Proposal: Screening Personality Disorders Through Everyday Conversations

Abhiram Rustagi, Feiyang Xie, Mrunal Madhav Hole

Abstract

This project aims to classify major personality disorders using daily conversations of people. Due to the lack of a pre-existing dataset, we will build our dataset entirely on synthetic data generation. The dataset is curated and evaluated using AI models and statistical metrics, and the next step is to reach out to people and psychotherapists to collect real data for evaluation purpose which is left out of this class project. This labeled dataset is used to two classification models, with the goal of enabling the detection of mental disorder patterns through natural dialogue rather than traditional questionnaires.

1 Problem Statement

The focus of this project is on personality disorder. "Personality disorders are defined as deeply ingrained and enduring patterns of behavior that typically manifest during childhood or adolescence and persist throughout adulthood. These patterns represent extreme deviations from the usual ways that individuals perceive, think, feel and relate to others. Such disorders complicate clinical practice and are associated with an increased risk of developing other mental health conditions. In addition, personality disorders are generally lifelong, significantly affecting an individual's ability to function in various domains of life" (World Health Organization, 1992).

"Personality disorders have been documented in approximately 9 percent of the general US population." (Angstman & Rasmussen, 2011). The population of the US was 311.6 million in 2011, and this means that around 30 million people suffer from personality disorder. Considering the fact that personality disorders are very complicated and enduring, and the fact that significant numbers of people are impacted, there will be a great benefit if AI can facilitate the early diagnosis of these personality disorders and so that people can take steps afterwards.

To be specific, this project aims to do is to help get an early-stage diagnosis of the person with daily conversation texts with the person on their everyday interactions, feelings and experiences with the society and other people. Regarding how to carry out the early diagnosis, we can do a classification to see if a person matches any type of personality disorder. Both the ICD-10 and DSM-IV-TR classify 10 major personality disorders which are divided into three clusters. The types are annotated with the statistics of population percentage mentioned in Angstman & Rasmussen, 2011.

- Cluster A (Odd or Eccentric Disorders): Paranoid(0.5-2.5%), Schizoid (0.5-7%), Schizotypal(3%)
- Cluster B (Dramatic, Emotional, or Erratic Disorders): Antisocial(1%), Borderline (1.6%), Histrionic(1-3%), Narcissistic (1% with all features met, 2-16% with partial)
- Cluster C (Anxious or Fearful Disorders): Avoidant(5.2%), Dependent (0.6%), Obsessive-Compulsive (2.4%)

To simplify our project, we pick the top three types: Schizoid from the Cluster A, Narcissistic from the Cluster B, and Avoidant from the Cluster C.

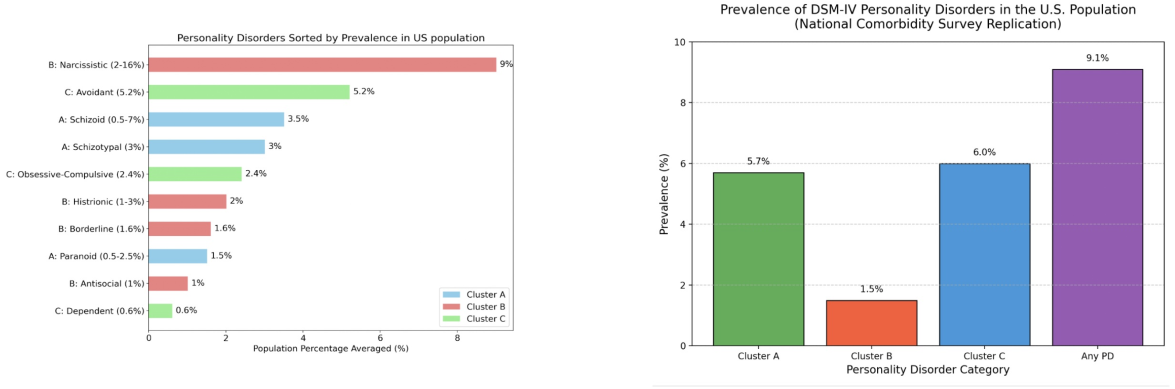


Figure 1: Statistics of Personality Disorder in US Population

Another survey on the comorbidity of personality disorder showed that the overlapping of disorders is not rare. (Lenzenweger et al. (2007)),so it is more reasonable if we assign different labels to people instead of just one. For example, people can have the true label of [A=1, B=1, C=0] which are diagnosed with both Schizoid and Narcissistic. And if our prediction has [A=0.7, B=0.6, C=0.1], our prediction is consistent with the true label. Since we aim to only have very early diagnosis of a person, our threshold doesn't have to be over 50% and we can set it to be 30% and give the person suggestion to take more sophisticated and clinical diagnosis based on our AI's daily observations.

2 Data Set Building

LLMs are the promising solution to build a dataset for our project with the following reasons. First, LLMs, through extensive pretraining, have acquired a vast repository of knowledge and demonstrate exceptional linguistic comprehension (Kim et al., 2022; Ding et al., 2023a), which forms the foundation for generating faithful data. In addition, the profound instruction follow-up capabilities of LLMs allow better controllability and adaptability over the generation process, facilitating the creation of tailored datasets for specific applications with more flexible process designs (El Dan and Li, 2023).

2.1 The Framework of Synthetic Data Generation

Our project follows the pipeline summarized in Long et al. (2025) as is shown in Figure2. Simply put, the pipeline is made up of three major steps: data generation mainly through prompt engineering, data curation and data evaluation.

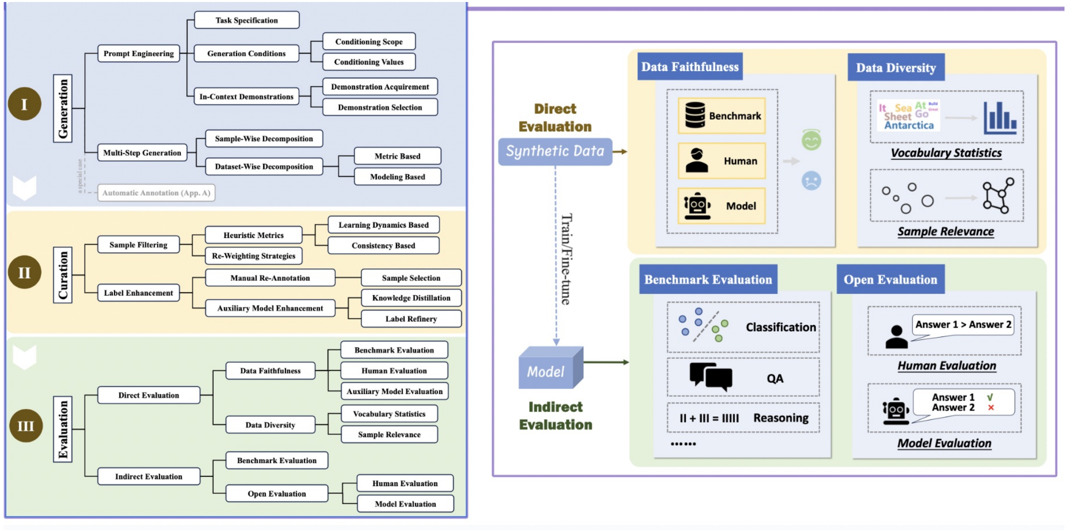
To apply this pipeline into our project, we first design prompts to generate small amounts of data, and then we apply data evaluation to the small datasets we have generated, and if they quality evaluated are not good enough we will improve our prompts accordingly, and we will also filter out the sub-datasets that are of bad quality.

2.2 Prompt Engineering

We have iterated two versions of the prompts for data generation, and the second version follows the design of prompt summarized in Long et al. (2025)

$$p(T, D) \leftarrow E(e_{\text{task}}, e_{\text{condition}}, e_{\text{demo}}).$$

Figure 2: The pipeline of building a synthetic dataset.



The task specification is crucial for setting the right context for LLMs- driven data generation, which can also include role-play (Li et al., 2023c), format clarification, knowledge augmentation (Xu et al., 2023b; Sudalairaj et al., 2024), etc.

The pivotal challenge in using LLMs for synthetic data generation is ensuring sufficient diversity, as directly prompting the LLMs to produce data for certain tasks often results in highly repetitive outputs, even with a high decoding temperature (Gandhi et al., 2024; Liu et al., 2024). Addressing this problem, a widely adopted strategy is conditional prompting, which explicitly and concretely communicates to the LLMs the specific type of data desired.The core of conditional prompting involves delineating the targeted data through the formulation of a series of condition-value pairs:

$$e_{\text{condition}} = \{(c_1, v_1), (c_2, v_2), \cdots, (c_n, v_n)\}$$

The first condition we used for our prompt is the detailed description of the criteria for the diagnosis of personality disorder. And the second condition is the topic list of possible daily conversations people will get involved in.

An example of the prompt: Generate a conversation between A and B about the topic about travel. A has the following traits: [”Believes that he or she is ’special’ and unique”].And only generate the conversation with the A: and B: at the beginning of each sentence, no other labels and boilerplate are needed,and put the data in one line

Analysis of the prompt:

- Task specification: generate a conversation between A and B about
- Task conditions: topic and following traits
- Generation format

2.3 The Synthetic Data Set

Using this prompt, we finally have built a dataset of over 1 million tokens mainly from Gemini-2.0. And one example of the data set is:

The Label:

- The precise label: [0, 1, 0]
- The binary label: True (derived from the precise label)

The conversation: (generated by Gemini-2.0) "A: Oh, travel. It is something I approach with a certain… distinction, unlike most, I find. B: Do you travel often? A: When the destination aligns with my particular sensibilities, yes. I don't simply go anywhere, you understand. It must offer something… exceptional. ... (the original text is around 3000 characters long)"

2.4 Data Curation

- By Length: filtered out texts that are too short, the Gemini example contains 2689 Ascii characters, and we can take this as a good length and take like 50
- Balanced labels: the label combinations are generated randomly to obtains a uniforma distribution of all types kinds (but the normal and abnormal);
- Remove repetitive texts: we currently use similarity as curation standard and have dropped datasets that contain very similar texts;

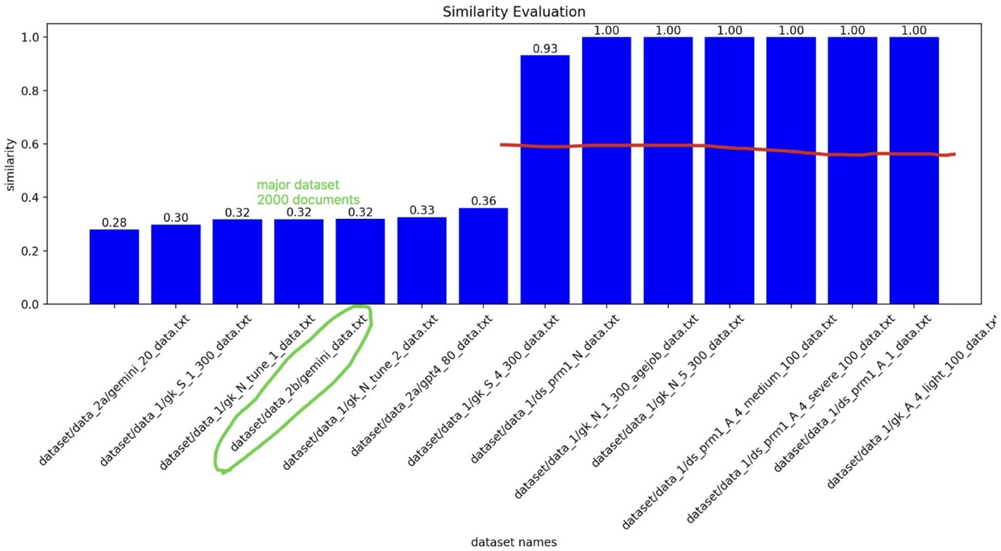
2.5 Data Evaluation

Diversity and faithfulness are the two major criteria to evaluate synthetic data.

2.5.1 Diversity

To gauge similarity, we use the average value of the pair-wise similarities of documents calculated using the MiniLM-L6-H384-uncased of SentenceTransformer (SBert). This model is trained with contrastive learning dedicated to encode short passages and compare similarity. And from Figure-3, we can see Gemini2.0 wins in small sampling and the diversity doesn't degrade when scaling up. The sub data sets that are of high similarity are filtered out in our final data set.

Figure 3: Diversity Evaluation (the smaller the better)



2.5.2 Faithfulness

The methodology of evaluating faithfulness is to cross-check a small sample of the data set generated by LLM-A and ask another LLM-B to annotate the personality disorders using the conversation. For example, take a few samples from Gemini-2.0 and ask GPT-4 for the labels, and the conversations generated by Gemini-2.0 can be correctly annotated by GPT-4o.

The basic assumption here is that the data quality generated with the same prompts and same LLM should be consistent, so we only need to cross check a small sample instead of a majority of them.

3 Prior Work

We learned from several therapy-related papers and haven't exhausted our paper list. Our reading for this phase mainly focuses on how to collect a dataset, and how to evaluate the model.

3.1 CaiTI, a Conversational AI Therapist

Nie et al. (2024) introduce CaiTI, a Conversational AI Therapist that will both do mental health screening and treatment. CaiTI aims the same with our initial plan, and it also encompasses a similar screening task as this project. CaiTI uses large language models (GPT and Llama) and reinforcement learning to create personalized conversation flows. CaiTI can provide psychotherapeutic interventions (CBT or MI) when needed.

The CaiTI dataset, developed with input from collaborating psychotherapists, comprises 6,950 therapist-labeled user response samples and 300 general responses (Yes, No, Maybe, Question, Stop). Although our course-based project lacks access to a dataset of this scale, we get valuable insights into the size of labeled data utilized by a more extensive study, and we may be too ambitious to expect more real data.

In terms of data format, CaiTI primarily screens users through a questionnaire-like method augmented with LLM. CaiTI's Questioner drives the conversation using Epsilon-Greedy Q-learning and a GPT-based ”Rephraser.” When a user responds, CaiTI segments the response into individual sentences and classifies each into predefined dimensions and scores. In connection to our project, We will learn from their dimensions which was designed with help of therapists, but we adopt a brand new approach other than questionnaires, focusing on detecting patterns of personality disorders by analyzing daily conversations and behavioral cues, moving away from predefined prompts to capture more natural indicators of mental health.

3.2 Other Related Work

Baihan Lin et al. (2023) built a framework to tune LLMs to have non-toxic conversations. In their study, they generated narcissistic-oriented conversations using in-context learning for their agents. This design inspired our idea of creating data with in-context learning and other prompt techniques to obtain real-world therapeutic data. And they have mentioned that their framework completely depends on high-quality training data for the AI agents, and this highlights our efforts to generate almost-real, high-quality synthetic data.

Hodson and Williamson (2024) concluded that large language models (LLMs) demonstrate potential in providing reasonable suggestions for identifying and reframing unhelpful thoughts, as evaluated by two independent CBT therapists currently practicing within the UK's National Health Service. Their findings serve as industry evidence that language models can support psychotherapeutic screening, reinforcing the feasibility of our proposal.

4 Classification on a minilm: MiniLM

We also wanted to test out the classification on a mini/small language model in order to know if you really require more compute resources to get better results. To test this out, we leveraged the paper on MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. MiniLM. In a nutshell, the authors propose a deep self-attention distillation framework designed to reduce the size of pre-trained Transformer models while preserving most of their performance across various NLP tasks. This approach helps deploy effective transformer-based models in resource-constrained environments without compromising accuracy.

We utilized the pretrained model microsoft/MiniLM-L12-H384-uncased which is available on Hugging Face. link.

Now to adapt the miniLM to our classification, we fine-tuned the model using the following hyperparameters:

- Tokenization sequence length: 512 tokens
- Loss function: Binary Cross Entropy with Logits Loss (BCEWithLogitsLoss)
- Learning rate (LR): 1e-4
- Training epochs: 8
- Batch size: 16

4.1 Performance and Results

The model gave us the following results:

- **Cluster 1** demonstrated a precision of 0.71, recall of 0.81, and an F1-score of 0.76.
- **Cluster 2** performed really well, with precision at 0.96, recall at 0.95, and an F1-score of 0.95.
- **Cluster 3** demonstrated a precision of 0.74, recall of 0.84, and an F1-score of 0.79.

These results indicate that even a smaller models like MiniLM can deliver a good performance, especially when computation resources are constrained. Models like BERT provide a much greater contextual understanding and potentially a higher f1-score, however models like MiniLM can be an alternative for early stage diagnostic purposes.

5 Personality Disorder Classification Using BERT

Multi-label text classification with BERT fine-tuning and accuracy optimization.

I] Introduction

Objective: Classify conversational text into three personality disorders (Schizoid, Narcissistic, Avoidant) using BERT.

II] Methodology

a. Bidirectional Context Understanding

- BERT uses a bi-directional approach considering both the left and right context of words in a sentence, instead of analyzing the text sequentially, BERT looks at all the words in a sentence simultaneously. It captures contexts like: ”I feel empty” (Avoidant trait) vs. ”I feel superior” (Narcissistic trait).This approach detects subtle differences in self-referential language critical for disorder classification.

b. [CLS] Token for Classification

- The [CLS] token aggregates entire conversation context into a single vector, used for classification.This enables multi-label prediction (e.g., [1, 0, 1] for Schizoid + Avoidant) via sigmoid outputs.

c. Pre-Trained Knowledge Transfer

- As the model is Pre-trained on 3.3B tokens (Wikipedia, BookCorpus), it learns general language patterns. We need 1M task-specific tokens for fine-tuning, making the model efficient for limited mental health datasets.

BERT is fine-tuned using labeled data specific to the downstream tasks of interest. These tasks could include sentiment analysis, question-answering, named entity recognition, or any other NLP application. The model's parameters are adjusted to optimize its performance for the particular requirements of the task. Thus, BERT's bidirectional analysis, [CLS] token efficiency, and pre-trained adaptability make it ideal for nuanced personality disorder classification.

III] Why BERTForSequenceClassification for classification?

a. BERT captures bidirectional context, which helps analyze data in a **dialogue format** effectively. This means it can understand the context of words, which provides a deeper understanding of conversational data.

b. BERT is known for its classification accuracy, allowing us to leverage its precise prediction capabilities to correctly identify personality disorders. This reduces the risk of misclassification, which is crucial as customers will depend on us for reliable results.

c. BERT supports multi-label classification. For example, following scenarios can be considered:
1) A person exhibiting traits of schizoid, narcissistic, and avoidant personality disorders may have the label [1,1,1].
2) A person displaying only schizoid and narcissistic traits would have the label [1,1,0].
3) A healthy individual without any disorders(listed above) would be labeled [0,0,0]. This kind of multi-label classification is possible with BERTś architecture.

IV] The Data Pipeline and File Structure:

a. data_preprocessing.py is used for dataset creation, preprocessing and tokenizing the data and labels: [Schizoid, Narcissistic, Avoidant] as the probability vector.
b. model_training.py will be used for classification where we use bertforsequenceclassification which is a BERT based classifier. We also calculate metrics such as accuracy and F1 score. AdamW is used as the optimizer which is BERT's recommended optimizer.
c. prediction.py :
1) Tokenize new conversation
2) Generate logits with fine-tuned BERT
3) Apply sigmoid + threshold (50'%') for disorder detection
d. main.py will incorporate the entire pipeline

V] Optimizer (AdamW) & Hyperparameters:

a. AdamW Optimizer: To optimize our BERT model for personality disorder classification, we implemented several key hyperparameter adjustments. First, we switched to the AdamW optimizer, which handles weight decay differently than standard Adam—instead of adding it to the loss function, AdamW applies decay directly during parameter updates. This provides more stable regularization, preventing the model from over-relying on noisy patterns in our limited dataset while preserving useful pre-trained knowledge.

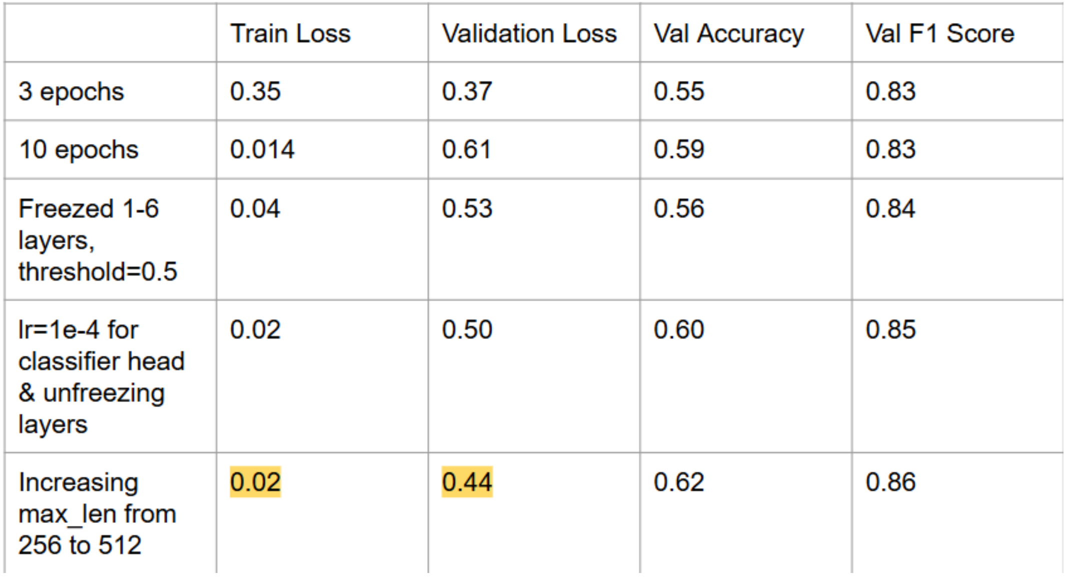
b. Layer-Wise Learning Rates: We adopted **layer-wise learning rates** to fine-tune the model more precisely. The later BERT layers (6–12) used a learning rate of **2e-5**, ensuring meaningful updates to their language understanding. Meanwhile, the classifier head—which starts from scratch—received a higher rate of **1e-4**. This was done to accelerate learning adaptation to our specific task.

c. Threshold Tuning: We also tuned the prediction threshold, initially set to 0.3 for balanced F1 performance. Raising it to 0.5 reduced false positives (e.g., mislabeling neutral statements as disordered) at the cost of slightly lower recall.

d. Layer Freezing and Gradual Unfreezing: To train the model effectively, we initially froze the first six layers of BERT, preserving its basic language understanding. Later, we gradually unfroze layers—starting with layers 4–5 at epoch 3 and layers 2–3 at epoch 5—so the model could adapt to new tasks step by step without losing its foundational knowledge. This strategy improved optimization, reducing validation loss by about 13 percent (from 0.61 to 0.53) and making the model more reliable.

VI] Evaluation Metrics

Figure 4: Results

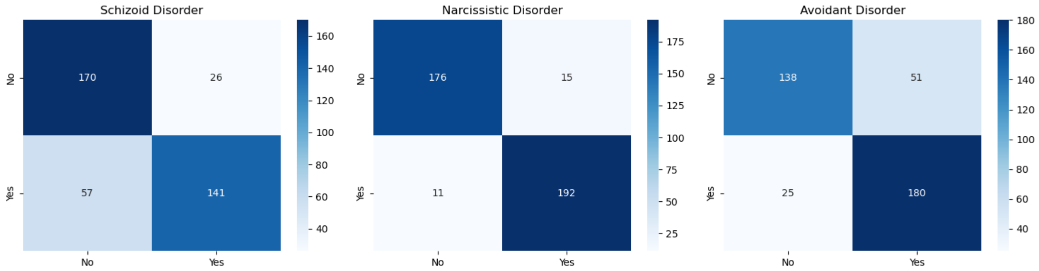


	Train Loss	Validation Loss	Val Accuracy	Val F1 Score
3 epochs	0.35	0.37	0.55	0.83
10 epochs	0.014	0.61	0.59	0.83
Freezed 1-6 layers, threshold=0.5	0.04	0.53	0.56	0.84
lr=1e-4 for classifier head & unfreezing layers	0.02	0.50	0.60	0.85
Increasing max_len from 256 to 512	0.02	0.44	0.62	0.86

Interpretation: As the validation loss is greater than the training loss, this implies that our model is overfitting on the training data.

VII] The Confusion Matrix:

Figure 5: Confusion Matrix



Interpretation: This confusion matrix evaluates the model's classification performance for three disorders: Schizoid, Narcissistic, and Avoidant. It highlights strong results for Narcissistic Disorder with minimum false positives and false negatives, while showing areas for improvement in handling false positives for Avoidant Disorder and false negatives for Schizoid Disorder.

6 Acknowledgement

The Section-5 (Personality Disorder Classification Using BERT) refers to geeks-for-geeks which makes this Section diagnosed as AI mixed by pangram.

7 Contribution

Feiyang Xie: Authored the entire Problem Statement section; the Data Set Building section (the Ciperd Data Set project); and collaborated with team members to summarize and restructure the Prior Work section.

Mrunal Madhav Hole: Designed and implemented the entire BERT-based classification pipeline (Section 5), including model training, hyperparameter tuning, and evaluation. (Code available here.)
Primarily authored Prior Work (Section 3 - Literature Review) by reading relevant research papers to summarize previous research done in this field.

Abhiram Rustagi: Designed and implemented the MiniLm (Small language model). Also looked into data modeling and generating data. (Link. (minilm link

8 References

Nie, J., Shao, H. (V.), Fan, Y., Shao, Q., You, H., Preindl, M., Jiang, X. (2025). LLM-based conversational AI therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices [Unpublished manuscript]. Columbia University, New York, NY, United States.

Hodson, N., Williamson, S. (2024). Can large language models replace therapists? Evaluating performance at simple cognitive behavioral therapy tasks. Journal of Mental Health Technology, 1(1), 23–34.

Jason R. D'Cruz, William Kidder, and Kush R. Varsh- ney. 2022. The empathy gap: Why ai can forecast behavior but cannot assess trustworthiness.

Lin, B., Bouneffouf, D., Cecchi, G., Varshney, K. R. (2023). Towards healthy AI: Large language models need therapists too. Icahn School of Medicine at Mount Sinai IBM TJ Watson Research Center.

Angstman, K. B., Rasmussen, N. H. (2011). Personality disorders: Review and clinical application in daily practice. American Family Physician, 84(11), 1253-1260.

World Health Organization. (1992). International Classification of Diseases (ICD-10): Mental and Behavioral Disorders. WHO.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.)

Lenzenweger, M. F., Lane, M. C., Loranger, A. W., Kessler, R. C. (2007). DSM-IV personality disorders in the National Comorbidity Survey Replication. Biological Psychiatry, 62(6), 553–564.