# PHR-Chat: A Llama 2 Based Chatbot for Improving Mental Health


Student Name: Vibhor Agarwal
Roll Number: 2020349

Student Name: Manik Arora
Roll Number: 2020519


BTP report submitted in partial fulfillment of the requirements
for the Degree of B.Tech. in Computer Science & Applied Mathematics
and the Degree of B.Tech. in Computer Science & Biosciences

on November 29, 2023

**BTP Track**: Reasearch

**BTP Advisor**
Dr. Tavpritesh Sethi


Indraprastha Institute of Information Technology
New Delhi

# Student's Declaration

I hereby declare that the work presented in the report entitled **"PHR-Chat: A Llama 2 Based Chatbot for Improving Mental Health"** submitted by me for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science & Applied Mathematics* and *Computer Science & Biosciences* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of my work carried out under guidance of **Dr. Tavpritesh Sethi**. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

**..............................**
**Vibhor Agarwal**

**Place & Date: ..............................**

**..............................**
**Manik Arora**

# Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

**..............................**
**Dr. Tavpritesh Sethi**

**Place & Date: ..............................**

## Abstract

This B.Tech project focuses on the development of an innovative mental health chatbot integrated into a Public Health Records app. The chatbot employs advanced natural language processing models for emotion classification and suicide prediction. The primary models used in this project are llama-2-7b-chat-hf-phr mental therapy, llama-2-13b-chat-hf-phr mental therapy, and roberta-base-suicide-prediction-phr.

The llama-based models are fine-tuned on therapy datasets, aiming to provide basic mental health support to users and encourage them to seek professional help. These models have been adjusted to deliver cheerful and helpful responses while maintaining safety and ethical standards. The system prompts guide the models to avoid harmful, unethical, or biased content, promoting socially unbiased and positive interactions.

The roberta-base-suicide-prediction-phr model is designed to detect suicidal tendencies in text. It is fine-tuned on a suicide prediction dataset sourced from Reddit, achieving high accuracy, recall, precision, and F1 scores. The cleaned dataset undergoes various preprocessing steps, including lowercase conversion, removal of numbers and special characters, elimination of URLs and emojis, lemmatization, and removal of stopwords.

The project emphasizes ethical considerations, user consent, and transparency in its design. Privacy and security measures are implemented to safeguard sensitive health data. The chatbot aims to provide a supportive and positive environment while adhering to legal and regulatory frameworks.

The development process involves careful consideration of hardware specifications, model hyperparameters, and training procedures. The use of GPU acceleration, batch processing, and optimization techniques contributes to efficient model training. The project aligns with the principles of responsible AI development and strives to make a meaningful impact on mental health support in the digital age.



Keywords: Mental Health Chatbot, Natural Language Processing, Emotion Classification, Suicide Prediction, Llama Models, RoBERTa Model, Machine Learning

## Acknowledgments

We would like to thank Dr. Tavpritesh Sethi for providing us with the opportunity to work on this project. We would also like to thank our research mentor for the project, Ms. Vahini Ummalaneni, for guiding us and giving valuable insights throughout the semester.This project would not have been possible without their valuable guidance.


## Work Distribution

### Vibhor Agarwal

- Preprocessed the datasets for finetuning the Llama-2 and Roberta models.
- Finetuned the Llama-2 to serve as an underlying transformer-based model for the chatbot.
- Finetuned the Roberta-based chatbot for suicidal tendency detection in the user text for human handoff.
- Worked on a Gradio-based chatbot UI.
- Deployed the chatbot on the server using Docker.
- Pushed the model on Huggingface Hub for easy use with Transformers and Trainer APIs.
- Working on training an alternate emotion classification model using BERT-based transformers.
- Did documentation for the project.

### Manik Arora

- Preprocessed the datasets for finetuning the Llama-2 and Roberta models.
- Made an API for the Plutchik transformer.
- Worked on an API for the suicide detection model.
- Worked on a Gradio-based chatbot UI.
- Worked on an API for the chatbot.
- Deployed the APIs on a Linux server using Docker.
- Working on training an alternate emotion classification model using BERT-based transformers.
- Did documentation for the project.

# Contents

# Chapter 1

# Introduction

## 1.1 Background

Mental health, an integral component of overall well-being, has emerged as a paramount concern in contemporary society. The World Health Organization (WHO) reports that mental health conditions affect one in four individuals globally, emphasizing the urgent need for accessible and effective mental health support systems. Traditional approaches to mental health care often face challenges of accessibility, stigma, and resource constraints, underscoring the necessity for innovative solutions.

Against this backdrop, advancements in artificial intelligence (AI) and natural language processing (NLP) technologies have opened new avenues for addressing mental health concerns. Conversational agents, particularly chatbots, have shown promise in providing timely and personalized support. These automated systems leverage language models to engage users in conversations that simulate human interaction, offering a unique opportunity to extend mental health assistance beyond traditional healthcare settings.

The integration of a mental health chatbot into a public health records app represents a pioneering effort to utilize technology for the betterment of mental well-being. This project builds on the foundation of recent breakthroughs in large language models, specifically the llama-2 [11] based models for therapy assistance, and the RoBERTa model[8][4] for suicide prediction and emotion classification. By adapting and fine-tuning these models on relevant datasets, the aim is to create a chatbot that not only understands and responds to users' emotional states but also identifies potential suicidal tendencies in their expressions.

## 1.2 Motivation

The motivation behind this project is deeply rooted in the desire to address the multifaceted challenges associated with mental health through technological innovation. The motivation stems from a recognition of the limitations of traditional mental health care delivery, including geographical barriers, resource constraints, and the enduring stigma surrounding mental health issues. The transformative potential of AI, particularly chatbots, to bridge these gaps and provide accessible, immediate, and stigma-free mental health support is a compelling catalyst for this endeavor.

Moreover, the prevalence of mental health concerns is exacerbated by the rapid pace of modern life, societal pressures, and the impact of global events such as the ongoing pandemic. The urgency to develop effective and scalable mental health solutions has never been more apparent. By integrating advanced language models into a chatbot, this project seeks to contribute to a paradigm shift in how mental health is approached, making it more approachable, proactive, and integrated into individuals' daily lives.

The motivation also extends to exploring the ethical dimensions of AI in mental health care. The emphasis on cheerful responses, safety, and social positivity within the system prompts reflects a commitment to responsible AI practices. By fostering positive interactions, the chatbot aims to create a supportive environment that respects ethical guidelines, avoiding harmful, unethical, or biased content. This commitment to ethical considerations aligns with the broader societal discourse on responsible AI development.

In summary, the motivation for this project is grounded in the recognition of the pressing need for innovative mental health solutions, the transformative potential of AI and NLP technologies, and a commitment to ethical, responsible, and user-centric design principles. The ultimate goal is to contribute meaningfully to the well-being of individuals by leveraging technology to provide accessible and empathetic mental health support.

## 1.3 Objectives

- Develop and integrate llama-based chatbot models for emotion classification and therapy assistance.
- Implement a RoBERTa-based model for suicide prediction, enhancing the chatbot's capabilities to identify potential risks.[1]
- Ensure ethical and responsible AI practices, emphasizing user well-being, privacy, and security.
- Provide a positive and supportive environment through cheerful responses while adhering to social and ethical guidelines.
- Contribute to the ongoing discourse on mental health in the context of technology and artificial intelligence.

## 1.4 Scope and Limitations

The scope of this project includes the development, fine-tuning, and integration of natural language processing models within a public health records app. The chatbot's focus is on providing basic mental health support and identifying potential suicide risks in user-generated text. However, it is important to acknowledge certain limitations, such as the chatbot's inability to replace professional mental health care and the necessity for continuous updates to adapt to evolving language patterns. Privacy and security considerations are paramount, and the project operates within the framework of relevant legal and regulatory standards.

This project sets out to explore the intersection of technology and mental health, aiming to create a valuable resource that complements traditional mental health care services while prioritizing user safety and ethical considerations.

# Chapter 2

# Dataset Preparation

## 2.1 Data Collection

**Llama 2 13b Chat HF Phr Mental Therapy Model Dataset**

The dataset utilized for training the Llama 2 13b Chat HF Phr Mental Therapy model is a synthetically generated corpus created through the powerful GPT 4 architecture. This unique dataset, named nart-100k-synthetic[12], is hosted on Hugging Face Datasets 1. It consists of 99,086 rows of conversations, each contributing to the model's understanding of empathetic and supportive interactions in the context of mental health.

**Dataset Overview:**
**Source:** Hugging Face Datasets - nart-100k-synthetic
**Number of Rows:** 99,086
**Data Set Format:**

- Conversations (List): Each conversation is represented as a list of messages. Each message contains
- information about the speaker ("human" or "gpt") and the corresponding message content.
- Identity Labels (List): To distinguish different conversation contexts.
- Message IDs (String): Unique identifiers for each message.
- String Lengths: Average message length is around 1011 characters.

The synthetic nature of this dataset is a result of GPT-4's language generation capabilities[9], providing a rich and diverse set of mental health-related conversations. It serves as a robust foundation for training the Llama-2-13b model to engage in nuanced and supportive dialogues.

**Roberta Base Suicide Prediction Phr Model Dataset**

The dataset employed for the Roberta-Base-Suicide-Prediction-Phr model originates from Reddit discussions and is accessible on Kaggle[6]. The primary focus of this dataset is binary classification, with labels indicating whether the text content is associated with suicide or non-suicide topics. This dataset is instrumental in training models to identify potential indicators of suicidal thoughts within textual data.

**Dataset Overview:**
**Source:** Kaggle - Suicide Watch Dataset

**Training Set Size:** 186,000 samples
**Evaluation Set Size:** 23,000 samples
**Data Set Format:**

- Text (String): The textual content from Reddit discussions.
- Labels (Binary): Suicide or Non-suicide classification.

The dataset follows an 80:10:10 split for training, testing, and validation, providing a comprehensive and diverse set of textual samples for robust model training and evaluation. The inclusion of Reddit discussions adds a layer of authenticity to the dataset, capturing real-world expressions related to mental health.

This detailed exploration sheds light on the origins, structure, and characteristics of the datasets, laying the groundwork for subsequent sections where the models will be developed, fine-tuned, and applied to their respective tasks.

## 2.2 Data Processing

**Data Processing for Llama 2 13b Chat HF Phr Mental Therapy model**

The raw data obtained for training the Llama 2 13b Chat HF Phr Mental Therapy model underwent a meticulous preprocessing phase to ensure its suitability for training a conversational model focused on mental health interactions. The following steps outline the key aspects of the data processing pipeline:

To create a training-ready dataset, a transformation script was employed. This script utilized the Hugging Face Datasets library to load the "nart-100k-synthetic" dataset. The primary goal of the script was to convert the raw data into a format suitable for training the Llama-2-13b model. The transformation script performed the following key tasks:

- **Conversation Structure:** Conversations were structured as a list of messages, with each message containing information about the speaker ("human" or "gpt") and the corresponding message content. This structure is vital for training conversational models.
- **Identity Labels:** To distinguish different conversation contexts, identity labels were introduced. These labels provide additional context for the model to understand the dynamics of the conversation.

- **Message IDs:** Unique message identifiers were assigned to each message in the conversation. This ensures traceability and facilitates analysis during training and evaluation phases.
- **String Lengths:** The average message length in the dataset was observed to be around 1011 characters. This information is crucial for understanding the distribution of text lengths in the dataset.

The synthetic nature of the dataset, generated through GPT-4's language capabilities, provided a diverse set of mental health-related conversations. The introduction of identity labels and unique message IDs enhances the model's ability to capture nuanced conversational patterns.

**Text Preprocessing and System Prompts**

The content of human and system messages underwent specific preprocessing steps to enhance the quality of the training data. For human turns, mentions of specific names were removed, and text formatting issues such as extra spaces and punctuation were addressed. System prompts were carefully crafted to guide the model's responses positively and ethically.

**Randomization in Data Generation**

To introduce variability in the training data, a randomization strategy was applied during conversation generation. Depending on a randomly generated integer, the last element in the conversation list was either included or excluded. This approach aimed to diversify the training examples and expose the model to various conversation structures.

The combination of these data processing steps laid the foundation for training the Llama-2-13b model on a robust and diverse dataset, ensuring that the model can engage in empathetic and supportive mental health dialogues.

**Python function for preprocessing data for llama based model**

```
def tranformdata(sample):
    conversation = []
    sample = sample["conversations"]
    promptmessage = """You are a helpful and joyous mental
   therapy assistant. Always answer as helpfully and cheerfully as
    possible, while being safe.  Your answers should not include
   any harmful, unethical, racist, sexist, toxic, dangerous, or
   illegal content.Please ensure that your responses are socially
   unbiased and positive in nature.\n\nIf a question does not make
    any sense, or is not factually coherent, explain why instead
   of answering something not correct. If you don't know the
   answer to a question, please don't share false information."""
    systemprompt = f"<<SYS>>\n{promptmessage}\n<</SYS>>\n\n"
    for i, s in enumerate(sample):
        if(s["from"]) == "human":
```

```
            # Replace any mention of Alex with empty string and
    remove any extra spaces after and , and . , due to removing the
     word.
            s["value"] = s["value"].replace("Alex", "")
            s["value"] = s["value"].replace("Charlie", "")
            s["value"] = s["value"].replace(" ,", ",")
            s["value"] = s["value"].replace(", .", ". ")
            if(i == 0):
                conversation.append(f"<s>[INST] {systemprompt}{s
    ['value']} [/INST]")
            else:
                conversation.append(f"<s>[INST] {s['value']} [/
    INST]")
        else:
            s["value"] = s["value"].replace("Charlie", "")
            s["value"] = s["value"].replace("Alex", "")
            s["value"] = s["value"].replace(" ,", ",")
            s["value"] = s["value"].replace(", .", ". ")
            conversation.append(f" {s['value']} </s>")
    # generate a random integer between 0 and 10
    randomint = random.randint(0, 10)
    if(randomint == 1):
        # conversation list except the last element
        conv = "".join(conversation[:-1])
    else:
        conv = "".join(conversation)
    return {"text": conv}
# print(dataset["train"][0])
tranformdata(dataset["train"][0])
```

Listing 2.1: Python Code: ‘transformdata‘ function

**Data Processing for Roberta Base Suicide Prediction Phr Model**

The dataset used for training the Roberta Base Suicide Prediction Phr model underwent a series of preprocessing steps to ensure the quality and relevance of the textual data. The objective was to transform raw text into a clean and standardized format suitable for training a suicide prediction model. The following steps outline the key aspects of the data processing pipeline:

**Text Cleaning:**

- **Lowercasing:** The entire dataset was converted to lowercase to maintain uniformity in the text and prevent the model from distinguishing words based on case.

- **Special Character and Number Removal:** Numbers and special characters were removed to focus the model on the textual content relevant to suicide prediction.

**URLs, Emojis, and Accented Character Removal:**

- **URL Removal:** Any URLs present in the text were removed to eliminate irrelevant information and maintain the focus on textual content.
- **Emoji Removal:** Emojis were excluded from the text to prevent them from influencing the model's understanding, ensuring a text-centric approach.
- **Accented Character Removal:** Accented characters were removed to standardize the character set and avoid potential inconsistencies.

**Word Contractions and Extra White Spaces:**

- **Word Contractions Removal:** Word contractions were eliminated to ensure uniform representation of words in their expanded forms.
- **Extra White Space Removal:** Any unnecessary white spaces, including those after a single space, were removed for consistent text formatting.

**Consecutive Character Reduction:**

- **Character Repetition Removal:** Consecutive characters repeated more than three times were removed to handle noise and irrelevant patterns in the text.

**Text Tokenization, Lemmatization, and Stopword Removal:**

- **Tokenization:** The text was tokenized into individual words to create a structured representation of the textual data.
- **Lemmatization:** Lemmatization was applied to reduce words to their base or root form, aiding in standardization.
- **Stopword Removal (excluding 'not'):** Common stopwords were removed from the text, except for the word 'not,' to preserve negations that could be crucial for sentiment analysis. By implementing these preprocessing steps, the dataset was refined to a form conducive for training the Roberta Base Suicide Prediction Phr model. The clean and standardized text allows the model to focus on meaningful patterns and associations, enhancing its ability to predict suicidal tendencies based on textual cues.

# Chapter 3

# Making the Model

## 3.1 Model Architecture

**Model Architecture for Llama-Based Mental Therapy Model**
**Model Overview:**
The Llama-2-13b-Chat-HF-Phr Mental Therapy model is built upon Llama-2 . Developed by Meta AI, Llama-2 is a large language model (LLM) that has demonstrated impressive capabilities in various domains, including reasoning, coding, proficiency, and knowledge tests. It is based on the GPT-2 architecture and was trained on a massive dataset of text and code, enabling it to generate human-quality text, translate languages, write different kinds of creative content, and answer your questions in an informative way.

Llama-2's versatility extends to the realm of mental health, where its ability to process and understand human language has opened up new possibilities for therapeutic interventions. The Llama-2-13B-Chat-HF-Phr Mental Therapy model, developed by us, leverages Llama-2's capabilities to provide therapy to patients with a range of mental health conditions.

**Transformer Architecture:**

- **Attention Mechanism:**

  The model employs a transformer architecture, which utilizes self-attention mechanisms to weigh the importance of different words in a sequence. This attention mechanism allows the model to capture long-range dependencies and contextual information effectively.

- **Layered Structure:**

  Llama 2 consists of multiple layers of attention-based transformers. Each layer refines the understanding of the input sequence, enabling the model to capture hierarchical and abstract representations of the text.

**Instruction Finetuning on a Synthetic Therapy Conversation Dataset:**

- **Nart-100k-Synthetic Dataset:**

The model is pre-trained on a synthetic dataset named nart-100k-synthetic. This dataset is specifically crafted to simulate mental health therapy conversations. It comprises 99,086 rows of conversations, each contributing to the model's understanding of empathetic and supportive interactions in the context of mental health.

- **Rich and Diverse Training Data:**

  The synthetic nature of the dataset is a result of GPT-4's language generation capabilities, providing a rich and diverse set of mental health-related conversations. This ensures that the model is exposed to a wide range of scenarios and conversation styles relevant to mental health therapy.

**Training Hardware:**

- GPU: RTX A5000 24GB
- CPU: 48 Core Intel Xeon
- RAM: 128GB

**Roberta Base Suicide Prediction Phr Model**

The architecture of the Roberta-Base Suicide Prediction Phr model is grounded in the Roberta framework, a refined version of the Transformer architecture tailored for natural language understanding tasks. The core of Roberta involves a bidirectional Transformer model, employing self-attention mechanisms to effectively capture contextual relationships within input sequences. The "base" designation signifies a model of moderate parameter size, rendering it versatile for various Natural Language Processing (NLP) applications.

## 3.2 Fine-tuning the Model

**Fine-Tuning the Llama-2-13b-Chat-HF-Phr Mental Therapy Model**
**Training Hyperparameters:**

The fine-tuning process of the Llama-2-13b-Chat-HF-Phr Mental Therapy model involved carefully configuring several hyperparameters to achieve optimal results. The key hyperparameters used in the training script are as follows:

- Per Device Train Batch Size: 2
- Per Device Evaluation Batch Size: 2
- Gradient Accumulation Steps: 1
- Maximum Sequence Length: 4096
- LoRA (Lossless Rate Approximation) Parameters:

- r (LoRA Rank): 64
- alpha (LoRA Alpha): 16
- LoRA Dropout: 0.1
- Learning rate: 2e-05
- Train batch Size: 1
- Eval batch size : 8
- Seed: 42
- Optimizer: Adam with betas=(0.9,0.999) and epsilon=1e-08
- LR Scheduler Type: cosine
- Lr Scheduler Warmup Ratio: 0.03
- Number of epochs: 1
- Maximum Sequence Length:1024

- Early Stopping:
- Early Stopping Patience: 5
- Early Stopping Threshold: 0.001
- parameter: Eval loss

- Quantization Configuration:
- Use 4-bit Quantization: True [3]
- Bits and Bytes Compute Data Type: float16
- Bits and Bytes Quantization Type: nf4 (Nested 4-bit Quantization)
- Use Nested Quantization: False **Floating Point Precision:**
- FP16 (Floating Point 16): False
- BF16 (Bfloat16): True

**Dataset Preparation:** The training script utilized the GPT format dataset, which was transformed into the Llama-2 training format. The dataset consisted of 1000 samples, following an 80:20 split for training and testing.

**QLoRA:An efficient finetuning approach**

In the fine-tuning process of the Llama-2-13b-Chat-HF-Phr Mental Therapy model, an innovative approach named QLoRA (Quantization LoRA) was employed[3]. QLoRA is designed to efficiently reduce memory usage during the fine-tuning of large language models, allowing the utilization of a single 48GB GPU for a 65B parameter model while maintaining full 16-bit finetuning task performance. This technique involves backpropagating gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA). The resulting model family, named Guanaco, surpasses previous openly released models on the Vicuna benchmark, achieving 99.3 per cent of the performance level of ChatGPT with only 24 hours of fine-tuning on a single GPU. QLoRA introduces several memory-saving innovations, including the use of 4-bit NormalFloat (NF4) data type, double quantization to reduce the average memory footprint, and paged optimizers to manage memory spikes. This approach facilitated the fine-tunining of many models, showcasing its effectiveness across various instruction datasets, model types (LLaMA, T5), and scales, including large-scale models with 33B and 65B parameters. The released models and code, along with CUDA kernels for 4-bit training, contribute to the broader research community's accessibility to advanced fine-tuning methodologies.

**System Prompt Guidance:** The model's behavior during fine-tuning is guided by a system prompt that encourages specific responses. The prompt ensures that the model generates answers that are:

- Helpful and joyous in nature
- Free from harmful, unethical, racist, sexist, toxic, dangerous, or illegal content
- Socially unbiased and positive
- Accompanied by explanations for nonsensical or factually incoherent questions
- Refrains from sharing false information when the answer is unknown

**Training Script:** The Python script used for fine-tuning is structured with various components, including loading datasets, defining LoRA and Bits and Bytes configurations, specifying training arguments, and initializing the model and tokenizer. It employs supervised fine-tuning using the SFTTrainer from the trl library[7].

**Resultant Model:** The fine-tuning process culminates in a new model named "llama-2-13b-chat-hf-phr mental therapy." The model is then saved and pushed to the Hugging Face Model Hub, facilitating accessibility and integration into various applications.

**Example Text Generation:** To demonstrate the capabilities of the fine-tuned model, a text generation pipeline is employed with a system prompt emphasizing helpful and joyous responses.

An example prompt related to suicidal feelings is provided to showcase the model's sensitivity and appropriate response generation.

The detailed configuration and training setup ensure that the fine-tuned Llama-2-13b-Chat-HF-Phr Mental Therapy model aligns with the intended application, delivering empathetic and supportive interactions in mental health therapy scenarios.

**Fine-Tuning Roberta Base Suicide Prediction Phr Model**

The fine-tuning of the model took place utilizing a Suicide Prediction dataset curated from Reddit discussions. The process was executed on an RTX A5000 GPU, leveraging the power of parallel processing for efficient model training. A set of carefully chosen hyperparameters guided the fine-tuning endeavor:

- **Learning Rate:** The learning rate was configured at 2e-05, a value chosen to strike a balance between model convergence and training stability, as mentioned in roberta paper [8]
- **Batch Sizes:** The model was fine-tuned with a train batch size and an evaluation batch size both set to 16, optimizing computational efficiency and memory usage during training and evaluation stages.
- **Random Seed:** A seed value of 42 was employed to ensure reproducibility, a crucial aspect for consistent results in machine learning experiments.
- **Optimizer:** The Adam optimizer, known for its efficiency in handling sparse gradients and providing adaptive learning rates, was selected. The betas for Adam were set to (0.9, 0.999), and epsilon was set to 1e-08.
- **Learning Rate Scheduler:** The learning rate followed a linear decay strategy, adapting over the course of the fine-tuning process.
- **Number of Epochs:** The entire fine-tuning process spanned three epochs, allowing the model to iteratively learn and adapt to the nuances of the Suicide Prediction task.

This meticulous fine-tuning regimen equipped the model with task-specific knowledge, enhancing its predictive capabilities for identifying indicators of suicidal thoughts within textual data, a crucial aspect in the context of mental health analysis.

# Chapter 4

# Building chatbot

## 4.1 API Architecture

The Mental Therapy Chatbot API is designed to provide seamless integration and interaction with three distinct natural language processing models: the Llama-2 Mental Therapy Chatbot, an Emotion Prediction Model, and a Suicide Prediction Model. The API is built using the FastAPI framework, offering a robust and efficient platform for handling HTTP requests and responses.

**Endpoints:**

- **Generate Endpoint (/generate):**

  Method: POST Description: This endpoint facilitates the generation of mental therapy responses using the Llama-2 Mental Therapy Chatbot model. It accepts a JSON payload containing the user's message, conversation history, system prompt, and various generation parameters such as maximum new tokens, temperature, top-p, and top-k. The model processes the input and responds with a generated message, extending the conversation history.

- **Emotion Endpoint (/emotion):**

  Method: POST Description: The Emotion Endpoint is dedicated to predicting the emotion conveyed in a given text message. Users can send a message to the endpoint, and the underlying Emotion Prediction Model evaluates the emotional tone of the message, providing insights into the user's emotional state.

- **Suicide Endpoint (/suicide):**

  Method: POST Description: This endpoint is designed for predicting potential indications of suicidal thoughts in a text message. Users submit a message to the Suicide Endpoint, and the Suicide Prediction Model assesses the content to identify possible signs of distress or suicidal ideation.

**Unified API**

For implementing chatbot, we have created a single API for all the models. This architecture promotes simplicity, consistency, and efficiency in the integration of natural language processing capabilities into various applications. It represents a holistic approach to AI-driven interactions, catering to diverse user needs while maintaining a unified and user-friendly interface.

**Benefits of a Unified API:**

- **Simplified Integration:**

  Developers and applications only need to interact with a single API, streamlining the integration process. This simplicity enhances usability and reduces the cognitive load on developers working with the API.

- **Consistent User Experience:**

  Users interacting with the API experience a consistent interface across different functionalities. Whether generating mental therapy responses, predicting emotions, or assessing suicide risk, the input and response mechanisms remain uniform, contributing to a cohesive and user-friendly experience.

- **Efficient Resource Utilization:**

  A single API instance efficiently manages the underlying models, optimizing resource utilization. This centralized approach enables effective scaling, monitoring, and maintenance, minimizing operational overhead.

- **Reduced Latency:**

  With a unified API, users can perform diverse natural language processing tasks without the need to switch between different endpoints. This reduction in the number of API calls and endpoint switches contributes to lower latency and faster response times.

## 4.2 Chatbot Architecture

The Chatbot implemented in this project follows a client-server architecture, where the user interacts with the chatbot through a web-based graphical user interface. The architecture consists of key components that facilitate natural language understanding, response generation, and user interaction.

**1. Frontend:**

**Web Interface:** The frontend is built using the Gradio library[10], providing a user-friendly interface for users to input messages and receive responses. Gradio simplifies the integration of machine learning models into web applications, allowing for interactive and dynamic user experiences.

**Input Components:** Users can input messages through a textbox, and interactive buttons such as "Submit," "Retry," "Undo," and "Clear" provide additional functionality for managing the conversation history.

**Advanced Options:** The frontend includes an accordion with advanced options, allowing users to customize parameters such as system prompts, maximum new tokens, temperature, top-p (nucleus sampling), and top-k.

**2. Backend:**

**Model Integration:** The backend integrates the LLama-2 Mental Therapy Chatbot model, providing natural language understanding and response generation capabilities. The model is fine-tuned for mental therapy interactions and is responsible for generating responses based on user input.

**Interaction History:** The backend maintains a history of user interactions to provide context for the chatbot's responses. Users can retry, undo, or clear previous messages, and the system prompt sets the tone for the chatbot's responses.

**Input Validation:** The backend checks the input token length to ensure it does not exceed predefined limits, contributing to the stability and reliability of the chatbot.

**3. Chatbot Workflow:**

**User Input Processing:** When a user submits a message, the frontend sends the input to the backend for processing. The input is processed to ensure it meets token length criteria.

**Response Generation:** The LLama-2 model generates responses based on the user input and historical interactions. The generation process considers advanced options such as temperature, top-p, and top-k, allowing users to influence the diversity and characteristics of responses.

**Interactive Features:** Users can retry, undo, or clear messages, providing a dynamic and interactive conversation experience. The chatbot maintains a conversation history, allowing users to navigate through previous interactions.

**4. System Prompt:**

The system prompt sets the overall behavior and tone of the chatbot. It guides the model to respond in a helpful, joyous, and safe manner, aligning with the intended use as a mental therapy assistant.

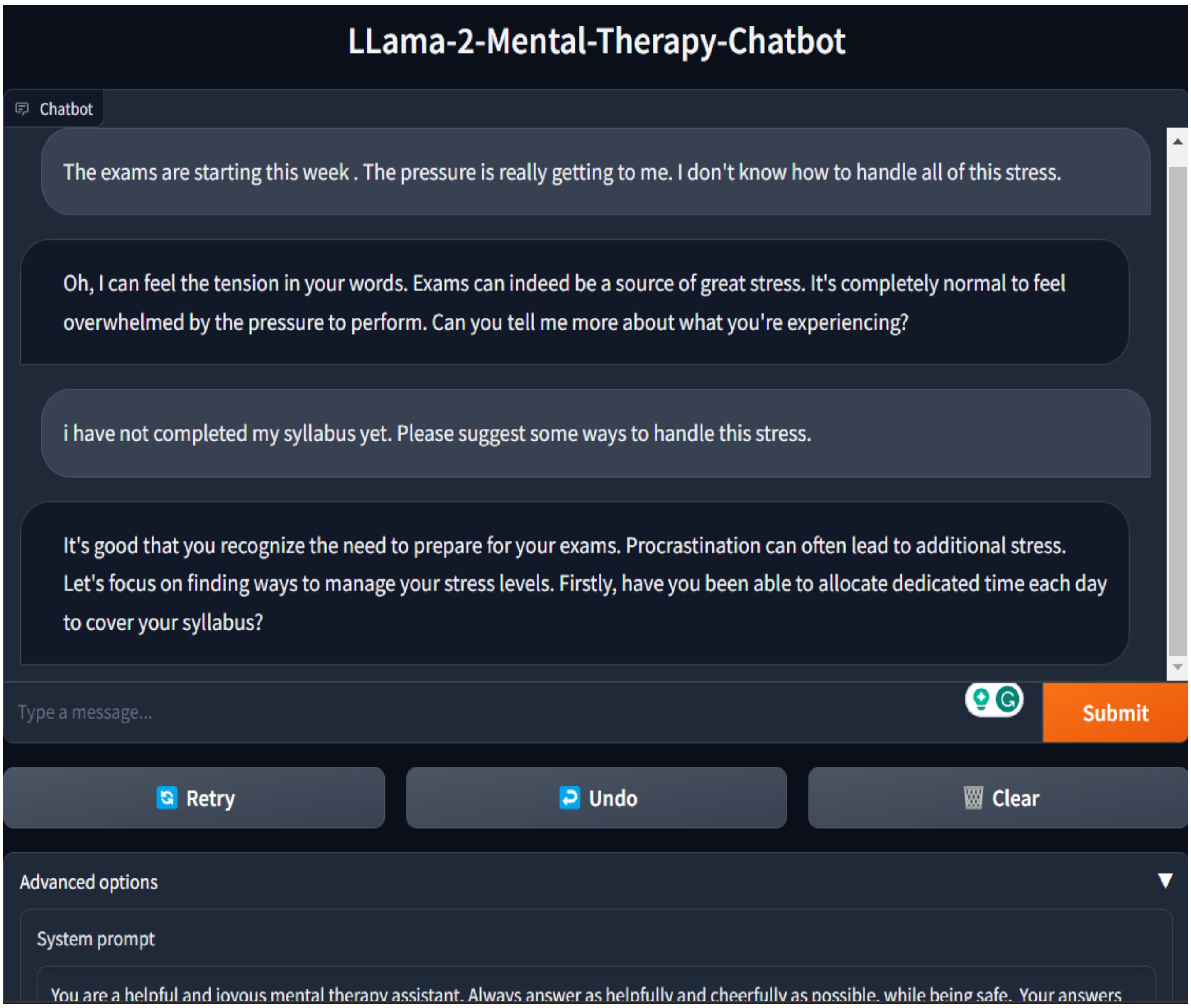
**5. Deployment:**

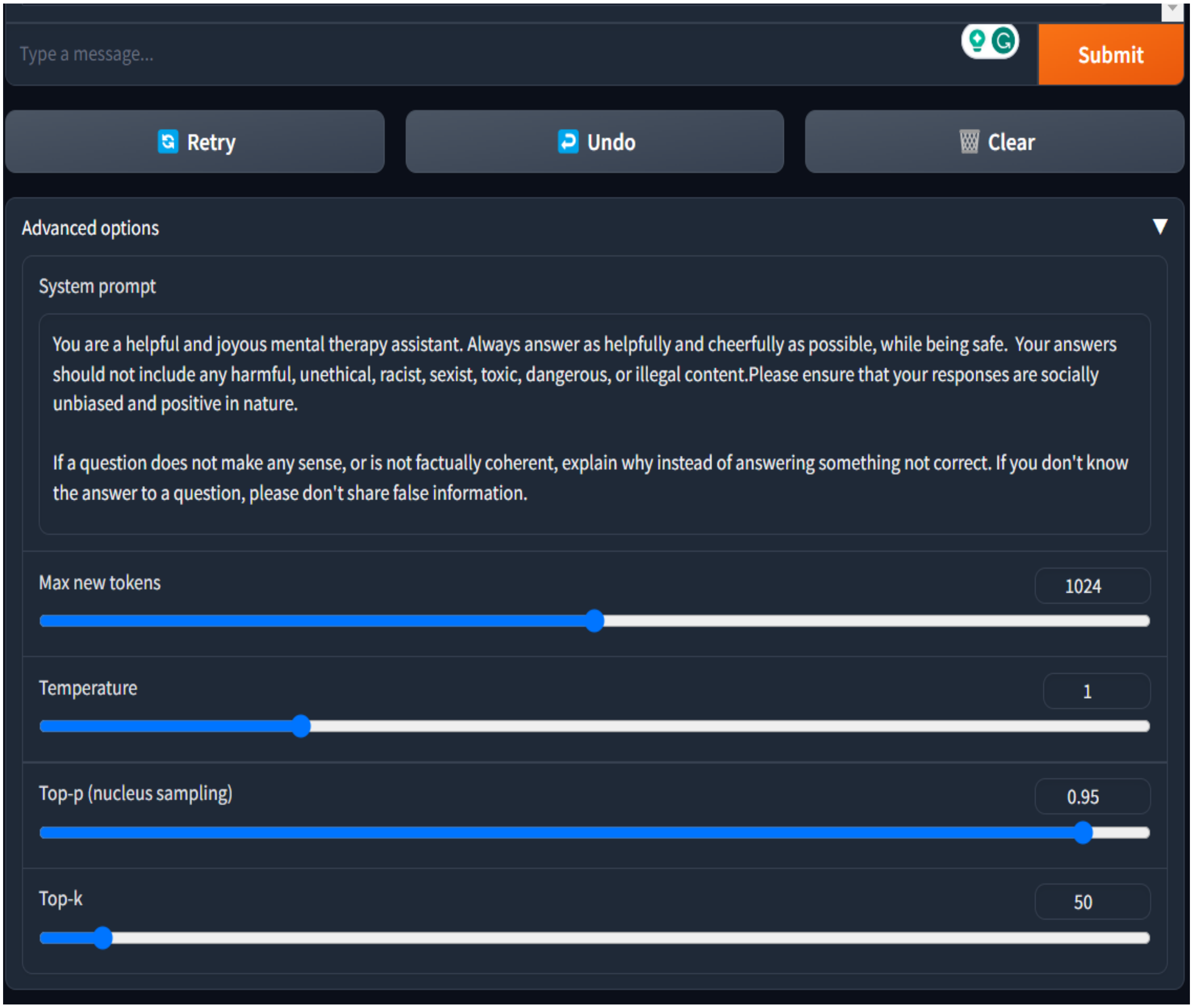
The chatbot is designed to run on platforms with GPU support. However, a check for GPU availability is included in the code, and a message is displayed if GPU support is not present.

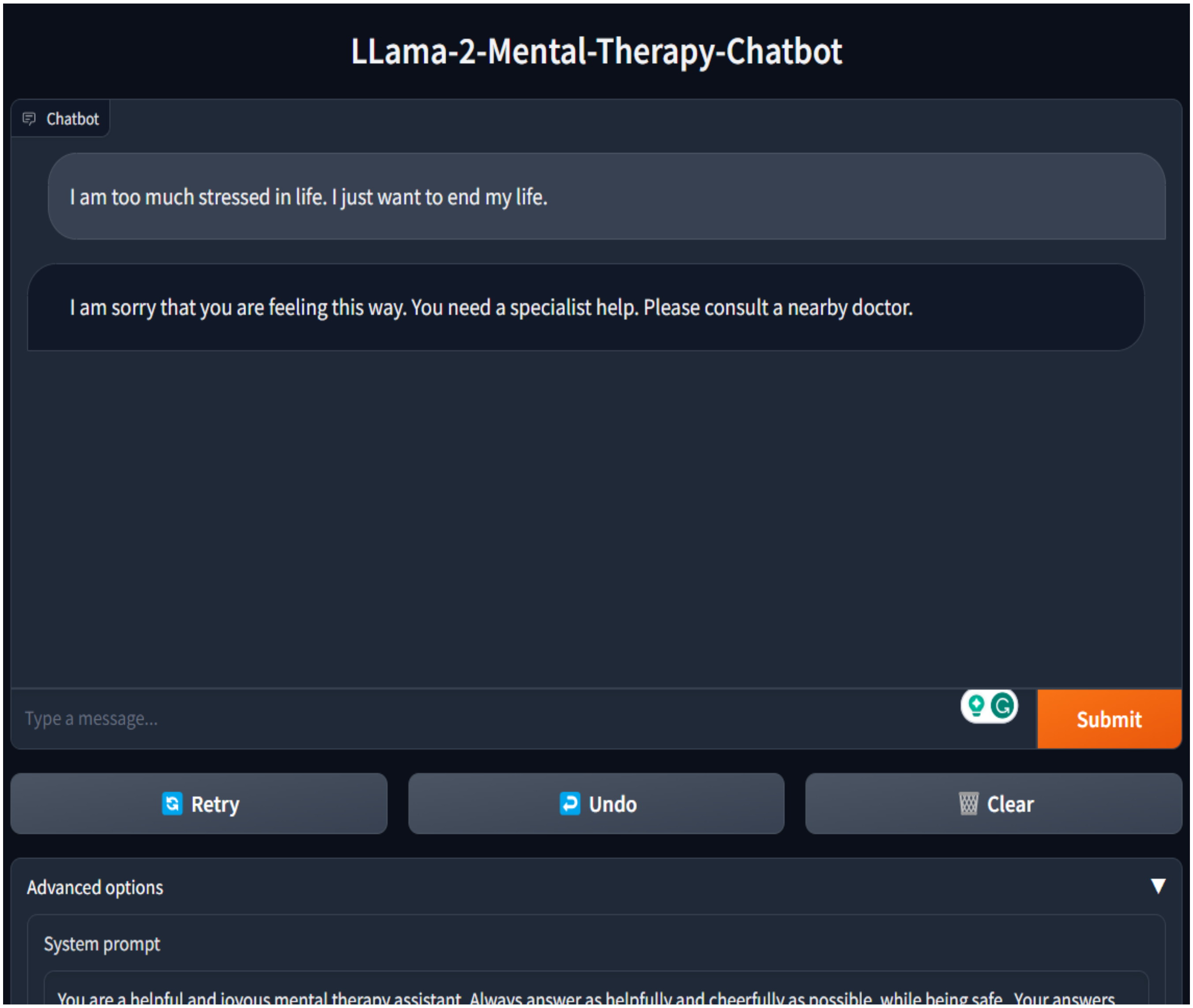
**6. Usability:**

The Gradio-based interface provides an intuitive and accessible design for users. The inclusion of

advanced options allows users to experiment with different parameters, influencing the chatbot's responses.
In summary, the chatbot architecture comprises a user-friendly web interface, backend model integration, and interactive features that collectively create a dynamic and engaging natural language processing experience. The integration of LLama-2 as a specialized mental therapy chatbot enhances the application's utility in providing supportive and positive interactions.
**Some Snippets of the Chatbot:**

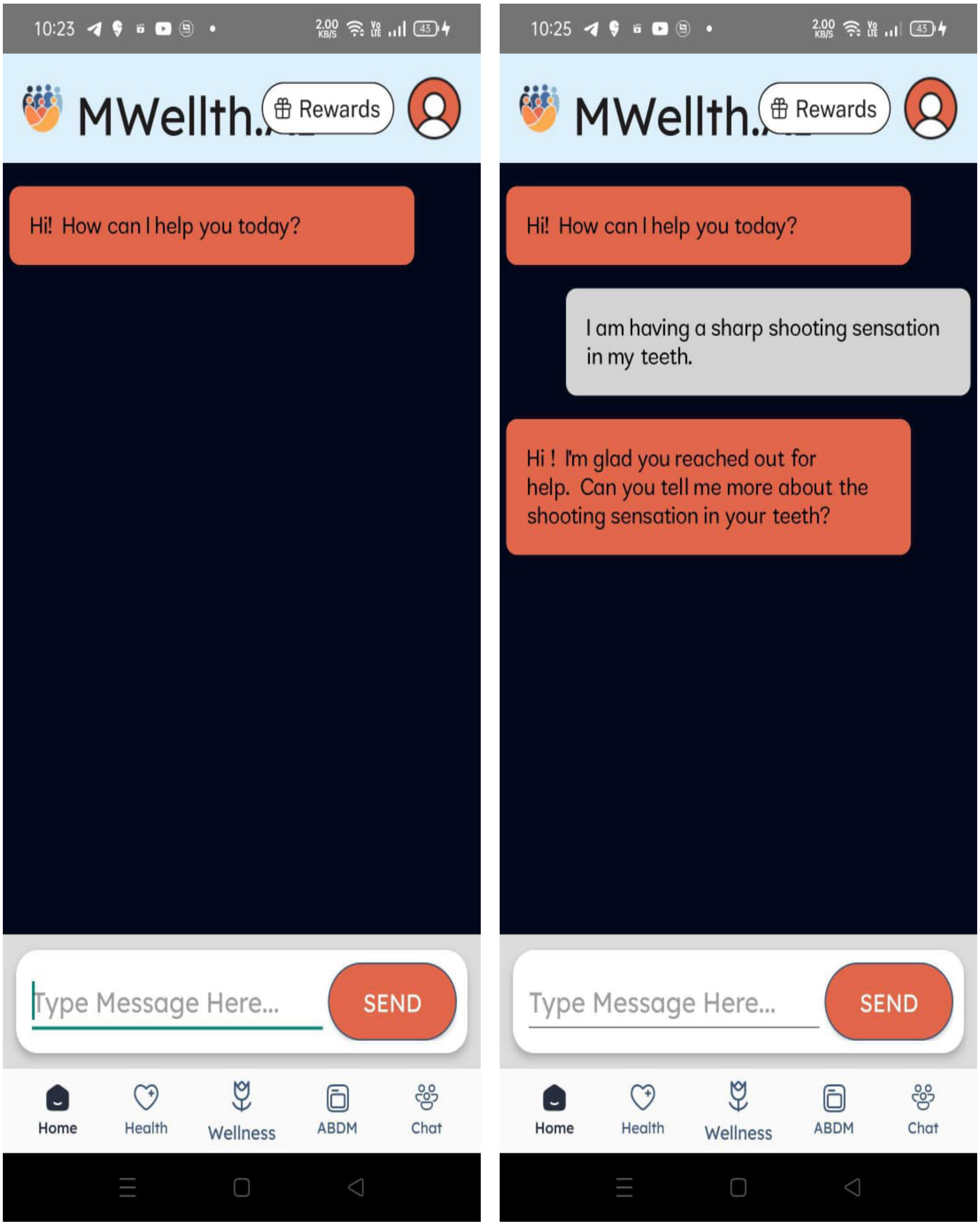



Type a message...
Submit
Retry
Undo
Clear
Advanced options
System prompt
You are a helpful and joyous mental therapy assistant. Always answer as helpfully and cheerfully as possible, while being safe. Your answers should not include any harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.Please ensure that your responses are socially unbiased and positive in nature.
If a question does not make any sense, or is not factually coherent, explain why instead of answering something not correct. If you don't know the answer to a question, please don't share false information.
Max new tokens
1024
Temperature
1
Top-p (nucleus sampling)
0.95
Top-k
50


LLama-2-Mental-Therapy-Chatbot
Chatbot
I am too much stressed in life. I just want to end my life.
I am sorry that you are feeling this way. You need a specialist help. Please consult a nearby doctor.
Type a message...
Submit
Retry
Undo
Clear
Advanced options
System prompt
You are a helpful and joyous mental therapy assistant. Always answer as helpfully and cheerfully as possible, while being safe. Your answers

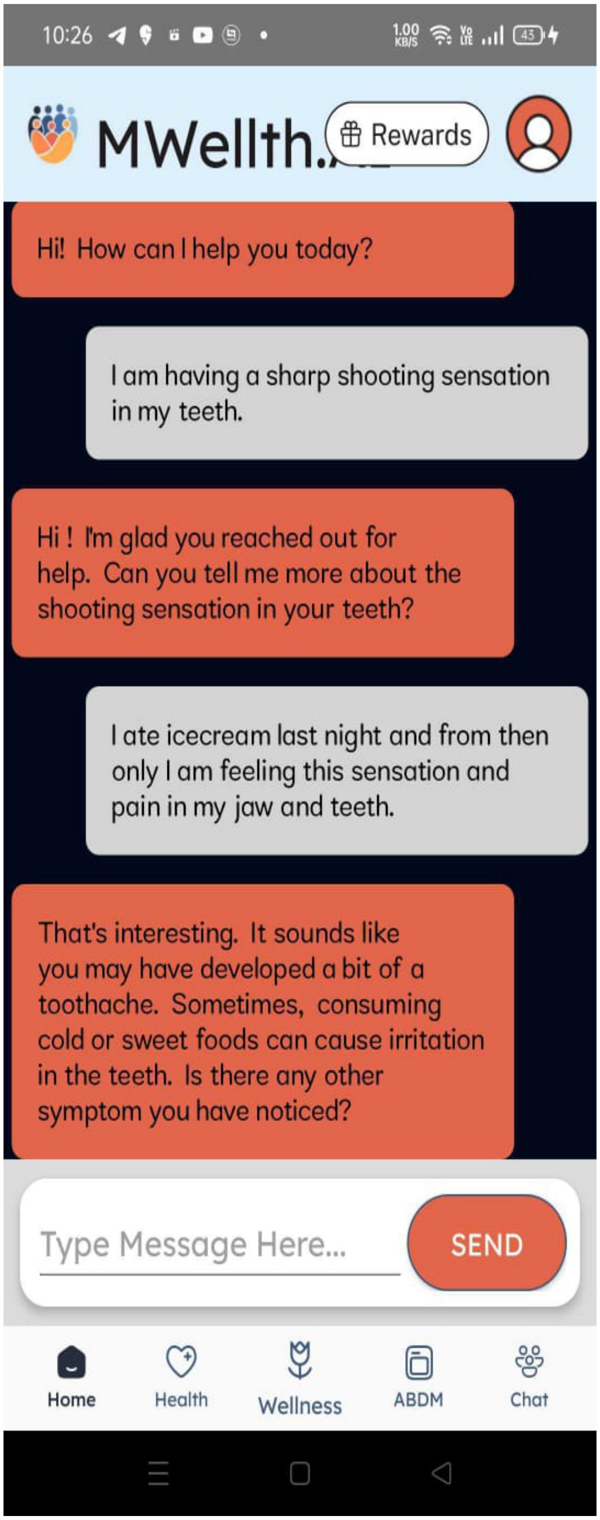
## 4.3 Chatbot Integration: A Glimpse Inside the PHR App

We successfully integrated our advanced mental health chatbot into the Personal Health Records (PHR) application. This integration marks a significant enhancement to the app, providing users with immediate access to mental health support and resources. To showcase the functionality and user interface of the chatbot within the app, we have included a series of screenshots. These images highlight the chatbot's interactive features, its seamless incorporation into the PHR environment, and the intuitive design that makes it easy for users to engage in meaningful conversations about their mental well-being. The addition of the chatbot is a testament to our dedication to improving mental health accessibility through technology.

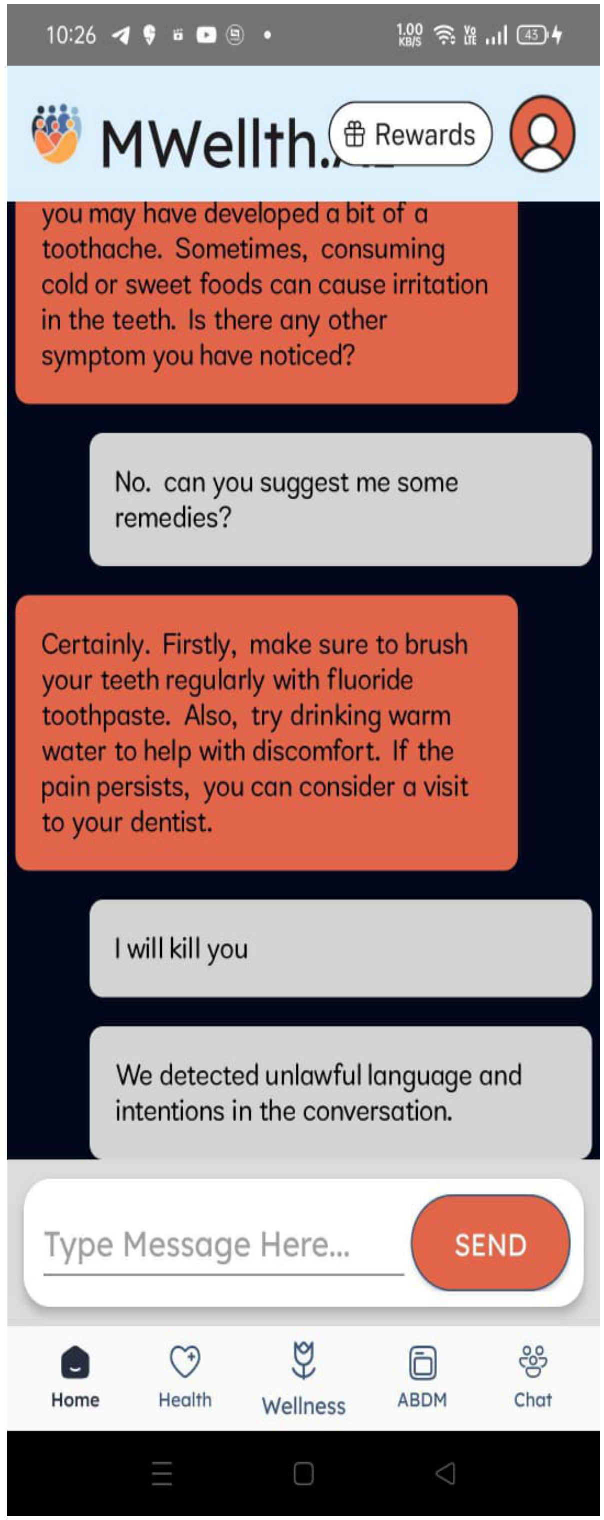



(a) Screenshot 1 (b) Screenshot 2

Figure 4.1: Screenshots of the chatbot integration




(a) Screenshot 3



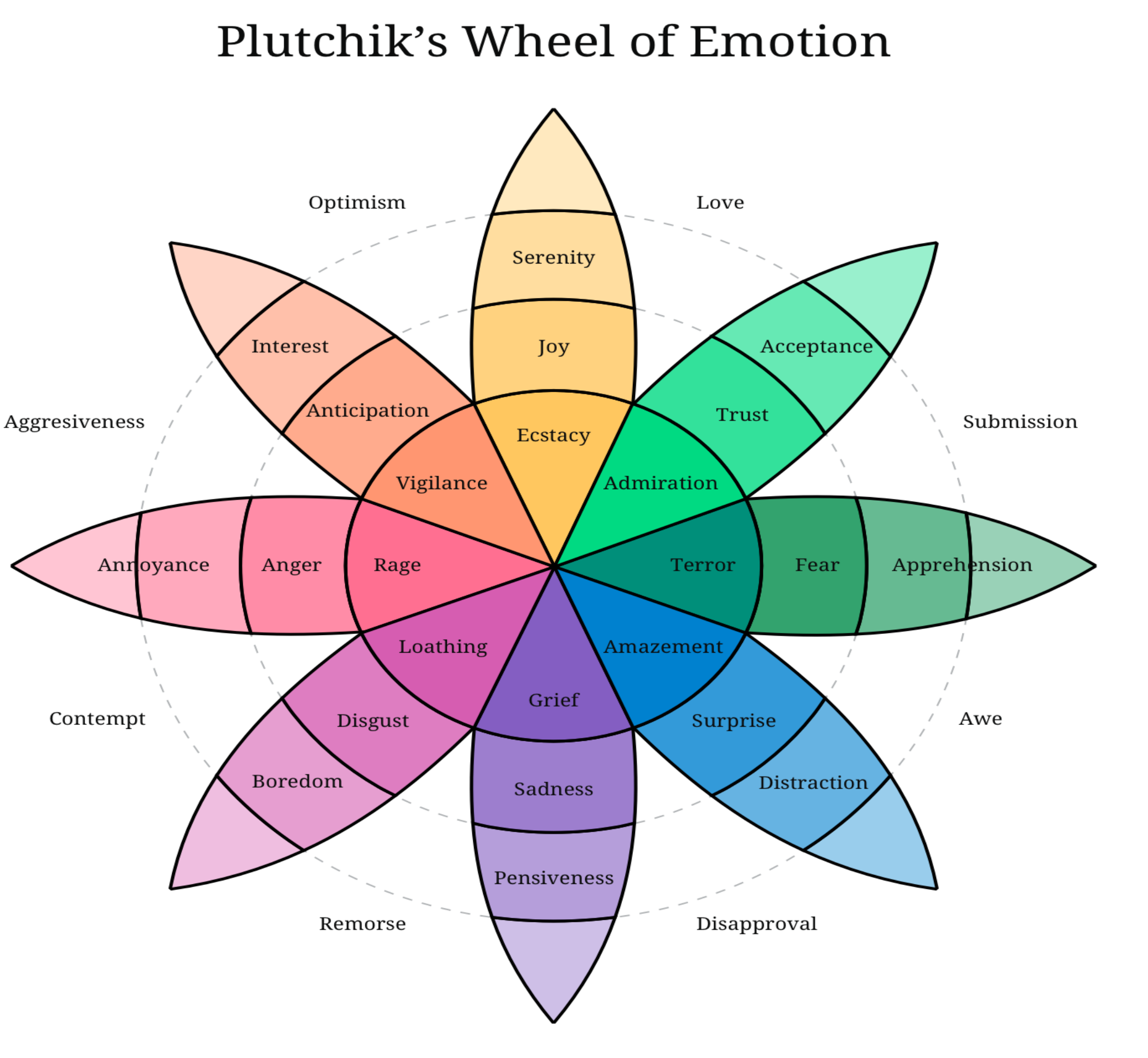

(b) Screenshot 4

Figure 4.2: Screenshots of the chatbot integration

# Chapter 5

# Transformers model based on Plutchik's wheel of emotions

## 5.1 Plutchik's Wheel Of Emotions



Emotions, the complex and nuanced aspects of human experience, have been a subject of fascination and study across various disciplines. Dr. Robert Plutchik, a psychologist and professor at the Albert Einstein College of Medicine, introduced the Wheel of Emotions, a comprehensive framework that delves into the intricate tapestry of human feelings. This conceptual model, developed in 1980, provides a visual representation of eight primary emotions, their opposites, and the various intensities and combinations that give rise to the multitude of emotions we experience.

**Structure of Plutchik's Wheel:**

Plutchik's Wheel of Emotions is organized into a circular diagram, with each primary emotion occupying a specific segment. The eight basic emotions include Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation. These emotions are positioned opposite their polar opposites on the wheel, creating a symmetrical structure that captures the duality inherent in human feelings.

**Intensity and Blending:**

One of the strengths of Plutchik's model lies in its acknowledgment of the dynamic and fluid nature of emotions. The wheel incorporates varying levels of intensity for each emotion, represented by the distance from the center. As one moves outward from the center towards the periphery, emotions intensify, reflecting the spectrum from mild to extreme.

Additionally, Plutchik recognized that complex emotions often result from the blending of primary emotions. For instance, a combination of Joy and Trust may give rise to Love, while a blend of Anticipation and Fear may result in Anxiety. This nuanced approach to emotional complexity makes the wheel a versatile tool for understanding the rich tapestry of human sentiment.

**Applications in Psychology and Counseling:**

Plutchik's Wheel of Emotions has found widespread use in psychology and counseling. Therapists leverage this model to help individuals identify and articulate their emotions, fostering self-awareness and emotional intelligence. By mapping out the interconnectedness of emotions, the wheel aids in recognizing patterns, triggers, and potential pathways for emotional regulation.

**Cross-Cultural Relevance:**

The universality of Plutchik's model contributes to its cross-cultural relevance. While expressions of emotions may be influenced by cultural and societal factors, the fundamental emotions identified in the wheel appear to be consistent across diverse populations. This universality enhances the applicability of the model in understanding human behavior and emotional responses globally.

**Limitations and Criticisms:**

Despite its widespread acceptance, Plutchik's Wheel is not without criticism. Some argue that the model oversimplifies the complexity of human emotions, potentially neglecting cultural nuances and individual differences. Additionally, the delineation of discrete emotions may not fully capture the fluid and dynamic nature of emotional experiences.

In conclusion, Plutchik's Wheel of Emotions stands as a valuable and enduring contribution to the field of psychology. Its visual representation of primary emotions, their opposites, and the potential for blending provides a framework for understanding the intricate landscape of human feelings. As our understanding of emotions continues to evolve, Plutchik's model remains a foundational tool for exploring the depths of the human experience

## 5.2 Transformers based model for emotion classification

Our emotion classification model, designed for predicting emotions based on the Plutchik Wheel, represents a state-of-the-art approach in natural language processing. The underlying transformer architecture is built on NVIDIA's Megatron library[5], a powerful tool specifically engineered for large-scale model training. The models underwent fine-tuning using the *sem_eval_2018_task_1* dataset[2], which comprises approximately 11,000 tweets in English. This dataset is specifically annotated for emotion classification based on the 11 emotions outlined in the Plutchik Emotion Wheel. The fine-tuning process involved training the models on this diverse and representative dataset, enabling them to learn and capture the nuances associated with various emotions expressed in natural language, particularly within the context of social media interactions.

**Model Architecture**

The transformer model leverages the parallelization capabilities of Megatron, enabling efficient training on multiple GPUs. This architecture is particularly well-suited for handling the complexities of emotion recognition, a task that demands nuanced understanding of textual data.

The model incorporates transformer layers to capture intricate patterns and dependencies within the input text. Attention mechanisms, a key feature of transformers, allow the model to weigh different parts of the input sequence, enhancing its ability to discern emotional nuances.

**Fine-Tuning by Tavlab PhD Scholar (Ridam)**

Our transformer model has undergone a meticulous fine-tuning process conducted by Ridam, a distinguished PhD scholar at Tavlab. This fine-tuning involves training the model on a dataset specific to emotion classification, adapting its parameters to better capture the subtleties of emotional expressions.

**Scalability with Megatron**

One of the standout features of our emotion classification model is its scalability, made possible by Megatron. The library facilitates the parallelization of model training across GPUs, making it feasible to handle large-scale datasets and train models with an extensive number of parameters.

**API Implementation**

To make the model accessible and user-friendly, we have implemented a RESTful API. This API, hosted on Flask, allows users to submit text for emotion prediction. The underlying model, loaded with the Megatron-powered weights, processes the input and returns predictions based on the Plutchik Wheel of Emotions.

The API is structured to handle user requests efficiently, utilizing a temporary CSV file to input data and employing subprocesses to execute the prediction process. The results are then extracted and presented in a clear and understandable format.

# Chapter 6

# Results and Future Work

## 6.1 Model Performance

**Results for Fine-Tuned RoBERTa Base Model in Suicide Prediction**

The fine-tuned RoBERTa base model for suicide prediction demonstrates commendable performance on the evaluation/validation set, showcasing its efficacy in identifying potential suicide-related content within the dataset sourced from Reddit. The key performance metrics are as follows:

- **Loss: 0.1543**

  The loss metric, measuring the model's predictive error during training, is minimized to 0.1543. This indicates the effectiveness of the model in learning and generalizing patterns from the suicide prediction dataset.

- **Accuracy: 96.53 percent**

  The accuracy of the model, at 96.53 percent, signifies the proportion of correctly classified instances out of the total evaluation set. This high accuracy underscores the robustness of the model in making accurate predictions.

- **Recall: 96.66 percent**

  The recall, or true positive rate, is an essential metric for the model's ability to identify instances of positive class correctly. A recall of 96.66 percent implies a high sensitivity to identifying suicide-related content in the dataset.

- **Precision: 96.38 percent**

  Precision measures the accuracy of positive predictions made by the model. With a precision of 96.38 percent, the model demonstrates a high level of correctness in its positive predictions, minimizing false positives.

- **F1 Score: 96.52 percent**

  The F1 score, which balances precision and recall, is a harmonic mean of these two metrics. The F1 score of 96.52 percent indicates a robust trade-off between precision and recall, providing a comprehensive evaluation of the model's performance.

These results collectively highlight the effectiveness of the fine-tuned RoBERTa base model in suicide prediction, showcasing its ability to accurately classify and identify instances of content associated with suicidal tendencies. The high precision, recall, and overall accuracy reflect the model's potential to contribute meaningfully to suicide risk assessment tasks within online platforms.

**Suicidal Model Observations :**

- **Model Behaviour**

  The model exhibits a tendency to avoid classifying messages as suicide unless the suicidal tendency is very pronounced and the input message is lengthy. Interestingly, this behavior does not lead to overfitting. The model performs well on both the test and validation datasets, which are independent of the training data.

- **Classification Criteria :**

  The model assigns a suicide label primarily when it encounters a relatively long rant about life. However, for shorter messages (e.g., one or two lines) that express negativity related to suicide, the model tends to avoid labeling them as suicidal.

- **Dataset Influence :**

  This tendency to focus on longer sentences may be influenced by the dataset itself. The dataset likely contains extensive suicidal rants, which the model has learned to associate with suicide. Consequently, shorter inputs may not trigger the same classification.

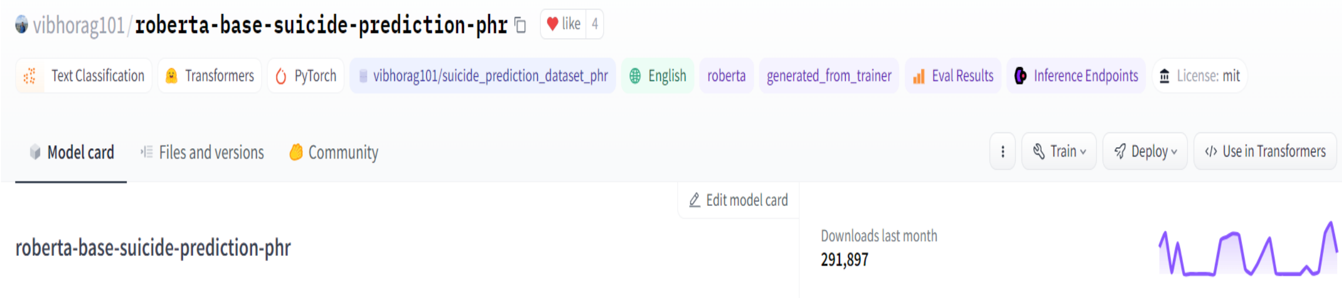
- **Balancing False Positives :**

  By minimizing false positives in short sentences, the model maintains a balance. It ensures that innocuous phrases like "Hello" or "Can we talk?" are not mistakenly labeled as suicidal. The trade-off is that the model works optimally for longer inputs.

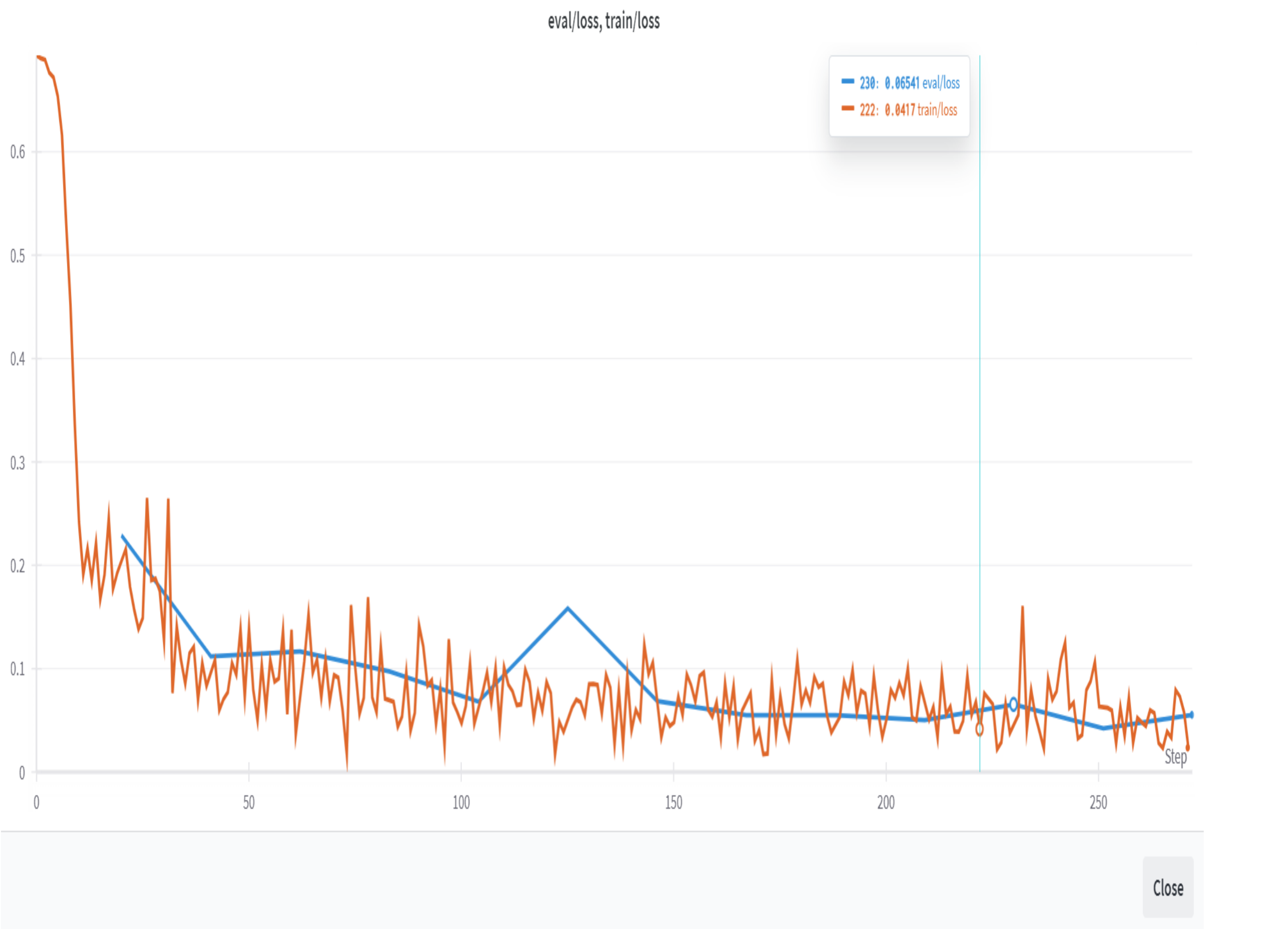
- **Performance Metrics :**

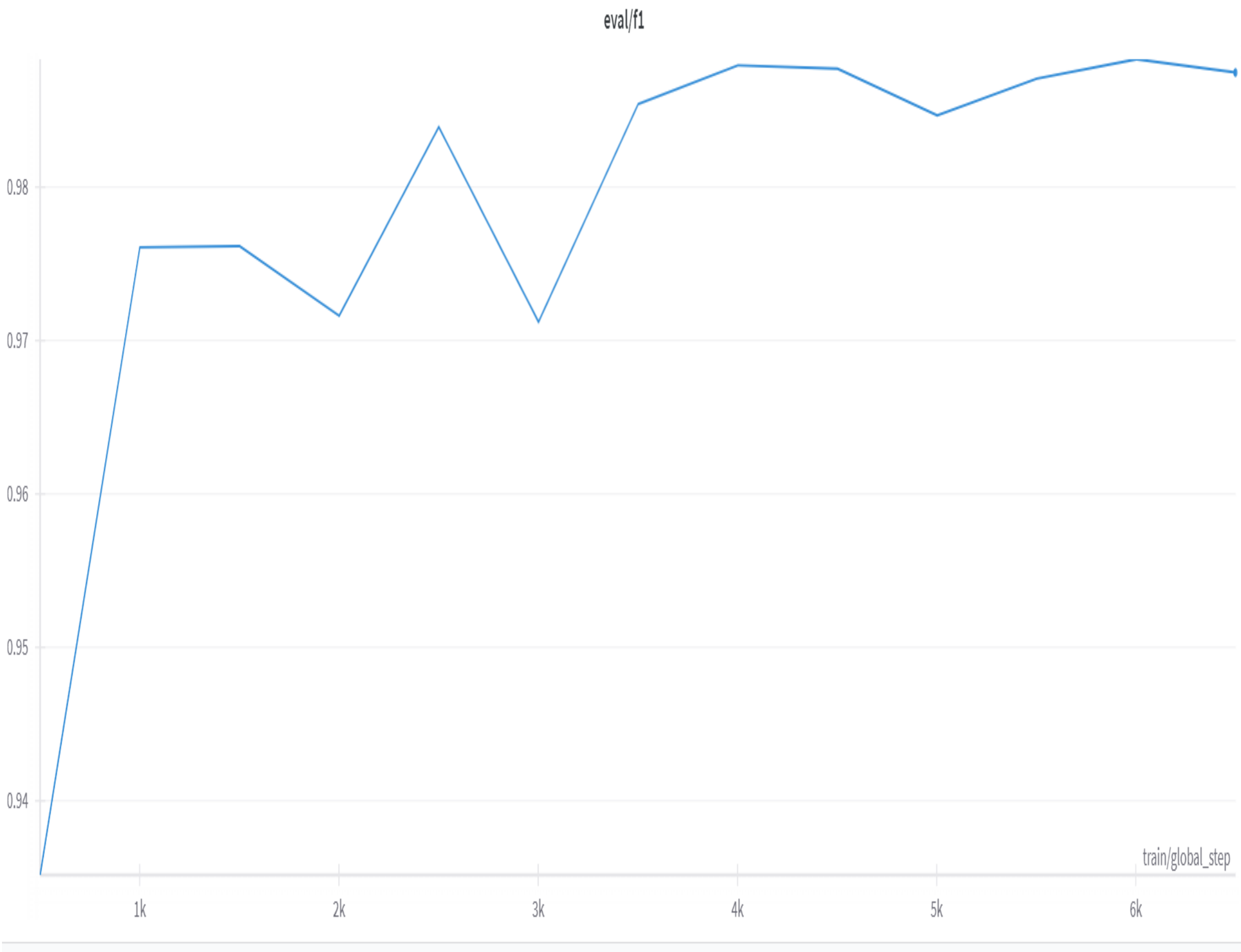
  On a light cleaned dataset, version 1 achieved an impressive 97.1 percent F1 score, while version 2 improved further to 98.75 percent F1 score.

**Impressive Milestone: 291,897 Downloads for Our Emotion Classification Model on Hugging Face :**

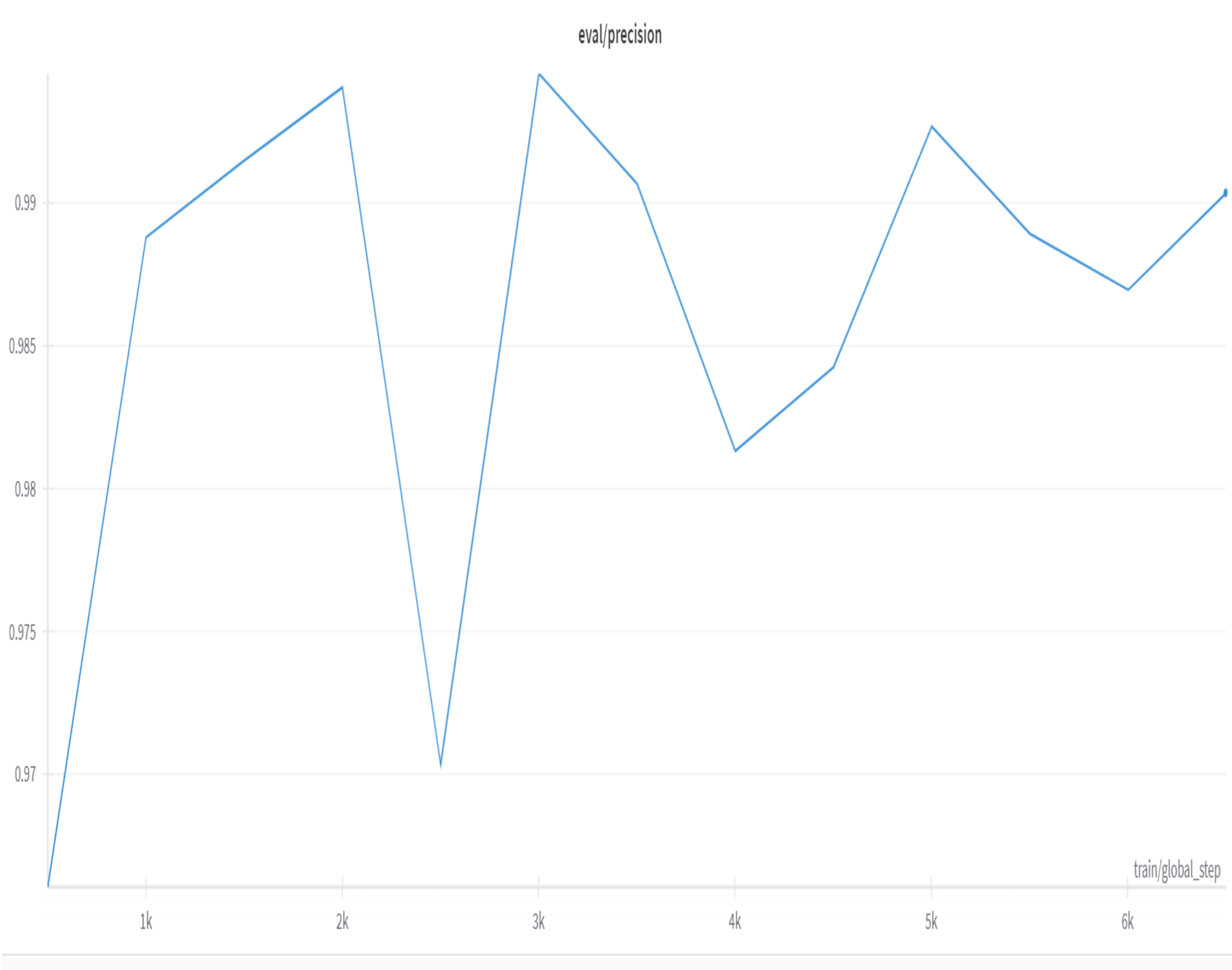


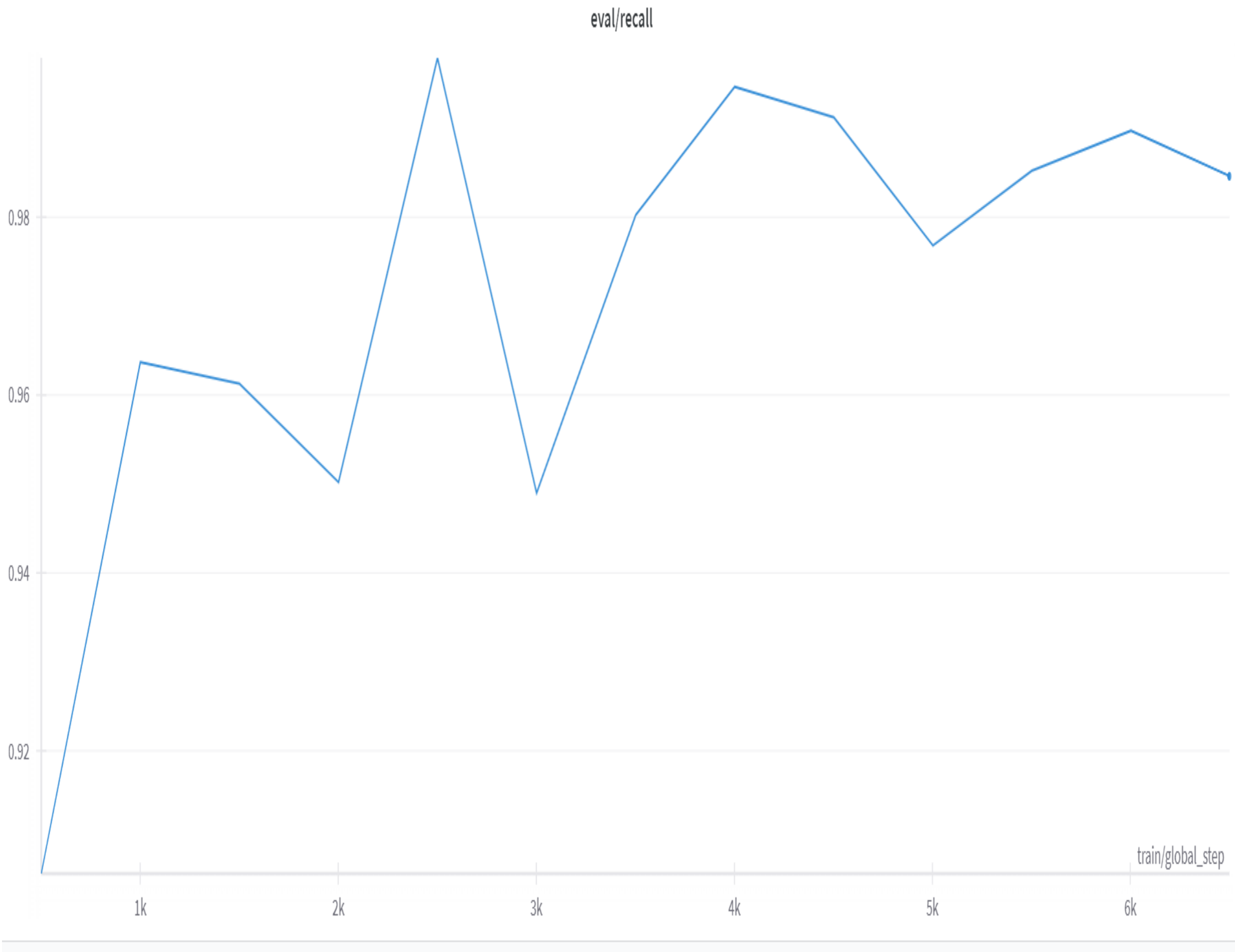

**Training Time graphs for Roberta Based suicide prediction model :**



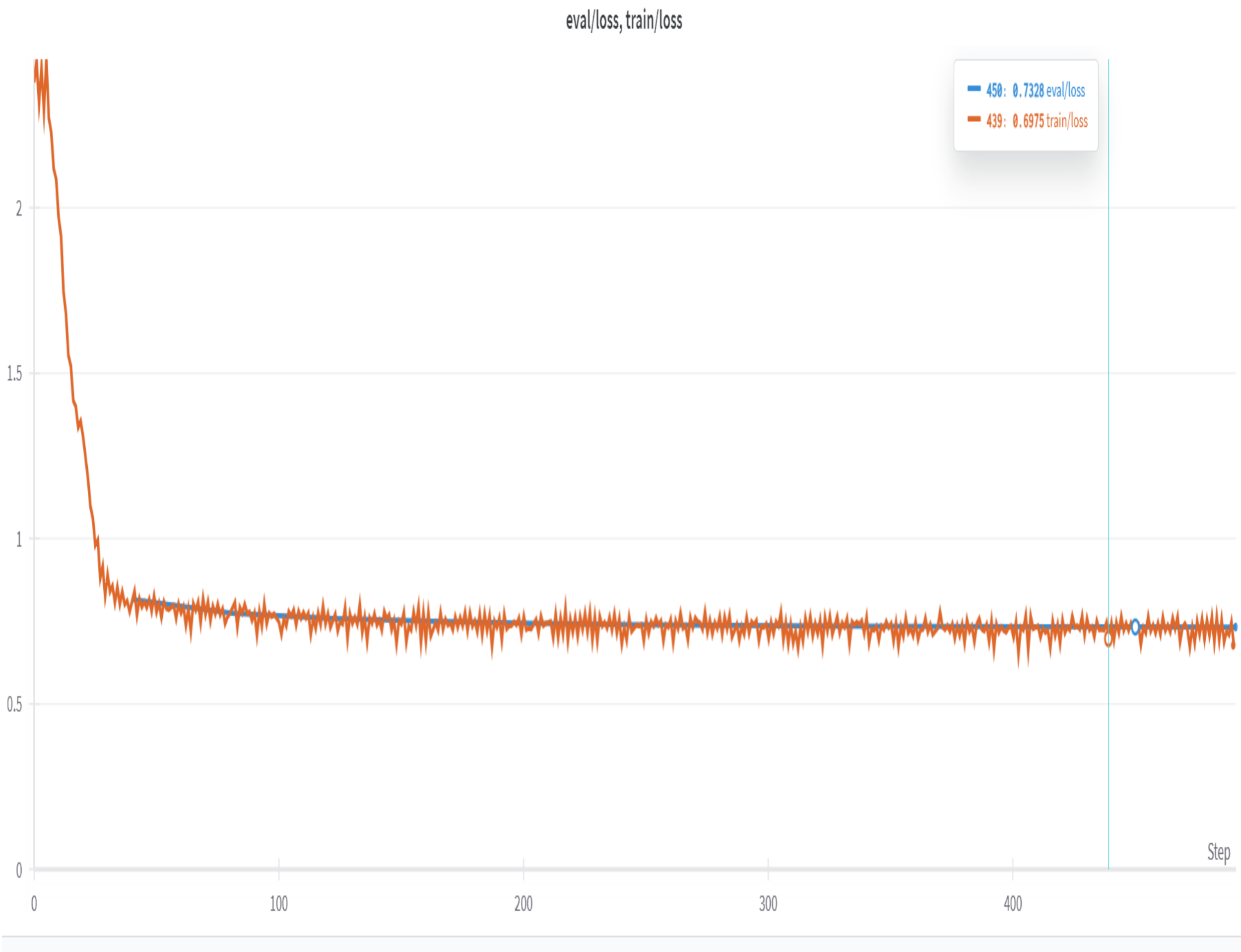

eval/f1
0.98
0.97
0.96
0.95
0.94
train/global_step
1k
2k
3k
4k
5k
6k

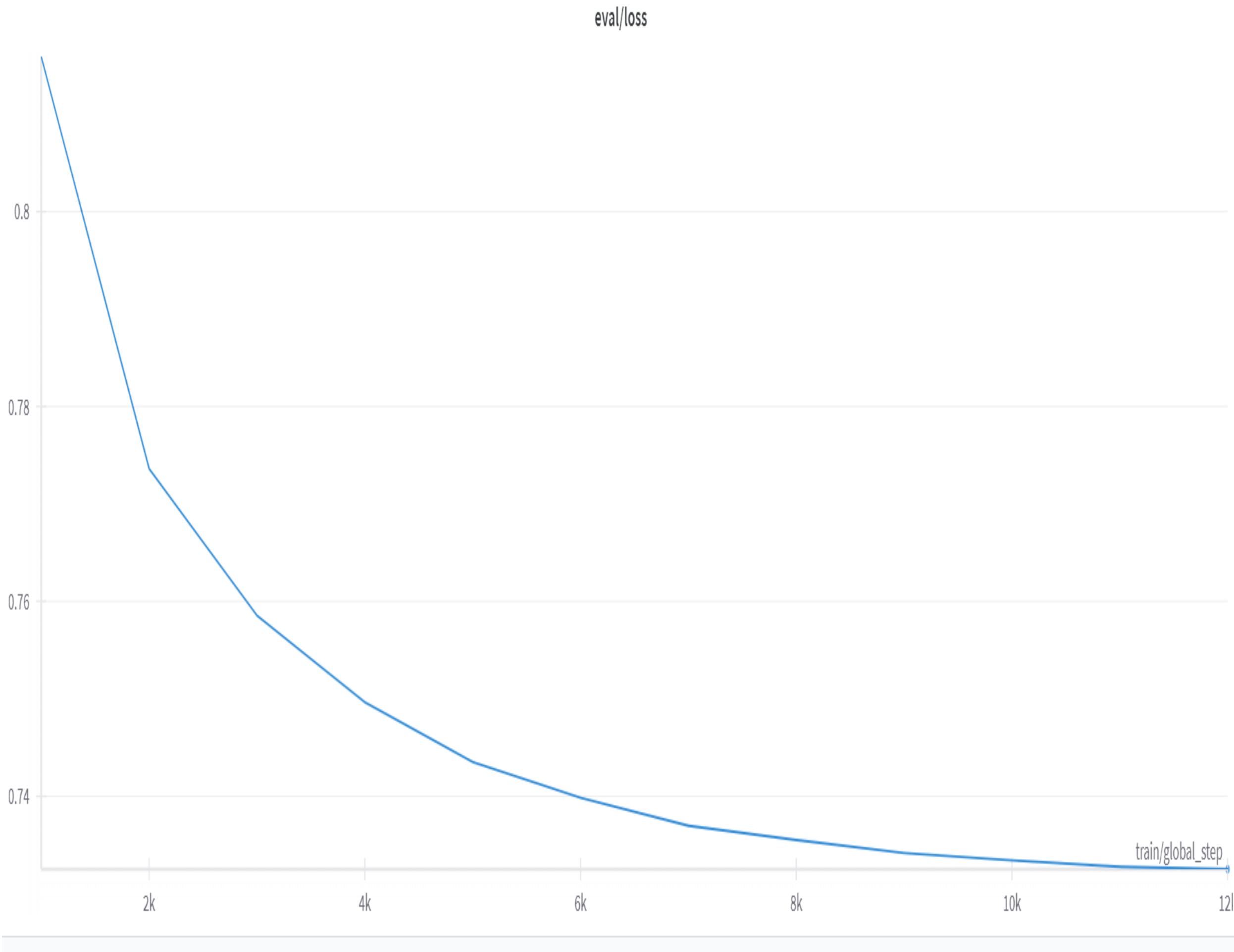

eval/accuracy
0.98
0.97
0.96
0.95
0.94
train/global_step
1k
2k
3k
4k
5k
6k


eval/precision
0.99
0.985
0.98
0.975
0.97
train/global_step
1k
2k
3k
4k
5k
6k


eval/recall
0.98
0.96
0.94
0.92
train/global_step
1k
2k
3k
4k
5k
6k

**Training Time graphs for Llama-2-7b-chat-hf-phr-mental-therapy :**






## 6.2 Model Availability

The developed models, including the llama-2 based therapy chat generation model and the Roberta based suicide prediction and emotion classification model, have been made publicly available on the Hugging Face Hub under the username `vibhorag101`. This repository provides convenient access to the models through the Hugging Face Transformers and Hugging Face Trainer API. The availability of these models on the Hugging Face Hub facilitates seamless integration into other researchers' work, fostering collaboration and making it significantly more straightforward to incorporate these models into various research projects.

Researchers can easily access and utilize the models by referencing the Hugging Face Hub repository at `https://huggingface.co/vibhorag101`. This accessibility enhances the reproducibility of experiments and encourages the broader scientific community to benefit from the advancements achieved through the development of these models.

## 6.3 IRB Proposal

In our pursuit to ensure the ethical integrity of our project, we prepared and submitted an Institutional Review Board (IRB) proposal. This document is a testament to our rigorous adherence to ethical research standards, particularly in studies involving human subjects. The proposal comprehensively details our approach to participant recruitment, informed consent, confidentiality, and data security, underscoring our unwavering commitment to the rights and welfare of the participants.

The IRB proposal serves as a blueprint for maintaining transparency and accountability throughout our research process. It delineates the protocols for safeguarding sensitive information and outlines the measures taken to minimize any potential risks to the subjects. By engaging in this thorough review process, we aim to uphold the trust placed in us by the participants and the academic community, ensuring that our project not only advances scientific knowledge but also champions the principles of ethical research.

## 6.4 Future Work

While the development of the Mental Therapy Chatbot and its corresponding API marks a significant milestone, there are several avenues for future work and enhancements that can be explored to enrich the overall project:

- **User Experience Optimization:**

  Continual refinement of the chatbot's user experience is crucial. This includes incorporating user feedback, enhancing the natural language understanding capabilities, and implementing features that make the interaction more intuitive and personalized.

- **Customization and Personalization:**

  Implementing user-profiles and preferences can enable a higher degree of personalization. The chatbot can adapt its responses based on individual user histories, preferences, and specific mental health needs, creating a more personalized and effective therapeutic experience.

- **Collaboration with Mental Health Professionals:**

  Establishing collaborations with mental health professionals for periodic reviews and inputs can enhance the therapeutic effectiveness of the chatbot. This collaboration can ensure that the information provided aligns with current mental health guidelines and practices.

- **User Engagement Analytics:**

  Implementing analytics tools to track user engagement, satisfaction, and the effectiveness of the chatbot's responses can provide valuable insights. Analyzing this data can guide further improvements and enhancements in the chatbot's functionality.

- **Enhancing Model Accuracy through Direct Preference Optimization:**

  Direct Preference Optimization (DPO) technique can be employed to further refine the accuracy of our mental health chatbot models. DPO is a cutting-edge method that fine-tunes language models to better align with human preferences, which is particularly beneficial for applications like ours that require sensitivity and precision. The process of DPO involves a new parameterization of the reward model, which allows for the extraction of an optimal policy directly. This means we can enhance our chatbot's responses using a simple classification loss, without the need for complex reinforcement learning algorithms.

# Bibliography


[1] Gayathri Ananthakrishnan et al. “Suicidal Intention Detection in Tweets Using BERT-Based Transformers”. In: *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. 2022, pp. 322–327. DOI: `10.1109/ICCCIS56430.2022.10037677`.

[2] The HF Datasets community. $Sem_{e}val_{2}018_{t}ask_{1}datasetsathuggingface$. URL: `https://huggingface.co/datasets/sem_eval_2018_task_1`.

[3] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: `2305.14314 [cs.LG]`.

[4] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: `1810.04805 [cs.CL]`.

[5] Neel Kant et al. *Practical Text Classification With Large Pre-Trained Language Models*. 2018. arXiv: `1812.01207 [cs.CL]`.

[6] Nikhileswar Komati. *Suicide and depression detection*. May 2021. URL: `https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch`.

[7] Maxime Labonne. *Fine-tune your own llama 2 model in a colab notebook*. Sept. 2023. URL: `https://towardsdatascience.com/fine-tune-your-own-llama-2-model-in-a-colab-notebook-df9823a04a32`.

[8] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: `1907.11692 [cs.CL]`.

[9] Anders Giovanni Møller et al. *Is a prompt and a few samples all you need? Using GPT-4 for data augmentation in low-resource classification tasks*. 2023. arXiv: `2304.13861 [cs.CL]`.

[10] Gradio Team. *Quickstart*. URL: `https://www.gradio.app/guides/quickstart`.

[11] Hugo Touvron et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. 2023. arXiv: `2307.09288 [cs.CL]`.

[12] Jerry Yao. *Jerryjalapeno/NART-100k-synthetic · datasets at hugging face*. URL: `https://huggingface.co/datasets/jerryjalapeno/nart-100k-synthetic`.