreviation disambiguation, and real-time medical information delivery were achieved, ensuring that both healthcare professionals and patients can access critical information in an accessible manner. Furthermore, the use of Streamlit provided a user-friendly interface, allowing for seamless interaction with the chatbot, catering to both technical and non-technical users.

Through rigorous testing, the chatbot demonstrated low latency, efficient memory usage, and robust handling of long-sequence conversations, making it a scalable and practical solution for environments with limited computational resources. The integration of MeDAL, a specialized dataset for medical abbreviation disambiguation, ensured that the chatbot consistently delivered accurate and contextually relevant responses.

Overall, the project successfully addresses the critical gap in multilingual medical assistance by delivering a scalable, efficient, and accurate medical chatbot that can serve patients, healthcare professionals, and medical students alike, enhancing both the accessibility and understanding of complex medical information..

# **References**

[1] “McGill-NLP/medal · Datasets at Hugging Face.” Accessed: Oct. 14, 2024. [Online]. Available: https://huggingface.co/datasets/McGill-NLP/medal

[2] “GitHub - unslothai/unsloth: Finetune Llama 3.2, Mistral, Phi & Gemma LLMs 2-5x faster with 80% less memory.” Accessed: Oct. 14, 2024. [Online]. Available: https://github.com/unslothai/unsloth