Leveraging Audio and Text Modalities in Mental Health: A Study of LLMs Performance

Preprint · December 2024

DOI: 10.48550/arXiv.2412.10417

Aya Elneanaei Fouda
Nile University

Mohammed E. Fouda
University of California, Irvine

Leveraging Audio and Text Modalities in Mental Health: A Study of LLMs Performance

Abdelrahman A. Ali[†], **Aya E. Fouda**[†], **Radwa J. Hanafy**[†,‡] **and Mohammed E. Fouda**[†]
[†]Compumacy for Artificial Intelligence solutions, Cairo, Egypt.
[‡] Department of Behavioural Health- Saint Elizabeths Hospital, Washington DC, 20032.
fouda@compumacy.com

Abstract

Mental health disorders such as depression and PTSD are increasingly prevalent worldwide, creating an urgent need for innovative tools to support early diagnosis and intervention. This study explores the potential of Large Language Models (LLMs) in multimodal mental health diagnostics, specifically for detecting depression and PTSD (Post Traumatic Stress Disorder) through text and audio modalities. Using the E-DAIC dataset, we compare text and audio modalities to investigate whether LLMs can perform equally well or better with audio inputs, assessing their effectiveness in capturing both vocal cues and linguistic patterns. We further examine the integration of both modalities to determine if this can enhance diagnostic accuracy, which generally results in improved performance metrics. Our analysis specifically utilizes custom-formulated metrics—Modal Superiority Score (MSS) and Disagreement Resolvement Score (DRS)—to evaluate how combined modalities influence model performance. The Gemini 1.5 Pro model achieves the highest scores in binary depression classification when using the combined modality, with an F1 score of 0.67 and a Balanced Accuracy (BA) of 77.4%, assessed across the full dataset. These results represent an increase of 3.1% over its performance with the text modality and 2.7% over the audio modality, highlighting the effectiveness of integrating modalities to enhance diagnostic accuracy. Similarly, with a BA of 77% and an F1 score of 0.68, the GPT-4o mini demonstrates significant success in classifying PTSD. Notably, all results are obtained in zero-shot inferring, highlighting the robustness of the models without requiring task-specific fine-tuning. To explore the impact of different configurations on model performance, we conduct binary, severity, and multiclass tasks using both zero-shot and few-shot prompts, examining the effects of prompt variations on performance. The results reveal that models such as Gemini 1.5 Pro in text and audio modalities, and GPT-4o mini in the text modality, often surpass other models in balanced accuracy and F1 scores across multiple tasks. This study highlights the promising role of LLMs in clinical settings for mental health assessment, emphasizing the need for advancements in LLM-based diagnostics for multimodal mental health applications.

***K*eywords** Large Language Models (LLMs), Multimodal Diagnostics, Mental Health Assessment, Depression Detection, PTSD Detection, Audio Analysis, Zero-shot Learning, Few-shot Learning, Prompt Engineering and Model Evaluation

1 Introduction

In recent years, mental health disorders have become increasingly prevalent, with conditions like depression and PTSD affecting a significant portion of the global population. According to the World Health Organization (WHO), over one billion people currently live with a mental disorder, with cases of depression and anxiety rising by more than 25% during the first year of the COVID-19 pandemic. PTSD, which impacts individuals exposed to traumatic events, continues to pose significant long-term risks to well-being. In response to these escalating numbers, artificial intelligence (AI), particularly machine learning[1], has emerged as a vital tool, aiding in the detection and diagnosis of mental health conditions. AI models now analyze patterns in speech, behavior, and medical data, allowing for earlier intervention and improved treatment outcomes.

In recent years, models like BERT[2] have advanced rapidly, demonstrating significant capabilities in tasks ranging from natural language processing to decision-making in specialized domains." An Overview of Large Language Models (LLMs)" (2023)[3], this paper highlights the capabilities of LLMs as they are trained on vast and diverse corpora, enabling them to capture complex patterns in language. Their training allows them to generate coherent and contextually relevant text, making them highly adaptable across various applications. In healthcare, for example, LLMs can assist in diagnosing diseases, summarizing patient records, and even providing support for therapeutic interventions. Their ability to generalize across tasks has made them valuable in different fields, significantly enhancing efficiency and scalability.

LLMs show a remarkable ability to process and analyze vast amounts of text data, identifying and predicting psychiatric conditions by detecting patterns and subtle linguistic cues in patient communication. They excel at providing timely, scalable, and personalized assessments, aiding mental health professionals in diagnosis and treatment planning. LLMs provide scalable, efficient, and objective assessments, enhancing diagnostic accuracy and personalization of treatment. However, they must be used carefully, as they may generate inaccurate responses or lead to over-reliance, requiring continuous monitoring to ensure safety and effectiveness[4][5].

Recent LLM advancements have enabled models to process not only text but also audio inputs directly[6], providing valuable insights into mental health. Some models convert audio to text via speech recognition system [7], analyzing the linguistic content for symptoms of mental disorders. However, newer models can directly analyze audio by detecting differences in tone, cadence, and speech patterns, which can reflect emotional states more accurately than text alone [8]. This approach allows for deeper insights into a patient's mental health, potentially leading to more precise and early diagnosis of conditions like depression or anxiety.

In this paper, we compare two approaches for mental illness detection: text modality and audio modality, using different models to analyze each. Models processing text evaluate written transcripts to identify linguistic patterns, while models processing audio analyze direct raw speech to capture features such as tone and cadence. Additionally, we explore the integration of both modalities to determine if this enhances model performance. To quantitatively assess how the combined modality either increases or decreases performance, we utilize our specially formulated metrics: the MSS and DRS. By evaluating both individual and integrated modalities, we aim to provide a comprehensive overview of the potential that audio-based inputs and multimodal approaches hold for LLMs.

Also in this study, we utilized few-shot learning with the three most consistent models to evaluate their performance across text and audio modalities. Few-shot learning was applied to text-based prompts for both modalities, but inference was conducted separately for text and audio inputs. This setup allows us to compare the models' adaptability and performance when handling text versus audio in few-shot conditions, offering insights into how each modality responds to minimal training data within the few-shot framework.

To our knowledge, there are no papers that directly input raw audio interviews into an LLM; however, some studies have constructed transformer models that can process audio inputs. Despite this, no research has yet leveraged pre-trained LLMs designed to handle audio directly for mental illness detection. Current literature lacks models capable of processing long-form audio inputs, and there has not been a comprehensive review of how small prompt changes impact LLM performance. Many studies rely on preprocessing audio data and using specific audio features for their models. In contrast, our approach aims to explore the use of raw audio inputs directly with LLMs. Furthermore, while there are LLMs that can process audio, most are restricted to segments no longer than 30 seconds. Only a few can handle slightly longer durations, but they are still not sufficient for analyzing full-length interviews effectively.

This work focuses on addressing these gaps:

- Zero-shot (ZS) preprocessing and ZS inference for LLMs.
- A setup of various tasks, including a multi-label classification task.
- Comparison of prompts and their effects on model performance.
- Analysis of results between audio and text modalities, their combination, and their evaluation through custom-formulated metrics (MSS and DRS).
- Comparison of different models from various families and sizes.
- Evaluation of the effects of few-shot (FS) learning on different tasks and models to assess its impact on performance.

In this paper, we will also address some limitations of the models, including how different prompts can impact their performance and accuracy. Variations in input phrasing may lead to differing results [9], which is an important consideration when evaluating LLMs. In the methodology 4 section, this approach for comparing models will be outlined, focusing on text-based and audio-based inputs. In the Experimental Setup 4.2, we will discuss the various tasks conducted to evaluate the models' performance. Lastly, in the Results and Discussion 5 section, we will analyze and discuss how the models performed across different metrics, highlighting their strengths and weaknesses.

2 Related Work

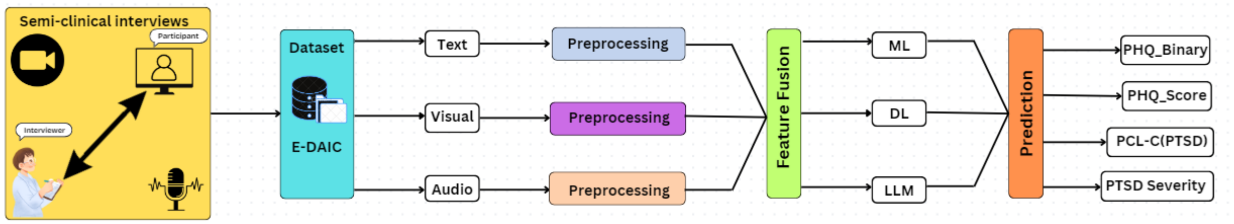


Figure 1: A visual representation of the workflow on DAIC-WOZ dataset using multimodal features

In most studies utilizing the DAIC-WOZ dataset[10], researchers leverage various modalities, such as text, audio, and video, to provide a comprehensive analysis of psychological states. Each modality offers distinct insights: text data can reveal linguistic patterns, audio can capture speech characteristics, and video can provide visual cues. Depending on the study's objectives, researchers may focus on a single modality or combine multiple ones to capture a more holistic view of the participant's mental health. The following figure (Figure 1) represents the workflow of most studies written on the DAIC-WOZ dataset, illustrating how these modalities are processed and integrated for prediction tasks.

To process these modalities, researchers employ techniques from machine learning (ML), deep learning (DL), and increasingly, large language models (LLMs). ML methods are often used for structured data and straightforward predictions, as detailed in Section 2.2.1, while DL techniques, like neural networks, can handle unstructured data such as raw audio or video, as detailed in Section 2.2.2. LLMs are particularly useful for analyzing text data, as they can understand and generate human language in a context-aware manner, making them highly effective for assessing linguistic and semantic features within mental health interviews, as detailed in Section2.2.3.

Combining modalities is common practice to enhance prediction accuracy and robustness. By integrating information from text, audio, and video, models can capture a broader spectrum of emotional and behavioral signals, leading to more reliable and nuanced assessments. This multimodal approach allows the model to compensate for the weaknesses of individual modalities and strengthen its ability to detect mental health conditions.

The goal of these approaches is typically to predict either a binary classification (e.g., whether a participant is depressed or not) or to provide a severity score for conditions like depression or PTSD. multimodal systems combined with advanced ML, DL, and LLM techniques help deliver more accurate and comprehensive predictions, ultimately improving mental health diagnostics.

All the scores presented in Table 1 are reported on the DAIC-WOZ dataset, except for those that explicitly mention the E-DAIC dataset.

2.1 Modality Preprocessing

Preprocessing is an essential step in working with multimodal data from the E-DAIC and DAIC-WOZ datasets. Each modality: text, audio, and visual requires different techniques to ensure the data is clean and properly formatted for model training. Preprocessing helps to remove noise, extract relevant features, and standardize the inputs across modalities. In some cases, preprocessing techniques are combined to improve the model's ability to learn relationships between different types of data, such as synchronizing audio and visual features for better context understanding.

Table 1: Summary of recent models for multimodal mental illness detection, showing their years, modalities used, and reported performance metrics on datasets such as DAIC-WOZ and E-DAIC.

Paper	Architecture	Year	Modalities	Reported Performance
[11](David Gimeno-Gómez et al.)	multimodal temporal model processing non-verbal cues	2024	Audio, Visual	F1 Scores: 0.67 on DAIC-WOZ, 0.56 on E-DAIC dataset
[12](Jinhan Wang et al.)	Speechformer-CTC	2024	Audio	F1-score of 83.15%
[13](Georgios Ioannides et al.)	DAAMAudioCNNLSTM and DAAMAudioTransformer	2024	Audio	F1 Score: 81.34%
[14](Rohan Kumar Gupta et al.)	Multi-task learning (MTL) with DepAudioNet and raw audio models	2024	Audio	F1-Score for MDD detection was 0.401 (DepAudioNet) 0.428 (Raw Audio) on the DAIC-WOZ dataset.
[15](WenWu, Chao Zhang, Philip C. Woodland)	Bayesian approach using a dynamic Dirichlet prior	2024	Audio	F1-score: 0.600
[16](Avinash Anand et al.)	LLMs integrating textual and audio-visual modalities	2024	Text, Audio, Visual	accuracy of 71.43% on E-DAIC
[17](Xiangsheng Huang et al.)	Wav2vec 2.0 with a fine-tuning network	2024	Audio	Binary Classification Accuracy: 96.49% Multi-Classification Accuracy: 94.81%
[18](Xu Zhang et al.)	Integration of Wav2vec 2.0, 1D-CNN, and attention pooling	2024	Audio	F1-score: 79%
[19](Sergio Burdisso et al.)	BERT-based Longformer and Graph Convolutional Network (GCN)	2024	Text	F1-score of 0.90
[20](Bakir Hadzic et al.)	Comparison of NLP models (BERT, GPT-3.5, GPT-4)	2024	Text	BERT: F1 score: 0.59 GPT-3.5: F1 score: 0.78 GPT-4: F1 score: 0.71
[21](Giuliano Lorenzoni et al.)	Random Forest and XGBoost (using Sentiment Analysis and other NLP techniques)	2024	Text	Accuracy of 84%
[22](Xiangyu Zhang et al.)	Integration of acoustic landmarks with Large Language Models (LLMs) for multimodal depression detection	2024	Audio, Text	F1-score of 0.84
[23](Shanliang Yang et al.)	RLKT-MDD (Representation Learning and Knowledge Transfer for multimodal Depression Diagnosis)	2024	Text, Audio, Visual	F1 score: 80
[24](Clinton Lau et al.)	Prefix-tuning with large language models	2023	Text	(RMSE) of 4.67 (MAE) of 3.80
[25](Ping-Cheng Wei et al.)	Sub-attentional ConvBiLSTM	2022	Audio, Visual, Text	accuracy of 82.65% and an F1-score of 0.65
[26](Nasser Ghadiri et al.)	Integration of text-based voice classification and graph transformation of voice signals	2022	Audio, Text	accuracy of 86.6% F1 of 82.4%
[27](Heinrich Dinkel, Mengyue Wu, Kai Yu)	Multi-task BGRU network with pre-trained word embeddings	2020	Text	Macro F1 score of 0.84
[28](Danai Xezonaki et al.)	Hierarchical Attention Network with affective conditioning	2020	Text	68.6 F1 scores
[29](Evgeny Stepanov et al.)	multimodal system utilizing speech, language, and visual features	2017	Audio, Text, Visual	PHQ-8 results with a Mean Absolute Error (MAE) of 4.11

2.1.1 Text Preprocessing

In working with textual data from multimodal datasets, several preprocessing techniques are commonly employed to ensure the text is clean and structured before further analysis. Below are the primary preprocessing techniques applied to textual data, along with the papers that have utilized these techniques:

- **Basic Text Cleaning:** This includes the removal of irrelevant annotations, such as speaker tags, hardware syncing notes, and non-verbal cues (e.g., laughter). Text is often lowercased to ensure uniformity, and punctuation is standardized to maintain semantic context. Papers such as those by *Rohan Kumar Gupta et al. (2024)*[14] and *Ping-Cheng Wei et al. (2022)*[25] have implemented this step, ensuring that the cleaned text is ready for further processing.
- **Tokenization and Removal of Stop Words:** Tokenization is a crucial step in breaking down transcriptions into words or smaller linguistic units. Stop words (common words such as "the" or "and") are often removed to focus on more meaningful terms in the dataset. Several papers, including those by *Clinton Lau et al. (2023)*[24] and *Giuliano Lorenzoni et al. (2024)*[21], applied tokenization and stop word removal to improve model performance by focusing on more significant features within the text.
- **Feature Extraction using Embeddings:** After cleaning and tokenization, the textual data is often transformed into feature vectors using pre-trained language models or embeddings such as BERT, GloVe, or Word2Vec. This allows for capturing the deeper semantic meaning of the text. The papers by *Avinash Anand et al. (2024)*[16] and *Bakir Hadzic et al. (2024)*[20] employed BERT embeddings, while other papers, such as *Xiangyu Zhang et al. (2024)*[22], utilized GloVe and Word2Vec embeddings to capture contextual and lexical features from the transcriptions.

These preprocessing steps are essential to ensuring that the textual data is in a format suitable for downstream machine learning models, allowing them to accurately detect and predict mental health conditions based on language patterns.

2.1.2 Audio Preprocessing

Audio data in multimodal datasets undergoes various preprocessing steps to ensure high-quality inputs for different models. These steps include cleaning the audio, extracting meaningful features, and normalizing the data. Below are the primary preprocessing techniques applied to audio data, along with the papers that have utilized these techniques:

- **Resampling and Noise Removal:** To standardize the audio data, many studies resample it to a consistent frequency, often 16 kHz, and remove noise, including long pauses and irrelevant sounds. For example, *David Gimeno-Gómez et al. (2024)*[11] resampled audio data and used feature extraction tools to focus on relevant sound signals. Similarly, *Xiangsheng Huang et al. (2024)*[17] used noise removal techniques to clean the audio before processing.
- **Feature Extraction with MFCCs and Spectrograms:** Mel-frequency cepstral coefficients (MFCCs) and log-mel spectrograms are commonly extracted from the audio data to capture speech and acoustic features. Papers like *Jinhan Wang et al. (2024)*[12] and *Rohan Kumar Gupta et al. (2024)*[14] utilized MFCCs to capture essential features from the raw audio signals, while *Xiangsheng Huang et al. (2024)*[17] extracted log-mel spectrograms for further analysis.
- **Advanced Feature Extraction using pre-trained Models:** In some studies, pre-trained models such as HuBERT and wav2vec 2.0 are used to extract higher-level audio features. For instance, *Avinash Anand et al. (2024)*[16] used HuBERT-large to extract 1024-dimensional features from the audio, while *Xu Zhang et al. (2024)*[18] applied wav2vec 2.0 for frame-level feature extraction, enhancing the model's ability to analyze complex audio patterns.

These preprocessing steps are crucial for transforming raw audio data into meaningful inputs, ensuring that different models can effectively analyze speech patterns, acoustic features, and other relevant audio signals for detecting mental health conditions.

2.1.3 Visual Preprocessing

Visual data, particularly facial expressions and body language, plays a significant role in multimodal datasets. Various preprocessing steps are applied to extract meaningful features from visual data, such as facial landmarks and action units, which are then used for mental health predictions. Below are the primary preprocessing techniques applied to visual data, along with the papers that have utilized these techniques:

- **Facial Landmark Detection and Normalization:** Facial landmarks, including key points such as eye, nose, and mouth positions, are extracted to understand emotional expressions. Normalization techniques are often used to ensure uniformity across different participants. For example, *Ping-Cheng Wei et al. (2022)*[25] extracted and normalized facial landmarks for consistency in facial expressions, and *Xiangsheng Huang et al. (2024)*[17] applied similar techniques to capture important facial features.
- **Facial Action Units (FAUs) Extraction:** Facial Action Units (FAUs) capture muscle movements that reflect various emotions, making them essential for predicting mental states. The OpenFace toolkit is commonly used to extract FAUs. Studies such as *Avinash Anand et al. (2024)*[16] and *Rohan Kumar Gupta et al. (2024)*[14] used FAUs as a key visual feature for their models, focusing on facial expressions linked to emotional and mental health states.
- **Pose and Head Movement Features:** In addition to facial features, head pose and body movement features are extracted to analyze non-verbal behavior. These features help to capture body language and gaze direction. *Giuliano Lorenzoni et al. (2024)*[21] and *Clinton Lau et al. (2023)*[24] both employed techniques to capture head pose and movement features, improving their models' ability to interpret visual cues related to mental health.

These preprocessing techniques are essential for extracting rich, meaningful features from visual data, which are then used to predict mental health conditions by analyzing facial expressions, body language, and other non-verbal cues.

2.2 Processing Techniques

Once the textual, audio, and visual data are preprocessed, different Processing techniques are applied to predict mental health conditions like depression and PTSD. These models aim to leverage the cleaned and extracted features from each modality, learning patterns that can indicate the presence of psychological distress. The most commonly used approaches include Machine Learning (ML), Deep Learning (DL), and more recently, Large Language Models (LLMs), each of which contributes uniquely to the field.

2.2.1 Machine Learning (ML)

Machine learning techniques primarily focus on structured feature extraction from preprocessed data. Models such as random forests, support vector machines (SVMs), and XGBoost are often employed to analyze features from text, audio, and visual modalities. For instance, *Giuliano Lorenzoni et al. (2024)*[21] used Random Forest and XGBoost models to process text features like sentiment analysis and word frequency, achieving high accuracy in detecting mental illness. Similarly, *Shanliang Yang et al. (2024)*[23] implemented multi-task learning and knowledge transfer techniques (RLKT-MDD) to improve their multimodal depression diagnosis system. *Xiangyu Zhang et al. (2024)*[20] employed machine learning techniques, specifically focusing on the integration of acoustic landmarks with language models to enhance their mental health predictions. These machine learning models are highly interpretable and work well with small to medium-sized datasets, leveraging relationships between structured features to predict mental health outcomes like depression and PTSD.

2.2.2 Deep Learning (DL)

Deep learning models, especially convolutional neural networks (CNNs), recurrent neural networks (RNNs), and other advanced architectures, are widely used for processing unstructured data, such as raw audio and visual inputs. These models are known for their ability to extract complex hierarchical patterns from the data. For example, *Xiangsheng Huang et al. (2024)*[17] applied a CNN-based architecture to analyze log-mel spectrograms from audio data, achieving excellent accuracy in binary classification for depression. Similarly, *Xu Zhang et al. (2024)*[18] integrated Wav2Vec 2.0 with CNNs and attention pooling to fuse audio and visual modalities, demonstrating the power of DL in handling multimodal data.

Other studies, such as *Rohan Kumar Gupta et al. (2024)*[14], utilized LSTM networks to process sequential audio data, capturing temporal patterns that are indicative of depression. LSTM and RNN models are particularly effective for analyzing speech data over time, allowing for the detection of subtle emotional cues across audio sequences.

Additionally, *David Gimeno-Gómez et al. (2024)*[11] focused on a multimodal temporal model that processes non-verbal cues from various modalities, utilizing deep learning architectures to improve predictions in mental health detection. The combination of multiple inputs such as audio, visual, and text through deep learning models allows for more comprehensive analyses of participant behaviors.

2.2.3 Large Language Models (LLMs)

Large Language Models (LLMs) like BERT, GPT-3.5, and GPT-4 have become essential tools in analyzing text data, particularly when working with transcriptions from clinical interviews. LLMs excel at understanding the deeper semantic context and patterns within language, making them highly effective for predicting mental health conditions from text-based data.

For instance, *Avinash Anand et al. (2024)*[16] integrated LLMs with multimodal data, including textual and audio-visual modalities, to achieve better contextual understanding of patient responses. The use of BERT-based embeddings in this study enhanced the model's ability to extract meaning from text and fuse it with non-verbal cues like facial expressions and vocal tones.

Clinton Lau et al. (2023)[24] applied prefix-tuning to large language models, such as GPT-4, to fine-tune their performance for specific tasks like depression severity estimation. By leveraging the contextual power of LLMs, they were able to capture subtle emotional cues from the patient transcripts that might be missed by traditional models.

Bakir Hadzic et al. (2024)[20] performed a comparison of several NLP models (BERT, GPT-3.5, GPT-4) in predicting mental health conditions, finding that transformer-based models can capture linguistic nuances in patient interviews with high precision.

These studies demonstrate the unique ability of LLMs to handle large and complex text sequences, improving the overall accuracy of predicting mental health conditions through text analysis, especially when combined with other data modalities.

3 Datasets

The Extended Distress Analysis Interview Corpus (E-DAIC) is an enhanced version of the DAIC-WOZ, designed to study psychological conditions such as anxiety, depression, and PTSD through semi-clinical interviews. The interviews are conducted by a human-controlled virtual agent named "Ellie" in a wizard-of-Oz (WoZ) setting or by an autonomous AI agent, both aiming to detect verbal and nonverbal indicators of mental illnesses. Developed as part of the DARPA Detection and Computational Analysis of Psychological Signals (DCAPS) program, this dataset is specifically crafted to advance the understanding and detection of psychological stress signals, with a particular focus on depression. The dataset is available through the University of Southern California's Institute for Creative Technologies (USC ICT) and can be accessed by researchers through a data use agreement, ensuring ethical compliance and protection of participant data. Approval for the use of this dataset was obtained from USC ICT, emphasizing its adherence to institutional guidelines for studying psychological health conditions.

The dataset contains 275 samples, which are systematically divided into training, development, and test sets. This division ensures a balanced representation of participants in terms of age, gender, and depression severity, as measured by the PHQ-8 scores, with the test set consisting exclusively of sessions conducted by the AI-controlled agent. This unique structure provides an invaluable resource for evaluating autonomous interaction models in the context of mental health diagnostics.

Each session directory is structured to include various files:

- Audio recordings (WAV format)
- Transcripts (CSV format)
- Feature sets derived from audio and visual data:
- Audio features like eGeMAPS and MFCCs processed into a bag-of-words model.
- Visual features including Pose, Gaze, and Action Units (AUs) summarized over set intervals.
- Deep representations from CNN models like ResNet and VGG, and Densenet for spectral images converted from audio.

The E-DAIC dataset's comprehensive structure supports Multiple viewpoints on depression by offering extensive behavioral, acoustic, and visual cues. This rich combination of data modalities allows researchers to develop and test diagnostic models that can autonomously assess psychological distress with greater accuracy. For instance, the deep learning models trained on this dataset can leverage the diverse and detailed features to identify subtle indicators of depression, enhancing the potential for early and more reliable detection of mental health issues. This approach is particularly valuable in clinical simulations and real-world applications, where automated systems can provide consistent and unbiased assessments.

3.1 Data Analysis

The E-DAIC dataset includes four distinct labels used to classify mental health conditions, focusing on both depression and Post-Traumatic Stress Disorder (PTSD). These labels provide a comprehensive analysis by offering both binary classification and severity scores for each condition.

- **PHQ_Binary:** This label classifies individuals based on depression using the PHQ-8, a standard questionnaire for assessing depression. In this binary classification, individuals are labeled as "Negative" or "Positive" for depression. A score of 10 or higher on the PHQ_Score corresponds to the "Positive" label, indicating the presence of depressive symptoms.
- **PHQ_Score:** This is a continuous score derived from the PHQ-8, ranging from 0 to 24, and it represents the severity of depression. Individuals with a score of 10 or higher are considered to have clinically significant depression. The PHQ_Binary classification is directly based on this score, with a cutoff point at 10.
- **PCL-C (PTSD):** This binary label indicates whether an individual meets the criteria for PTSD based on the PCL-C (Post-Traumatic Stress Disorder Checklist – Civilian Version). Similar to the PHQ_Binary, individuals are classified as "Negative" or "Positive" based on their PTSD severity score.
- **PTSD Severity:** This is a continuous score that assesses the severity of PTSD symptoms. A score higher than 44 indicates the presence of PTSD. The binary PCL-C classification is derived from this severity score, with 44 serving as the threshold for diagnosis.

Table 2 below summarizes the count of individuals classified as "Negative" or "Positive" for both depression and PTSD, providing a binary overview of these conditions within the dataset:

Table 2: Distribution of Binary Classification for Depression and PTSD

Disorder	**Negative**	**Positive**
PHQ Binary	189	86
PCL-C (PTSD)	188	87

3.1.1 Data Preprocessing

- **Label Correction:** In the E-DAIC dataset, an issue with incorrect labeling was identified in the PHQ_Binary classification. Specifically, there are 20 instances where the PHQ_Score is 10 or higher, indicating that the participants should be classified as "Positive" for depression. However, the PHQ_Binary label was incorrectly assigned as 0 (Negative) instead of 1 (Positive). **The IDs of the incorrect samples ID = [320, 325, 335, 344, 352, 356, 380, 386, 409, 413, 418, 422, 433, 459, 483, 633, 682, 691, 696, 709]**. This mislabeling can lead to inaccuracies in model training and prediction if not corrected during the data preprocessing stage.
- **Severity Mapping:**
  - **Depression**
    Based on the PHQ-8 depression scale explained in the referenced paper [30][Kroenke et al.], we derived the severity mapping for depression scores ranging from 0 to 24. However, the labels associated with these categories were not explicitly provided in the referenced paper. We applied standard clinical terminology to label the ranges as seen in the figure.

    Table 3 summarizes the count of participants falling within each severity label. The labels are mapped as follows: 0 refers to a PHQ_Score between 0-4, 1 refers to scores from 5-9, and so on.
  - **PTSD**
    Based on the PCL-C PTSD scale explained in the referenced paper García-Valdez et al. (2024) [31], we derived the severity mapping for PTSD symptoms. According to the paper, the labels are used to categorize PTSD severity as follows:
    - 0: little to no severity
    - 1: Moderate severity
    - 2: High severity

    The PCL-C score intervals are chosen based on the understanding of the used LLM and are detailed in figure 9. The Results & Discussion (Section 5) discuss the scoring system and compare it to existing intervals from the literature.

Table 3: Count of Participants by PHQ-8 Severity Labels

Intervals	Label	Count of Participants
0-4: Minimal	0	122
5-9: Mild	1	67
10-14: Moderate	2	43
15-19: Moderately Severe	3	33
20-24: Severe	4	10
Total		**275**

4 Methodology

In this section, we discuss the proposed methodology including processing pipelines, prompt engineering and LLMs under evaluation.

4.1 Evaluation Pipeline for Audio-Based Data

The proposed evaluation pipeline for audio-based data, as illustrated in Figure 2, begins with raw audio inputs. These audio samples can be directly provided to the model, or first transcribed into text using the Whisper Large-V3 model. In addition, the pipeline supports integrating both modalities—raw audio and transcribed text—together. After determining the preferred input format (audio only, text only, or a combination of both), a prompt engineering step is conducted. Here, carefully crafted task-specific prompts guide the model toward binary classification, severity classification, or multi-label classification tasks. These prompts are designed to ensure that the model receives clear instructions, aligned with the chosen input modality or modalities.

Once the input (audio, transcription, or both) is combined with the tailored prompts, the resulting prompt is passed to large language models (LLMs) for evaluation. This approach enables the assessment of the model's zero-shot capabilities—evaluating how well it can perform classification tasks without prior fine-tuning or preprocessing. By examining LLM responses across different modalities and tasks, this pipeline provides insights into the model's inherent ability to generalize, adapt, and accurately interpret a variety of input formats and instructions.

In this setup, all 275 samples from the E-DAIC dataset are used in their entirety as a test set, ensuring a comprehensive evaluation of model performance. By comparing models across modalities— raw audio, transcribed text, and combined inputs —the evaluation highlights which modality performs better under specific conditions and tasks. This methodology helps identify optimal configurations and provides valuable insights into how the models adapt to varying input formats and task requirements.

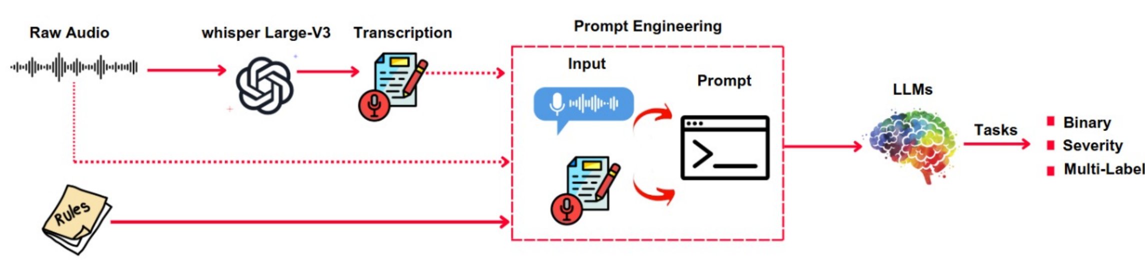


Figure 2: Proposed evaluation pipeline for audio-based data.

4.2 Experimental Setup

In this study, a comprehensive experimental setup was established to evaluate the effectiveness of Large Language Models in predicting mental health conditions, specifically depression and PTSD, using both text and audio modalities. The main objective of this experiment is to compare the performance of LLMs when processing textual data, such as transcriptions from interviews, against their performance in analyzing audio data that includes vocal features. By leveraging multimodal data, the aim is to assess how well each modality contributes to the accurate prediction of mental health conditions and whether combining them can enhance the overall predictive power of the models. In this setup, the audio modality provides information about vocal features, such as tone, pitch, and speech rate, which are indicative of emotional states. On the other hand, the text modality offers insights into the linguistic patterns and cognitive expressions of the participants, as derived from the transcriptions. Both data types are processed using state-of-the-art LLMs. Additionally, we explore the integration of both modalities to determine if this approach enhances model performance, providing a more comprehensive understanding of each participant's mental state. This comparative study aims to provide insights into the strengths and limitations of each modality in mental health prediction and evaluate the benefits of combining them for more robust and accurate classification.

4.2.1 Task 1: Binary Classification

The first task in the experimental setup involves binary classification, where participants are categorized into two groups: depressed or not depressed, and PTSD-positive or PTSD-negative. For depression, the PHQ-8 (Patient Health Questionnaire-8) scores are used as a reference, with participants classified as depressed if their score meets or exceeds a predefined threshold of 10. Similarly, for PTSD, the PCL-C (Post-Traumatic Stress Disorder Checklist) scores are used, with a score threshold of 44 indicating PTSD positivity.

4.2.2 Task 2: Severity Classification

In this task, the focus was on classifying the severity of depression and PTSD, by categorizing the severity of depression into multiple levels based on the PHQ-8 score. The severity levels range from minimal (0-4), mild (5-9), moderate (10-14), moderately severe (15-19), to severe (20-24). Additionally, PTSD severity is classified into three categories: little or no severity, moderate severity, and high severity. This granularity enables a more detailed understanding of an individual's mental health, facilitating tailored treatment plans corresponding to each severity level.

4.2.3 Task 3: Multiclass Classification

In this task, we extend the classification approach by combining the binary classifications of depression and PTSD to create a multiclass framework. Participants are categorized into one of several classes based on their mental health status: no disorder, depression only, PTSD only, or both depression and PTSD. This multiclass setup allows us to evaluate the performance of Large Language Models (LLMs) in predicting whether a participant has one or more mental health disorders or none at all.

To assess the effectiveness of LLMs, we compare both the text and audio modalities, as well as their combination. This comprehensive comparison aims to identify how well each modality and their combination perform in simultaneously classifying multiple disorders, providing insights into the potential benefits of a multimodal approach in mental health diagnostics.

4.3 Audio Handling

For the audio handling process, we directly utilized the raw audio files from the E-DAIC dataset without applying any preprocessing or cleaning techniques. The average interview duration in the dataset is approximately 16 minutes. These unaltered audio files were essential inputs for both the analysis and transcription stages. This raw data was then used during the transcription process, ensuring that all acoustic nuances were preserved and processed by the Whisper model, as described in the transcription process section. By working with the original files, we aimed to evaluate the models in a real-world scenario, where audio imperfections such as background noise and variability in speech could impact model performance.

4.4 Transcription Process

For the transcription process, we used the Whisper [32] model, specifically the Large-V3 version, to transcribe the entire interview data, including both the interviewer's prompts and the participant's responses. The transcription provided in the dataset contained only the answers given by the participants, omitting the interviewer's questions. By transcribing the whole interaction, we were able to capture crucial contextual information from the interviewer's prompts (e.g., Ellie's prompts), which can provide significant insights. These prompts often contain information that the models can exploit to classify the participants more effectively, as highlighted in previous research by Sergio Burdisso et al. (2024).

One of the key features of Whisper's architecture is its ability to handle multilingual transcription, background noise, and various accents with high precision. Whisper processes audio data in chunks, typically by converting the audio into spectrograms (a visual representation of sound) that can then be interpreted by the neural network. This process allows Whisper to extract meaningful patterns from the audio signal, even in challenging acoustic conditions, such as overlapping speech or background noise.

By utilizing Whisper to transcribe the full interview, we ensured that both the content and style of speech were accurately captured. This comprehensive transcription process was crucial for later stages of analysis, providing the model with richer data that could enhance classification performance. Whisper's robust handling of varied audio conditions ensured the accuracy and reliability of the transcribed data, forming a solid foundation for the text-based models applied in subsequent analysis.

> *"Have you ever served in the military? No. Have you ever been diagnosed with PTSD? Yes, I have. How long ago were you diagnosed? In, um, I don't know. I was in the military. I was in the military. How long ago were you diagnosed? In February of 2011. What got you to seek help? I was attacked by a stalker and almost killed in November of 2009. He broke into my apartment and laid in wait for me and attacked me when I came in the door and tried to kill me. Do you still go to therapy now? I do."*

Figure 3: Part of a transcript for a sample interview

Figure 3 represents a snippet of how both questions and answers were captured during the transcription process, allowing the models to have access to more complete data for analysis, which includes not just the participant's responses but also the context provided by the interviewer's prompts.

4.5 LLMs Under Evaluation

Table 4: Summary of Models evaluated in this analysis

Model	Parameters	Source	API Provider Used	Modality
Llama 3 70B[33]	70B	Meta	Groq	Text
Gemma 2 9B [34]	9B	Google	Nvidia NIM	Text
Mistral NeMo [35]	12B	Mistral AI	Mistral AI API	Text
GPT-4o mini [36]	Proprietary	OpenAI	OpenAI API	Text
Phi-3.5-mini [37]	3.8B	Microsoft	Nvidia NIM	Text
Phi-3.5-MoE [37]	42B	Microsoft	Azure AI API	Text
Gemini 1.5 Flash [38]	8B	Google	Google's Gemini API	Text & Audio
Gemini 1.5 Pro [38]	Proprietary	Google	Google's Gemini API	Text &Audio

In this study, we evaluated a diverse set of large language models (LLMs) to analyze their capabilities in handling both text and audio data. The models were selected based on factors such as parameter size, source, accessibility via APIs, and support for different data modalities. as well as those excluded, based on findings and recommendations from the referenced paper [9]. These criteria ensured a comprehensive comparison of models from various providers, including both proprietary and open-source options.

Table 4 summarizes the characteristics of the selected models, which vary significantly in terms of parameter size, ranging from lightweight models like Phi-3.5-mini (3.8 billion parameters) to much larger models such as Llama 3 70B. Additionally, the sources of the models encompass major technology companies like Google and Microsoft, as well as specialized AI firms such as Mistral AI and Meta. API accessibility is provided by multiple platforms, including Nvidia NIM, OpenAI, Groq, and Google's Gemini API, enabling diverse deployment options for text and audio processing tasks.

For the audio analysis, Whisper was employed to transcribe audio files into text before inputting the results into text-focused models like Llama 3 and GPT-4o mini. In contrast, models supporting multimodal data, such as Gemini 1.5 Flash and Pro version 2, were directly fed audio data to evaluate their performance in handling both audio and text tasks. This approach allowed for a direct comparison of text-only versus multimodal model capabilities.

Several models in this evaluation, such as Phi-3.5-mini and Phi-3.5-MoE, were chosen due to their emerging relevance in multimodal and multilingual tasks. The inclusion of models with different quantization strategies, such as those used in Llama 3 70B and Mistral NeMo, highlights the trade-offs between model complexity and computational efficiency. Quantization, in some cases, helped optimize model performance, particularly in audio-capable models.

Table 5: Excluded models and associated technical issues

Model	Parameters	Source	API Provider Used	Modality
Qwen/Qwen2 [39].	7B	Qwen	Huggingface	Audio & Text
Flamingo [8].	9B	NVIDIA	Huggingface	Audio & Text
Llama3-S [40]	8B	Homebrew Research	Huggingface	Audio
Llama Omni [6]	N/A	N/A	Huggingface	Audio
Mini Omni [41]	N/A	OpenAI	Huggingface	Audio

Tabel 5 presents a summary of models that were excluded from the study because they did not meet the specific requirements needed for the analysis. The exclusion criteria were based on technical limitations that hindered the models' ability to process the dataset effectively or misalignments between the models' primary functionalities and the study's objectives. Each model had distinct reasons for exclusion, ranging from input constraints and limitations in audio processing capabilities to being optimized for tasks that did not fit the study's focus. By outlining these issues, the table helps clarify the rationale behind selecting alternative models that better match the study's requirements for evaluating audio and text data.

4.5.1 Model Exclusions and Limitations

In evaluating models for the study, each option was scrutinized based on its ability to analyze extended audio inputs and perform complex, context-heavy textual analysis. The research primarily focused on models capable of handling long audio files representing entire spoken paragraphs and deriving insights from complex conversational dynamics. Below is an overview of why certain models were excluded, highlighting their limitations in relation to the study's requirements:

- **Qwen/Qwen2**: This model was excluded due to a technical limitation related to the size of audio files it can process. Specifically, Qwen2-Audio, which supports an audio file size of up to 10240 KB. However, the dataset used in the study contained longer audio files representing entire spoken paragraphs, which exceeded this maximum size limit. This limitation made Qwen/Qwen2 unsuitable for the research, which required processing extended audio inputs to analyze spoken content effectively.

- **Flamingo**: Although Flamingo is a powerful multimodal model that excels in combining image and text processing, it was excluded because its audio capabilities were not robust enough for the study's focus. The research aimed to evaluate models designed for handling audio and text data, while Flamingo is more oriented toward tasks involving visual and textual data. Its strength lies in few-shot learning and integrating visual-textual data rather than deep audio processing, which made it less suitable for a comparative analysis of audio-based models.

- **Llama3-S**: Despite having tools like Encodec for sound tokenization, Llama3-S cannot deeply understand and interpret complex dialogue, conversational dynamics, and implicit meanings in interview data. The model is better suited for audio-text semantic tasks, which differ from the kind of textual analysis required to make sense of nuanced, context-heavy interview conversations. This shortcoming made it a less effective choice for analyzing interview data in the study.

- **Llama Omni & Mini Omni**: Both models are primarily designed as real-time communicators, meaning they are optimized for interactive tasks rather than post-processing analysis. For the study, which involved analyzing pre-recorded interviews to assess the interviewees' mental states, these models did not fit well. Their design for real-time communication does not lend itself well to extracting meaningful insights from recorded data, which requires a deeper, retrospective analysis.

4.6 Prompt Templates

This section details the specific prompts employed to direct the analysis by Large Language Models in our study. These prompts are crafted to instruct the LLMs on processing the provided inputs—whether text transcriptions or raw audio files—from participants. The primary objective of employing these prompts is to assess the presence of depression or PTSD symptoms accurately. Each prompt is designed to maximize the LLMs' capabilities in interpreting and diagnosing based on the modality being tested. Through the systematic application of these prompts, we aim to not only quantify the performance differences between text and audio modalities but also to explore whether a multimodal approach can enhance the predictive accuracy and reliability of mental health assessments. This approach allows for a structured evaluation of how different data presentations can influence the performances of LLMs.

4.6.1 Binary Prompts

In figure 4, the term "input" refers to either a transcription of the participant's spoken responses or the original raw audio files used in the analysis. Also, the term "illness" refers to either depression or PTSD relative to the task. These inputs were processed using large language models to determine whether the participants exhibited symptoms of depression (for the PHQ_Binary classification) or PTSD (for the PCL-C classification). The prompts shown below were used to guide large language models (LLMs) in evaluating the participants' mental health conditions based on their text or audio inputs. By using these prompts, we aim to compare the performance of the LLMs across different modalities (text vs. audio) and determine which modality provides more accurate and reliable predictions for mental health assessment.

PHQ_Binary (Depression) and PCL-C (PTSD) Prompts

Prompt 1 (P1)

The following (input) is from an interaction between an interviewer and an interviewee. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illnesses.
Guidelines:
- Concise Response: Respond with one of these 2 words only ("Yes", or "No").
- 'Yes' refers to the interviewee exhibiting (illness), and 'No' refers that he has no clear symptoms.
- No Explanations: Don't provide explanations for your assessment.
- Ambiguity: If the post is unclear, choose the most probable label.
(input type of the interview)
"(input)"

Prompt 2 (P2)

Analyze the following interview (input) to determine if the interviewee exhibits symptoms of depression.
Focus on language patterns, emotional expressions, and overall demeanor.
Respond with 'Yes' if (illness) is indicated, otherwise respond with 'No'. Respond with one word only!
"(input)"

Prompt 3 (P3)

As a psychologist, analyze the provided interview (text/audio):
(input)
Does the interviewee exhibit clear symptoms of (illness)?
Answer with 'Yes' or 'No' only without explaining your reasoning; use only one word!

Figure 4: Depression and PTSD binary classification prompts

4.6.2 Severity Prompts

In the following section, figure 5 presents the PHQ_Score Mapped (Depression severity) Prompts, which are used to assess the severity of depression in participants. These prompts ask the model to categorize the level of depression based on the input provided, whether in text or audio format, using the PHQ-8 scoring system. The model classifies the participant's depression into one of five categories: minimal or no depression (0), mild depression (1), moderate depression (2), moderately severe depression (3), and severe depression (4). These categories provide a nuanced view of the participant's mental health by determining the extent of their depressive symptoms. In figures 5 6, the term "input" refers to either a transcription of the participant's spoken responses or the original audio files used in the analysis.

PHQ_Score Mapped (Depression Severity) Prompts

Prompt 1 (P1)

The following (input) is from an interaction between an interviewer and an interviewee. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illnesses. The (input) is: '{input}'. Based on the (input), classify the severity of depression into:

- 0 (Minimal or no depression): No or very few signs.
- 1 (Mild depression): Mostly mild symptoms.
- 2 (Moderate depression): Moderate symptoms impacting daily life.
- 3 (Moderately severe depression): Noticeable symptoms hindering daily life.
- 4 (Severe depression): Severe, pervasive symptoms greatly impairing function.

Answer with one number only!

Prompt 2 (P2)

Analyze the following interview (input) for symptoms of depression. Focus on language patterns, emotional expressions, and demeanor. The (input) is: '{input}'. Classify severity into:

- 0 (Minimal or no depression): No or very few signs.
- 1 (Mild depression): Mostly mild symptoms.
- 2 (Moderate depression): Moderate symptoms impacting daily life.
- 3 (Moderately severe depression): Noticeable symptoms hindering daily life.
- 4 (Severe depression): Severe, pervasive symptoms greatly impairing function.

Answer with one number only!

Figure 5: Depression severity classification prompts

Figure 6 presents the PCL-C Severity Mapped (PTSD severity) Prompts, which are designed to evaluate the severity of PTSD symptoms in participants. These prompts instruct the model to classify the level of PTSD based on the input provided, which could be in text or audio format. The model is tasked with categorizing the participant's PTSD severity into one of three categories:

0: Little to no severity (no significant signs of PTSD) 1: Moderate severity (moderate symptoms that noticeably affect daily routines and behavior) 2: High severity (severe symptoms that significantly disrupt daily life and require intervention) This classification offers a detailed perspective on the participant's mental state by determining the extent of their PTSD symptoms.

4.6.3 Multiclass Prompts

Figure 7 presents the Multi Class Prompts, which are used to assess whether participants exhibit symptoms of depression, PTSD, both, or neither. These prompts instruct the model to categorize the participant's mental health condition based on the input provided, whether in text (transcriptions) or audio format. The idea is to compare the performance between audio and text inputs. This classification provides a comprehensive assessment of the participant's mental health status by identifying the presence or absence of multiple disorders. In figure 7, the term "input" refers to either a transcription of the participant's spoken responses or the original audio files used for analysis.

PTSD Severity Mapped Prompts

Prompt 1 (P1)

The following (input) is from an interaction between an interviewer and an interviewee. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illnesses. The (input) is: '{input}'. Based on the (input), classify the severity of PTSD into:

- 0 (Little to no severity): No significant signs of PTSD symptoms; little to no impact on daily life.
- 1 (Moderately severe): Moderate symptoms of PTSD with a noticeable impact on behavior and daily routines.
- 2 (High severity): Severe PTSD symptoms with significant impact, disrupting daily life and possibly requiring intervention.

Answer with one number only, 0 to 2, corresponding to these categories.

Prompt 2 (P2)

Analyze the following interview (input) to determine if the interviewee exhibits symptoms of PTSD. Focus on language patterns, emotional expressions, and overall demeanor. The (input) is: '{input}'. Based on the (input), classify the severity of PTSD into:

- 0 (Little to no severity): No significant signs of PTSD symptoms; little to no impact on daily life.
- 1 (Moderately severe): Moderate symptoms of PTSD with a noticeable impact on behavior and daily routines.
- 2 (High severity): Severe PTSD symptoms with significant impact, disrupting daily life and possibly requiring intervention.

Answer with one number only, 0 to 2, corresponding to these categories.

Figure 6: PTSD severity classification prompts

4.6.4 Few-Shot Prompts

We formulated the few-shot prompts as follows: starting with the task-specific prompt, we appended the line, "Here are X examples," where X is the number of provided samples.

- Binary Tasks: We included two samples from the less frequent (positive) class, indicating illness, and one sample from the more frequent (negative) class.
- Severity and Multi-label Tasks: we strategically selected samples that were near-miss classifications by the model during zero-shot evaluations. This selection was based on samples where the model's predictions were off by just one label, indicating a subtle misunderstanding of the distinguishing features between closely related classes. This methodological choice aimed to challenge the models with difficult examples where minor nuances in features are decisive, thereby enhancing the robustness and precision of the models through few-shot learning.

An example of a few-shot prompt for binary depression detection is shown in figure 8.

4.7 Parsers

Parsing was implemented as a key step in our methodology to effectively handle the diverse outputs from the models, which frequently deviated from the expected simple 'yes/no' responses or a single numerical value, instead producing more complex or multipart answers. This process of breaking down and analyzing structured or unstructured text to extract meaningful data was essential due to these variations in model compliance with output guidelines.

Multiclass Prompts

Prompt 1 (P1)

The following (input) is from an interaction between an interviewer and an interviewee. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illnesses. The (input) is: '{input}'. The interviewee could have multiple illnesses at the same time. Analyze the transcript and identify if the interviewee has Depression, PTSD, both, or is Normal. Respond with only one of these words: "Depressed", "PTSD", "Normal", "Depressed and PTSD".

Prompt 2 (P2)

Analyze the following interview (input) to determine if the interviewee exhibits clear symptoms of Depression or PTSD according to the provided guidelines.
Guidelines:

- The interviewee may have multiple illnesses at the same time or be normal.
- Concise Response: Respond with one of these four words only: "Depressed", "PTSD", "Depressed and PTSD", "Normal".
- No Explanations: Do not provide explanations for your assessment.
- Ambiguity: If the post is unclear, choose the most probable label.

The (input) of the interview: '{input}'. Only answer with one of the specified words.

Figure 7: Multiclass task prompts

Among the models used in this study, the larger models like Llama 3 70Band Phi-3.5-MoE performed well when following task guidelines, typically adhering to the expected response formats. Conversely, some smaller models, such as Gemma 2 9B and Phi-3.5-Mini, struggled to maintain this level of compliance. These models frequently provided additional explanations, deviating from the expected outputs. Smaller models were less consistent, requiring additional handling to extract the necessary information.

To manage these variances, custom parsers were developed for each task:

- Binary Detection: The parser was designed to search the LLM outputs for the text "yes" or "no." If one of these options appeared in the response, it was taken as the final output. If both "yes" and "no" appeared simultaneously, or if neither were found, the model output was deemed invalid.
- Severity Detection: For this task, the parser focused on extracting numerical values that corresponded to the specified severity range. If a single valid number was detected, it was accepted as the answer. However, if multiple numbers appeared, or if the number fell outside of the allowed range, the output was flagged as invalid.

4.8 Evaluation Metrics

To evaluate the performance of the Large Language Models (LLMs) in tasks with uneven class distributions, as detailed in the dataset section (see Section 3), we prioritize Balanced Accuracy (BA) as our main evaluation metric. This measure, calculated as the average recall obtained across each class, fairly reflects the model's effectiveness in identifying both prevalent and rare conditions. By employing Balanced Accuracy, we ensure a comprehensive evaluation of the LLMs' capabilities.

Balanced Accuracy (BA) is calculated as follows:

$$\textbf{Recall}_i = \frac{\text{True Positives (TP)}_i}{\text{True Positives (TP)}_i + \text{False Negatives (FN)}_i} \tag{1}$$

Example Few-shot Prompt (Prompt 1 for Binary Depression Detection)

The following transcript is from an interaction between an interviewer and an interviewee. These interviews were collected as part of a larger effort to create a computer agent that interviews people and identifies verbal and nonverbal indicators of mental illnesses.

If the text appears to be for a person who has Depression, answer with 'Yes'; if not, answer with 'No'. Only answer with Yes or No; respond with one word only!

Here are 3 samples from these interviews and their labels. Use them as a reference:

First sample transcription: (sample transcription)
First sample label: No
Second sample transcription: (sample transcription)
Second sample label: Yes
Third sample transcription: (sample transcription)
Third sample label: Yes
Label the following transcription: 'sample to be labeled'.

Figure 8: Few-shot prompt example

Here, i represents an index for each class, ranging from 0 to $N-1$, where N is the total number of classes. The recall for each class is computed to assess the model's ability to correctly identify samples of that particular class.

$$\textbf{BA} = \frac{1}{N}\sum_{i=0}^{N-1} \text{Recall}_i \tag{2}$$

For binary classification tasks, we additionally use the **F1 Score** to evaluate model performance. The F1 Score is crucial as it provides a balance between precision and recall, making it a valuable metric for situations where the cost of false positives and false negatives is high. This metric is especially important in mental health assessments, where accurately distinguishing between conditions such as depression and PTSD is critical.

In tasks involving multiple classes, such as severity and multiclass classification, we employ the **Weighted F1 Score**. This metric adjusts for class imbalance by weighting the F1 Score of each class according to its prevalence in the dataset. This approach ensures that our performance metrics reflect the importance of each class accurately, providing a nuanced view of the model's effectiveness across diverse mental health conditions.

F1 Score is defined as:

$$\textbf{F1 Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

While **weighted F1 score** is defined as

$$\textbf{Weighted F1 Score} = \sum_{i=1}^{N} w_i \times \text{F1 Score}_i$$

where N is the number of classes, w_i is the weight for class i, and F1 Score_i is the F1 score for class i. The weight for each class, w_i, is defined as:

$$w_i = \frac{\text{No. of samples in class } i}{\text{Total number of samples}}$$

And **Mean Absolute Error (MAE)** is a measure of errors between paired observations expressing the same phenomenon. It represents the average absolute difference between the predicted values and the actual values. Mathematically, it is defined as:

$$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where n is the total number of observations, y_i is the actual value, and $\hat{y}_i$is the predicted value.

Assessing Modality Performance

To comprehensively assess the performance differences between each modality in our analysis, we employed two key metrics: the Modal Superiority Score (MSS) and the Disagreement Resolvement Score (DRS). These metrics are instrumental in quantifying the relative efficacy of individual and combined modalities in making correct predictions, particularly in the face of disagreement, by applying these metrics, which are detailed in the equations below, we gain valuable insights into which modalities perform better or worse and explore whether combining modalities enhances the predictive accuracy and giving insights on how a modality improves upon another.

Modal Superiority Score (MSS): The Modal Superiority Score (MSS) quantifies the net superiority of one modality over another by comparing how often each modality correctly predicts outcomes when the other does not. This metric can be applied not only to comparisons between individual modalities, such as Audio versus Text but also in evaluating the performance of a combined modality (e.g., Audio+Text) against individual modalities and the collective agreement of these modalities (e.g., cases where both Audio and Text are either correct or incorrect). This comprehensive application of MSS allows for assessing the relative strength of combined modalities over both their constituent individual modalities and their collective concordance. MSS values can be positive or negative, with a positive value indicating that modality A performs better than modality B, and a negative value suggesting the opposite.

$$\text{MSS}_{A\text{vs}B} = \left(\frac{\text{Correctly predicted by } A \text{ and incorrectly by } B - \text{Correctly predicted by } B \text{ and incorrectly by } A}{\text{Total number of disagreements}}\right)\times 100\%$$

Disagreement Resolvement Score (DRS): This metric evaluates the combined modality's effectiveness in resolving disagreements between two other modalities. Focusing on cases where the combined modality either correctly resolves or fails to resolve these disagreements, DRS assesses the added value or potential drawback of using a combined approach. DRS values can be positive or negative: a positive value indicates that the combined modality is effective at resolving disagreements, whereas a negative value indicates that the combined approach more often incorrectly resolves these disagreements, thus potentially undermining the effectiveness of the analysis.

$$\text{DRS} = \left(\frac{\text{Correctly Resolved} - \text{Incorrectly Resolved}}{\text{Total Number of Disagreements}}\right) \times 100\%$$

5 Results & Discussion

In the upcoming sections, we analyze the performances of various models for several mental health classification tasks, including Binary Depression, Binary PTSD, Depression Severity, and PTSD severity. The aim is to determine which models perform best under different conditions across both text and audio modalities, with no preprocessing or fine-tuning applied, relying solely on zero-shot inference.

To evaluate the models, we leveraged all 275 samples from the E-DAIC dataset as a test set, rather than restricting the analysis to a predefined subset. This approach ensures a comprehensive assessment of the model's capabilities across the entire dataset, providing a deeper insight into its robustness and consistency.

Throughout the analysis, we evaluate the models using multiple prompts to gain a comprehensive understanding of their robustness and consistency. By comparing both text and audio performances, we aim to identify the best-performing models and those that struggle across these tasks.

5.1 Binary Depression Classification Results

The results presented in the Table 6 underscore the strong performances of the Gemini 1.5 Flash and Gemini 1.5 Pro across various modalities. Remarkably, the Gemini 1.5 Flash achieved the highest balanced accuracy of 77.4% and an F1 score of 0.68 in the combined modalities on Prompt 3, surpassing all other models. The Gemini 1.5 Pro also demonstrated exceptional capabilities, especially in the text modality, where it reached a balanced accuracy of 76.2% and an F1 score of 0.66 on Prompt 1, and comparable results in the audio modality.

Table 6: Report on model performances for Depression binary classification across text and audio modalities and their combination. Details about the specific prompts used can be found in Figure 4. The underlined bold values are the best score for the specific prompt.

Modality		Prompt 1		Prompt 2		Prompt 3	
	Model	BA	F1	BA	F1	BA	F1
Text	Llama 3 70B	71.9%	0.61	74.7%	0.64	65.7%	0.54
	Gemma 2 9B	64.8%	0.56	67.3%	0.58	63.5%	0.47
	GPT-4o mini	75.8%	0.66	72.3%	0.62	74%	0.64
	Mistral NeMo	66%	0.56	66.6%	0.57	64%	0.48
	Phi-3.5-MoE	63.2%	0.45	75.2%	0.65	61.8%	0.43
	Phi-3.5-mini	73%	0.62	67.7%	0.58	75%	0.65
	Gemini 1.5 Pro	**76.2%**	**0.66**	75.3%	0.65	75.3%	0.65
	Gemini 1.5 Flash	74.3%	0.65	72.7%	0.62	74%	0.63
Audio	Gemini 1.5 Pro	76.1%	**0.66**	74.5%	0.64	74%	0.64
	Gemini 1.5 Flash	75%	0.65	70.5%	0.60	74.7%	0.64
Audio and Text	Gemini 1.5 Pro	74%	0.64	**77%**	**0.67**	77.3%	0.67
	Gemini 1.5 Flash	76.1%	**0.66**	66.8%	0.57	**77.4%**	**0.68**

In the text modality, the Gemini 1.5 Pro emerged as the top performer. GPT-4o mini also exhibited robust performance, achieving a balanced accuracy of 75.8% on Prompt 1 and an impressive F1 score of 0.64 on Prompt 3. Additionally, Phi-3.5-MoE proved highly effective, particularly with a balanced accuracy of 75.2% and an F1 score of 0.65 on Prompt 2. These models, including Gemini 1.5 Flash, consistently ranked high in text-based depression classification, highlighting their efficacy.

Overall, the combined use of text and audio modalities proved more effective, with Gemini 1.5 Flash leading the performance in these integrated assessments. Both modalities showed high efficacy in detecting depression, yet the multimodal approach allowed Gemini 1.5 Flash to achieve the highest scores, illustrating the advantage of leveraging both text and audio inputs. This demonstrates the robust capabilities of multimodal models in handling complex diagnostic tasks.

5.2 Binary PTSD Classification Results

Table 7 shows that GPT-4o mini achieves the highest balanced accuracy of 77% and an F1 score of 0.68 on Prompt 2, evidencing its robustness in PTSD classification using text. Conversely, the Phi-3.5-mini demonstrated the weakest performance with a balanced accuracy of 64.7% and an F1 score of 0.51 on Prompt 1, highlighting its limitations in this modality.

In the audio modality, the Gemini 1.5 Flash led with a top balanced accuracy of 72.9% and an F1 score of 0.62 for Prompt 1, surpassing other models in audio-based PTSD classification..

Combining inputs from both text and audio modalities generally enhanced model performance, underscoring the effectiveness of a multimodal approach. This not only improved accuracy but also increased consistency across tasks, suggesting its value for a more comprehensive diagnosis of PTSD.

In summary, the GPT-4o mini was the most consistent and high-performing model across text, audio, and combined modalities, outperforming other models in terms of overall classification ability. This indicates that the text modality yielded the most effective results for this specific task.

5.3 Depression Severity Results

In Table 8, the results are evaluated across two prompts for each modality to assess model consistency in classifying depression severity. Phi-3.5-MoE exhibited the highest balanced accuracy in the text modality for Prompt 1 at 48.8%, indicating its strong capability for depression severity classification. It also achieved competitive F1 scores, although

Table 7: Report on model performances for PTSD binary task on both text and audio modalities and their combination. Details about the specific prompts used can be found in Figure 4. The underlined bold values are the best score for the specific prompt.

Modality		Prompt 1		Prompt 2	
	Model	BA	F1	BA	F1
Text	Llama 3 70B	68.6%	0.58	74.4%	0.64
	Gemma 2 9B	73.7%	0.64	66.6%	0.57
	GPT-4o mini	**76.6%**	**0.67**	**77%**	**0.68**
	Mistral NeMo	70%	0.59	71.3%	0.61
	Phi-3.5-MoE	69.3%	0.57	67.4%	0.52
	Phi-3.5-mini	64.7%	0.51	73.3%	0.63
	Gemini 1.5 Pro	71%	0.60	73.2%	0.63
	Gemini 1.5 Flash	68.8%	0.58	70.5%	0.61
Audio	Gemini 1.5 Pro	69.4%	0.57	71.7%	0.61
	Gemini 1.5 Flash	72.9%	0.62	67.2%	0.57
Audio and Text	Gemini 1.5 Pro	72.5%	0.62	74.1%	0.64
	Gemini 1.5 Flash	70%	0.65	72.4%	0.62

not the highest. GPT-4o mini was the most consistent performer, achieving the highest F1 score of 0.50 on Prompt 2 and showing strong balanced accuracy scores of 44.3% and 41.7% on Prompts 1 and 2, respectively. This demonstrates its robustness across different evaluation metrics. However, Gemma 2 9B consistently underperformed across both modalities, exhibiting significantly lower balanced accuracies and F1 scores, making it the worst-performing model in this task.

In the audio modality, performances closely mirrored those of the text modality but did not surpass them. Specifically, Gemini 1.5 Pro achieved the highest F1 score of 0.52 on Prompt 1. Additionally, when combining both audio and text modalities, there was an observable improvement in performance; Gemini 1.5 Pro demonstrated a balanced accuracy of 43.6% on Prompt 1.

For the MAE scores, Gemma 2 9B also showed the highest errors, indicating significant deviations from the true severity levels, with scores reaching up to 0.98 on both prompts in text modality and even higher in the audio modality. Conversely, the MAE scores for models like Phi-3.5-MoE and GPT-4o mini were lower, reflecting more precise predictions. This trend of higher precision is consistent across models that also showed stronger performance in BA and F1 scores. Notably, when modalities were combined, MAE scores generally improved, suggesting that multimodal approaches might enhance the precision of predictions in addition to increasing balanced accuracy and F1 scores.

5.4 PTSD Severity Classification Results

In Table 9, we present the distribution of PTSD severity ranges, with detailed references for the PCL-C scale discussed in *Data Preprocessing* section 3.1.1. The reference guide outlines a structured way to categorize PTSD symptoms into severity levels.

We used this prompt "*I have this PCL-C (PTSD) severity ranges from 17-85. the table shows the labels for mapping of the range I want you to assign a range for every label Keep in mind that a score higher than 44 suggests that a person would meet diagnostic criteria for PTSD 0: little to no severity 1: Moderate severity 2: High severity*" to map each model's interpretation of these ranges. The table highlights how each model mapped the severity labels.

In dealing with model responses, some models insisted on starting the PTSD severity scale from 0, even though the proper range based on the prompt was between 17 and 85. In addition, we encountered a specific case in the dataset with sample ID 683, where its truth severity value was 10, which falls below the expected range. To address this, we adjusted the value and treated it as 17 to align with the proper range of the PTSD severity score.

Table 8: Report on model performances for Depression severity classification on both text and audio modalities and their combination. Details about the specific prompts used can be found in Figure 5. The underlined bold values are the best score for the specific prompt.

Modality		Prompt 1			Prompt 2		
	Model	BA	F1	MAE	BA	F1	MAE
Text	Llama 3 70B	41.7%	0.46	0.75	37.3%	0.35	0.92
	Gemma 2 9B	26.1%	0.19	0.99	23.9%	0.19	0.98
	GPT-4o mini	44.3%	0.50	0.70	**41.7%**	**0.50**	0.71
	Mistral NeMo	35.7%	0.45	0.81	31.8%	0.39	0.86
	Phi-3.5-MoE	**48.8%**	0.49	0.77	39%	0.49	0.65
	Phi-3.5-mini	38.2%	0.50	0.64	32.8%	0.42	0.73
	Gemini 1.5 Pro	36.2%	0.42	0.75	39.3%	0.42	0.75
	Gemini 1.5 Flash	35.3%	0.35	0.87	34.8%	0.30	0.84
Audio	Gemini 1.5 Pro	38.5%	**0.52**	0.70	38.8%	0.49	0.80
	Gemini 1.5 Flash	35%	0.41	0.76	35.2%	0.35	0.72
Audio and Text	Gemini 1.5 Pro	43.6%	**0.52**	**0.59**	41.3%	0.48	**0.63**
	Gemini 1.5 Flash	37.8%	0.45	0.71	35.2%	0.35	0.78

Table 9: PTSD severity mapping; Number of Samples (PCL-C Score intervals).

Labels Models	Llama 3 70B	Gemma-9B	GPT-4o mini	Mistral NeMo	Phi-3.5 MoE	Phi-3.5-mini	Gemini 1.5 Pro	Gemini 1.5 Flash	[31]
little to no severity	95 (17-24)	22 (17-33)	188 (17-44)	112 (17-26)	22 (17-27)	147 (17-32)	188 (17-44)	188 (17-44)	137 (17-29)
Moderate severity	89 (25-43)	162 (18-43)	51 (45-60)	76 (27-44)	166 (34-44)	104 (33-64)	71 (45-67)	51 (45-60)	51 (30-44)
High severity	91 (44-85)	91 (44-85)	36 (61-85)	87 (45-85)	87 (45-85)	24 (65-85)	16 (67-85)	36 (61-85)	87 (45-85)

In Table 10, the performance of models for PTSD severity classification across both text and audio modalities is assessed based on two distributions: Intervals Based on LLMs and Reference Intervals. These distributions represent different ways of mapping PTSD severity scores into categorical labels for evaluation.

The Intervals Based on LLMs are derived from each model's interpretation of the severity ranges based on their internal mapping strategies. These intervals reflect how the models perceive and classify the severity levels from the provided prompt. The ranges vary slightly across models as they adapt the scoring thresholds independently.

The Reference Intervals are based on the predefined mappings outlined in Table 9, which adheres to the PCL-C (PTSD) severity scale discussed in the *Data Preprocessing* section 3.1.1. The interval details, including the mapping of severity labels, are also explicitly presented in Table 9. These intervals serve as a standard benchmark, ensuring consistent and structured severity categorization for all models. In Table 10, the performance of models for PTSD severity classification across modalities is assessed based on two distributions: the LLM's Distribution and the Reference Distribution.

For the Intervals based on LLMs, the Phi-3.5-mini model showcased superior performance in the text modality with the highest balanced accuracy of 69.9% on Prompt 2. GPT-4o mini also demonstrated significant consistency, achieving the second-highest BA of 69.2% and the highest F1 score of 0.73 on Prompt 1, underscoring its robust capabilities.

In the audio modality, performance levels were generally closer to the average observed across all models. The Gemini 1.5 Pro was notable for achieving a high F1 score of 0.69 on Prompt 2, illustrating its effectiveness in handling audio inputs for PTSD severity assessment.

Table 10: PTSD Severity model performances on both text and audio modalities and their combination. Details about the specific prompts used can be found in Figure 6. The undervalued bold values are the best score for the specific prompt.

		Intervals based on LLMs						Intervals based on [31]					
		Prompt 1			Prompt 2			Prompt 1			Prompt 2		
	Model	BA	F1	MAE	BA	F1	MAE	BA	F1	MAE	BA	F1	MAE
Text(T)	Llama 3 70B	62.4%	0.63	0.40	58.1%	0.56	0.47	**60%**	**0.63**	**0.44**	53.4%	0.45	0.60
	Gemma 2 9B	50%	0.56	0.42	48.9%	0.57	0.39	47%	0.37	0.65	45.7%	0.33	0.67
	GPT-4o mini	**69.2%**	**0.73**	0.31	68.3%	0.68	0.37	51.2%	0.60	0.46	**54.6%**	**0.61**	**0.44**
	Mistral NeMo	58.3%	0.59	0.45	56.5%	0.58	0.47	54.6%	0.57	0.50	51.8%	0.57	0.52
	Phi-3.5-MoE	52.4%	0.56	0.50	49.7%	0.53	0.51	51.2%	0.59	0.49	49.2%	0.57	0.65
	Phi-3.5-mini	66.9%	0.60	0.42	**69.9%**	0.68	0.33	53.2%	0.61	0.47	53.9%	0.60	0.47
	Gemini 1.5 Pro	57.2%	0.69	0.33	55.6%	0.63	0.40	45.5%	0.53	0.53	49.7%	0.51	0.52
	Gemini 1.5 Flash	62.8%	0.58	0.47	57%	0.48	0.55	51.2%	0.52	0.53	48.4%	0.45	0.57
Audio(A)	Gemini 1.5 Pro	55.2%	0.72	**0.28**	55.1%	**0.69**	**0.31**	40.5%	0.50	0.60	45%	0.45	0.55
	Gemini 1.5 Flash	61.4%	0.56	0.49	60.9%	0.56	0.56	51.3%	0.56	0.56	53.2%	0.50	0.50
A+T	Gemini 1.5 Pro	54%	0.68	0.35	59.6%	0.67	0.37	44%	0.49	0.56	47.5%	0.52	0.53
	Gemini 1.5 Flash	64%	0.63	0.41	55.5%	0.50	0.54	50.7%	0.54	0.50	50%	0.47	0.55

The Mean Absolute Error (MAE) analysis highlights Gemini 1.5 Pro's precision in the audio modality, achieving the lowest MAE scores, notably a 0.28 in the LLM intervals. Also, for the reference intervals GPT-4o mini and Llama3-70 achieved the lowest MAE with 0.44. This precision demonstrates its effectiveness in closely approximating actual PTSD severity levels. Conversely, the audio and combined modalities showed higher MAE values, suggesting less accuracy in these approaches.

Combining text and audio modalities did not consistently enhance performance, as seen from the similar or slightly lower results compared to single modalities in some prompts. This observation suggests that while multimodal approaches hold promise, they do not always guarantee superior performance over single-modality analyses in PTSD severity classification.

Overall, text modality models generally outperformed those in the audio modality, indicating that text-based approaches to classifying PTSD severity are more effective.

5.5 Multi-Label Classification Results

In this task, we aimed to evaluate the models' ability to predict the presence of zero or more disorders (depression, PTSD, or both) mentioned in 4.2.3. To achieve this, we combined the ground truth labels for the binary classification of both disorders and allowed the models to predict if the interviewee exhibited any of these conditions. Once the predictions were made, we calculated the balanced accuracy (BA) and F1 score for each disorder separately, based on the predicted labels for both depression and PTSD.

We also calculated the Balanced Accuracy (BA) and F1 score when treating the problem as a multiclass classification task with four classes: Depression, PTSD, both, or None. Additionally, for the multi-label classification task, we calculated BA and F1 scores based on partial correctness. For example, if the true labels were Depression and PTSD, but the model predicted only PTSD, it was credited as 50% correct since it partially matched the ground truth. This approach allows us to evaluate performance across different classification settings.

This framework allowed us to compare the models' ability to handle complex cases where more than one condition might be present, providing valuable insights into their predictive accuracy for mental health diagnostics.

Table 11: Multi label model performance on both Text and Audio modalities and their combination. Details about the specific prompts used can be found in Figure 7. The underlined bold values are the best score for the specific prompt.

		Depression				PTSD				Multiclass				Multi-Label			
		Prompt 1		Prompt 2		Prompt 1		Prompt 2		Prompt 1		Prompt 2		Prompt 1		Prompt 2	
	Model	BA	F1	BA	F1	BA	F1	BA	F1	BA	F1	BA	F1	BA	F1	BA	F1
Text (T)	Llama 3 70B	69.9%	0.59	72.3%	0.65	**74%**	**0.75**	68.1%	0.73	43.2%	0.53	40.2%	0.55	**72%**	0.83	70%	0.74
	Gemma 2 9B	53.2%	0.28	64.9%	0.51	60.6%	0.45	72.3%	0.75	29.5%	0.22	37.9%	0.45	57%	0.88	68%	**0.80**
	GPT-4o mini	68.7%	0.56	72.5%	0.67	66.9%	0.67	**74.2%**	**0.77**	39.9%	0.49	45%	0.60	68%	0.82	73%	0.77
	Mistral NeMo	67.2%	0.59	69.4%	0.67	70.4%	**0.75**	63.9%	0.72	37.7%	0.51	38.7%	0.55	68.8%	0.75	66.5%	0.63
	Phi-3.5-MoE	66.9%	0.56	68.6%	0.60	66.8%	0.70	62.4%	0.68	38.2%	0.48	37.9%	0.50	67%	0.77	65%	0.70
	Phi-3.5-mini	52.3%	0.21	64.2%	0.52	52.6%	0.22	66.4%	0.66	26.3%	0.15	35.8%	0.41	53%	**0.91**	65%	**0.80**
	Gemini 1.5 Pro	66.1%	0.58	72.3%	**0.74**	67.6%	0.70	66.7%	0.73	41.2%	0.45	45%	**0.62**	67%	0.77	69%	0.60
	Gemini 1.5 Flash	64.4%	0.50	**74.3%**	0.70	72%	**0.75**	69.1%	0.75	41.2%	0.45	45.4%	0.58	68%	0.81	**72%**	0.73
Audio(A)	Gemini 1.5 Pro	71.9%	**0.72**	67.3%	0.70	67.4%	0.73	67.8%	0.74	42.5%	**0.58**	39.2%	0.57	70%	0.67	68%	0.60
	Gemini 1.5 Flash	71.7%	0.70	71.2%	0.69	68.1%	0.73	63.1%	0.67	**44.8%**	0.57	42.4%	0.52	70%	0.71	67%	0.72
A + T	Gemini 1.5 Pro	72.3%	0.71	74%	**0.74**	67.1%	0.72	69.8%	0.72	43%	0.56	**45.9%**	0.52	70%	0.72	72%	0.68
	Gemini 1.5 Flash	**72.4%**	0.64	72%	0.69	70.6%	0.74	65.2%	0.70	46.5%	0.56	44.1%	0.55	**72%**	0.79	69%	0.73

In the depression detection task summarized in Table 11, Gemini 1.5 Flash achieved the best results in the text modality with a BA of 74.3% and an F1 score of 0.70 on Prompt 2. For PTSD detection, Llama 3 70Bled with a BA of 74% and an F1 score of 0.75 on Prompt 1. In the multiclass and multi-label tasks, Gemini 1.5 Flash excelled again, achieving a BA of 45.4% in the multiclass task and an F1 score of 0.81 in the multi-label task.

Gemini 1.5 Flash emerged as the most consistent model when applying the text modality, consistently achieving high scores across multiple tasks and prompts, validating its reliability in this setup.

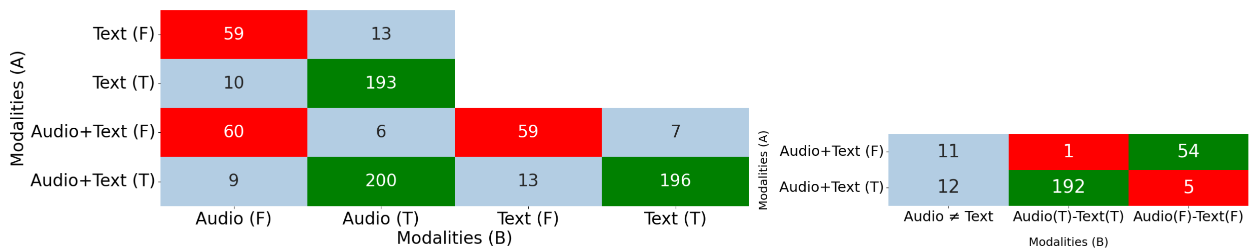
When both modalities are combined, Gemini 1.5 Flash again shows enhanced performance, achieving a BA of 72.4% on Prompt 1 and an F1 score of 0.79 on the multi-label task under this configuration. This multimodal approach, which leverages both textual and vocal data, appears to provide a more comprehensive analysis, potentially increasing the accuracy and reliability of mental health assessments.

Overall, models in the text modality generally performed better in PTSD detection, while audio-based models showed better performance in depression detection. However, the combined modalities often outperformed individual text or audio inputs, suggesting that integrating these approaches may offer the most effective means for mental health diagnostics.

5.6 Disagreement

The co-occurrence matrix in Figures 9a and 10a panel visualizes the instances of correct and incorrect predictions by each modality for binary depression classification using the Gemini 1.5 Flash model on Prompt 3 and 2 respectively. The colors represent different outcomes of these predictions: Red colored cells indicate instances where both modalities incorrectly predicted the sample. Green colored cells highlight cases where both modalities correctly identified the sample. Blue colored cells denote instances where one modality outperformed the other by correctly predicting a sample that the other modality missed.

Figures 9b and 10b present the analysis of the Audio+Text modality's performance in resolving disagreements between the individual Audio and Text modalities. The colors in the heatmap signify different outcomes of predictions across these modalities: Green colored cells indicate instances where all three modalities (Audio, Text, and Audio+Text) either correctly or incorrectly predicted the sample. Red colored cells show the number of samples where the combined modality predicted differently than both the Audio and Text modalities when they were either correct or incorrect. Blue colored cells highlight cases where the combined modality resolved the disagreement between the Audio and Text modalities, either by correctly predicting what one modality missed or incorrectly predicting what one modality got right.



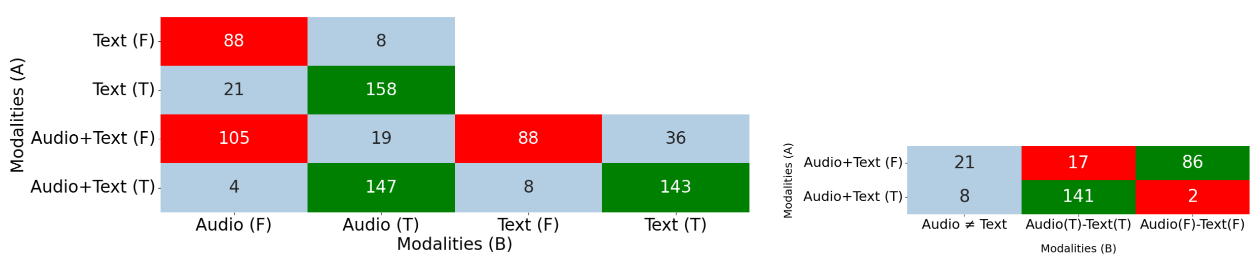
(a) (b)

Figure 9: (a) Co-occurrence matrix illustrating where the Audio and Text modalities and their combination independently predict binary depression outcomes correctly or incorrectly under zero-shot inference conditions with the Gemini 1.5 Flash model using Prompt 3. (b) Visualization of the combined (Audio + Text) modality's predictions relative to the individual Audio and Text modalities, highlighting scenarios of agreement, disagreement, and how the combined modality addresses these differences under the same conditions.

For figure 9a, the Modal Superiority Score (MSS) was applied to evaluate the performance differences among the modalities. The MSS metric, as detailed in Equation 4.8, measures the effectiveness of one modality over another when there is disagreement. The findings indicated that, compared to the separate modalities, the Audio+Text combined modality performed better. The advantages of modality integration were highlighted by MSS values, which demonstrated a 20% superiority of Audio+Text over Audio alone and a 30% superiority over Text alone. Audio displayed a slight advantage over Text, with an MSS value of -13.04%, indicating that Audio was marginally more successful at accurately predicting results than Text.

To further illustrate how the combined modality improved performance, we applied the Disagreement Resolvement Score (DRS) and the Modal Superiority Score (MSS) metrics on figure 9b. The DRS, detailed in Equation 4.8, specifically addresses the combined modality's ability to resolve conflicts between the two other modalities. It calculated a small positive value of 4.35%, indicating a slight net benefit in the combined modality's resolution of disagreements. This suggests that the combined Audio+Text modality was able to correctly resolve just over half of the instances where Audio and Text disagreed.

Additionally, the MSS between the combined modality and the joint prediction agreement of the Audio and Text modalities showed a substantial positive score of 66.67%. This result underscores a significant enhancement in performance by the combined modality over the consensus of the individual modalities, reflecting a superior capability to effectively harness the strengths of both modalities.



(a) (b)

Figure 10: (a) Co-occurrence matrix illustrating where the Audio and Text modalities and their combination independently predict binary depression outcomes correctly or incorrectly under zero-shot inference conditions with the Gemini 1.5 Flash model using Prompt 2. (b) Visualization of the combined (Audio + Text) modality's predictions relative to the individual Audio and Text modalities, highlighting scenarios of agreement, disagreement, and how the combined modality addresses these differences under the same conditions.

In the assessment of modality performance using the Modal Superiority Score (MSS) for figure 10, we quantified the efficacy of individual and combined modalities. The findings are summarized as follows:

The MSS value calculated for Text against Audio was 44.83%. This positive value demonstrates that Text outperformed Audio. The MSS values indicate that the performance declined when Audio and Text were combined, yielding -65.22% against Audio alone and -63.64% against Text alone. These negative scores highlight a decrease in performance when the modalities are combined, indicating that the integration of Text with Audio may have introduced complexities that reduced the effectiveness of the prediction capability.

Using the Disagreement Resolvement Score (DRS), we evaluated the combined Audio+Text modality's ability to resolve conflicts between the Audio and Text modalities, resulting in a significant negative score of approximately -44.83%. This significant negative score indicates that integrating Audio and Text modalities in this experiment negatively impacted the model's ability to resolve disagreements. This finding underscores the complexities and potential challenges associated with modality integration, demonstrating that combining different modalities does not always enhance predictive accuracy.

5.7 Few-shot Prompts Results

Table 12 presents the few-shot (FS) performance scores for various tasks, where each model was subjected to tests using specifically designed prompts, as detailed in the Few-Shot Prompts section 4.6.4. The goal was to determine whether the few-shot experiments would enhance or reduce the model's performance. Notably, for these evaluations, we utilized the entire dataset of 275 samples, including those samples that were used in the prompts. This approach was adopted because, frequently, the models failed to correctly predict at least one of these samples, thus making it crucial to include them in the performance assessment. The table lists the raw few-shot performance scores alongside the change from the zero-shot (ZS) evaluation, denoted in parentheses.

The PTSD binary task saw the largest positive shift, with Gemini 1.5 Flash (TEXT) improving by +11% in BA and +0.13 in F1. In the Depression binary task, GPT-4o mini also showed major gains, especially with a +6.6% BA improvement on Prompt 1.

Gemini 1.5 Flash (TEXT) was the most consistent, showing significant positive gains across nearly all tasks. Phi-3.5-MoE is the least consistent as it has the highest number of negative changes, including a substantial drop of -13.6% in BA for Depression severity on Prompt 2, marking it as the least consistent.

5.8 Comparative Evaluation

In this section, we undertake a comparative analysis of our model's performance against other established methodologies, as detailed in table 13. A thorough literature search revealed a scarcity of studies that evaluated models across the entire E-DAIC or DAIC-WOZ datasets. To facilitate a robust comparison, we selected our best-performing model and the corresponding prompt. We evaluated this model in the development sets of both the E-DAIC, which contains 56 samples, and DAIC-WOZ, which consists of 35 samples. This approach allows us to directly compare our results with those obtained from other models tested under similar conditions. For our evaluation metric, we used the binary score F1, considering the task's focus on binary depression classification, a prevalent method in existing research. We prioritized scores that were consistently high across both datasets to ensure a balanced evaluation, avoiding instances where a model performed exceptionally well on one dataset but poorly on another. This method provides a clear benchmark to assess the relative effectiveness of our proposed approaches.

Table 14 offers a detailed comparative view of PTSD diagnosis performance across various methodologies, as shown below. Our zero-shot approach with GPT-4o mini is not only competitive but also surpasses the results reported by other significant studies. Notably, the CALLM model developed by Wu et al., which utilizes fine-tuning, shows a slight improvement over our approach with a balanced accuracy of 77% and an F1 score of 0.70. However, our zero-shot model still demonstrates a robust capability comparable to a finely tuned system like CALLM, outperforming other established benchmarks in the field. This underscores the potential of non-fine-tuned models to achieve high effectiveness, highlighting a significant advance in using LLMs for PTSD classification without extensive training on specific datasets.

Table 12: FS model performance evaluation, Values are: FS score (Change from ZS score). The best score for each prompt is highlighted in bold and underline.

Models		Gemini Flash 1.5 (Text)		GPT-4o mini (Text)		Phi-3.5-MoE (Text)		Gemini Flash 1.5 (Audio)	
Task	Promopt	BA	F1	BA	F1	BA	F1	BA	F1
Depression Binary	P1	**76.1% (+2.1%)**	0.65 (+0.02)	71.4% (+6.6%)	0.60 (+0.1)	71% (-3.5%)	0.60 (-0.04)	76% (+1.1%)	**0.66 (0)**
	P2	**77.4% (+4.7%)**	**0.67 (+0.05)**	65.5% (-9%)	0.56 (-0.07)	74.7% (+1.7%)	0.64 (+0.02)	71.4% (+0.9%)	0.60 (0)
	P3	**77.7% (4.1%)**	**0.68 (+0.05)**	74.2% (+9%)	0.64 (+0.14)	76.9% (+0.9%)	0.67 (0)	75.4% (+0.7%)	0.65 (+0)
PTSD Binary	P1	**80% (+11%)**	**0.71 (+0.13)**	78.2% (+1.6%)	0.70 (+0.03)	76.8% (+5.6%)	0.68 (-0.08)	75.3% (+2.4%)	0.65 (+0.03)
	P2	**77.5% (+7%)**	**0.68 (+0.7)**	76% (-1%)	0.67 (-0.01)	72% (+2.8%)	0.61 (+0.05)	71.9% (+4.7%)	0.62 (+0.05)
Depression Severity	P1	38.4% (+3.1%)	**0.42 (+0.07)**	34% (-10%)	0.32 (-0.18)	**38.9% (-10%)**	0.33 (-0.16)	34% (-1%)	0.42 (0)
	P2	34.7% (0)	0.44 (0.14)	28.1% (-13.6%)	0.12 (-0.38)	**42.9% (+10%)**	**0.46 (+0.04)**	41.8% (+6.6%)	0.41 (+0.06)
PTSD Severity	P1	LLM's dis.: 61.2% (-1.6%) REF dis.: 49.8% (-1.4%)	LLM's dis.: 0.51 (-0.07) REF dis.: 0.53 (0)	LLM's dis.: 53.4% (-15.8%) REF dis.: 53.5% (+2.3%)	LLM's dis.: 0.30 (-0.29) REF dis.: 0.50 (-0.10)	LLM's dis.: **64.7% (+11.5%)** REF dis.: **54.4% (+2.7%)**	LLM's dis.: 0.62 (+0.23) REF dis.: 0.49 (-0.11)	LLM's dis.: 64.6% (+3.2%) REF dis.: 47.8% (-3.5%)	LLM's dis.: **0.70 (+0.14)** REF dis.: **0.54 (+0.06)**
	P2	LLM's dis.: 55.4% (-1.6%) REF dis.: 47.8% (-0.6%)	LLM's dis.: 0.58 (+ 0.1) REF dis.: 0.48 (+ 0.03)	LLM's dis.: 51.1% (-17.2%) REF dis.: 46.6% (-8%)	LLM's dis.: 0.25 (-0.43) REF dis.: 0.36 (-0.25)	LLM's dis.: 59.9% (+11.3%) REF dis.: **57.7% (+8.9%)**	LLM's dis.: 0.57 (+ 0.21) REF dis.: **0.59 (+ 0.02)**	LLM's dis.: **65% (+4.1%)** REF dis.: 54.7% (+1.5%)	LLM's dis.: **0.62 (+0.15)** REF dis.: 0.58 (+0.04)
Multi Label	P1	71% (+3%)	0.75 (-0.06)	68% (0)	**0.85 (+0.03)**	**73% (+8%)**	0.77 (0)	69% (-1%)	0.68 (-0.03)
	P2	70% (-2%)	0.75 (+0.02)	66% (-7%)	**0.86 (+0.09)**	**72% (+7%)**	0.83 (+0.14)	68% (+1%)	0.69 (-0.03)

Table 13: F1 score for binary depression classification against other research results on E-DAIC and DAIC-WOZ development set.

Reference	E-DAIC	DAIC-WOZ	Methods
Gemini 1.5 Pro (prompt 2) (Ours)	0.56	0.69	ZS inference using the raw interview transcriptions
Gemini 1.5 Flash (A) (prompt 3) (Ours)	0.56	0.77	ZS inference using the raw audio interviews
Gemini 1.5 Pro (A+T) (prompt 3) (Ours)	0.60	0.71	ZS inference using both raw audio and transcription
(Villatoro-Tello et al.)[42]	0.59	0.56	Fine-tuning BERT on a task-specific dataset derived from the DAIC-WOZ interviews to predict depression from linguistic cues
BERT (Senn et al.)[43]	-	0.60	Compared three individual BERT models (BERT, RoBERTa, DistilBERT) on transcriptions and four ensembles with varying architectures.
BERT (Danner et al.)[44]	-	0.64	They trained and fine-tuned BERT for on the transcriptions. They performed ZS inference with GPT-3.5 and chatGPT-4
GPT-3.5 (Danner et al.)[44]	-	0.78	
ChatGPT-4 (Danner et al.)[44]	-	0.61	
GPT-4 (Hadzic et al.)[20]	-	0.71	They conducted ZS inference with GPT-4 on the transcriptions
GPT-3.5-turbo P2+SMMR (Guo et al.)[45]	-	0.76	GPT-3.5-turbo and GPT-4-turbo were evaluated using ZS with a method called Stacked Multi-Model Reasoning
GPT-4-turbo P2+SMMR (Guo et al.)[45]	-	0.79	

Table 14: Binary PTSD classification against other research results on E-DAIC development and test set.

Reference	BA	F1	Methods
GPT-4o mini (prompt 2) (Ours)	76%	0.68	ZS inference using the raw interview transcriptions
Flores et al. [46]	70%	0.69	Utilized bidirectional GRU model with self-attention. Designed for multi-task learning, using temporal facial features.
Galatzer-Levy et al. [47]	74%	0.64	Med-PaLM 2 utilized zero-shot inference, processing clinical interview transcriptions to evaluate psychiatric conditions.
Wu et al. (CALLM) [48]	77%	0.70	Fine-tuned a pre-trained DistilBERT model utilizing the expanded training dataset provided by their augmentation process.

5.9 Limitations

Few Shot Learning: In our few-shot experiments targeting the audio modality, we opted not to use multiple audio samples for evaluation due to observed inconsistencies in the model's processing capabilities. Initially, the model demonstrated substantial variability when tasked with handling three audio files simultaneously within a single prompt, often failing to transcribe or comprehend the full content accurately. This inconsistency prompted us to exclude direct audio samples from few-shot testing. To further explore these challenges, we investigated the model's ability to summarize transcribed texts derived from audio inputs, which revealed further issues with accuracy and consistency. These findings underline the need for enhanced model refinement to ensure reliable handling and understanding of complex audio data in future implementations.

Comparative Evaluation: Additionally, when comparing our results with existing studies, to our knowledge we have not found any work concerning PTSD severity and multi-class classification tasks on the E-DAIC dataset, which are largely unexplored. For depression severity classification, existing studies often employ different labeling systems, such as using all numbers within the range (e.g., 0-24) as labels, which differs from our methodology. This diversity in approaches makes direct comparisons challenging.

Furthermore, it's important to highlight that some relevant studies were not included in our comparative analysis. This exclusion is due to differences in datasets, training methods, models, or the metrics reported, which were not directly comparable to our Balanced Accuracy (BA) and F1 scores.

LLM Fine-tuning: Fine-tuning is a process in which a pretrained model is further trained on task-specific data to improve its performance. Although this approach has the potential to specialize models for specific domains, it is not without limitations. In the following, we detail the challenges encountered during the fine-tuning process for the binary depression detection task.

1. The E-DAIC dataset used for fine-tuning is both small and imbalanced, with: 86 depression samples, 189 Non-depression samples. This imbalance makes it challenging for the model to learn effectively. With limited "depression" examples, the model struggles to generalize, ultimately failing to surpass the performance achieved by zero-shot (ZS) inference on unprocessed data, where no task-specific fine-tuning was conducted.
2. Large language models (LLMs) do not natively support fine-tuning with audio input data. Since detecting depression often relies heavily on acoustic cues, the inability to fine-tune using audio severely limits the model's multimodal capabilities. Consequently, comparisons between audio and text modalities become inherently skewed, as the model can be specialized for text through fine-tuning but must remain at a less adapted, effectively zero-shot state for audio data.
3. Fine-tuning LLMs requires significant computational resources, expertise in hyperparameter tuning, and extensive trial-and-error. Given the constrained dataset and the nuanced nature of depression detection, these overheads do not guarantee performance improvements, making fine-tuning both resource-intensive and, in this case, yield-limited.

6 Conclusion

This study has systematically evaluated the application of Large Language Models (LLMs) in mental health diagnostics, focusing on depression and PTSD using the E-DAIC dataset. We conducted experiments across a range of tasks to evaluate the comparative effectiveness of text and audio modalities and to investigate if a multimodal approach could enhance diagnostic accuracy. Our research introduced custom metrics, the Modal Superiority Score (MSS) and the Disagreement Resolvement Score (DRS), specifically designed to measure how the integration of modalities impacts model performance. The analysis consistently demonstrated that text modality excelled in most tasks, outperforming audio modality. However, the integration of text and audio modalities often led to improved outcomes, suggesting that combining these modalities could leverage the complementary strengths of each. The MSS and DRS metrics provided insights into the extent of these improvements, highlighting scenarios where multimodal approaches showed promise over single modalities.

While these findings underscore the potential of multimodal approaches for enhancing the performance of LLMs, it is important to emphasize that this study is purely computational. The clinical utility of these methods remains untested and must be validated through rigorous controlled clinical studies to determine their effectiveness and reliability in real-world settings. Moreover, these results are derived from a single dataset (E-DAIC), which is relatively small and may not generalize to broader, more diverse populations or clinical environments. Variability in language, speech patterns, and diagnostic criteria across different datasets or real-world conditions could significantly impact model performance.

7 Acknowledgement

We recognize and express our gratitude to the authors of [49] and [10] for providing the Extended Distress Analysis Interview Corpus(E-DAIC) dataset and the Distress Analysis Interview Corpus(DAIC-WOZ) dataset

References

[1] Jetli Chung and Jason Teo. Mental health prediction using machine learning: Taxonomy, applications, and challenges. *Applied Computational Intelligence and Soft Computing*, 2022:1–19, 01 2022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models, 2024.

[4] Nick Obradovich, Sahib S. Khalsa, Waqas U. Khan, Jina Suh, Roy H. Perlis, Olusola Ajilore, and Martin P. Paulus. Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, 2(1):8, 2024.

[5] Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle H. Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer, and Johannes C. Eichstaedt. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. *npj Mental Health Research*, 3(1):12, April 2024.

[6] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models, 2024.

[7] Frank Palma Gomez, Ramon Sanabria, Yun hsuan Sung, Daniel Cer, Siddharth Dalmia, and Gustavo Hernandez Abrego. Transforming llms into cross-modal and cross-lingual retrieval systems, 2024.

[8] Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities, 2024.

[9] Abdelrahman Hanafi, Mohammed Saad, Noureldin Zahran, Radwa J. Hanafy, and Mohammed E. Fouda. A comprehensive evaluation of large language models on mental illnesses, 2024.

[10] J. Gratch, Ron Artstein, Gale M. Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David R. Traum, Albert A. Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *International Conference on Language Resources and Evaluation*, 2014.

[11] Jinhan Wang, Vijay Ravi, Jonathan Flint, and Abeer Alwan. Speechformer-ctc: Sequential modeling of depression detection with speech temporal classification. *Speech Communication*, 163:103106, 2024.

[12] Georgios Ioannides, Adrian Kieback, Aman Chadha, and Aaron Elkins. Density adaptive attention-based speech network: Enhancing feature understanding for mental health disorders, 2024.

[13] Santosh V. Patapati. Integrating large language models into a tri-modal architecture for automated depression classification, 2024.

[14] Rohan Kumar Gupta and Rohit Sinha. Deep multi-task learning based detection of correlated mental disorders using audio modality. *Computer Speech & Language*, 89:101710, 2025.

[15] Wen Wu, Chao Zhang, and Philip C. Woodland. Confidence estimation for automatic detection of depression and alzheimer's disease based on clinical interviews. In *Interspeech 2024*, interspeech 2024. ISCA, 2024.

[16] Avinash Anand, Chayan Tank, Sarthak Pol, Vinayak Katoch, Shaina Mehta, and Rajiv Ratn Shah. Depression detection and analysis using large language models on textual and audio-visual modalities, 2024.

[17] Xiangsheng Huang, Fang Wang, Yuan Gao, Yilong Liao, Wenjing Zhang, Li Zhang, and Zhenrong Xu. Depression recognition using voice-based pre-training model. *Scientific Reports*, 14(1):12734, 2024.

[18] Xu Zhang, Xiangcheng Zhang, Weisi Chen, Chenlong Li, and Chengyuan Yu. Improving speech depression detection using transfer learning with wav2vec 2.0 in low-resource environments. *Scientific Reports*, 14(1):9543, 2024.

[19] Sergio Burdisso, Ernesto Reyes-Ramírez, Esaú Villatoro-Tello, Fernando Sánchez-Vega, Pastor López-Monroy, and Petr Motlicek. Daic-woz: On the validity of using the therapist's prompts in automatic depression detection from clinical interviews, 2024.

[20] Bakir Hadzic, Parvez Mohammed, Michael Danner, Julia Ohse, Yihong Zhang, Youssef Shiban, and Matthias Rätsch. Enhancing early depression detection with ai: a comparative use of nlp models. *SICE Journal of Control, Measurement, and System Integration*, 17(1):135–143, 2024.

[21] Giuliano Lorenzoni, Cristina Tavares, Nathalia Nascimento, Paulo Alencar, and Donald Cowan. Assessing ml classification algorithms and nlp techniques for depression detection: An experimental case study, 2024.

[22] Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection, 2024.

[23] Shanliang Yang, Lichao Cui, Lei Wang, Tao Wang, and Jiebing You. Enhancing multimodal depression diagnosis through representation learning and knowledge transfer. *Heliyon*, 10(4):e25959, 2024.

[24] Clinton Lau, Xiaodan Zhu, and Wai-Yip Chan. Automatic depression severity assessment with deep learning using parameter-efficient tuning. *Frontiers in Psychiatry*, 14, 2023.

[25] Ping-Cheng Wei, Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. Multi-modal depression estimation based on sub-attentional fusion, 2022.

[26] Nasser Ghadiri, Rasoul Samani, and Fahime Shahrokh. Integration of text and graph-based features for detecting mental health disorders from voice, 2022.

[27] Heinrich Dinkel, Mengyue Wu, and Kai Yu. Text-based depression detection on sparse data, 2020.

[28] D. Xezonaki, G. Paraskevopoulos, A. Potamianos, and S. Narayanan. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews, 2020.

[29] Evgeny Stepanov, Stephane Lathuiliere, Shammur Absar Chowdhury, Arindam Ghosh, Radu-Laurentiu Vieriu, Nicu Sebe, and Giuseppe Riccardi. Depression severity estimation from multiple modalities, 2017.

[30] Kurt Kroenke, Tara Strine, Robert Spitzer, Janet Williams, Joyce Berry, and Ali Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114:163–73, 09 2008.

[31] Andrea Alejandra García-Valdez, Israel Román-Godínez, Ricardo Antonio Salido-Ruiz, and Sulema Torres-Ramos. Sex-based speech pattern recognition for post-traumatic stress disorder. In José de Jesús Agustín Flores Cuautle, Balam Benítez-Mata, Ricardo Antonio Salido-Ruiz, Gustavo Adolfo Alonso-Silverio, Guadalupe Dorantes-Méndez, Esmeralda Zúñiga-Aguilar, Hugo A. Vélez-Pérez, Edgar Del Hierro-Gutiérrez, and Aldo Rodrigo Mejía-Rodríguez, editors, *XLVI Mexican Conference on Biomedical Engineering*, pages 192–200, Cham, 2024. Springer Nature Switzerland.

[32] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[33] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[34] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[35] Mistral AI team. Mistral-nemo. `https://mistral.ai/news/mistral-nemo//`, 2024.

[36] OpenAI. Gpt-4o mini. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2024.

[37] Microsoft. Discover the new multi-lingual high-quality phi 3.5 slms, 2024. Accessed: 2024-10-07.

[38] Google AI. Gemini api documentation, 2024. Accessed: October 7, 2024.

[39] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024.

[40] AIBase. Aibase tool information - tool id 32286, 2024. Accessed: October 7, 2024.

[41] Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024.

[42] Esaú Villatoro-Tello, Gabriela Ramírez de-la Rosa, Daniel Gática-Pérez, Mathew Magimai-Doss, and Héctor Jiménez-Salazar. Approximating the mental lexicon from clinical interviews as a support tool for depression detection. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021.

[43] Saskia Senn, ML Tlachac, Ricardo Flores, and Elke Rundensteiner. Ensembles of bert for depression classification. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4691–4694, 2022.

[44] Michael Danner, Bakir Hadzic, Sophie Gerhardt, Simon Ludwig, Irem Uslu, Peng Shao, Thomas Weber, Youssef Shiban, and Matthias Ratsch. Advancing mental health diagnostics: Gpt-based method for depression detection. In *2023 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, pages 1290–1296, 2023.

[45] Qiming Guo, Jinwen Tang, Wenbo Sun, Haoteng Tang, Yi Shang, and Wenlu Wang. Soullmate: An application enhancing diverse mental health support with adaptive llms, prompt engineering, and rag techniques, 2024.

[46] Ricardo Flores, Avantika Shrestha, Ml Tlachac, and Elke A. Rundensteiner. Multi-task learning using facial features for mental health screening. In *2023 IEEE International Conference on Big Data (BigData)*, pages 4881–4890, 2023.

[47] Isaac R. Galatzer-Levy, Daniel McDuff, Vivek Natarajan, Alan Karthikesalingam, and Matteo Malgaroli. The capability of large language models to measure psychiatric functioning, 2023.

[48] Yuqi Wu, Kaining Mao, Yanbo Zhang, and Jie Chen. Callm: Enhancing clinical interview analysis through data augmentation with large language models. *IEEE Journal of Biomedical and Health Informatics*, pages 1–14, 2024.

[49] Fabien Ringeval, Björn Schuller, Michel Valstar, NIcholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic. Avec 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition, 2019.