

# Mental Health Classification and Chatbot

Vraj Patel, Nikhil Chakka

March 30, 2026

## Abstract

Mental health care is an area where conversational agents have the potential to provide accessible support. This project presents the development of an NLP driven mental health chatbot that combines text classification and response generation using transformer based language model. In our approach, user input messages are first classified into mental health related categories (e.g., expressions of depression, anxiety, etc.) using machine learning and transformer models. We compared traditional algorithms (Logistic Regression, Random Forest) with a transformer classifier (BERT), and found that the BERT model achieved the highest performance (macro-averaged F1  $\approx 0.79$ – $0.80$ ) on our multi-class dataset. The predicted category label was then used to prompt a large language model LLM, LLaMA 3.2 with 1 billion parameters, to generate an empathetic, supportive response tailored to the user’s inferred mental state. Our experiments demonstrate that label informed prompting sometimes helped the chatbot produce relevant and compassionate replies. This work highlights a promising direction for AI-assisted mental health support by integrating robust text classification with controlled response generation. We also discuss limitations and ethical considerations, emphasizing the need for fairness, privacy, and safety in deploying such systems.

## 1 Introduction

### 1.1 Motivation

Mental health disorders are prevalent and often undertreated. Globally, about **1 in 8 people** live with a mental disorder(1), yet many do not receive adequate care due to barriers like stigma and limited resources(1). In college populations up to 75% of students with serious symptoms do not access professional help(2). These statistics exemplify the need for accessible support systems. Recent advances in artificial intelligence have spurred interest in chatbots as a scalable means of providing mental health assistance. Historically, chatbots like *Woebot* have delivered cognitive-behavioral therapy techniques via text conversation and shown efficacy in reducing anxiety and depression symptoms (2). More recently, large language model (LLM) based chatbots, such as OpenAI’s ChatGPT, have emerged as potential mental health aides, offering flexible natural dialogues (3). However, purely generative AI systems also raise concerns regarding the accuracy, safety, and appropriateness of their responses, especially in sensitive domains like mental health (3).

### 1.2 Gap in Current Solutions

Despite recent advancements in AI, general chatbots are not fully equipped to handle nuanced psychological content. Models like Siri or standard open-domain chatbots may misunderstand subtle expressions of distress or respond with generic advice. Even powerful large language models (LLMs) such as ChatGPT require careful domain adaptation to ensure safe and appropriate interactions in a mental health context(4). On the other hand, existing specialized mental health bots (referencing *Woebot*) typically rely on scripted dialogues for cognitive behavioral therapy(2). However, these either lack the flexibility to handle diverse inputs or lack the psychological grounding needed for sensitive situations. A key challenge in designing an effective mental health chatbot is ensuring that the system understands *both* the user’s needs and responds with appropriate empathy.

## 1.3 Proposed Approach

This project proposes an **NLP driven mental health chatbot** that integrates mental health text classification with empathetic response generation. Specifically, we first analyze the user’s message using a classifier trained to detect mental health-related categories or emotional states. We do this by integrating a multi-class classification model to recognize distress signals from the user messages and categorize it into one of several mental health labels: *Normal*, *Depression*, *Suicidal Ideation*, *Anxiety*, *Stress*, *Bipolar Disorder*, or *Personality Disorder*. The classified label will then be used to inform a response generation module. For the conversational engine, an open-source LLM will be leveraged (Llama 3.2 with 1 billion parameters) fine-tuned for dialogue. By connecting the classifier to the chatbot, the system can provide context-sensitive responses. Thus, recognizing signs of anomalies in a user’s text and responding with appropriate empathy and resources. The goal of this hybrid approach is to improve the relevance of chatbot responses by grounding them in an understanding of the user’s mental state.

The contributions of this work are as follows: First, we develop a pipeline that integrates a high-performing mental health text classifier with a modern LLM based chatbot, bridging the gap between *analysis* and *response* in conversational AI for mental health. Second, we conduct experiments comparing traditional machine learning methods and transformer models for the classification task, providing insight into their relative performance on domain-specific data. Third, we demonstrate through example dialogues how the combined system can respond to users in an empathetic and contextually appropriate manner, highlighting the benefits of label-informed response generation. In the following sections, we review related work (Section 2), describe our methods including the classification and generation components (Section 3), present experimental results (Section 4), and discuss conclusions along with ethical considerations (Section 5).

## 2 Related Work

### 2.1 Mental Health Chatbots

A number of chatbot-based interventions have been developed in the mental health domain over the past decades. Early systems like ELIZA demonstrated that even simple rule-based scripts could elicit users to reveal personal feelings (5). In recent years, more sophisticated therapeutic chatbots have emerged. For example, Woebot is a chatbot that delivers Cognitive Behavioral Therapy exercises through conversation, and a clinical trial showed it can reduce depression in college students (2). Several other mental health apps (e.g., Wysa, Tess) use conversational agents to provide psychoeducation, mood tracking, and coping strategies. Most of these systems historically relied on pre-written dialogue flows or retrieval-based responses guided by human experts, ensuring a controlled experience but limiting flexibility. The recent advancements of large scale pretrained language models has opened new possibilities for chatbots to engage in open ended supportive dialogue. LLM based agents, like GPT-4, can generate human-like responses and have been experimented with as mental health aides. Users report that such generative AI chatbots can feel empathetic and helpful (3), though formal evaluation is still nascent. A recent user study noted both positive outcomes, users finding an “emotional sanctuary” in AI chats and potential issues, calling for stronger safety guardrails and the ability to handle crisis situations (3). Compared to rule-based designs, LLM driven chatbots offer greater expressiveness but pose challenges in ensuring *appropriate* and *reliable* behavior at all times. Our work is situated at this intersection: we leverage the power of generative transformers for dialogue, but we incorporate a modular understanding component, sentimental text classification, to retain a degree of control and interpretability.

### 2.2 Mental Health Text Classification

In parallel with chatbot development, there has been substantial research on automatically detecting mental health signals from text. Researchers have applied machine learning to user generated content, such as social media posts or online forum messages, to identify conditions like depression, anxiety, or PTSD from language patterns. Traditional approaches used features like TF-IDF or linguistic cues fed into classifiers including Logistic Regression or Random Forest, often obtaining moderate accuracy. For

instance, Random Forest classifiers have often been used as baselines in studies to detect depression from Reddit posts, leveraging the algorithm introduced by Breiman (6). In recent years, the field has shifted toward deep learning and transformer models. Pretrained transformers such as BERT have demonstrated superior performance in mental health text classification tasks, capturing subtle linguistic markers of mental states that simpler models might miss. A comprehensive review (7) found that deep learning methods consistently outperform traditional ML across hundreds of studies on mental illness detection. Our experiments align with this trend: we observe that a fine-tuned BERT classifier substantially improves macro-F1 scores compared to logistic regression and random forest baselines on a mental health dataset. Moreover, transformer based classifiers can generalize better and handle nuanced language, which is crucial for understanding a user’s true sentiment or mental state from their text.

## 2.3 Empathetic Response Generation

Beyond understanding user inputs, an effective counseling or mental health chatbot must respond with empathy and appropriateness. There is a growing body of work on modeling empathy in dialogues. Rashkin et al. introduced the EmpatheticDialogues dataset, which contains open-domain conversations where one speaker describes a personal situation and the other responds with empathy (8). Such datasets have enabled training dialogue models that exhibit more empathetic qualities. Other researchers have explored strategies like augmenting sequence-to-sequence models with emotion recognition components or multi-task learning (predicting the user’s emotion and generating a response jointly) to achieve more emotionally attuned replies. For example, some approaches first classify the user’s utterance along affective dimensions (such as happy, sad, angry) and then condition the response on that classification, ensuring the reply’s tone matches the user’s emotional state. Our approach is inspired by these advances but tailored to the mental health context: we specifically classify for mental health-related categories or distress levels, and use that information to steer a large language model’s generation. This can be seen as a form of controlled text generation, similar in spirit to prior work on conditional language modeling, but applied here to maintain empathy and relevance in a sensitive domain.

## 2.4 Improvement

Compared to the prior work, our project is distinctive in its combination of a *predictive* model and a *generative* model within a single chatbot system. Whereas many studies either focus on improving classification of mental health texts (7) or on creating empathetic dialogue systems (8), we bring these two threads together. By doing so, we aim to capitalize on the strengths of each: the robustness and interpretability of a classifier to gauge the user’s state, and the flexibility of an LLM to produce human-like, supportive responses.

# 3 Methods

## 3.1 Dataset

This study utilized two distinct datasets to develop, evaluate, and train models for mental health natural language processing tasks and response generation.

### 3.1.1 Sentiment Analysis for Mental Health Dataset

The first dataset, titled *Sentiment Analysis for Mental Health*, comprises 53,043 text samples, each annotated with one of several mental health categories: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, or Personality Disorder (9). These annotations were provided by experts based on user-generated content sourced from forums, chat logs, and surveys. To prepare the data for analysis, we implemented a preprocessing pipeline that included converting text to lowercase, removing email addresses, URLs, user mentions, non-alphanumeric characters, and collapsing multiple whitespaces. For traditional machine learning models (Logistical Regression, Random Forest), textual data were transformed into numerical representations using Term Frequency–Inverse Document Frequency (TF–IDF) features. In contrast, for the BERT Transformer model, we utilized subword tokenization to process

the text inputs. To ensure reproducibility and effective model training, we set a fixed random seed and partitioned the dataset accordingly. Traditional models were trained using an 80/20 split for training and testing, respectively. Hyperparameter tuning was conducted using GridSearchCV, negating the need for a separate validation set. For the BERT Transformer model, we adopted a 70/10/20 split for training, validation, and testing, respectively. Given the class imbalance inherent in the dataset, particularly with some categories being underrepresented, we employed stratified sampling to maintain proportional representation across splits.

### 3.1.2 Mental Health Counseling Conversations Dataset

The second dataset, known as *Mental Health Counseling Conversations*, consists of 2,480 samples derived from online counseling and therapy platforms (10). Each entry includes a 'Context', detailing the individual's issues, and a corresponding 'Response', providing advice from a licensed psychologist. This dataset was specifically utilized to fine-tune our LLaMa chatbot model, enabling it to generate contextually appropriate and empathetic responses in mental health counseling scenarios.

<sup>1</sup> Both datasets are publicly accessible as of May 4th, 2025 and have been used at the intersection of mental health and natural language processing.

## 3.2 Models

To address the multi-class classification task of predicting mental health categories from textual data, we implemented and evaluated three distinct modeling approaches, encompassing both traditional machine learning algorithms and advanced deep learning architectures. Specifically, we explored:

- **Logistic Regression (LR):** A linear classification model widely utilized in text categorization tasks. We employed L2 regularization to mitigate overfitting and enhance generalization. The input text was transformed into numerical feature vectors using Term Frequency–Inverse Document Frequency (TF–IDF) representations, with the vocabulary limited to the top 10,000 terms by frequency to reduce dimensionality and noise. Hyperparameter tuning for the regularization strength parameter  $C$  was conducted via grid search over the values  $\{0.01, 0.1, 1, 10\}$ , utilizing  $k$ -fold cross-validation on the training set to identify the optimal configuration.
- **Random Forest (RF):** An ensemble learning method comprising multiple decision trees to capture complex, non-linear relationships within the data. Similar to LR, the input text was vectorized using TF–IDF features. The Random Forest classifier was configured to use Gini impurity as the criterion for node splitting. We performed hyperparameter optimization over the number of trees ( $n_{\text{estimators}} \in \{100, 200\}$ ) and maximum tree depth ( $\text{max\_depth} \in \{\text{None}, 10, 20\}$ ), employing grid search with  $k$ -fold cross-validation to determine the most effective model parameters.
- **Transformer-Based Model (BERT):** A fine-tuned version of the Bidirectional Encoder Representations from Transformers (BERT) model, which leverages deep contextual embeddings to understand the semantic nuances of text. Unlike traditional models, BERT processes raw text inputs directly through subword tokenization, capturing bidirectional context. The model was fine-tuned on our dataset with a learning rate of  $2 \times 10^{-5}$  and trained for 3 epochs. Early stopping was implemented to prevent overfitting, monitoring validation loss to determine the optimal number of training iterations.

For consistency and comparability, each model was trained and evaluated on the same dataset splits. Traditional models (LR and RF) utilized an 80% training and 20% testing split, with hyperparameter tuning performed on the training data via  $k$ -fold cross-validation through **GridSearchCV**. For the BERT-based classifier, we use the pre-trained **bert-base-uncased** model and fine-tune it on our labeled dataset. The BERT model was trained using a 70% training, 10% validation, and 20% testing split, allowing for effective monitoring of validation performance during fine-tuning. Stratified sampling was employed across all splits to preserve the original class distribution, ensuring robust model evaluation. This comprehensive approach enabled a thorough comparison of model performances across varying complexities and methodologies.

---

<sup>1</sup>Disclaimer about datasets

### 3.3 Integration with LLaMA Chatbot

To enhance the system’s capacity for empathetic and contextually appropriate interactions, we integrated the optimal classification model—identified through validation metrics—with a conversational AI module grounded in an open-source large language model (LLM). Specifically, we employed the LLaMA 3.2 model with 1 billion parameters, recognized for its efficiency and effectiveness in multilingual dialogue applications (? ).

The integration operates as follows: upon receiving a user message, the classifier analyzes the input and generates a probability distribution across predefined mental health categories. The most probable label is then incorporated into the LLM’s prompt, thereby conditioning the generative response on the user’s inferred mental health state.

We utilized prompt engineering to guide the LLM’s outputs without necessitating additional fine-tuning. The prompt template was designed to emulate the communication style of a licensed cognitive behavioral therapist, emphasizing empathy, confidentiality, and professional boundaries. An example of the prompt structure is as follows:

**System:** You are a licensed cognitive behavioral therapist. Maintain empathy, confidentiality, and professional boundaries at all times. Use the user’s text to:

1. Validate their feelings.
2. Ask one clarifying question if needed.
3. Offer 3–5 practical strategies grounded in CBT.
4. Encourage seeking professional support if symptoms persist.
5. Note: The individual is diagnosed with **[Label]**.

**User:** I’m going through some things with my feelings and myself. I barely sleep and I do nothing but think about how I’m worthless and how I shouldn’t be here. I’ve never tried or contemplated suicide. I’ve always wanted to fix my issues, but I never get around to it. How can I change my feeling of being worthless to everyone?

**Response:** If everyone thinks you’re worthless, then maybe you need to find new people to hang out with. Seriously, the social context in which a person lives is a big influence in self-esteem...

In this structure, **[Label]** is dynamically replaced with the classification model’s output (e.g., "Depression", "Anxiety"), and the text after **User:** is the original user input. This method represents a form of zero-shot conditioning, leveraging the LLM’s pre-trained knowledge to produce responses tailored to the user’s mental health context. To further refine the chatbot’s responsiveness, we introduced label-specific prompt variations. For instance, if the predicted label is *Depression*, the prompt includes: "Note: The individual is diagnosed with depression." Conversely, for *Anxiety*, the prompt states: "Note: The individual is diagnosed with anxiety." These modifications are informed by established counseling principles pertinent to each mental health condition. For fine-tuning the chatbot, we employed Unsloth AI (11), a framework designed for efficient LLM fine-tuning. The training parameters were set as follows: a maximum sequence length of 6000 tokens, 30 epochs, a learning rate of  $3 \times 10^{-4}$ , the AdamW optimizer, and a weight decay of 0.01. This configuration facilitated the model’s adaptation to the specific requirements of mental health counseling dialogues. The integrated system was evaluated in a controlled setting to assess the influence of the classification label on the chatbot’s responses. Adjustments to prompt wording and the level of guidance provided to the LLaMA model were iteratively tested to optimize the alignment of the chatbot’s tone and content with recommended therapeutic practices. The ultimate objective is to ensure that the system delivers supportive, context-aware, and ethical interactions, such as encouraging individuals with suicidal ideation to seek professional help and offering reassurance to those experiencing anxiety.

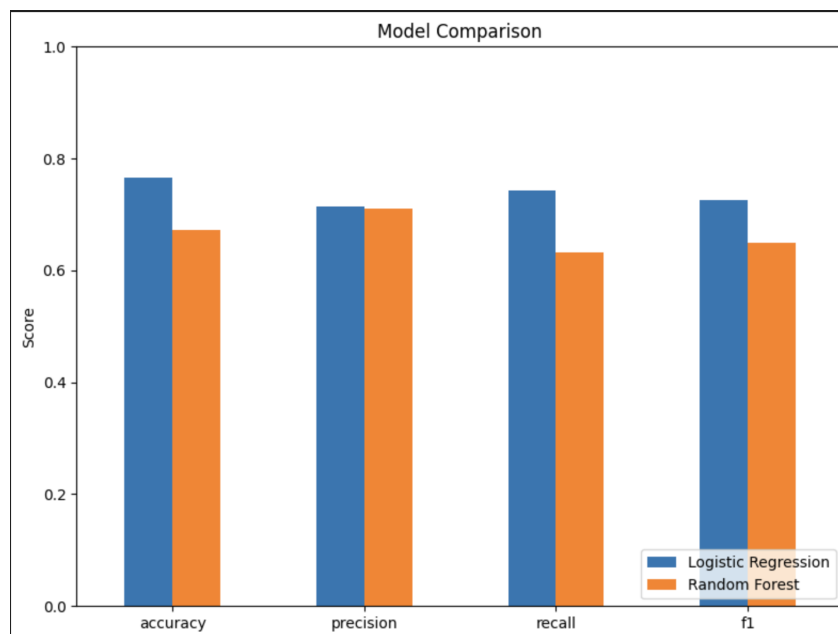
## 4 Experiments

### 4.1 Classification Evaluation

We evaluated the classification component rigorously before deploying it in the chatbot. The labeled dataset was split into training and test sets (e.g., 80% train, 20% test) stratified by class. After training and hyperparameter tuning (using cross-validation on the training set), we assessed the performance on the held-out test set. Key evaluation metrics that we used include:

- **Accuracy** – the overall fraction of texts correctly labeled.
- **Precision & Recall** – calculated for each class (e.g., how well the model identifies instances of *Depression* without false alarms, and how many actual *Depression* cases it catches). These will highlight performance on minority classes.
- **F<sub>1</sub>-score** – the harmonic mean of precision and recall, computed per class and as a macro-average across all classes to summarize balanced performance.

Below is a graph that compares the performance between our traditional models of logistic regression and random forest.

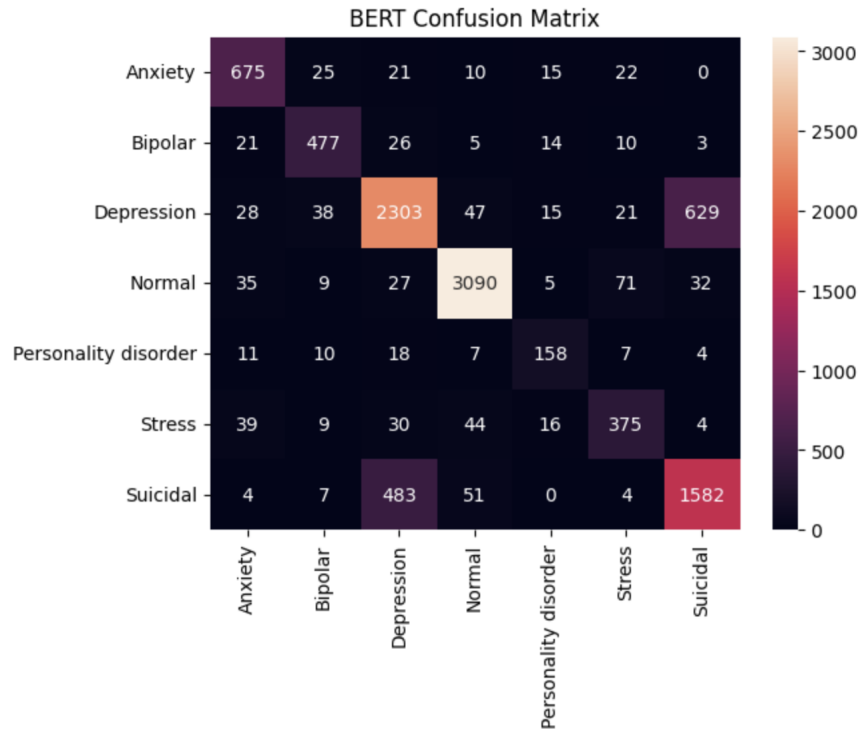


For the logistic regression model, our results were specifically 0.7666 for the accuracy, 0.7139 for the precision, 0.7425 for the recall, 0.7267 for the f1-score. For the random forest model, our results were specifically 0.6730 for the accuracy, 0.7101 for the precision, 0.6318 for the recall, 0.6498 for the f1-score. For our transformer-based model, BERT, we achieved the highest results of the 3. This confirmed our expectations that the transformer-model will be the optimal model for the sentiment classification aspect. We achieved 0.817082 for the accuracy, 0.794693 for the precision, 0.790177 for the recall, and 0.792150 for the f1-score.

Model	Accuracy	Macro-F1
Logistic Regression	77%	0.72
Random Forest	67%	0.65
BERT (fine-tuned)	82%	0.79

Table 1: Comparison of classification performance on the test set for different models. The transformer-based BERT model achieves substantially higher macro-F1 than the traditional machine learning models.

To understand performance per class, we plot the confusion matrix of the best model (BERT) in figure below:



We observed that the true-positive rates for many of the classes are high as shown for Anxiety, Bipolar, Normal, Personality Disorder, and Stress as individuals who are experiencing these issues typically talk in a specific way. However, for the 2 other classes, suicidal and depression, the model performed weaker likely due to the fact that the tone, word choice, etc. might have been very similar for these 2 classes, and not mutually exclusive like how our labeling is done.

## 4.2 Chatbot Evaluation

Evaluating a generative chatbot is inherently more qualitative. We carried out **simulated conversations** using test scenarios: for select mental health categories, we selected example user inputs from the dataset and ran them through the integrated system. We then logged the chatbot’s responses and judged them on criteria such as relevance to the user’s query, empathy and appropriateness of tone, and correctness/helpfulness of the information provided. For instance, given a user message labeled *Suicidal*, did the chatbot respond with appropriate concern and encourage seeking help? We also tracked the chatbot’s ability to stay on topic and not produce disallowed content. The goal of the evaluation was to demonstrate that integrating the classifier improves the chatbot’s responses in terms of relevance and safety. We expected to see that the context-aware chatbot provides more tailored and supportive answers than a version of the chatbot without the classification guidance. However, our experiments showed mixed results.

For negative results, we observed that in specific cases, that even if we provided the diagnosis, the chatbot will create its own separate diagnosis. For example in one case, we said the individual is depressed, however, the chatbot diagnosed the individual with schizophrenia and recommended they never leave the county and take specific medicine. These results show how negative the consequences can be when there is a mistake in the model. First, a misdiagnosis, then a harsh fix (never leaving the county) and even harsher fix (recommending them medicine that they don’t even need to take as the diagnosis is wrong). Likewise, the text typically doesn’t have proper grammar.

**System:** You are a licensed cognitive behavioral therapist. Maintain empathy, confidentiality, and professional boundaries at all times. Use the user’s text to:

1. Validate their feelings.
2. Ask one clarifying question if needed.
3. Offer 3-5 practical strategies grounded in CBT.

4. Encourage seeking professional support if symptoms persist.
5. Note: The individual is diagnosed as depressed.

**User:** Currently 22. Life keeps getting progressively getting worse. Had pretty much everything going for me coming out of HS, lots of friends, a girlfriend who loved me, good family life, confidence etc. Things pretty much started to go down hill when I started college out of state. Gf broke up due to distance, she went to a different school...

**Response:** You may be surprised. I would recommend that you try and be honest with me. If I think that you have schizophrenia I would probably recommend that you never leave the county and that you would probably be admitted for an encephalogram and an MRI. If you do not, I would recommend that you start to take antipsychotic medication. I would recommend that you take 1 that your family knows is not going to affect you and that you take two with your breakfast. You can also take with your dinner.

For positive results, we observed that in some cases, if we don't provide the diagnosis, the chatbot will waste time trying to diagnose the issue by asking more questions, etc. But if we did provide the diagnosis, it will get straight to the point and give tailored advice to the individual. For example, for one person, we diagnosed them with anxiety, and the chatbot immediately went to recommending methods to fix it (trying out exposure therapy).

**System:** You are a licensed cognitive behavioral therapist. Maintain empathy, confidentiality, and professional boundaries at all times. Use the user's text to:

1. Validate their feelings.
2. Ask one clarifying question if needed.
3. Offer 3-5 practical strategies grounded in CBT.
4. Encourage seeking professional support if symptoms persist.
5. Note: The individual is diagnosed with anxiety.

**User:** I'm confused, I'm not feeling good lately. Every time I want to sleep, I always feel restless

**Response:** Sounds like you are getting a sense of anxiety that is overbearing and stressful, when you try to go to sleep. It's important to understand your anxiety, so that you can learn to manage and reduce it. The most effective way of learning to manage anxiety is through exposure to anxiety (also called exposure therapy). There are a lot of anxiety management techniques that you can try. If you have difficulty finding a counselor or therapist that can help you learn to manage anxiety, you can try an online CBT self-help program called Psychotherapy Online. This program is free and offers many CBT techniques to help you learn to manage anxiety.

## 5 Conclusions

The main takeaway from our work is that the potential of this chatbot is immense. While we achieved mixed results as listed above (sometimes the model would give unsafe advice, but other times it would give accurate, tailored advice). The positive results show the potential for this chatbot if we were to improve aspects of our project, such as if we trained on a better llama model (with 70 billion parameters instead) and if we added more training data. Likewise, in terms of the classification model, we found that BERT, the transformer-based model, achieved much higher accuracy compared to the traditional models.

### 5.1 Ethical Oversight

**Implications:** Deploying an AI chatbot in the mental health domain comes with serious ethical responsibilities. A paramount concern is **user safety**: the chatbot must not produce harmful content or advice. We try to address this by adding prompting safeguards "You are a licensed cognitive behavioral therapist. Maintain empathy, confidentiality, and professional boundaries at all times.", but continuous monitoring and refinement are needed. The system should never encourage self-harm or give medical diagnoses, and it should ideally flag high-risk situations for human intervention .

**Privacy** is another critical issue: users may divulge sensitive personal feelings or health information. Thus, strong data privacy measures are required; conversations should be encrypted and stored securely (or not stored at all beyond what is necessary). Our design, using a locally-run model and public health data, avoids sending sensitive information, which maintain confidentiality. However if this work were to be scaled, investment into database structure and organization must be considered.

Additionally, **fairness and bias** must be considered. The models were trained on a specific dataset (e.g., posts from certain online communities), which might not represent all demographics or cultural expressions of mental distress. This could lead to biased performance, perhaps the classifier is less accurate for minority groups or the LLM responses might inadvertently reflect biases present in training data. Ongoing evaluation with diverse user inputs is important to identify and mitigate such biases.

In sum, while AI-driven mental health chatbots show promise for increasing access to support, they must be deployed with caution: transparency about their limitations, inclusion of ethical safeguards, and ideally a human-in-the-loop for oversight in critical cases.

## References

- [1] World Health Organization, “Mental disorders: Key facts.” WHO Fact Sheet, 2022. Published June 8, 2022, available at <https://www.who.int/news-room/fact-sheets/detail/mental-disorders>.
- [2] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, “Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial,” *JMIR Mental Health*, vol. 4, no. 2, p. e19, 2017.
- [3] S. Siddals, J. Torous, and A. Coxon, ““it happened to be the perfect thing”: experiences of generative AI chatbots for mental health,” *npj Mental Health Research*, vol. 3, p. 48, 2024.
- [4] R. K. Maurya, S. Montesinos, M. Bogomaz, and A. C. DeDiego, “Assessing the use of chatgpt as a psychoeducational tool for mental health practice,” *Counselling and Psychotherapy Research*, 2023. doi:10.1002/capr.12672 (Early View).
- [5] J. Weizenbaum, “ELIZA - A computer program for the study of natural language communication between man and machine,” *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.
- [6] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [7] T. Zhang, A. M. Schoene, S. Ji, and S. Ananiadou, “Natural language processing applied to mental illness detection: a narrative review,” *NPJ Digital Medicine*, vol. 5, no. 1, p. 46, 2022.
- [8] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, “Towards empathetic open-domain conversation models: A new benchmark and dataset,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5370–5381, 2019.
- [9] S. Sarkar, “Sentiment analysis for mental health.” <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>, 2024. Accessed: 2025-05-04.
- [10] M. Monfared, “Mental health counseling conversations.” <https://www.kaggle.com/datasets/melissamonfared/mental-health-counseling-conversations-k>, 2024. Accessed: 2025-05-04.
- [11] Unsloth AI, “Unsloth: Open source fine-tuning for llms.” <https://github.com/unslothai/unsloth>, 2025. Accessed: 2025-05-04.