DepSy: A Dataset and Benchmark for *Dep*ression *Sy*mptoms Detection with Hierarchical Transformers and Fine-Tuned LLMs

Anonymous ACL submission

Abstract

Early detection of symptoms of depression can help minimise its impact on people suffering from depression. Social media, where users often share emotions and life experiences, offers a valuable resource for NLP-driven mental health research. We posit that mining social media posts enables researchers to identify clinically significant depressive symptoms. This paper introduces: a) the DepSy dataset, a novel resource annotated by psychologists for depressive symptoms, containing over 40k posts; and b) the DepSy model, a fine-tuned model trained to identify and extract depressive symptoms. We conducted comparative experiments between BERT-based models and large language models (LLMs) for symptom extraction. Our results show that both BERT-based models and LLMs demonstrated comparable performance, with BERT achieving the highest overall f-1 score of 0.522.

1 Introduction

Mental health issues are rising globally, with WHO estimating that one in four people will experience a condition in their lifetime (World Health Organization, 2001). These conditions significantly impact the quality of life and contribute to disability and high suicide rates (DSM5, 2013). Understanding symptoms is essential for advancing mental health research and developing effective models to support both individuals with mental health conditions and those with similar experiences.

Most NLP research on mental health monitoring has focused on electronic health records and diagnostic assessments (Kim et al., 2020; Pradier et al., 2021; Mesbah et al., 2021; de Oliveira et al., 2021; Ignashina et al., 2025). While these provide valuable insights, social media offers an alternative by capturing large-scale, real-time expressions of emotions and mental states. With billions of users sharing experiences online, social media data presents a rich resource for NLP-driven mental health research. However, identifying high-quality datasets for training models to monitor depressive symptoms remains a significant challenge (Gadzama et al., 2024).

According to recent surveys (Montejo-Ráez et al., 2024; Garg, 2023), most existing studies utilizing NLP on social media data have centred around detecting the presence or absence of depression and identifying shifts from depression to suicidal ideation (De Choudhury et al., 2016; Gong et al., 2019; Sawhney et al., 2020; Kour and Gupta, 2022; Baghdadi et al., 2022; Khafaga et al., 2023; Adarsh et al., 2023). However, to date, little attention has been given to examining the occurrence of depressive symptoms. This study aims to fill this identified gap by enhancing early detection and intervention strategies through improved datasets and model development. Our key contributions can be summarized as follows:

- DepSy Dataset: The first English dataset of depression symptoms in textual posts of users who self-reported being diagnosed with depression that is **fully annotated by psychologists**[1].
- Hier-DepSy: a BERT-based hierarchical model architecture for depression symptom classification from social media data.
- DepSyLlama: a fined-tuned LLM for identifying depression symptoms
- Empirical work comparing multiple predictive models (based on BERT, RoBERTa, MentalBERT, GPT, Llama 2, Llama 3, MentaLlama) built using our dataset for the task of classifying/extracting depression symptoms from posts.

[1]The DepSy dataset, DepSy model, and code will be made available upon paper acceptance

2 Related Work

2.1 Datasets for Depression Monitoring on Social Media

Several studies have developed social media datasets for depression analysis, emphasising the need for labelled data. Kabir et al. (2023) introduced a dataset categorising tweets as "non-depressed" or "depressed," with severity levels. However, reliance on symptom-related keywords may exclude relevant posts, and crowdworker annotations can lack contextual depth. The PRIMATE dataset (Gupta et al., 2022), based on PHQ-9 responses from Reddit, raises similar concerns. The lack of expertise among crowdworkers may lead to inaccuracies. Milintsevich et al. (2024) later re-annotated PRIMATE, finding errors and false positives, highlighting the need for more rigorous annotation processes to ensure high-quality mental health datasets. To address these limitations, our work involves constructing a large dataset fully annotated by expert psychologists, with the goal of achieving a high level of annotation agreement.

2.2 Monitoring Depression Symptoms through Social Media

Several studies have focused on monitoring depression through social media platforms such as X (Twitter) and Reddit (Aragon et al., 2023; Shah et al., 2020; Tavast et al., 2022; Zogan et al., 2022; Chiong et al., 2021a,b; Wang et al., 2022). These studies typically annotate posts using matching terms, which can result in data loss or inaccurate labelling, as the context in which these terms are used may often be sarcastic (Ezerceli and Dehkharghani, 2024; Pavlova and Berkers, 2022). This limitation highlights the challenges inherent in relying solely on keyword matching for sentiment analysis in sensitive mental health contexts. Several studies have adopted the PHQ-9 framework to detect depressive symptoms in social media. Early work by Mowery et al. (2016) identified three PHQ-9 symptoms using a two-stage classifier, laying groundwork for symptom-level annotation. Yazdavar et al. (2017) used a labelled LDA model on tweets from keyword-identified users to detect nine symptoms, but relied on non-expert annotations and showed performance disparities across symptoms. More recent work by Yadav et al. (2020) proposed a BERT-based multi-task model incorporating figurative language, though keyword-based data collection limited generalisability. Yadav et al. (2023) introduced RESTORE, a multimodal dataset of annotated memes, showing improvements with orthogonal constraints across modalities. To improve cross-domain robustness, Nguyen et al. (2022) grounded predictions in PHQ-9 descriptions, enhancing interpretability, though performance may be limited for contextually subtle symptoms. In a large-scale study, Liu et al. (2023) used Reddit data and RoBERTa to detect 13 expert-validated symptoms, achieving strong results on an external benchmark, though subreddit-based labelling may limit clinical validity. Other studies employed the Beck Depression Inventory (BDI) for symptom estimation, including DepressMind (Fernández-Iglesias et al., 2024), which uses sentence similarity, and a prompting-based method by Aragón et al. (2024) that maps user posts to BDI questions via ChatGPT. While promising, both rely on surface-level or model-generated inferences rather than clinical annotation. Accurate symptom detection requires modelling symptoms and involving domain experts in annotation. However, few existing works offer clinically annotated, large-scale, multi-symptom datasets with strong modelling baselines, leaving a gap that this study aims to address.

3 DepSy Dataset

Recognizing the critical importance of precise and expert-driven annotations in the study of mental health through natural language processing (NLP), we undertook a rigorous annotation process guided by a robust annotation scheme. We annotated a dataset originally collected by Alhamed et al. (2024b) of users who self-reported being clinically diagnosed with depression on X (formerly Twitter). We annotated the posts of users in the class "after" being diagnosed with depression. The annotation process followed the annotation scheme in (Alhamed et al., 2024a) that was designed for annotating posts for depressive symptom and severity, and where high annotation agreement was reported.

The depressive symptoms in this scheme were synthesised from well-established and validated depression assessment tools, including the PHQ-9 (Kroenke et al., 2001), BDI (Beck et al., 1961), and CES-D (Radloff, 1977) questionnaires; the final symptoms list is shown in Table 1. Experienced psychologists meticulously annotated each post in the dataset for the mentioned list of depression symptoms. This annotation process has resulted in the creation of the first and largest dataset fully an-

Depression Symptoms
Poor appetite or eating disturbances
Feeling down and depressed
Crying
Concentration problems
Feeling tired or having little energy
Feelings of failure
Sleep disturbances
Loss of interest
Self-blame and shame
Loneliness
Suicidal thoughts

Table 1: List of depression symptoms

notated by professional psychologists, comprising over 40,000 posts annotated for depression symptoms. Each post is annotated to indicate whether it contains no symptoms, one symptom, or multiple symptoms, with the corresponding symptom name(s) included in the annotation. This resource is intended to support further research and model development in the detection and analysis of depression within NLP applications[2].

3.1 Data Annotation

The annotation task was carried out by five psychologists (co-authors), each with at least three years of specialised experience in diagnosing depression and/or anxiety disorders. Their involvement in this task was entirely voluntary and driven by a shared commitment to advancing mental health research. The annotators completed consent forms, which are securely stored on the college's OneDrive servers, and were provided with information sheets and detailed annotation guidelines. They were asked to select all (if any) depressive symptoms that existed in a post from an existing list of symptoms. We used Labelstudio [3] as a labelling interface for all experiments in this work. Within Label Studio, we designed a custom labelling interface to meet the specific needs of our task, as none of the available templates offered a suitable match. To evaluate the consistency of our annotations, we utilised Cohen's kappa (κ), a widely recognised statistic for assessing inter-rater reliability on nominal data (Cohen, 1960). Cohen's kappa accounts for agreement occurring by chance, thus offering a more rigorous measure of concordance than simple percentage agreement. We obtained a pairwise kappa score of 0.67 across 10% of annotated posts. According to the interpretation scale proposed by Landis and Koch (1977), this score falls within the "substantial" agreement range. The 10% subset was selected to ensure robust and reliable results, exceeding the smaller subsets (20–100 posts) commonly used in similar studies (e.g.,Chancellor et al. (2021); Harrigian and Dredze (2022)).

[2]The dataset will be made available upon paper acceptance

[3]https://labelstud.io/

4 Data Analysis

This section presents an analysis of depression symptom prevalence within the corpus of social media posts. By examining the frequency distribution of symptoms, we aim to identify the most commonly occurring depressive symptoms that users are willing to expose online. This analysis enables us to gain deeper insights into the dynamic nature of depressive symptoms and potential trajectories of the condition.

4.1 Symptoms Co-morbidity

Number of Symptoms	Number of Posts
0 (No Symptoms)	37,335
1	2,080
2	407
3	78
4	9
5	1

Table 2: Distribution of symptom co-occurrence in DepSy dataset.

4.2 Depressive Symptoms Frequency

When we analysed the frequency distribution of depressive symptoms within the users' posts, feeling down or depressed emerged as the most prevalent symptom, followed by feeling tired or having little energy, and crying. Symptoms such as self-blame and suicidal ideation were reported to be the least frequent. Figure 1 illustrates the frequency distribution of symptoms in the posts.

We further analysed the distribution of depressive symptoms in the dataset to understand their prevalence and co-occurrence. As shown in Table 2, the dataset is highly imbalanced, with the majority of posts (37,335) containing no symptoms and only 2,575 posts labelled with one or more symptoms. Most of the symptom-labelled posts contain a single symptom, while posts with multiple symptoms are increasingly rare. This imbalance is expected to pose a challenge for model training.

Figure 2 shows the co-morbidity matrix of depressive symptoms, capturing how often symptom

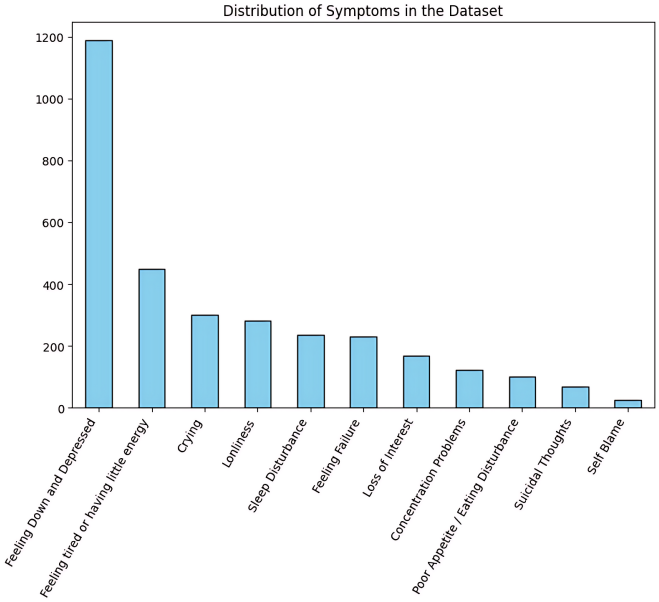

Figure 1: depressive symptom frequency in posts

pairs co-occur. Diagonal values reflect individual symptom frequency, while off-diagonal values indicate joint occurrences. "Feeling Down and Depressed" is the most frequent symptom and co-occurs frequently with others such as "Feeling Failure" and "Feeling tired or having little energy," suggesting a core cluster. In contrast, symptoms like "Self Blame" and "Suicidal Thoughts" appear less often and with weaker co-morbidity. These patterns highlight inter-symptom dependencies relevant for symptom-level classification.

5 Task Definition

Given a user-generated post P_i, the objective is to predict a binary label for each depressive symptom from the predefined set of 11 symptoms. The task is multi-label multi-class classification and is formulated as:

$$\begin{aligned} Y_i &= f(P_i; \theta), \\ Y_i &= \{y_{i,1}, y_{i,2}, \ldots, y_{i,11}\}, \quad y_{i,j} \in \{0, 1\} \end{aligned} \tag{1}$$

Where, P_i represents an individual user-generated post, and $f(\cdot)$ denotes the classification model parameterized by θ. The output Y_i corresponds to the predicted set of binary labels for the 11 depressive symptoms. Each element $y_{i,j}$ in Y_i indicates the binary label for symptom j in post P_i.

6 Hier-DepSy: A Hierarchical Model for Classifying Unbalanced Data

We introduce *Heir-DepSy*, a hierarchical model architecture for automated depression symptom classification from social media data. When examined more thoroughly, the task of symptom classification can be naturally divided into two sequential components. The first component involves determining whether a post expresses any depressive symptom. The second, conditional on the first, involves identifying which specific symptoms are present. Heir-DepSy explicitly models this structure using two successive classification stages describe in Section 6.2. In addition to aligning with the task's inherent structure, this approach also addresses the significant class imbalance in the dataset—where a large proportion of posts are non-symptomatic and several symptoms are underrepresented.

6.1 Model Selection and Training

Multiple pre-trained transformer models, including BERT, RoBERTa, and MentalBERT, were evaluated independently for each classification stage. For each task, models were trained and evaluated separately, and the best-performing model for each stage was selected to construct the final Heir-DepSy architecture. The models performance on each stage can be found in Table 3

Based on the results, **BERT** was selected for the binary classification task and **RoBERTa** was selected for multi-label classification, as they achieved the best micro-averaged F1-score and handled symptoms more effectively.

Figure 2: depressive symptom frequency in posts

	Model	Accuracy	Precision	Recall	F-1
Binary Classification	BERT	0.83	0.80	0.80	0.80
	RoBERTa	0.83	0.79	0.80	0.79
	MentalBERT	0.82	0.78	0.77	0.78
Multi-Label Symptom Classification	BERT	0.47	0.71	0.59	0.64
	RoBERTa	0.49	0.67	0.65	0.66
	MentalBERT	0.47	0.70	0.59	0.64

Table 3: Results for models on classifying depressive symptoms, preparing for Hier-DepSy model. Precision, recall, and F1 are micro-average scores.

6.2 Hier-DepSy Model Architecture

Hier-DepSy is implemented as a two-stage hierarchical classification model composed of two independently trained transformer-based classifiers. As we mentioned earlier, the first stage detects whether a post expresses any depressive symptom (binary classification), and the second stage, triggered only for positive cases, identifies the specific symptoms present (multi-label classification).

7 DepSyLlama Model

7.1 DepSy Instruction Dataset

The DepSy Instruction dataset is constructed using all posts from the raw dataset introduced in Section 3, along with the selected prompts. The symptoms listed in the "symptoms" columns are directly utilized as responses to the corresponding questions. To create the training portion of the DepSy Instruction dataset, we merge the prompt, post, and response into a single text. For optimal model selection, a test set of 10% of the DepSy dataset developed employing the same methodol-

Algorithm 1: Heir-DepSy Model Architecture

Input: Input post x
Output: Predicted symptom vector $\hat{y}_{\text{symptom}} \in \{0,1\}^{11}$
// Stage 1: Binary Classification
$x_{\text{tok}} \leftarrow$ BERT_Tokenizer(x, max_len=512);
$[CLS]_{\text{bert}} \leftarrow$ BERT_Encoder(x_{tok});
$z_{\text{binary}} \leftarrow$ LinearLayer_Binary($[CLS]_{\text{bert}}$);
$b \leftarrow \arg\max(z_{\text{binary}})$; // Binary prediction
if $b = 0$ **then**
  $\hat{y}_{\text{symptom}} \leftarrow [0, 0, \ldots, 0]$; // No symptoms
else
  // Stage 2: Multi-label Classification
  $x_{\text{tok}} \leftarrow$ RoBERTa_Tokenizer(x, max_len=512);
  $[CLS]_{\text{roberta}} \leftarrow$ RoBERTa_Encoder(x_{tok});
  $z_{\text{multi}} \leftarrow$ LinearLayer_MultiLabel($[CLS]_{\text{roberta}}$);
  $p \leftarrow \sigma(z_{\text{multi}})$; // Apply sigmoid
  $\hat{y}_{\text{symptom}}[i] \leftarrow 1$ if $p[i] \geq 0.5$ else $0, \quad \forall i \in \{1, \ldots, n\}$
return $\hat{y}$

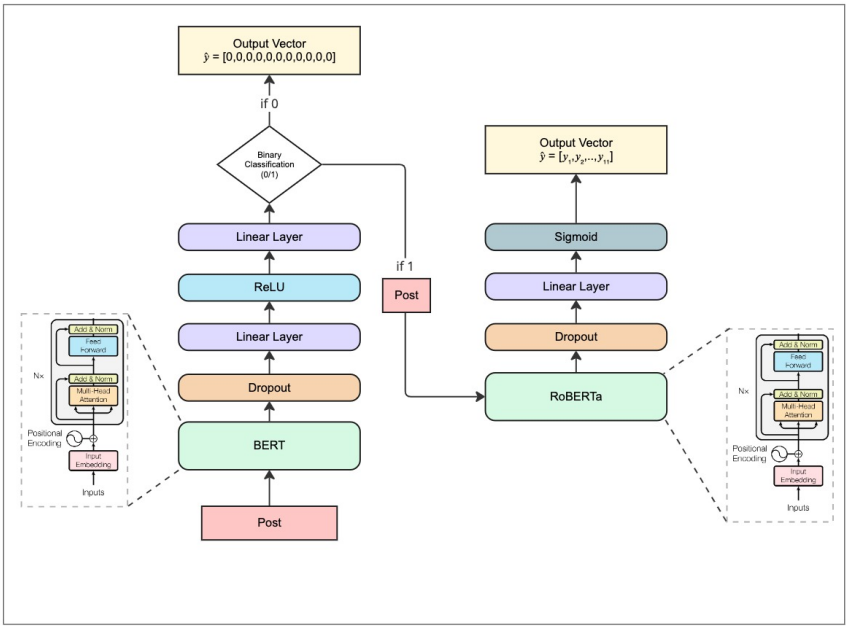

Figure 3: Architecture of the Hier-DepSy model for depression symptoms classification. The model operates in two independent stages without joint training.

ogy was used to test the model.

7.2 DepSyLlama Model Training

We fine-tune the LLaMA models on the DepSy Instruction dataset to develop our DepSyLlaMA model. Specifically, we constructed 2 versions of DepSy by training LLaMA2-7B and Llama3-8B on the DepSy training set for `16 epochs`, selecting the optimal model based on performance on the DepSy validation set. The training process employs a `batch size of 8` and is `optimized using the AdamW` optimizer, with a maximum `learning rate of 1e-5`. The maximum input length for the model is set to `4096 tokens`. To expedite the training process, we utilize QLoRA, configuring the `LoRA rank to 8` and the `alpha parameter to 16`. All models are trained on Nvidia Tesla A100 GPUs, each with 40GB of memory. [4] For model inference we used Depsy with the settings `(temperature = 0.2, max_new_tokens = 30)`

8 Experiments and Results

We evaluate a range of models for our classification tasks, including BERT, RoBERTa, MentalBERT, and large language models (LLMs). Full model details are provided in the appendix. We explore the use of LLMs in zero-shot, zero-shot with labels, and few-shot settings using various prompt formats. The specific prompts used in our experiments are detailed in Appendix A, Table 6.

8.1 Evaluation Metrics

For BERT-based models, we used 5-fold cross-validation to evaluate performance based on accuracy, micro-averaged precision, recall, and F1 scores. Our implementation utilizes Scikit-learn (Pedregosa et al., 2011). To evaluate the performance of large language models (LLMs), we assessed the presence of each symptom label within the generated responses by identifying the corresponding text spans. For the "no symptoms" label, we verified the presence of the phrase "does not contain symptoms" as this was introduced in the prompt when no symptoms were indicated in the post. After this step, we got 11 predicted columns (11 symptoms) and 11 true labels. For evaluating the model's output, we used Scikit-learn's multilabel confusion matrix and extracted micro-averaged precision, recall, F1-score, and accuracy. Additionally, we performed 1,000 bootstrap resamples to compute 95% confidence intervals (CI) for these metrics. DepSyLlama was evaluated using 3-fold cross-validation. [5]

8.2 Results

Table 4 presents the performance comparison of various models on the task, evaluating accuracy,

[4]The model will be made publicly available on Hugging Face upon paper acceptance

[5]The code will be made available upon paper acceptance

	Model	Accuracy	Precision	Recall	F-1	CI 95% F-1
	BERT	0.711	0.569	0.482	0.522	[0.514, 0.529]
	RoBERTa	0.686	0.505	0.501	0.501	[0.488, 0.512]
	MentalBERT	0.701	0.546	0.483	0.512	[0.500, 0.524]
	Hier-DepSy	**0.845**	0.668	**0.679**	**0.673**	[0.614, 0.702]
Zero shot	GPT-4o	0.527	**0.714**	0.033	0.064	[0.038, 0.091]
	GPT-3.5	0.527	0.704	0.032	0.061	[0.037, 0.086]
	MentalLlama_7b	0.526	0.418	0.068	0.118	[0.086, 0.149]
	DepSy Llama2	0.526	0.419	0.66	0.486	[0.47, 0.502]
	DepSy Llama3	0.313	0.370	0.494	0.423	[0.411, 0.435]
Zero shot with labels	GPT-4o	0.564	0.634	0.185	0.287	[0.245, 0.325]
	GPT-3.5	0.53	0.342	0.432	0.382	[0.352, 0.411]
	Llama2	0.479	0.116	0.099	0.107	[0.082, 0.132]
	Llama3	0.521	0.305	0.066	0.108	[0.079, 0.141]
	MentalLama	0.446	0.099	0.15	0.12	[0.099, 0.142]
	Depsy Llama 2	0.196	0.14	0.62	0.226	[0.208, 0.244]
	Depsy Llama 3	0.425	0.138	0.543	0.198	[0.177, 0.218]
Few shots	GPT-4o	0.612	0.527	0.492	0.509	[0.475, 0.543]
	GPT-3.5	0.476	0.309	0.574	0.402	[0.375, 0.429]
	Llama2	0.397	0.119	0.135	0.126	[0.103, 0.15]
	Llama3	0.422	0.13	0.197	0.157	[0.132, 0.181]
	MentalLama	0.168	0.121	0.457	0.191	[0.173, 0.21]
	DepSy Llama2	0.360	0.259	0.379	0.308	[0.29, 0.325]
	DepSy Llama3	0.168	0.121	0.457	0.191	[0.173, 0.21]

Table 4: Results for models on classifying depressive symptoms using DepSy dataset. Precision, recall, and F1 are micro-average scores.

Symptom	BERT			RoBERTa			MentalBERT			Hier-DepSy		
	P	R	F1	P	R	F1	P	R	F1	P	R	**F1**
Feeling Down and Depressed	0.57	0.54	0.55	0.53	0.60	0.56	0.55	0.53	0.54	0.64	0.76	**0.70**
Feeling tired or having little energy	0.49	0.46	0.47	0.64	0.50	0.56	0.56	0.43	0.48	0.60	0.60	**0.60**
Crying	0.67	0.75	0.71	0.62	0.72	0.67	0.64	0.89	0.74	0.74	0.91	**0.82**
Lonliness	0.61	0.23	0.33	0.66	0.37	0.47	0.46	0.27	0.34	0.65	0.71	**0.68**
Sleep Disturbance	0.69	0.72	0.71	0.67	0.72	0.70	0.67	0.70	0.68	0.78	0.90	**0.84**
Feeling Failure	0.29	0.11	0.16	0.18	0.20	0.19	0.38	0.13	0.20	0.48	0.52	**0.50**
Loss of Interest	0.27	0.23	0.25	0.14	0.16	0.15	0.17	0.16	0.17	0.36	0.33	**0.35**
Concentration Problems	0.32	0.26	0.29	0.15	0.13	0.14	0.28	0.22	0.24	0.50	0.55	**0.52**
Poor Appetite / Eating Disturbance	0.56	0.65	0.60	0.46	0.52	0.49	0.58	0.61	0.60	0.70	0.78	**0.74**
Suicidal Thoughts	1	0.33	0.50	1	0.50	0.67	0.62	0.42	0.50	0.62	0.80	**0.70**
Self Blame	0	0	0	0	0	0	0	0	0	0	0	0

Table 5: Per-symptom precision, recall, and F1-score across four models. Symptoms are sorted by their frequency in the dataset.

precision, recall, and F1-score with 95% confidence intervals. Among all models, our Hier-Depsy model achieved the highest F1-score (0.673) and accuracy (0.845), outperforming BERT, RoBERTa, and MentalBERT and LLMs. In the LLMs zero-shot setting, GPT-4o showed the highest precision (0.714), but with extremely low recall (0.033), resulting in a low F1-score (0.064). When provided with labels, GPT-4o's performance improved substantially, achieving an F1-score of 0.385, suggesting the benefit of added contextual guidance for general-purpose LLMs. Among few-shot models, GPT-4o again outperformed others with an F1-score of 0.475 and the highest accuracy (0.612), approaching the performance of fine-tuned models.

Overall, LLaMA models showed limited effectiveness across all settings, as reflected by low precision, recall, and F1-scores. This trend supports prior findings (Ignashina et al., 2025), indicating their difficulty in recognising nuanced mental health symptoms. Given the importance of recall in depression screening—where missing symptoms can have serious implications—models with higher recall are more appropriate for this task (Ren et al., 2015). Depsy Llama models consistently demonstrated strong recall across settings. In the zero-shot setting, Depsy Llama 2 achieved the highest recall (0.66), while in the few-shot setting, Depsy Llama 3 reached a recall of 0.494. Although these models had low precision, their ability to detect a

wider range of depressive symptoms makes them useful for early-stage mental health assessment.

Table 5 reports per-symptom precision, recall, and F1-scores across four models, with symptoms sorted by frequency. Notably, higher frequency does not always correspond to better performance. Symptoms like *Crying* and *Sleep Disturbance* achieved the highest F1-scores (up to 0.82 and 0.84 with Hier-DepSy), likely due to their more explicit linguistic expression. In contrast, abstract symptoms such as *Loss of Interest* or *Feeling Down and Depressed* showed lower performance, despite being more frequent. Rare symptoms like *Self-Blame* remained difficult to detect across all models. Hier-DepSy consistently outperformed baselines, particularly on less frequent symptoms, highlighting its effectiveness in handling imbalanced and subtle symptom classes.

9 Discussion

The Hier-DepSy model outperformed all other approaches in our symptom classification task, demonstrating the value of a hierarchical structure for handling task complexity and class imbalance. By decoupling binary detection from multi-label classification, the model was better able to capture symptom-specific patterns. Similar layered strategies have proven effective in related domains, reinforcing the broader applicability of hierarchical approaches in psychological NLP tasks.

Large language models (LLMs), in contrast, underperformed compared to fine-tuned BERT-based models. This aligns with previous findings in health-related NLP and may be partly due to embedded safety alignment, which can lead to conservative outputs when handling sensitive content. While we did not observe explicit refusals or safety prompts, implicit alignment may still have influenced prediction behaviour. However, per-symptom confusion matrix analysis (Appendix C) did not reveal systematic avoidance of sensitive symptoms such as suicidal ideation.

DepSyLLaMA, fine-tuned on our dataset using QLoRA, did not outperform general LLMs. This may be due to limitations in parameter-efficient fine-tuning, which updates only a subset of parameters, reducing the model's ability to adapt fully to the task. Additionally, the limited size of posts with symptoms and class imbalance likely affected generalisation. Interestingly, DepSyLLaMA performed best in the pure zero-shot setting, suggesting that fine-tuning may reduce reliance on in-context prompts and thus reduce the size of inference.

Finally, domain-specific models like MentalLLaMA and MentalBERT did not outperform general-purpose models. This suggests that pretraining on mental health data alone is insufficient for accurate symptom classification. These models may have been tuned for empathetic dialogue rather than multi-label detection. Our findings underscore the need for task-specific fine-tuning strategies that directly optimise for distinguishing nuanced symptom categories, beyond domain relevance alone.

To further investigate model limitations, we conducted detailed error analysis across symptoms and models. Table 5 shows that symptoms such as Crying, Sleep Disturbance, and Feeling Down and Depressed achieved consistently high F1-scores, likely due to their prevalence and clearer linguistic cues. In contrast, low-frequency symptoms like Loss of Interest and Self Blame were rarely detected, with some models failing to predict them at all. This highlights the effect of data imbalance and the need for symptom-specific learning strategies. Confusion matrix analysis of LLMs (Appendix C) showed no strong evidence of safety-driven suppression of sensitive symptoms, though underprediction was observed for abstract or less explicitly stated symptoms such as Loneliness and Feeling Failure. These findings reinforce the strengths of hierarchical models in managing imbalance and the limitations of LLMs in capturing subtle symptom expressions without direct optimisation.

10 Conclusions

This paper addressed the task of detecting depressive symptoms in social media posts as a multi-label classification problem. We evaluated transformer-based models, large language models (LLMs), and a hierarchical architecture. The proposed Hier-DepSy model improved detection of low-frequency symptoms by mitigating class imbalance through a two-stage structure. We also introduced DepSyLLaMA, a domain-adapted LLM, which outperformed general-purpose open-source LLMs in zero-shot settings and showed competitive results compared to proprietary models. Despite these advances, symptom detection remains a challenging task, with the best model achieving an F1-score of 0.673, highlighting the need for further research in this area.

11 Limitations

Despite the novel contributions of this study, several limitations should be acknowledged. First, while our model was developed using the DepSy dataset, which is substantial in size (over 40,000 entries) compared to existing datasets and believed to be robust, we aimed to further evaluate its generalizability by testing it on additional datasets. However, our request for access to the SAD depressive symptoms dataset (Mowery et al., 2017), which would have enabled broader validation, did not receive a response from its owners. This limited our ability to assess the model's generalizability across diverse data sources. Secondly, while the DepSy dataset is carefully annotated by expert psychologists, it relies on publicly available social media posts, which may not fully capture the diversity of individuals experiencing depression. Social media content often reflects self-presentation biases, potentially affecting symptom reporting and model generalizability.

Ethical Consideration

This study has received ethics approval from XXXXXX [6] (Reference: 21IC7222). The dataset contains only publicly available posts from X, and we are committed to following ethical practices to protect the privacy and anonymity of the users. To ensure this, the author's usernames, which could contain sensitive information related to the names or locations of the user, are not saved or used. Instead, the information was pre-processed and replaced with user IDs. Social media data is often sensitive, particularly when it is related to mental health, and we take great care to ensure that our dataset is handled responsibly. Since the dataset is related to mental disorders, it might trigger some people, thus, annotators were advised to take breaks during annotation and were given plenty of time.

References

V Adarsh, P Arun Kumar, V Lavanya, and GR Gangadharan. 2023. Fair and explainable depression detection in social media. *Information Processing & Management*, 60(1):103168.

Falwah Alhamed, Rebecca Bendayan, Julia Ive, and Lucia Specia. 2024a. Monitoring depression severity and symptoms in user-generated content: An annotation scheme and guidelines. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 227–233.

Falwah Alhamed, Julia Ive, and Lucia Specia. 2024b. Classifying social media users before and after depression diagnosis via their language usage: A dataset and study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3250–3260.

Mario Aragon, Adrián Pastor López Monroy, Luis Gonzalez, David E Losada, and Manuel Montes. 2023. Disorbert: A double domain adaptation model for detecting signs of mental disorders in social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15305–15318.

Mario Ezra Aragón, Javier Parapar, and David E Losada. 2024. Delving into the depths: evaluating depression severity through bdi-biased summaries. In *9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 12–22. Association for Computational Linguistics.

Nadiah A Baghdadi, Amer Malki, Hossam Magdy Balaha, Yousry AbdulAzeem, Mahmoud Badawy, and Mostafa Elhosseini. 2022. An optimized deep learning approach for suicide detection through arabic tweets. *PeerJ Computer Science*, 8:e1070.

Aaron T Beck, Charles H Ward, Morris Mendelsohn, John Mock, and James Erbaugh. 1961. An inventory for measuring depression. *Archives of General Psychiatry*, 4(6):561–571.

Stevie Chancellor, Steven A Sumner, Corinne David-Ferdon, Tahirah Ahmad, and Munmun De Choudhury. 2021. Suicide risk and protective factors in online support forum posts: annotation scheme development and validation study. *JMIR mental health*, 8(11):e24471.

Raymond Chiong, Gregorious Satia Budhi, and Sandeep Dhakal. 2021a. Combining sentiment lexicons and content-based features for depression detection. *IEEE Intelligent Systems*, 36(6):99–105.

Raymond Chiong, Gregorius Satia Budhi, Sandeep Dhakal, and Fabian Chiong. 2021b. A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, 135:104499.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. *Conference on Human*

[6] masked for anonymity

Factors in Computing Systems - Proceedings, pages 2098–2110.

Joseigla Pinto de Oliveira, Karen Jansen, Taiane de Azevedo Cardoso, Thaíse Campos Mondin, Luciano Dias de Mattos Souza, Ricardo Azevedo da Silva, and Fernanda Pedrotti Moreira. 2021. Predictors of conversion from major depressive disorder to bipolar disorder. *Psychiatry Research*, 297(January):113740.

DSM5. 2013. *Diagnostic and statistical manual of mental disorders : DSM-5*, 5th ed. edition. American Psychiatric Association Arlington, VA.

Özay Ezerceli and Rahim Dehkharghani. 2024. Mental disorder and suicidal ideation detection from social media using deep neural networks. *Journal of Computational Social Science*, 7(3):2277–2307.

Roque Fernández-Iglesias, Marcos Fernández-Pichel, Mario Aragon, and David E Losada. 2024. Depressmind: A depression surveillance system for social media analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 35–43.

Wadzani Aduwamai Gadzama, Danlami Gabi, Musa Sule Argungu, and Hassan Umar Suru. 2024. The use of machine learning and deep learning models in detecting depression on social media: A systematic literature review. *Personalized Medicine in Psychiatry*, 45:100125.

Muskan Garg. 2023. Mental health analysis in social media posts: a survey. *Archives of Computational Methods in Engineering*, 30(3):1819–1842.

Jue Gong, Gregory E. Simon, and Shan Liu. 2019. Machine learning discovery of longitudinal patterns of depression and suicidal ideation. *PLoS ONE*, 14(9):1–15.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to automate follow-up question generation using process knowledge for depression triage on Reddit posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 137–147, Seattle, USA. Association for Computational Linguistics.

Keith Harrigian and Mark Dredze. 2022. Then and now: Quantifying the longitudinal validity of self-disclosed depression diagnoses. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 59–75, Seattle, USA. Association for Computational Linguistics.

Mariia Ignashina, Paulina Bondaronek, Dan Santel, John Pestian, and Julia Ive. 2025. Llm assistance for pediatric depression. *Preprint*, arXiv:2501.17510.

Mohsinul Kabir, Tasnim Ahmed, Md Bakhtiar Hasan, Md Tahmid Rahman Laskar, Tarun Kumar Joarder, Hasan Mahmud, and Kamrul Hasan. 2023. Deptweet: A typology for social media texts to detect depression severities. *Computers in Human Behavior*, 139:107503.

D Sami Khafaga, Maheshwari Auvdaiappan, K Deepa, Mohamed Abouhawwash, and F Khalid Karim. 2023. Deep learning for depression detection using twitter data. *Intelligent Automation & Soft Computing*, 36(2):1301–1313.

Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *Preprint*, arXiv:2307.10234.

Hyewon Kim, Yuwon Kim, Ji Hyun Baek, Maurizio Fava, David Mischoulon, Andrew A. Nierenberg, Kwan Woo Choi, Eun Jin Na, Myung Hee Shin, and Hong Jin Jeon. 2020. Predictive factors of diagnostic conversion from major depressive disorder to bipolar disorder in young adults ages 19–34: A nationwide population study in South Korea. *Journal of Affective Disorders*, 265(December 2019):52–58.

Harnain Kour and Manoj K Gupta. 2022. An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bidirectional lstm. *Multimedia Tools and Applications*, 81(17):23649–23685.

Kurt Kroenke, Robert L Spitzer, and Janet B Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Tingting Liu, Devansh Jain, Shivani R Rapole, Brenda Curtis, Johannes C Eichstaedt, Lyle H Ungar, and Sharath Chandra Guntuku. 2023. Detecting symptoms of depression on reddit. In *Proceedings of the 15th ACM web science conference 2023*, pages 174–183.

Rahele Mesbah, Nienke de Bles, Nathaly Rius-Ottenheim, A. J.Willem van der Does, Brenda W.J.H. Penninx, Albert M. van Hemert, Max de Leeuw, Erik J. Giltay, and Manja Koenders. 2021. Anger and cluster B personality traits and the conversion from unipolar depression to bipolar disorder. *Depression and Anxiety*, (August 2020):1–11.

Kirill Milintsevich, Kairit Sirts, and Gaël Dias. 2024. Your model is not predicting depression well and that is why: A case study of primate dataset. *arXiv preprint arXiv:2403.00438*.

Arturo Montejo-Ráez, M Dolores Molina-González, Salud María Jiménez-Zafra, Miguel Ángel García-Cumbreras, and Luis Joaquín García-López. 2024. A survey on detecting mental disorders with natural

language processing: Literature review, trends and challenges. *Computer Science Review*, 53:100654.

Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, Mike Conway, et al. 2017. Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study. *Journal of medical Internet research*, 19(2):e6895.

Danielle L. Mowery, Albert Park, Craig Bryan, and Mike Conway. 2016. Towards automatically classifying depressive symptoms from Twitter data for population health. In *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pages 182–191, Osaka, Japan. The COLING 2016 Organizing Committee.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. *arXiv preprint arXiv:2204.10432*.

Alina Pavlova and Pauwke Berkers. 2022. "mental health" as defined by twitter: Frames, emotions, stigma. *Health communication*, 37(5):637–647.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Melanie F. Pradier, Michael C. Hughes, Thomas H. McCoy, Sergio A. Barroilhet, Finale Doshi-Velez, and Roy H. Perlis. 2021. Predicting change in diagnosis from major depression to bipolar disorder after antidepressant initiation. *Neuropsychopharmacology*, 46(2):455–461.

Lenore S Radloff. 1977. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401.

Yanping Ren, Hui Yang, Colette Browning, Shane Thomas, and Meiyan Liu. 2015. Performance of screening tools in detecting major depressive disorder among patients with coronary heart disease: a systematic review. *Medical science monitor: international medical journal of experimental and clinical research*, 21:646.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Ratn Shah. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. pages 7685–7697.

Faisal Muhammad Shah, Farzad Ahmed, Sajib Kumar Saha Joy, Sifat Ahmed, Samir Sadek, Rimon Shil, and Md Hasanul Kabir. 2020. Early depression detection from social network using deep learning techniques. In *2020 IEEE region 10 symposium (TENSYMP)*, pages 823–826. IEEE.

Mikke Tavast, Anton Kunnari, and Perttu Hämäläinen. 2022. Language models can generate human-like self-reports of emotion. In *27th International Conference on Intelligent User Interfaces*, IUI '22 Companion, page 69–72, New York, NY, USA. Association for Computing Machinery.

Wei-Yao Wang, Yu-Chien Tang, Wei-Wei Du, and Wen-Chih Peng. 2022. Nycu_twd@ lt-edi-acl2022: Ensemble models with vader and contrastive learning for detecting signs of depression from social media. In *Proceedings of the second workshop on language technology for equality, diversity and inclusion*, pages 136–139.

World Health Organization. 2001. The world health report 2001: Mental disorders affect one in four people. Accessed: 2024-10-05.

Shweta Yadav, Cornelia Caragea, Chenye Zhao, Naincy Kumari, Marvin Solberg, and Tanmay Sharma. 2023. Towards identifying fine-grained depression symptoms from memes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8890–8905.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. *arXiv preprint arXiv:2011.06149*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, volume 35 of *WWW '24*, page 4489–4500. ACM.

Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. 2017. Semi-Supervised Approach to Monitoring Clinical Depressive Symptoms in Social Media. *Proceedings of the ... IEEE/ACM International Conference on Advances in Social Network Analysis and Mining. IEEE/ACM International Conference on Advances in Social Network Analysis and Mining*, 2017:1191–1198.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1):281–304.

A Prompts Used for LLMs

Table6 presents the prompts used for LLMs to extract depression symptoms from posts. The selected prompts are highlighted in bold, indicating those chosen for the final analysis

Zero-shot	Identify depression symptoms in this post
	Can you extract depression symptoms from this text
	Extract depression symptoms from this text, return a list of symptoms
	Extract depression symptoms from this text, return a list of symptoms, return "no symptoms" if none existed
	Extract depression symptoms from the post below, if there is no depression symptom, return "this post does not contain depression symptoms"
Zero-shot with labels	Extract depression symptoms from the post below , the symptoms could be one or more from this list [poor appetite or eating disturbances, feeling down and depressed, crying, concentration problems, feeling tired or having little energy, feelings of failure, sleep disturbances, loss of interest, self-blame and shame, loneliness, suicidal thoughts] if there is no depression symptoms, return "this post does not contain depression symptoms."
	Extract depression symptoms from the post below , the symptoms could be one or more from this list only [poor appetite or eating disturbances, feeling down and depressed, crying, concentration problems, feeling tired or having little energy, feelings of failure, sleep disturbances, loss of interest, self-blame and shame, loneliness, suicidal thoughts] if there is no depression symptoms, return only "this post does not contain depression symptoms."
Few shots	**Extract depression symptoms from the post below, the symptoms could be one or more from this list only [poor appetite or eating disturbances, feeling down and depressed, crying, concentration problems, feeling tired or having little energy, feelings of failure, sleep disturbances, loss of interest, self-blame and shame, loneliness, suicidal thoughts] if there is no depression symptoms, return only "this post does not contain depression symptoms." Examples: - post: "I Received an unexpected surprise it's been an emotional afternoon. I'm feeling sentimental about someone. This season has been challenging, but this moment has lifted my spirits. I'm feeling nostalgic" symptoms: feeling down and depressed - post: "others have commented on my appearance, saying I seem more toned, and that makes me thrilled. Personally, I don't notice the difference, but it's obvious that my new eating habits are paying off." Symptoms: this post does not contain depression symptoms - post: "i've noticed a discrepancy between my online presence and the response I get when I share my thoughts on a critical issue. It seems that despite having a large following, my words often fall on deaf ears" symptoms: feeling failure, loneliness POST: {post} symtoms:**

Table 6: Prompts used for LLMs in extracting depression symptoms from a post. The chosen prompts are bold-faced

B Models Hyper-parameters for Extracting Depressive Symptoms

We performed hyperparameter tuning for BERT-based models to optimize performance. The best results, obtained after extensive experimentation, are presented below.

BERT .

```
Model_card: "bert-base-uncased"
Epochs: 64
Batch_size: 8
Learning_rate:5e-5
Hidden_size:128
Optimizer:Adam
Loss: BCEWithLogitsLoss
```

RoBERTa .

```
Model_card: "roberta-base"
Epochs: 128
Batch_size: 32
Learning_rate:1e-05
Hidden_size:128
Optimizer:Adam
Loss: BCEWithLogitsLoss
```

MentalBERT .

```
Model_card: "mental-bert-base-uncased"
Epochs: 64
Batch_size: 32
Learning_rate:1e-05
Hidden_size:128
Optimizer:Adam
Loss: BCEWithLogitsLoss
```

GPT. We used the GPT-3.5 "gpt-3.5-turbo" and GPT-4 "gpt-4-turbo" versions, as these have shown a strong ability to understand human-like emotional context (Tavast et al., 2022), and in sentiment analysis (Kheiri and Karimi, 2023). We used the Official OpenAI Python library[7] to collect responses with the settings (`temperature = 0.2, max tokens = 30`).

Llama 2. The Hugging Face library is used for MentalBERT tokenization and fine-tuning, namely the `'meta-llama/Llama-2-7b-hf'` model card.

Llama 3. The Hugging Face library is used for MentalBERT tokenization and fine-tuning, namely the `meta-llama/Meta-Llama-3-8B'` model card.

MentalLama. MentaLLama is a specialized large language model built on Llama designed

[7] https://platform.openai.com/docs/libraries/python-library

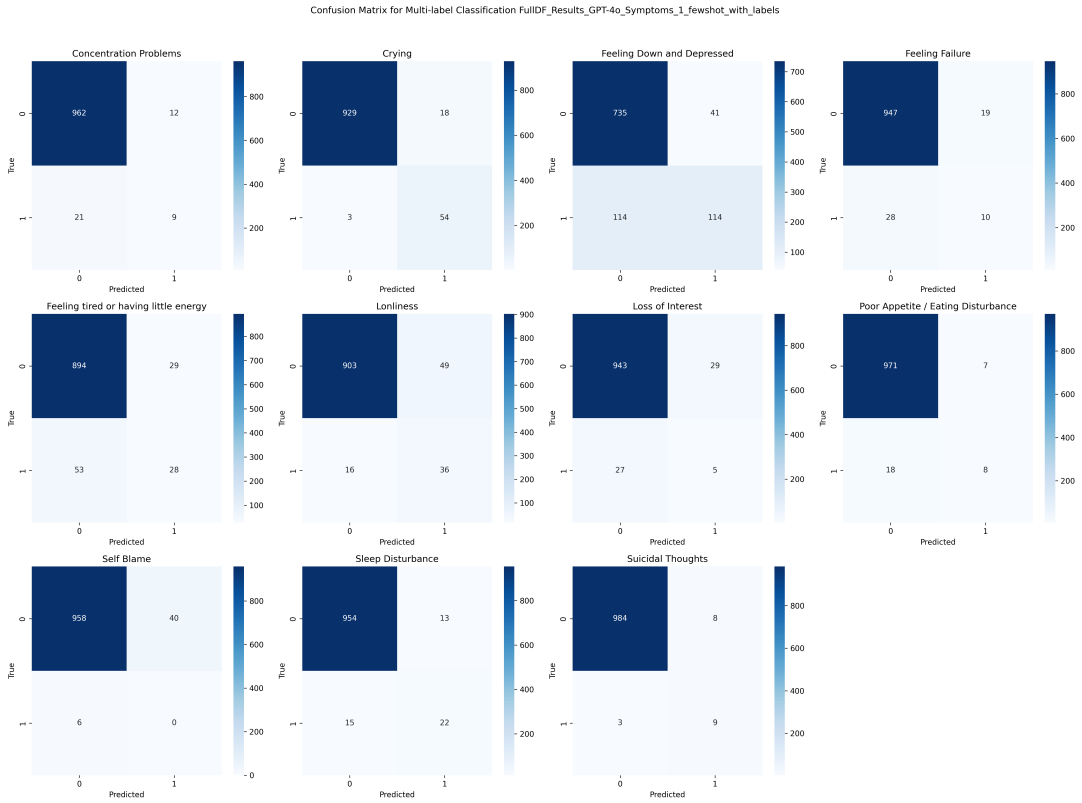

Figure 4: Confusion matrices for GPT-4o in the few-shot setting for symptom-level classification.

to excel in the domain of mental health (Yang et al., 2024). MentaLLama incorporates domain-specific training data and techniques to enhance its ability to understand, process, interpret, and generate text related to mental health. ‘`klyang/MentaLLaMA-chat-7B`’

C Appendix: Per-Symptom Confusion Matrices for LLMs

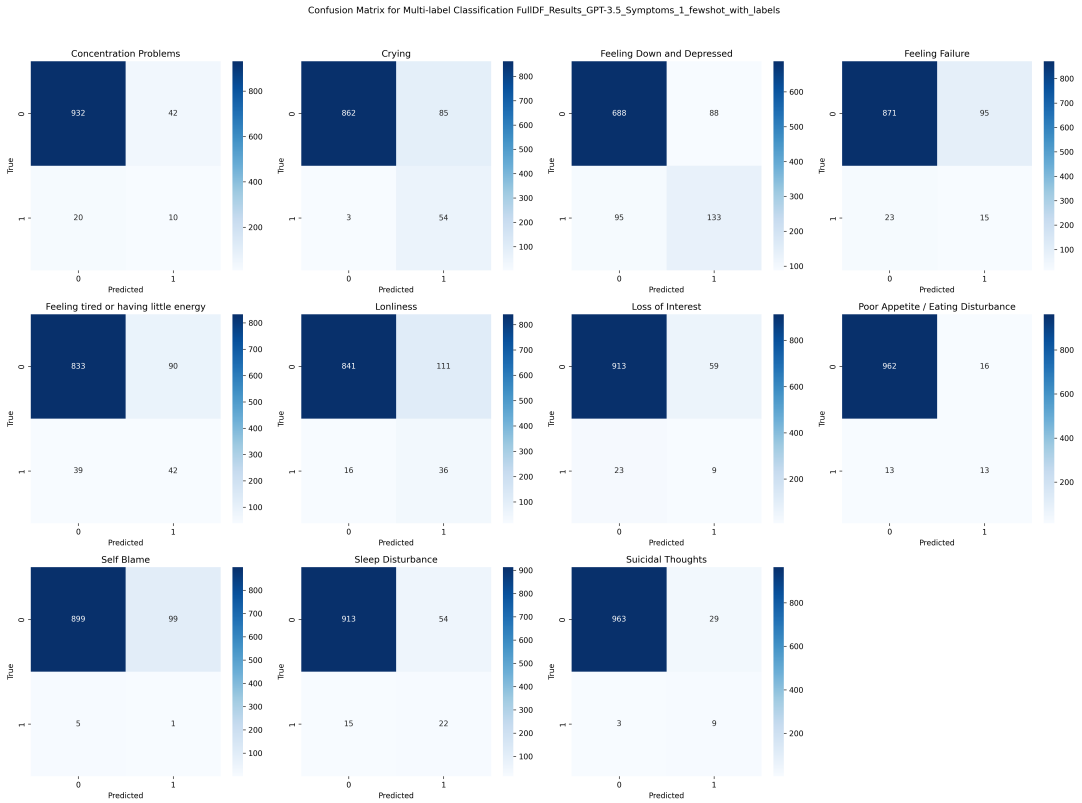

Figure 5: Confusion matrices for GPT-3.5 in the few-shot setting for symptom-level classification.

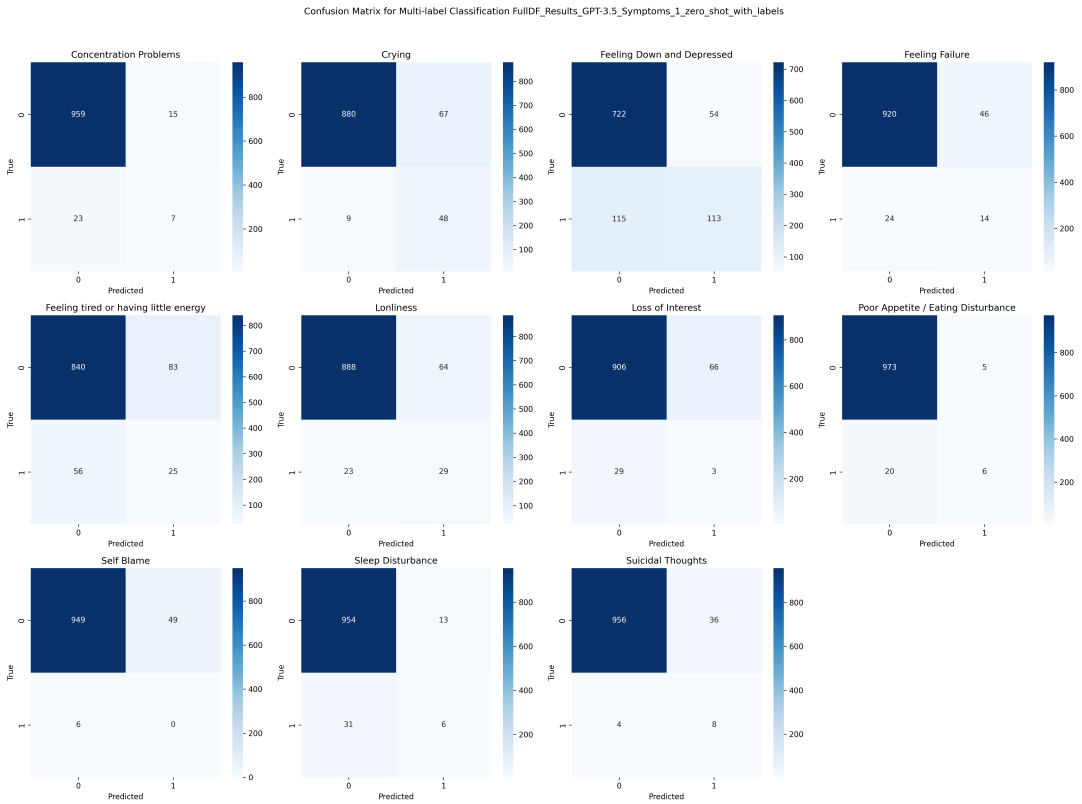

Figure 6: Confusion matrices for GPT-3.5 in the zero-shot with labels setting.

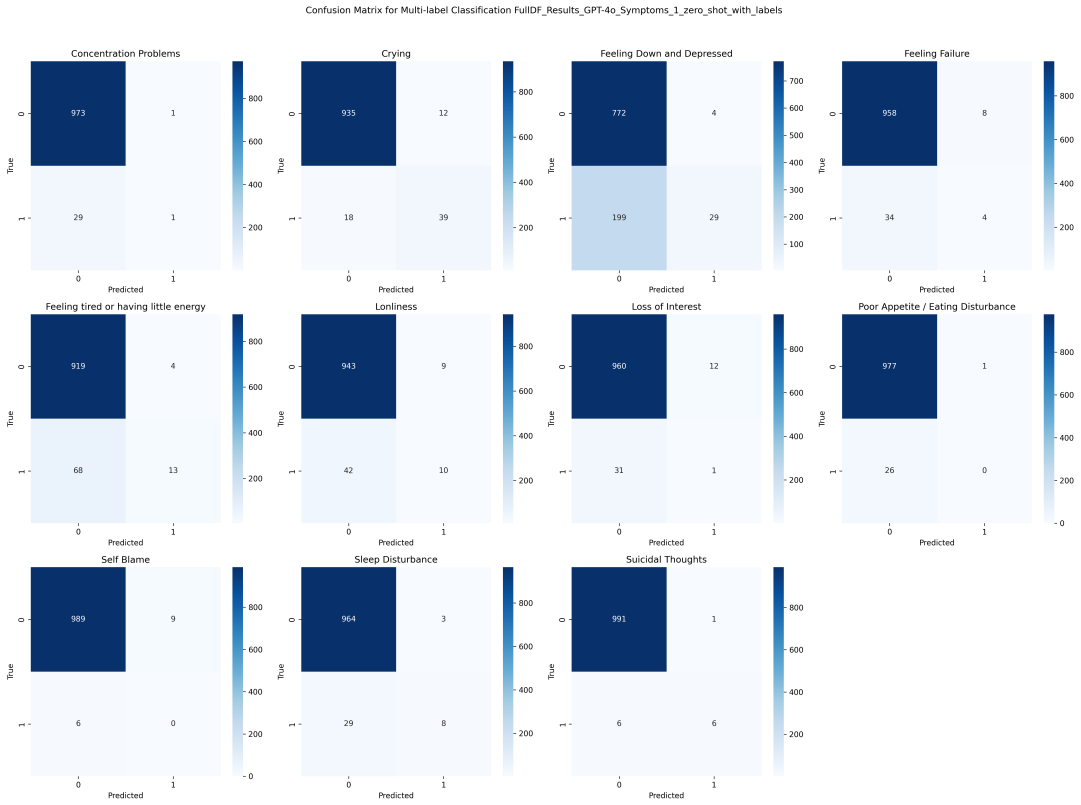

Figure 7: Confusion matrices for GPT-4o in the zero-shot with labels setting.

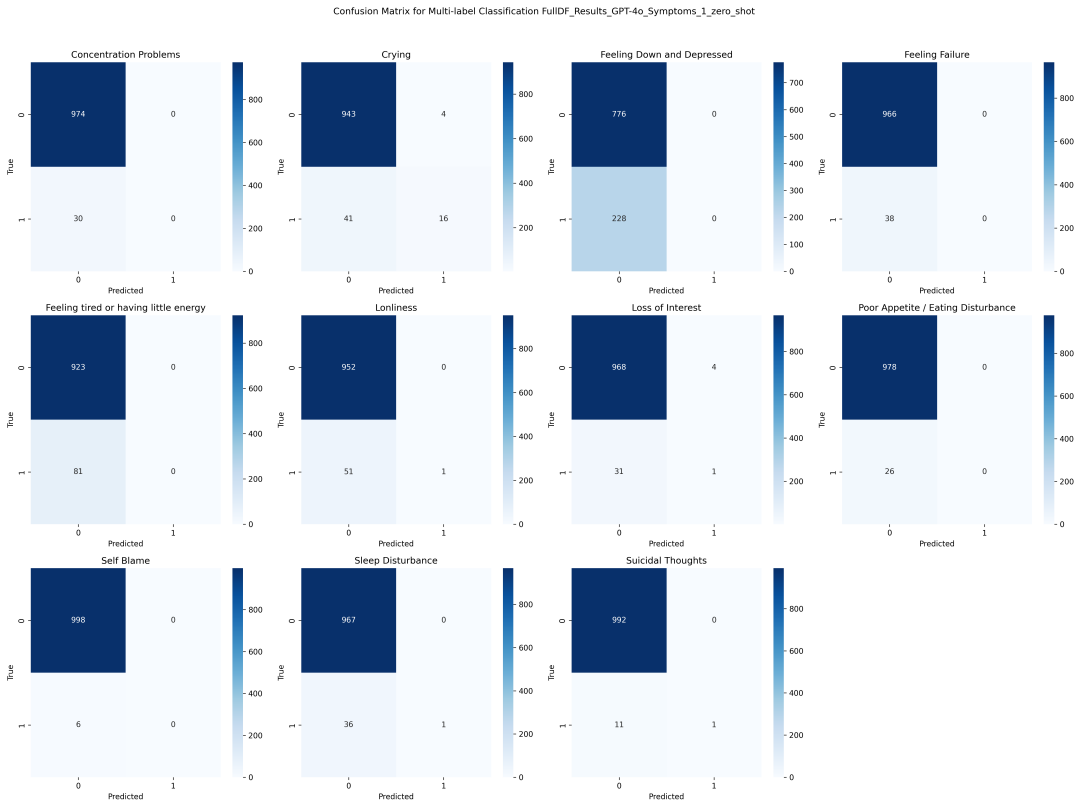

Figure 8: Confusion matrices for GPT-4o in the pure zero-shot setting (no labels or examples provided).

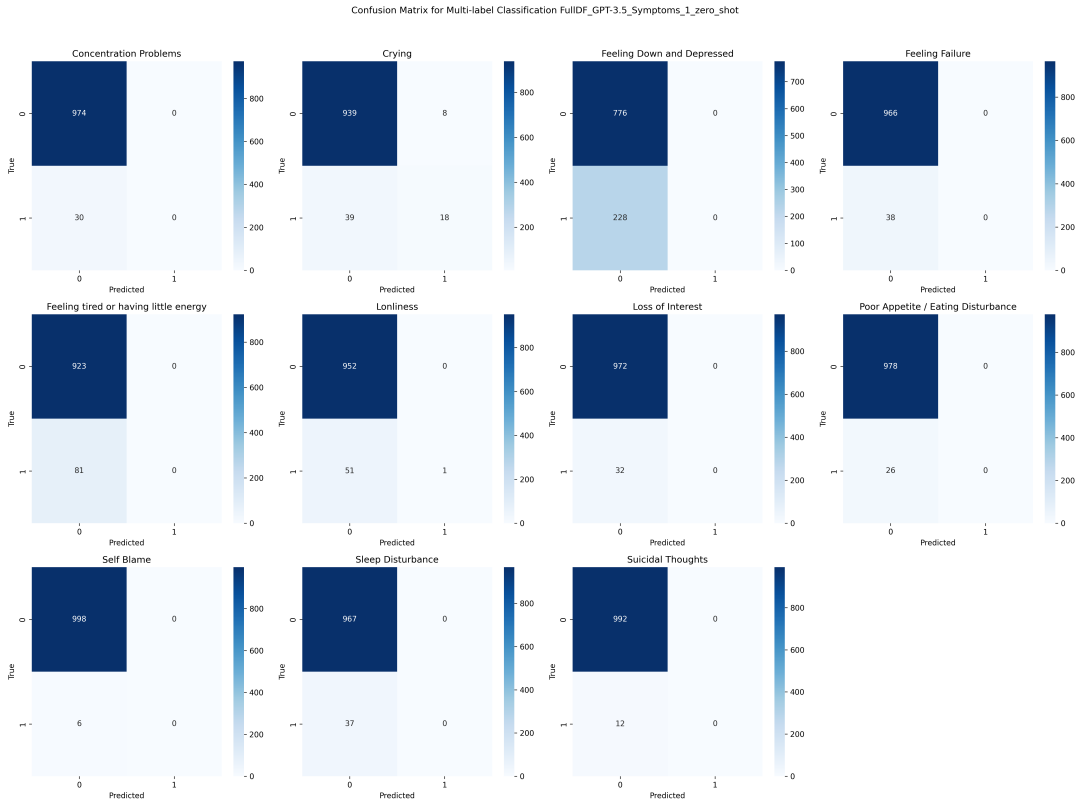

Figure 9: Confusion matrices for GPT-3.5 in the pure zero-shot setting (no labels or examples provided).

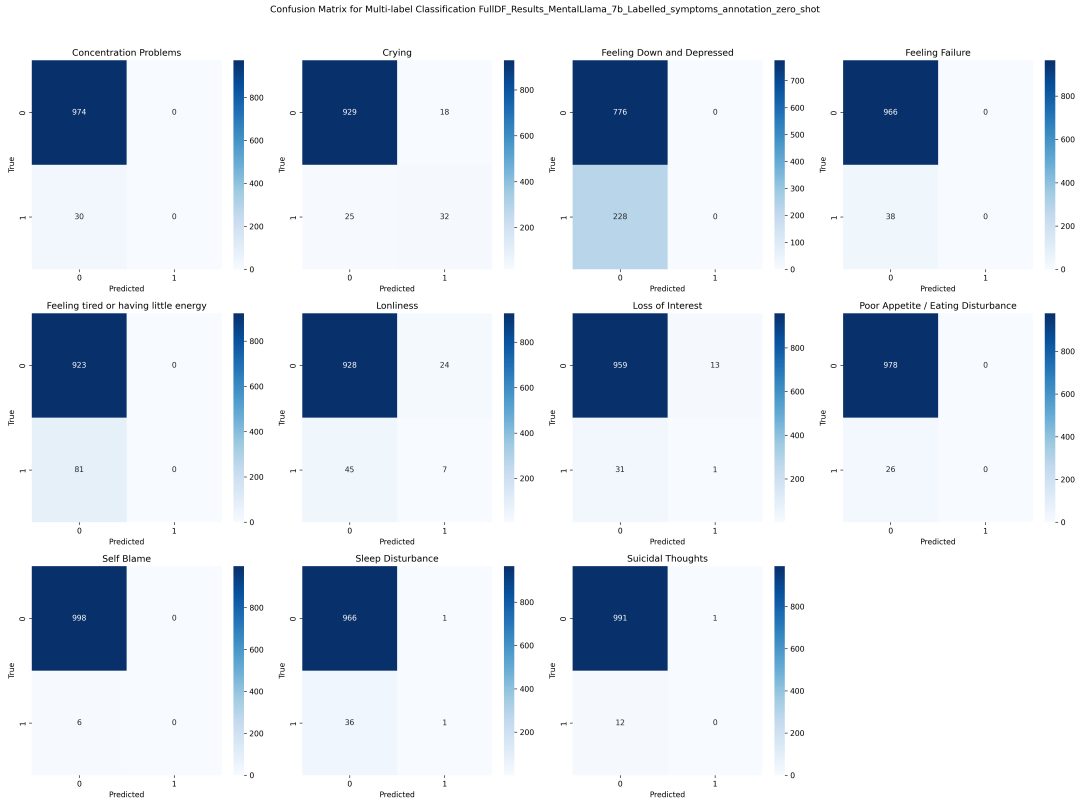

Figure 10: Confusion matrices for MentalLLaMA-7B in the zero-shot setting.

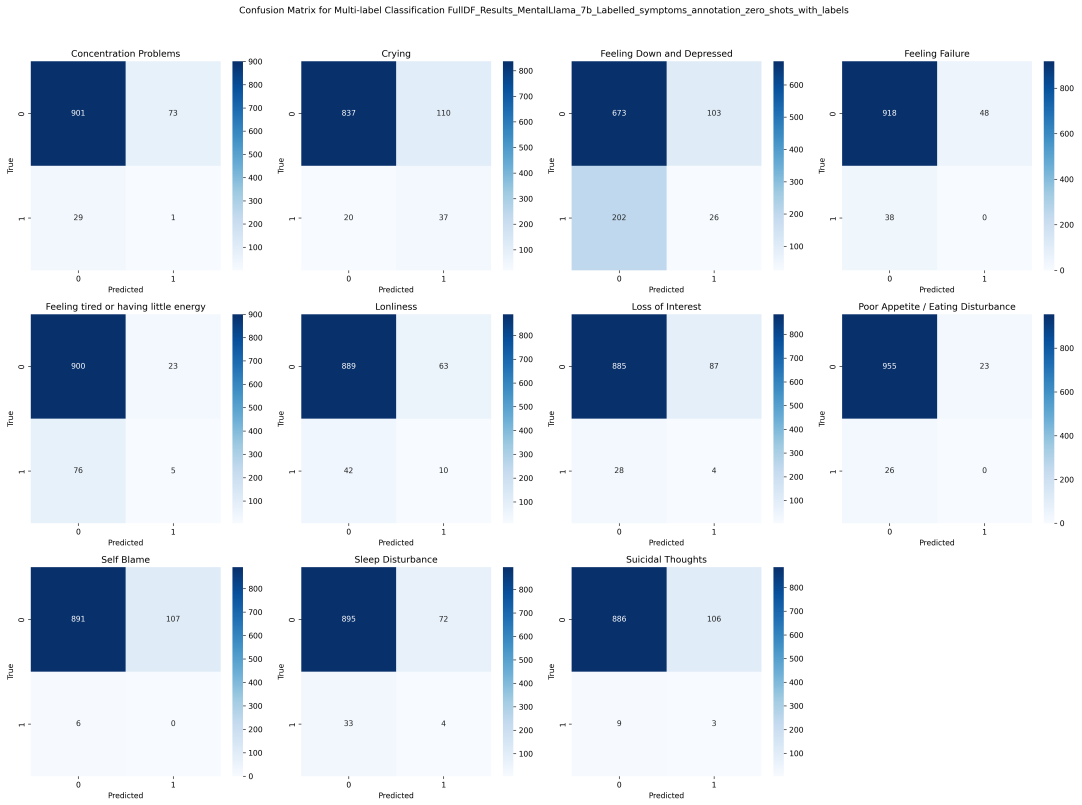

Figure 11: Confusion matrices for MentalLLaMA-7B in the zero-shot with labels setting.

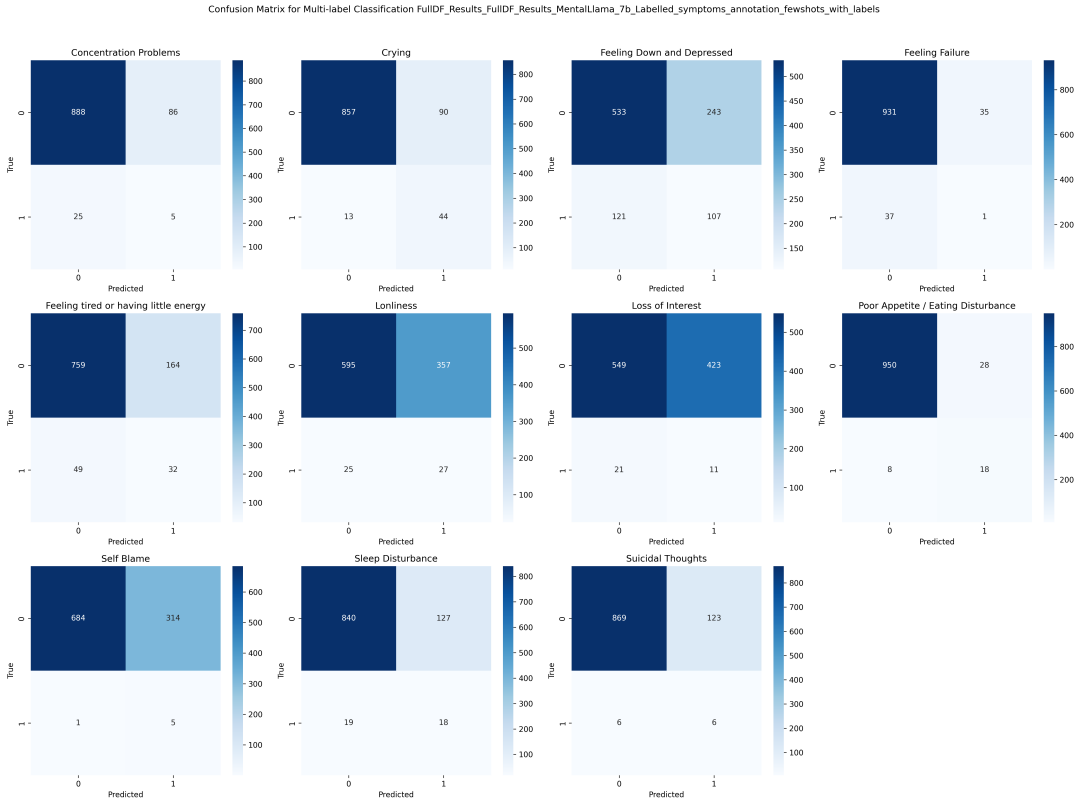

Figure 12: Confusion matrices for MentalLLaMA-7B in the few-shot setting.

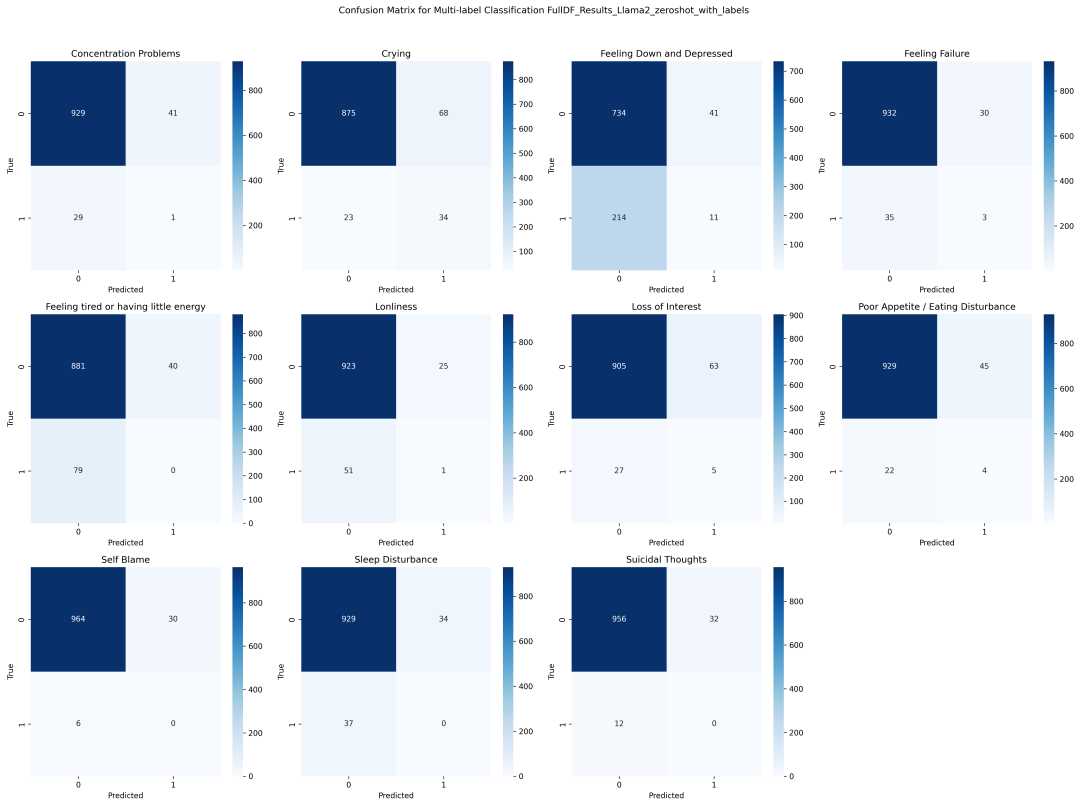

Figure 13: Confusion matrices for Llama2-7B in the zero-shot with labels setting.

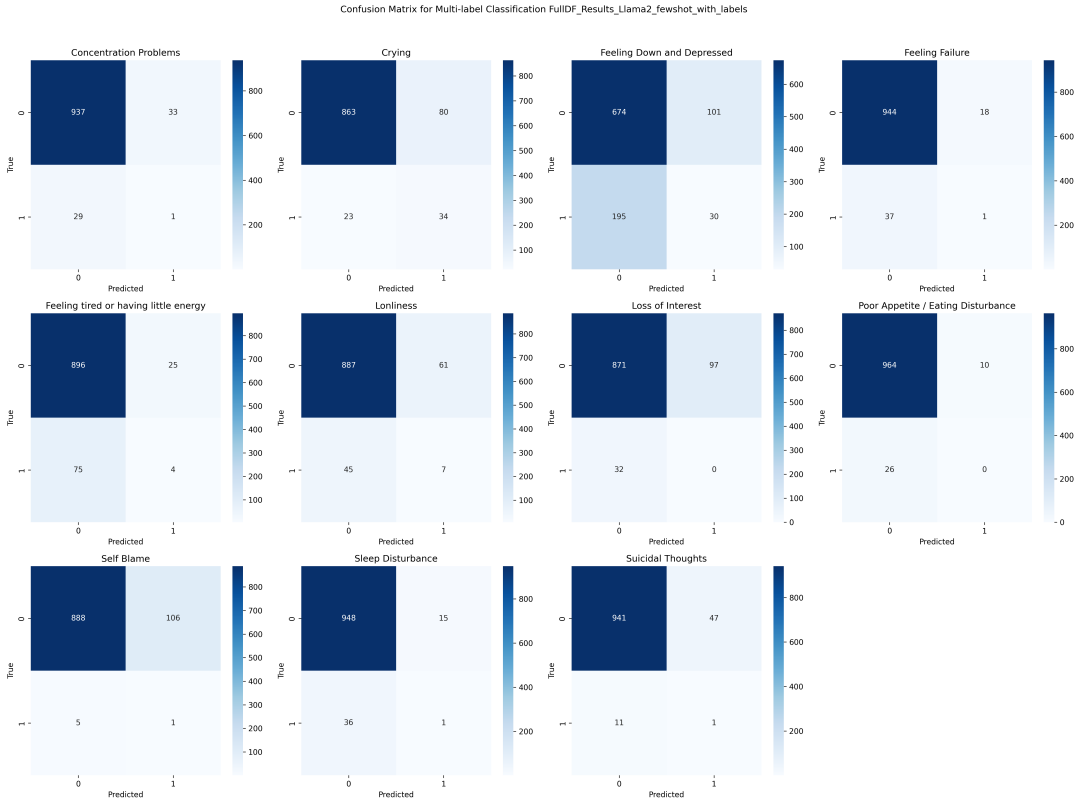

Figure 14: Confusion matrices for LLaMA-7B in the few-shot setting.

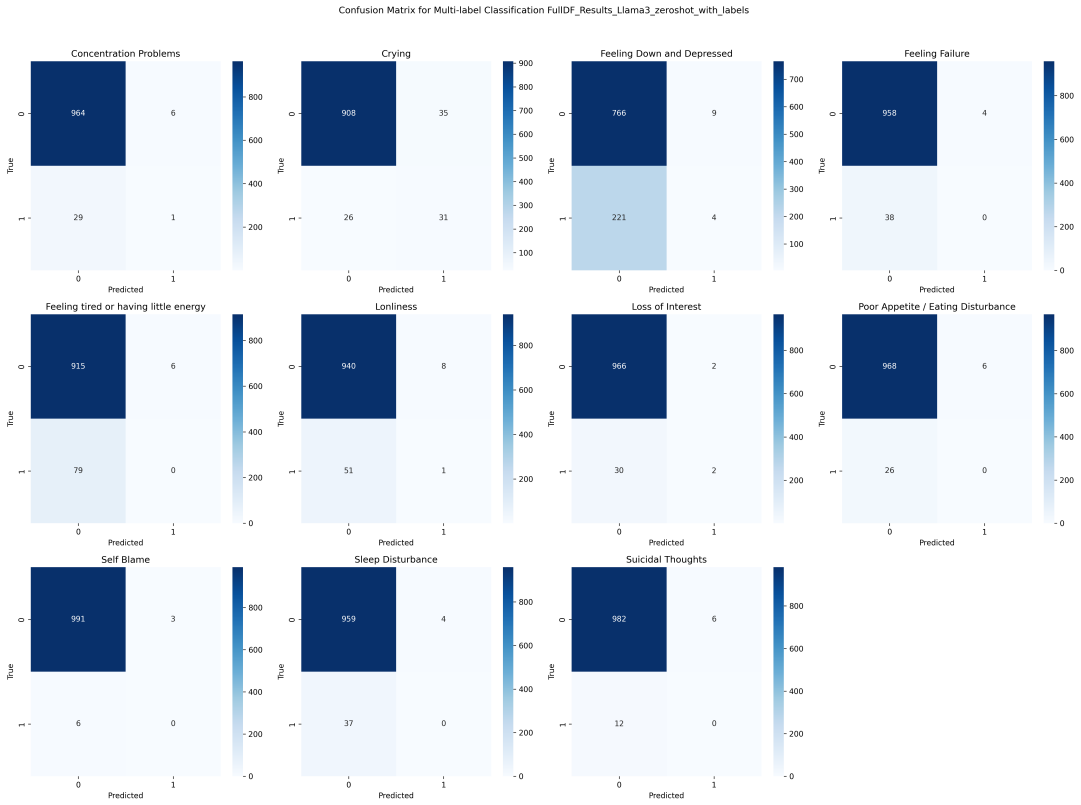

Figure 15: Confusion matrices for LLaMA3-8B in the zero-shot with labels setting.

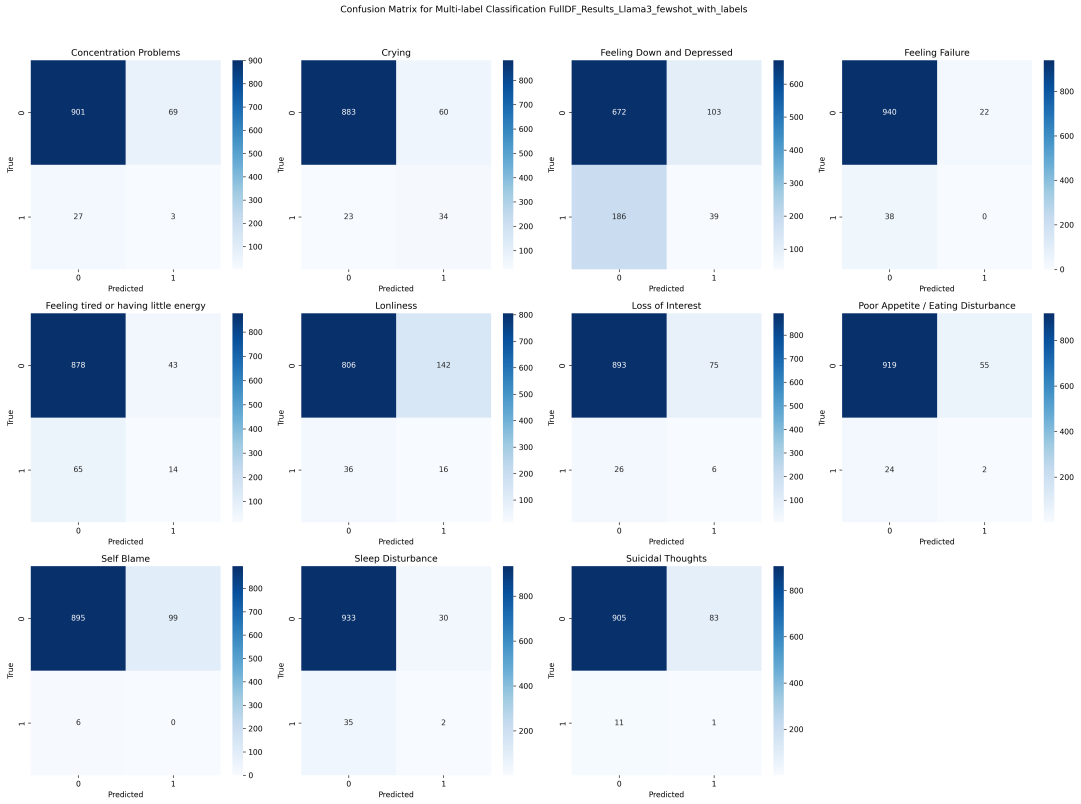

Figure 16: Confusion matrices for LLaMA3-8B in the few-shot setting.