Still Not Quite There! Evaluating Large Language Models for Comorbid Mental Health Diagnosis

Amey Hengle ♥*   **Atharva Kulkarni**♠*   **Shantanu Patankar** ♣
Madhumitha Chandrasekaran♡   **Sneha D'Silva**◇   **Jemima Jacob**♧   **Rashmi Gupta**◇
♥Skit.ai   ♠University of Southern California   ♣Georgia Institute of Technology
♡Sion Hospital   ◇Clinical Psychotherapist   ♧Sophia College for Women
◇Indian Institute of Technology Bombay
Correspondence: ameyhengle@gmail.com

Abstract

Warning: This paper includes examples displaying symptoms of mental health disorders for contextual understanding.

In this study, we introduce ANGST, a novel, first of its kind benchmark for depression-anxiety comorbidity classification from social media posts. Unlike contemporary datasets that often oversimplify the intricate interplay between different mental health disorders by treating them as isolated conditions, ANGST enables multi-label classification, allowing each post to be simultaneously identified as indicating depression and/or anxiety. Comprising 2876 meticulously annotated posts by expert psychologists and an additional 7667 silver-labeled posts, ANGST posits a more representative sample of online mental health discourse. Moreover, we benchmark ANGST using various state-of-the-art language models, ranging from Mental-BERT to GPT-4. Our results provide significant insights into the capabilities and limitations of these models in complex diagnostic scenarios. While GPT-4 generally outperforms other models, none achieve an F1 score exceeding 72% in multi-class comorbid classification, underscoring the ongoing challenges in applying language models to mental health diagnostics.

1 Introduction

With the advancement of web technologies, social media platforms such as Reddit (Gkotsis et al., 2016; Gaur et al., 2018), Twitter (De Choudhury et al., 2013a; Coppersmith et al., 2014a), and ReachOut (Shandley et al., 2010; Kahl et al., 2020), become popular hubs for mental health support and information exchange. They offer an anonymous, safe space that fosters a sense of community and empowerment that begets in-depth mental health

discussions (De Choudhury and De, 2014; Berry et al., 2017). Consequently, there is a growing body of research on "*digital psychiatry*" (Tsugawa et al., 2015) that analyzes the mental health discourse and language usage on these platforms to enhance the discovery, understanding, and detection of mental health concerns. However, despite significant efforts, there exist numerous concerns regarding the curation of these resources.

Drawbacks of Existing Data Resources. The prevalent modus operandi for sourcing mental health-related datasets involves crawling social media posts that are either intrinsic to specific online communities pertaining to mental health (Turcan and McKeown, 2019) or those that bear certain attributive words (Mowery et al., 2016), hashtags (Berry et al., 2017), or self-reported diagnoses (Coppersmith et al., 2015; Yates et al., 2017). The binary indicator of the presence or absence of these signals is used to determine the positive class, while the negative class[1] constitutes randomly crawled posts. This strategy engenders highly biased data and skewed annotations incongruous with the real-world distribution. Furthermore, it exacerbates the semantic divergence between the positive and negative classes, simplifying the task and diminishing its utility in real-world settings. We term this issue as the **'*Data Source Bottleneck*'**.

Moreover, these '*proxy diagnostic signals*' of affiliation behavior (hashtags/community membership) or self-reports of diagnosis provide an easy and inexpensively means of collecting silver-labeled data, obviating the need for hiring professional annotators. However, as elucidated by Ernala et al. (2019), they lack clinical grounding, theoretical contextualization, and psychometric

---

*Equal contribution

[1] In this study, we refer to the 'control group' more from a machine learning perspective – as the 'negative class,'. That is, 'control group' represents posts that do not contain the target labels (depression or anxiety).

validity. They presents a rather myopic approach to labeling data as it does not consider the text's semantics and focuses solely on attribution. As unearthed by Resnik et al. (2013) and Kulkarni et al. (2021), people post a variety of content in mental health-related communities, ranging from trauma episodes and emotion regulation to general advice and therapy experiences. Thus, conglomerating them under a single label leads to noisy annotations, spurious correlations, and unwanted biases. We term this quandary the **'*Annotation Bottleneck*'**.

Lastly, most existing works treat mental health disorder identification as binary classification tasks – differentiating posts exhibiting symptoms of a particular mental illness from those of the control users. While some endeavors have been made towards multi-class classification of multiple mental health disorders (Cohan et al., 2018; Souza et al., 2020; Garg et al., 2022; Ji et al., 2022a), they fail to capture the main narrative – Mental health conditions are not mutually exclusive. In reality, conditions like depression and anxiety often manifest concurrently, and a social media post may reflect both conditions simultaneously, necessitating multiple labels. For example, Borba de Souza et al. (2022) identify persistent traits of anxiety in the Reddit posts of individuals experiencing depression. This phenomenon, known as comorbidity, is well-documented in psychology research (Hirschfeld, 2001; Lamers et al., 2011). By overlooking this crucial aspect, current frameworks oversimplify the complex interplay of mental health disorders, thereby limiting our ability to develop comprehensive solutions for this multifaceted challenge. We term this drawback the **'*Task Bottleneck*'**.

Proposed Dataset. To this end, we introduce **ANGST** (**AN**xiety-Depression Comorbidity Dia**G**nosis in Reddit Po**ST**), a novel corpus that addresses the *data source*, *annotation*, and *task bottleneck* issue discussed above. Specifically, we frame comorbidity identification as a multi-label classification task, wherein a post can be labeled as indicative of depression, anxiety, both, or neither. Likewise, ANGST can be used for traditional binary classification of control vs mental health group. ANGST comprises 2876 Reddit posts meticulously annotated by 3 expert psychologists. Our scrupulous data acquisition and filtering pipeline ensures that ANGST represents data proximate to the real-world distribution of online

mental health discourse. The depression, anxiety, and control class samples exhibit reduced semantic divergence, rendering them more challenging to distinguish than in contemporary datasets and, thus, better reflecting the nuances of real-world data. Furthermore, we also compile ANGST-SILVER, a silver-labeled corpus of 7667 Reddit posts for comorbidity classification. This supplementary dataset is sourced from the same distribution as ANGST, but its annotations are derived through a carefully tuned prompting technique using GPT-3.5-turbo (Ouyang et al., 2022). It provides an ancillary resource for exploring semi-supervised and few-shot learning paradigms in the domain of mental health comorbidity analysis.

Benchmarking Methods. We benchmark ANGST, on a host of discriminative pretrained models (PLMs) like Mental-BERT, Mental-RoBERTa, Mental-XLNet, and Mental-LongFormer (Ji et al., 2022b, 2023), along with large generative language models (LLMs) like LLama-2 (Touvron et al., 2023), GPT-3.5-turbo (Ouyang et al., 2022), and GPT-4 (Achiam et al., 2023). Our results unveil intriguing findings. While GPT-4 emerges as the top performer overall, the pretrained BERT-style models fine-tuned on ANGST-SILVER are formidable competitors. Notably, PLMs almost always outperform GPT-3.5-turbo in binary classification setups and Llama-2 in both comorbidity and binary classification tasks. However, none of the models achieve an F1 score exceeding 72% in the comorbid multi-class classification or surpass 69% and 75% in depression and anxiety binary classification, respectively. These insights prompt reconsideration on the suitability of current language models for sensitive and nuanced tasks like mental health diagnosis (De Choudhury et al., 2023; Timmons et al., 2023).

Contributions. In summary, we make the following contributions: [2]
- **Novel Dataset** – We curate ANGST, a meticulously crafted, gold-labeled, and neutrally-seeded Reddit corpus tailored for depression-anxiety comorbidity classification. ANGST can be employed in both multi-label and binary classification setups. Section 3 details the data sourcing, filtering, and annotation processes.

---

[2]The source code and dataset of this study are available at https://github.com/AmeyHengle/ANGST

- **Silver Labels** – We compile `ANGST-SILVER`, a silver-labeled dataset, which can complement `ANGST` for few-shot learning or supervised fine-tuning (Section A.1).
- **Cross-dataset Analysis** – We conduct a comprehensive examination of `ANGST` across various facets in comparison to existing mental health corpora, highlighting its unique characteristics (Section 4).
- **Large-Scale Benchmarking** – We evaluate numerous pretrained language models (PLMs) and large language models (LLMs) on `ANGST`, delineating their respective advantages and shortcomings (Section 6).
- **Qualitative and Quantitative Evaluation** – We thoroughly analyze the results across different metrics, complemented by an in-depth error analysis, providing a holistic assessment of model performance (Section 7).

2 Related Work

Corpora related to Mental Health Disorders. Since the past decade, social media platforms have been actively used for compiling datasets for various mental health issues (De Choudhury et al., 2013a,b; Tsugawa et al., 2015; Pedersen, 2015). For instance, CLPsych15 (Coppersmith et al., 2015), RSDD (Yates et al., 2017), Depression Reddit (Pirina and Çöltekin, 2018), and Dreaddit (Turcan and McKeown, 2019) are commonly used corpora for depression analysis. Similarly, DATD (Owen et al., 2020) and the Anxiety on Reddit corpus (Shen and Rudzicz, 2017) are popular benchmarks for anxiety issues. UMD (Shing et al., 2018) and T-SID (Ji et al., 2021) are compiled exclusively for identifying suicide ideation.

The CLPsych16 dataset (Milne et al., 2016) supports multi-label classification, covering depression, control, and PTSD categories. Additionally, the CAMS (Garg et al., 2022) and SAD (Mauriello et al., 2021) datasets offer more granular classifications, categorizing the causes of depression into six and nine categories, respectively. On a broader scale, SMHD (Cohan et al., 2018), SWMH (Ji et al., 2022a), and CAMS (Garg et al., 2022) encompass a more comprehensive coverage of several mental health disorders. Yang et al. (2024) aggregated multiple mental health-related datasets into a unified benchmark named IMHI. More recently, Jin et al. (2023) proposed PsyEval, a suite of mental health-related tasks designed to evaluate the performance of large language models.

Computational Models for Mental Health Disorder Identification. Early approaches for identifying mental health disorders from social media platforms relied heavily on feature engineering and traditional machine learning classifiers (De Choudhury et al., 2013a,b; Coppersmith et al., 2014b; Mitchell et al., 2015; Tsugawa et al., 2015). Subsequent research endeavors aimed to obviate the need for hand-crafted features by employing neural network architectures such as LSTMs and CNNs, enabling more accurate identification of mental health conditions (Sawhney et al., 2018; Tadesse et al., 2020). Borba de Souza et al. (2022) propose a deep learning ensemble method that effectively classifies anxiety, depression, and their comorbidity using Reddit posts. Another line of research explores the use of mental health questionnaires (Nguyen et al., 2022), multi-task learning (Sarkar et al., 2023), and hierarchical attention networks (Han et al., 2022) to augment existing architectures. In recent years, numerous transformer-based models pre-trained on mental health-related data, such as Mental-BERT, Mental-RoBERTa, Mental-XLNet, and Mental-LongFormer (Ji et al., 2022b, 2023), have been developed and released. More recently, several open-source mental health-focused LLMs, including Mental-Flant-T5, Mental-Alpaca (Xu et al., 2024), and Mental-LLama (Yang et al., 2024) have been introduced. Furthermore, endeavors such as that of Chen et al. (2023), Yang et al. (2023a), and Yang et al. (2023b) provide in-depth analyses of proprietary LLMs like chatGPT, GPT-3, and GPT-4, evaluating their performance in identifying various mental health conditions.

3 Curating `ANGST`

3.1 Data Collection and Filtering

`ANGST` is compiled from publicly available Reddit posts. As illustrated in Table 1, each post in `ANGST` is labeled in a multi-label fashion to indicate depression and/or anxiety. We adhere to established protocols for data collection as described in prior research on constructing mental health datasets (Yates et al., 2017; Cohan et al., 2018). To begin with, we compile a list of relevant mental-health-related Subreddits informed by prior investigations into depression and anxiety on Reddit (Pavalanathan and De Choudhury, 2015; Yates et al., 2017; Cohan et al., 2018). We employ PRAW (Python Reddit

Example	Depression	Anxiety
I no longer hate myself in the same way. I feel somewhat better now that I've accepted my ugly appearance, shy personality, and few mental health issues. Before, I used to despise myself for these things. Although I am aware that I will never be loved, have a successful career, be a good friend, or find a girlfriend, I no longer feel horrible about myself because I accept how boring and ugly I am. The only thing that worries me is that I am more foolish than I care to acknowledge.	✗	✗
I'm in pain all the time. Really, what's the purpose of existence? I literally experience daily, unrelenting emotional pain. and the best i can do is distract myself from it. I literally experience constant emotional pain every single day. Most of the time, all I want to do is curl up into a ball and cry my eyes out, but I'm so disconnected that I can't cry anymore.. I mean, I won't kill myself, but when every day is agony, how is it possible for anyone to really think that I would want to live? Nothing ever seems to improve the situation, and I can feel my condition gradually deteriorating every day. I wish I could just vanish.	✓	✗
What just I experienced. I suddenly experienced a wave of extreme anxiety around 9:40 p.m. I had to lock every door, but I swear I saw someone when I peered outside. I swear I saw someone, but they would have to be taller than seven feet. My dad hurried to tell my parents after he went outside and saw nobody. This should normally calm people down, right? It didnt ease me. I still feel as though I'm going to die or am in danger. I couldnt stop shaking and I was only able to stop crying at 10 p.m. My chest and head hurt really bad now. Was this an extreme case of anxiety, or something else?	✗	✓
Stuck after graduation. August was when I graduated, and I still haven't found employment in my field. I continue to work a few hours a week at my menial retail job, and because I am so severely depressed and lonely, I am afraid to apply to as many jobs as I should because I feel incompetent and afraid, literally frozen with fear. I'll die over it. I spend everyday in a constant cycle of loneliness and anxiety and worry and panic and emptiness. I feel like I'm burdening my family more and more every day because I can't seem to get a job and get my life started. I simply want to put an end to this years-long suffering because it keeps getting worse. I despise myself for it, Due to family issues, I've literally cut off from my friends and am dreading Christmas. and talk about why Im not working yet.	✓	✓

Table 1: Examples of depression, anxiety, comorbidity, and control posts in ANGST. Red signifies signs of depression and Blue highlights anxiety symptoms. All of the posts in this example have been paraphrased in order to protect the user's identity.

API Wrapper) [3] to scrape all publicly available Reddit posts in these Subreddits over five years from January 2018 to December 2022. This resulted in approximately 400,000 anonymous posts.

Next, as outlined by Yates et al. (2017), we identified authors who had self-disclosed their experiences with depression, anxiety, or related disorders using high-precision diagnostic patterns (Cohan et al., 2018). We identified 8,083 unique authors who had made at least one self-disclosure post. Subsequently, we compiled all posts made by these authors, resulting in a corpus of 185,064 posts. Further refining this dataset, we excluded posts shorter than 75 words and those authored by individuals with fewer than 10 posts, producing a more concentrated corpus of 72,583 posts. From this filtered corpus, we selected the top 25,000 posts that exhibited the highest linguistic alignment with depression and anxiety cues. Specifically, we identified and ranked posts based on their Empath (Fast et al., 2016) and NRC (Mohammad and Turney, 2013, 2010) sentiment scores. We use the Empath and NRC scores to prioritize posts that were more likely to contain pertinent emotional and psychological indicators associated with depression and/or anxiety, in light of prior research that highlights their effectiveness (Liu et al., 2022; Castiglioni et al., 2023; Zhang et al., 2023). Appendix Table 7 provides examples of posts with low, moderate, and high Empath and NRC scores, respectively. As illustrated in the table, posts with higher scores show a stronger alignment with the linguistic cues of depression and anxiety. Finally, we randomly sample 3000 posts from this pool to form our final annotation corpus.

3.2 Data Annotation

We employed two primary and one secondary annotator, each a trained professional psychologist. To preserve objectivity and prevent potential biases, we established a protocol ensuring complete anonymity between the annotators. Specifically, no additional information or context beyond the post's content was shared with them. Moreover, the annotators worked in isolation, ensuring their annotations were performed independently, precluding any collaborative influence. The primary role of the annotators was twofold. First, they discerned and labeled if a post signified depression or anxiety, indicating their decision with a clear "yes" or "no." Second, highlight specific statements in the posts to support the validity of their judgment (see Table 1). The secondary annotator is used to resolve conflicting annotations. Additionally, to protect the identity of users, the annotators were asked to flag any posts that might

---

[3] https://praw.readthedocs.io/en/stable

Dataset	Label 1	Label 2	JSD		MMD	
			Unigram	Bigram	Mean	Median
ANGST	control	comorbid	0.036	0.154	0.047	0.043
	control	depression	0.027	0.142	0.047	0.043
	control	anxiety	0.034	0.150	0.048	0.045
	depression	anxiety	0.014	0.070	0.041	0.038
	depression	comorbid	0.002	0.011	0.042	0.038
	anxiety	comorbid	0.013	0.068	0.041	0.038
DATD	control	anxiety	0.104	0.182	0.083	0.056
Dep Reddit	control	depression	0.060	0.165	0.082	0.073
Dreaddit	control	Stress	0.056	0.174	0.067	0.063
SDCNL	depression	Suicide	0.031	0.150	0.070	0.060

Table 2: Inter-class similarity in different mental health datasets measured via Jensen–Shannon Divergence (JSD) and Maximum Mean Discrepancy (MMD). The lower the values of these metrics, the harder it will be to separate the classes.

inadvertently disclose user details, including names, demographics, or other personal identifiers. This resulted in some posts being flagged, and the final remaining corpus consisted of 2,876 posts, forming our ANGST dataset.

We report the inter-annotator agreement between our primary annotators for both depression and anxiety labels using Krippendorff's alpha (α) (Krippendorff, 2011), and Fleiss kappa (κ) (Cohen, 1960). For depression, the agreement scores are $\alpha = 0.622$ and $\kappa = 0.624$, while for anxiety, the scores are $\alpha = 0.423$ and $\kappa = 0.444$. These metrics indicate a moderate level of agreement, within the acceptable range of 0.4 to 0.7. Notably, the agreement on depression labels is higher compared to anxiety, reflecting a more consistent identification of depressive symptoms among the annotators.

4 Yet Another Mental Health Dataset?

In this section, we conduct a comprehensive cross-sectional analysis of ANGST. We scrutinize it across various facets and juxtapose it with existing mental health corpora, thereby underlining its unique characteristics and distinguishing features. Specifically, we compare ANGST against SDCNL (Haque et al., 2021), Depression Reddit (Pirina and Çöltekin, 2018) Dreaddit (Turcan and McKeown, 2019), and DATD (Owen et al., 2020). We employ the corpora comparison strategy as outlined by Kulkarni et al. (2023).

4.1 Inter-Class Similarity

Hypothesis – The more similar the inter-class samples, the more difficult it becomes to classify the dataset. We hypothesize that due to its neutral-seeding policy, ANGST would exhibit higher inter-

class similarity than its counterparts, thereby rendering it more challenging.

Metrics – We utilize the Jensen–Shannon divergence (JSD) (Dagan et al., 1997) and Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) as measures to quantify the proximity between different classes in a dataset. To calculate MMD, we use RoBERTa embeddings (Liu et al., 2019). The transformer weights were frozen, and a classification head with the required number of class labels was employed for each experiment. A lower value of JSD or MMD indicates that the class distributions are more similar, making it more challenging to classify accurately.

Experiments – we generate a Laplacian smoothed unigram and bigram distribution for each dataset for each class. Table 2 presents the Jensen–Shannon divergence (JSD) of the pairwise selection of labels. For ANGST, the JSD of depression, anxiety, and comorbidity versus the control group are 0.027, 0.034, and 0.036, respectively, suggesting that the comorbid and control examples exhibit the least similarity. The JSD of anxiety with depression and comorbidity is 0.014 and 0.013, while that of depression and comorbidity is 0.002. This low divergence is what renders the dataset arduous to classify and is the cause for the high disagreement in the non-control classes.

When extending the above experiment with other datasets, we observe that the pairwise JSD of ANGST is lower than that of DATD (0.104), Depression Reddit (0.060), and Dreaddit (0.056). The higher divergence in these datasets is expected, as their control group samples are sourced from non-mental health contexts. Thus, ANGST serves as a superior benchmark, as it captures the vital yet minute differences between anxiety and depression more effectively. Results for MMD are described in Section A.2.

4.2 Adversarial Validation

Hypothesis – Data drift (Lu et al., 2019) quantifies the change in the feature space between two datasets. All samples in the old (source) dataset are considered the negative class, and all the samples in the new (target) dataset are deemed the positive class. A simple classifier is trained for this binary classification task. A high performance suggests the presence of discriminatory features between the two datasets. We conjecture that due to its

meticulous data curation and filtering pipeline and its gold annotation scheme, `ANGST` will exhibit significant differences compared to existing mental health datasets.

Metrics – We analyze the performance based on accuracy, macro-F1, ROC-AUC scores, and Matthews Correlation Coefficient.

Experiment – The training split (T) from the source dataset $X_{\text{src}}^{(T)}$ is labeled as $(X_{\text{src}}^{(T)} = 0)$, and `ANGST`'s training split $X_{\text{tgt}}^{(T)}$ is labeled as $(X_{\text{tgt}}^{(T)} = 1)$. We utilize an n-gram-based TF-IDF logistic regression model to detect the data drift of `ANGST` concerning the chosen datasets. As shown in Table 8, none of the dataset comparisons result in a ROC-AUC less than 0.5 or an MCC approximately equal to 0, which would have suggested that our dataset is indistinguishable from existing datasets. On the contrary, the results demonstrate variation within a narrow range of 0.096 for ROC-AUC and 0.125 for MCC. The highest and lowest ROC-AUC scores are 1.0 and 0.904, obtained from SDCNL and DATD, respectively, while the corresponding MCC scores are 0.990 and 0.875. This indicates that `ANGST` exhibits an inherent difference from existing corpora, which can be attributed to its meticulous data curation, filtering pipeline, and gold annotation scheme.

5 Experimental Setup

5.1 Models

We benchmark `ANGST` using both discriminative and generative language models. The former include Mental-BERT$_{\text{base}}$, Mental-BERT$_{\text{large}}$, Mental-RoBERTa$_{\text{large}}$, Mental-XLNet$_{\text{base}}$, and Mental-LongFormer$_{\text{base}}$ (Ji et al., 2022b, 2023). These models have undergone continued pre-training (Gururangan et al., 2020) on extensive mental health-related data collected from Reddit, demonstrating state-of-the-art performance on various mental health corpora. Amongst generative language models, we report results on Llama-2 chat models (7B and 13B) (Touvron et al., 2023), GPT3.5-turbo (Ouyang et al., 2022), and GPT-4 (Achiam et al., 2023). We report results for both the zero and few-shot settings.

5.2 Evaluation Criteria

We benchmark `ANGST` on two tasks, providing alternative but complementary viewpoints for our analysis. They are as follows:

- **Multi-label classification** where a post can simultaneously be classified as depression, anxiety, both, or None.

- **Binary classification** for depression vs non-depression and anxiety vs non-anxiety.

We report the standard weighted precision, recall, and F1 scores for both multi-label and binary classification tasks. Additionally, we provide macro-F1 scores for both tasks. In the case of multi-label classification, we also report the Hamming loss (Tsoumakas and Katakis, 2007), which measures the fraction of labels for which the predicted and actual labels disagree.It is calculated as the average number of incorrect labels divided by the total number of labels.

6 Results and Discussion

6.1 Binary Classification.

The overall binary classification results for depression and anxiety are presented in Table 3. We observe that discriminative models exhibit consistent performance across both tasks, outperforming much larger models such as LLama-2 and GPT3.5-turbo. Apart from GPT-4, all other generative models show considerable variance in performance in both zero- and few-shot settings. Furthermore, there is a noticeable difference in how models handle depression versus anxiety.

Depression results – The fine-grained results for depression classification presented in Table 4, demonstrate a differential performance across labels. Models consistently exhibit higher recall and F1 scores for the depression label compared to the control group, indicating a stronger ability to identify true cases of depression while accepting lower precision. This trend signifies a model's bias towards minimizing false negatives, thus reducing the risk of missing actual depression cases. The comparison between discriminative models and GPT-4 shows only a slight difference. Both GPT-4 and Mental-XLNet achieve similar F1 scores; however, GPT-4 exhibits superior performance in classifying the control groups. This suggests that while the models are effective in identifying depression, there exists a precision trade-off that prioritizes avoiding missed diagnoses.

Anxiety results – Table 5 shows the fine-grained results for anxiety classification. GPT-4 significantly outperforms all other discriminative

Model	Type	Depression vs. Control					Anxiety vs. Control				
		Acc	Precision	Recall	F1	macro-F1	Acc	Precision	Recall	F1	macro F1
Mental-BERT$_{base}^{‡}$	SFT	$68.6_{1.7}$	$\mathbf{77.5_{1.3}}$	$68.6_{1.7}$	$64.3_{2.6}$	$62.8_{2.7}$	$\mathbf{78.3_{1.1}}$	$79.9_{1.8}$	$\mathbf{78.8_{1.1}}$	$78.3_{1.3}$	$72.2_{2.1}$
Mental-BERT$_{large}$	SFT	$68.7_{1.9}$	$76.6_{0.9}$	$68.7_{1.9}$	$64.6_{3.0}$	$63.1_{3.3}$	76.5_{1}	$77.6_{1.8}$	$76.5_{1.0}$	$76.9_{1.2}$	$69.2_{2.2}$
Mental-RoBERTa$_{large}$	SFT	$68.9_{1.5}$	$76.4_{1.3}$	$68.9_{1.5}$	$65.0_{2.3}$	$63.6_{2.4}$	$76.2_{0.9}$	$77.7_{1.8}$	$76.2_{0.9}$	$76.7_{1.1}$	$69.3_{2.2}$
Mental-XLNet$_{base}^{‡}$	SFT	$69.8_{1.0}$	$76.9_{1.5}$	$69.8_{1.0}$	$66.3_{1.5}$	$65.0_{1.6}$	$76.9_{0.8}$	$78.2_{1.6}$	$76.9_{0.8}$	$77.4_{1.0}$	$70.0_{1.9}$
Mental-LongFormer$_{base}$	SFT	$68.7_{1.1}$	$76.3_{1.1}$	$68.7_{1.1}$	$64.7_{1.7}$	$63.3_{1.8}$	$77.5_{1.1}$	$78.8_{2.0}$	$77.5_{1.1}$	$77.9_{1.3}$	$70.8_{2.2}$
LLama-2-chat-7B	zero-shot	$55.7_{0.5}$	$57.3_{2.2}$	$55.7_{0.5}$	$43_{1.2}$	$39.9_{1.3}$	$25.4_{0.4}$	$66.2_{3.5}$	$25.4_{0.4}$	$12_{0.9}$	$21.2_{0.6}$
	few-shot	$51.6_{0.1}$	$52_{0.1}$	$51.6_{0.1}$	$51.7_{0.1}$	$51.5_{0.1}$	43.8_{1}	$64.1_{0.8}$	43.8_{1}	$46.5_{1.1}$	$43_{0.9}$
LLama-2-chat-13B	zero-shot	$60.9_{0.2}$	$62.5_{0.2}$	$60.9_{0.2}$	$56.6_{0.4}$	$54.9_{0.4}$	$39.8_{0.5}$	$68.1_{0.3}$	$39.8_{0.5}$	$39.8_{0.7}$	$39.8_{0.5}$
	few-shot	$46.2_{0.9}$	$49.1_{1.5}$	$46.2_{0.9}$	41.1_{1}	$42.6_{0.9}$	54.9_{1}	$62.7_{0.8}$	54.9_{1}	$57.7_{0.9}$	48.2_{1}
GPT-3.5-turbo	zero-shot	$64.1_{0.1}$	$71.3_{0.1}$	$64.1_{0.1}$	$58.3_{0.1}$	$56.5_{0.1}$	$41.2_{0.1}$	$77.3_{0.1}$	$41.2_{0.1}$	$39.6_{0.1}$	$41.0_{0.1}$
	few-shot	$64.8_{0.0}$	$70.1_{0.1}$	$64.8_{0.0}$	$60.2_{0.0}$	$58.6_{0.0}$	$38.4_{0.1}$	$77.2_{0.1}$	$38.4_{0.1}$	$35.4_{0.2}$	$38.0_{0.2}$
GPT-4	zero-shot†	$\mathbf{70.7_{0.0}}$	$74.7_{0.2}$	$\mathbf{70.7_{0.0}}$	$\mathbf{68.4_{0.1}}$	$\mathbf{67.4_{0.1}}$	$77.6_{0.1}$	$\mathbf{83.3_{0.0}}$	$77.6_{0.1}$	$\mathbf{78.9_{0.1}}$	$\mathbf{74.0_{0.1}}$
	few-shot	$65.5_{0.1}$	$76.7_{0.1}$	$65.5_{0.1}$	$59.4_{0.1}$	$57.6_{0.1}$	$68.1_{0.2}$	$82.5_{0.1}$	$68.1_{0.2}$	$70.3_{0.2}$	$66.2_{0.2}$
$\Delta_{model^{†}-model^{*}}$		↑ 0.9	↓ 2.2	↑ 0.9	↑ 2.1	↑ 2.4	↓ 0.	↑ 3.4	↓ 1.2	↑ 0.6	↑ 1.8

Table 3: Results of various methods for depression and anxiety binary classification.

Model	Type	Depression			Control		
		Precision	Recall	F1	Precision	Recall	F1
Mental-BERT$_{base}$	SFT	$64.0_{1.3}$	$98.2_{0.8}$	$77.5_{0.9}$	$94.0_{2.5}$	$32.5_{4.2}$	$48.1_{4.6}$
Mental-BERT$_{large}$	SFT	$64.3_{1.7}$	$97.3_{1.6}$	$77.4_{0.8}$	$91.7_{3.2}$	$33.7_{5.7}$	$48.9_{5.8}$
Mental-RoBERTa$_{large}$	SFT	$64.1_{1.3}$	$97.2_{1.2}$	$77.5_{0.8}$	$91.1_{2.9}$	$34.4_{4.1}$	$49.7_{4.2}$
Mental-XLNet$_{base}^{*}$	SFT	$65.2_{0.9}$	$97.0_{1.7}$	$77.9_{0.6}$	$91.2_{3.6}$	$36.7_{3.3}$	$52.2_{2.9}$
Mental-LongFormer$_{base}$	SFT	$64.3_{0.9}$	$97.2_{1.0}$	$77.4_{0.6}$	$91.0_{2.7}$	$33.9_{3.1}$	$49.3_{3.2}$
LLama-2-chat-7B	zero-shot	$55.6_{0.3}$	$97.3_{0.2}$	$70.7_{0.2}$	$59.3_{4.4}$	$5_{1.4}$	$9.2_{2.4}$
	few-shot	$56.5_{0.1}$	$52.4_{0.5}$	$54.3_{0.3}$	$46.6_{0.1}$	$50.7_{0.5}$	$48.6_{0.2}$
LLama-2-chat-13B	zero-shot	$59.7_{0.2}$	$88.4_{0.1}$	$71.3_{0.1}$	$65.8_{0.3}$	$27.2_{0.9}$	$38.5_{0.9}$
	few-shot	$52.9_{2.3}$	$19.1_{0.8}$	$28.1_{1.2}$	$44.5_{0.6}$	$79.2_{1.1}$	$57_{0.7}$
GPT-3.5-turbo	zero-shot	$61.0_{0.0}$	$96.1_{0.0}$	$74.6_{0.0}$	$84.0_{0.2}$	$24.9_{0.1}$	$38.4_{0.1}$
	few-shot	$61.8_{0.0}$	$94.2_{0.1}$	$74.6_{0.0}$	$80.3_{0.1}$	$28.9_{0.1}$	$42.5_{0.1}$
GPT-4	zero-shot†	$\mathbf{66.7_{0.1}}$	$93.5_{0.3}$	$\mathbf{77.9_{0.0}}$	$84.5_{0.4}$	$\mathbf{42.9_{0.4}}$	$\mathbf{56.9_{0.2}}$
	few-shot	$61.6_{0.0}$	$\mathbf{98.9_{0.2}}$	$75.9_{0.0}$	$\mathbf{95.0_{0.2}}$	$24.7_{0.2}$	$39.3_{0.2}$
$\Delta_{model^{†}-model^{*}}$		↑ 1.5	↓ 3.5	0.0	↑ 6.	↑ 6.1	↑ 4.

Table 4: Detailed results per label for depression

Model	Type	Anxiety			Control		
		Precision	Recall	F1	Precision	Recall	F1
Mental-BERT$_{base}^{*}$	SFT	$\mathbf{54.7_{1.9}}$	$64.9_{7.7}$	$59.2_{3.7}$	$88.0_{2.1}$	$82.6_{2.1}$	$\mathbf{85.2_{0.8}}$
Mental-BERT$_{large}$	SFT	$51.8_{2.0}$	$58.0_{9.1}$	$54.3_{4.5}$	$86.0_{2.2}$	$82.5_{2.9}$	$84.1_{0.8}$
Mental-RoBERTa$_{large}$	SFT	$51.0_{1.7}$	$59.8_{8.9}$	$54.8_{4.3}$	$86.4_{2.2}$	$81.5_{2.7}$	$83.8_{0.7}$
Mental-XLNet$_{base}$	SFT	$52.5_{1.6}$	$59.8_{8.6}$	$55.6_{3.9}$	$86.5_{2.1}$	$82.5_{2.9}$	$84.4_{0.8}$
Mental-LongFormer$_{base}$	SFT	$53.3_{1.8}$	$61.7_{8.8}$	$56.9_{4.1}$	$87.1_{2.4}$	$82.6_{2.4}$	$84.7_{0.7}$
LLama-2-chat-7B	zero-shot	$24.5_{0.1}$	$98.8_{0.1}$	$39.2_{0.1}$	$79.7_{4.6}$	$1.7_{0.6}$	$3.2_{1.2}$
	few-shot	$25_{0.5}$	$65.2_{1.2}$	$36.2_{0.7}$	$76.7_{0.9}$	$36.9_{1.2}$	$49.8_{1.2}$
LLama-2-chat-13B	zero-shot	$26.4_{0.1}$	$81.6_{0.7}$	$39.9_{0.2}$	$81.6_{0.4}$	$26.3_{0.8}$	$39.8_{0.9}$
	few-shot	23.9_{1}	$38.7_{1.3}$	$29.5_{1.2}$	$75.2_{0.7}$	$60.1_{0.9}$	$66.8_{0.8}$
GPT-3.5-turbo	zero-shot	28.6_{0}	$94.4_{0.1}$	$43.9_{0.1}$	$93.0_{0.1}$	$24.0_{0.1}$	$38.1_{0.2}$
	few-shot	27.8_{0}	$95.5_{0.1}$	$43.1_{0.1}$	$93.2_{0.1}$	$20.0_{0.2}$	$32.9_{0.3}$
GPT-4	zero-shot†	$52.5_{0.1}$	$83.0_{0.0}$	$\mathbf{64.4_{0.1}}$	$93.3_{0.0}$	$75.8_{0.1}$	$83.6_{0.1}$
	few-shot	$42.8_{0.2}$	$90.6_{0.2}$	$58.1_{0.2}$	$95.3_{0.1}$	$60.9_{0.2}$	$74.3_{0.2}$
$\Delta_{model^{†}-model^{*}}$		↓ 2.2	↑ 18.1	↑ 5.2	↑ 5.3	↓ 6.8	↓ 1.6

Table 5: Detailed results per label for anxiety.

and generative models in both zero-shot and few-shot settings, achieving the highest F1-scores for both anxiety (64.4%) and control (83.6%). Generally, all models demonstrate better performance for the control label across precision, recall, and F1-score compared to the anxiety label, indicating a consistent pattern of effectively identifying non-anxiety cases. Furthermore, there is a notable discrepancy between the precision and recall for the anxiety label, with precision significantly lower than recall. This discrepancy suggests a high rate of false positives, indicative of Type I errors. In practical diagnostic settings, this could lead to over-diagnosis. A possible cause for this issue is the imbalance in ANGST where anxiety cases are underrepresented compared to control cases, potentially causing the models to predict anxiety more frequently to capture most true anxiety instances.

6.2 Multi-Label Classification.

Table 6 presents the results for multi-label classification. We observe a consistent trend of high recall and moderate precision. This indicates that while the models excel at identifying relevant labels and minimizing false negatives, they incurr high false positives. Notably, the models demonstrate superior performance in predicting depression compared to anxiety, as evidenced by consistently higher F1 scores for depression. This disparity could stem from the more pronounced or readily learnable language features associated with depression in the dataset. GPT-3.5 Turbo with few-shot prompting stands out as the most effective, achieving an overall balanced F1 score of 71%, with noteworthy scores of 53% for depression and 17.9% for anxiety. On the other hand, PLMs that are characterized by lower Hamming Loss, adopt a more conservative prediction approach, likely reducing false positives but potentially missing some true positives. In contrast, GPT-4, with its higher F1 scores and greater Hamming Loss, adopts a more liberal approach, enhancing its ability to

Method	Prompt	Hamming Loss	Precision	Recall	**F1**	macro-F1	depression F1	anxiety F1
Mental-BERT$_{large}$	SFT	$28.3_{7.7}$	$59.4_{8.1}$	$85.2_{15.7}$	$70.0_{8.8}$	$65.2_{13.4}$	$53.3_{2.6}$	$16.4_{7.9}$
Mental-RoBERTa$_{large}$	SFT	$28.3_{3.6}$	$59.4_{4.0}$	$84.5_{14.6}$	$69.7_{7.5}$	$64.6_{10.1}$	$53.5_{3.0}$	$16.0_{5.4}$
Mental-XLNet$_{base}$	SFT	$\mathbf{28.1_{4.5}}$	$\mathbf{59.6_{4.5}}$	$84.7_{15.0}$	$69.9_{7.7}$	$64.8_{12.0}$	$\mathbf{53.7_{1.6}}$	$16.0_{7.1}$
Mental-LongFormer$_{base}$	SFT	$28.4_{6.5}$	$59.3_{7.2}$	$84.7_{18.7}$	$69.8_{10.3}$	$64.8_{15.4}$	$53.4_{2.8}$	$16.1_{9.0}$
LLama-2-chat-7B	zero-shot	$52.5_{0.1}$	$51.3_{0.2}$	$96.9_{0.4}$	$65_{0.1}$	$58_{0.1}$	52.7_{0}	$12.3_{0.1}$
	few-shot	$52.9_{0.7}$	$48.6_{0.5}$	$81.4_{1.1}$	$59.1_{0.7}$	$53.3_{0.6}$	$47.4_{0.6}$	$11.7_{0.2}$
LLama-2-chat-13B	zero-shot	$50.4_{0.3}$	$52.2_{0.1}$	$95.1_{0.5}$	$65.3_{0.2}$	$58.5_{0.3}$	$52.8_{0.2}$	$12.5_{0.1}$
	few-shot	$44.3_{1.1}$	$47.5_{1.8}$	28.1_{1}	34.2_{1}	$32.4_{1.3}$	$25.6_{0.4}$	$8.6_{0.7}$
GPT-3.5-turbo	zero-shot	$46.1_{10.2}$	$51.4_{5.3}$	$92.1_{9.4}$	$67.5_{4.5}$	$60.5_{4.5}$	$52.3_{3.1}$	$14.0_{1.4}$
	few-shot	$30.3_{11.7}$	$57.8_{8.0}$	$\underline{\mathbf{98.4_{0.7}}}$	$\underline{\mathbf{71.0_{3.9}}}$	$\underline{\mathbf{67.4_{5.6}}}$	$53.0_{0.9}$	$\mathbf{17.9_{3.1}}$
GPT-4	zero-shot	$34.0_{5.2}$	$55.6_{2.2}$	$97.4_{0.3}$	$70.8_{2.2}$	$66.3_{3.9}$	$52.9_{0.5}$	$17.5_{2.5}$
	few-shot	$35.6_{15.1}$	$54.8_{5.7}$	$97.4_{3.5}$	$70.1_{5.1}$	$65.7_{8.6}$	$52.6_{0.3}$	$17.0_{5.4}$
$\Delta_{model^{\dagger}-model^{*}}$		↑ 2.2	↓ 1.8	↑ 1.0	↑ 0.02	↑ 0.7	↓ 0.7	↑ 0.4

Table 6: Results for depression and anxiety multi-label classification

capture true positives at the expense of increasing mis-classification rates.

7 Error Analysis

7.1 Few shot vs Zero Shot

From our qualitative analysis, we observe that zero-shot prompting outperforms few-shot learning. Out of 334 samples where the predictions of the two approaches differed, the zero-shot prompt correctly classified 241 samples that the few-shot prompt misclassified. Only 4/241 samples were actually labeled as depressed in the ground truth, while the remaining 237 were non-depressed instances. Notably, all 4 of these depressed examples contained self-diagnosis statements, which the few-shot approach failed to identify correctly (see Table 12). For anxiety classification, out of the 468 samples where the predictions differed, the zero-shot approach correctly classified 363 non-anxiety samples that the few-shot approach misclassified. We observe that slight anxious behavior mentioned in the posts is aggressively classified as an anxiety disorder by the few-shot approach, despite the ground truth indicating non-anxiety

Upon qualitative analysis of these misclassified samples, we hypothesize that the few-shot prompts may be inducing noise and bias by overgeneralizing from the limited examples provided in context. Specifically, many non-depressed posts contain mentions of depressive symptoms, which the few-shot model appears to heavily weight, leading to false positive predictions. This can also be attributed to the fact that the in-context learning examples are silver labeled, the retriever is not potent enough to understand the minutiae in the semantics, or LLM's ability to handle long context.

7.2 Error-Analysis – The Battle of the GPTs

Both GPT-3.5-turbo and GPT-4 show a limitation in assessing the temporal aspect of depression. Among the 761 samples misclassified by both models, only 42 samples with a true label of "depressed" were incorrectly labeled as "not depressed." The models struggled to differentiate between past and present states of depression (Refer Table 13). Focusing on zero-shot depression, a total of 334 samples differed for GPT-3.5-turbo vs GPT-4. Out of these, GPT-4 demonstrated superior performance, with 270 samples correctly classified, while GPT-3.5 provided incorrect classifications. Among the 270 samples where GPT-4 was accurate, 251 instances were correctly identified as non-depressed, contradicting GPT-3.5's classification as depressed. This finding highlights GPT-4's enhanced contextual understanding and ability to differentiate between genuine depressive symptoms and non-depressive states, thus avoiding the over-classification of non-depressive instances as depressed.

8 Conclusion

In this study, we propose ANGST a novel dataset aimed at bridging critical gaps in the comorbid diagnosis of depression and anxiety from social media posts. By embracing a multi-label classification approach, ANGST promotes research in early detection and understanding of comorbid depression and anxiety conditions, which are often co-occurring yet underserved by existing datasets.

Our benchmarking experiments with state-of-the-art language models, including both discriminative and generative methods, highlight the superior performance of GPT-4 in both binary and multi-label classification tasks, reflecting its advanced capabilities in symptom identification and context understanding. Conversely, LLMs such as GPT-3.5 and LLaMa underperform relative to domain-specific PLMs like Mental-XLM, underscoring the inherent challenges and limitations of these models in handling the complexity of mental health diagnosis from text. However, despite these advancements, all models demonstrated a tendency to misclassify complex cases, indicating a need for improved methods that can differentiate the subtleties of mental health language and symptom expressions, especially in settings where disorders co-occur. In the future, we aim to explore hybrid models that combine the strengths of both discriminative and generative approaches to improve the accuracy and reliability of mental health assessments based on social media data, ultimately contributing to the development of more effective tools for early diagnosis and intervention, thus improving outcomes for individuals with mental health needs.

9   Limitations

This study has several limitations that future research could address. Firstly, our data collection from Reddit was constrained by time limitations and a finite search space, leading to potential blind spots in identifying users' self-disclosure. There may be rare expressions of self-disclosure or posts from non-mental health related subreddits were we missed curating ANGST. Additionally, our dataset only includes text, omitting other potentially informative modalities. For instance, the timing of posts could indicate insomnia or social relationship issues, which are early signs of depression or anxiety. We also limit our study to two disorders - depression and anxiety, and include posts in English only. Moreover, as discussed in section 6, the performance of all baseline models in classification tasks is far from ideal. Both generative and discriminative models, including GPT-4, consistently exhibit low precision, which shows that they are prone to false diagnoses. Therefore, while these models could serve as preliminary tools for individuals unaware of their mental conditions or those unable to access mental health services, their predictions must be rigorously reviewed by professionals before confirming a diagnosis.

10   Ethical Considerations

Social media data can be highly sensitive, especially when pertaining to mental health. Thus, it is imperative to prioritize privacy and recognize the potential risks posed to individuals represented within the data (Hovy and Spruit, 2016; Šuster et al., 2017; Benton et al., 2017). Given this consideration, we assert that the risks tied to the data used in this study are minimal. Our assessment is corroborated by prior studies which have introduced similar datasets (Coppersmith et al., 2015; Milne et al., 2016; Losada and Crestani, 2016)

The ANGST dataset solely consists of publicly accessible Reddit posts. We diligently remove any information that could disclose an author's identity or demographics. We provided annotators only with anonymized posts and ensured their commitment to neither deanonymize nor contact the authors. Moreover, we extensively adhere to the ethical and privacy guidelines set forth by (Benton et al., 2017). We do not collect any identifiable information, and we securely store all data on protected servers, accessible solely through written agreements with the creators. The institutional review board (IRB) at our institution has classified our experiments using these datasets as exempt from additional review.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aaron T Beck, Norman Epstein, Gary Brown, and Robert Steer. 1993. Beck anxiety inventory. *Journal of consulting and clinical psychology*.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Natalie Berry, Fiona Lobban, Maksim Belousov, Richard Emsley, Goran Nenadic, and Sandra Bucci. 2017. #whywetweetmh: Understanding why people use twitter to discuss mental health problems. *J Med Internet Res*, 19(4):e107.

Vanessa Borba de Souza, Jeferson Campos Nobre, and Karin Becker. 2022. Dac stacking: A deep learning ensemble to classify anxiety, depression, and their comorbidity from reddit texts. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3303–3311.

Marco Castiglioni, Cristina Liviana Caldiroli, Attà Negri, Gian Mauro Manzoni, and Rossella Procaccia. 2023. Linguistic predictors of psychological adjustment in healthcare workers during the covid-19 pandemic. *International journal of environmental research and public health*, 20(5):4482.

Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, and Shafiq Joty. 2023. Chatgpt's one-year anniversary: Are open-source large language models catching up? *arXiv preprint arXiv:2311.16989*.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, Denver, Colorado. Association for Computational Linguistics.

Glen Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):579–582.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–63, Madrid, Spain. Association for Computational Linguistics.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, WebSci '13, page 47–56, New York, NY, USA. Association for Computing Machinery.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):71–80.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013b. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.

Wei Du, Laksh Advani, Yashmeet Gambhir, Daniel Perry, Prashant Shiralkar, Zhengzheng Xing, and Aaron Colak. 2023. Effective proxy for human labeling: Ensemble disagreement scores in large language models for industrial NLP. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 53–61, Singapore. Association for Computational Linguistics.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–16, New York, NY, USA. Association for Computing Machinery.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2016. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI'16. ACM.

Muskan Garg, Chandni Saxena, Sriparna Saha, Veena Krishnan, Ruchi Joshi, and Vijay Mago. 2022. CAMS: An annotated corpus for causal analysis of mental health issues in social media posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6387–6396, Marseille, France. European Language Resources Association.

Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. "let me tell you about your mental health!": Contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 753–762,

New York, NY, USA. Association for Computing Machinery.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 63–73, San Diego, CA, USA. Association for Computational Linguistics.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 94–104, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ayaan Haque, Viraaj Reddi, and Tyler Giallanza. 2021. Deep learning for suicide and depression identification with unsupervised label correction. In *Artificial Neural Networks and Machine Learning – ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V*, page 436–447, Berlin, Heidelberg. Springer-Verlag.

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. Annollm: Making large language models to be better crowdsourced annotators.

Robert MA Hirschfeld. 2001. The comorbidity of major depression and anxiety disorders: recognition and management in primary care. *Primary care companion to the Journal of clinical psychiatry*, 3(6):244.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Shaoxiong Ji, Xue Li, Zi Huang, and Erik Cambria. 2022a. Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, 34(13):10309–10319.

Shaoxiong Ji, Shirui Pan, Xue Li, Erik Cambria, Guodong Long, and Zi Huang. 2021. Suicidal ideation detection: A review of machine learning methods and applications. *IEEE Transactions on Computational Social Systems*, 8(1):214–226.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022b. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. Domain-specific continued pretraining of language models for capturing long context in mental health. *arXiv preprint arXiv:2304.10447*.

Haoan Jin, Siyuan Chen, Mengyue Wu, and Kenny Q Zhu. 2023. Psyeval: A comprehensive large language model evaluation benchmark for mental health. *arXiv preprint arXiv:2311.09189*.

Bianca Lorraine Kahl, Hilary May Miller, Kathryn Cairns, Hayley Giniunas, and Mariesa Nicholas. 2020. Evaluation of reachout.com, an unstructured digital youth mental health intervention: Prospective cohort study. *JMIR Ment Health*, 7(10):e21280.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Atharva Kulkarni, Amey Hengle, Pradnya Kulkarni, and Manisha Marathe. 2021. Cluster analysis of online mental health discourse using topic-infused deep contextualized representations. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 83–93, online. Association for Computational Linguistics.

Atharva Kulkarni, Sarah Masud, Vikram Goyal, and Tanmoy Chakraborty. 2023. Revisiting hate speech benchmarks: From data curation to system deployment. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 4333–4345, New York, NY, USA. Association for Computing Machinery.

Femke Lamers, Patricia van Oppen, Hannie C Comijs, Johannes H Smit, Philip Spinhoven, Anton JLM van Balkom, Willem A Nolen, Frans G Zitman, Aartjan TF Beekman, and Brenda WJH Penninx. 2011. Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the netherlands study

of depression and anxiety (nesda). *Journal of clinical psychiatry*, 72(3):341.

Rumeng Li, Xun Wang, and Hong Yu. 2023. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7129–7143, Singapore. Association for Computational Linguistics.

Hui Liu, Lin Zhang, Weijun Wang, Yinghui Huang, Shen Li, Zhihong Ren, and Zongkui Zhou. 2022. Prediction of online psychological help-seeking behavior during the covid-19 pandemic: An interpretable machine learning method. *Frontiers in Public Health*, 10:814366.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

David E. Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–39, Cham. Springer International Publishing.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. 2019. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363.

Wolfgang Maier, Raimund Buller, Michael Philipp, and Isabella Heuser. 1988. The hamilton anxiety scale: reliability, validity and sensitivity to change in anxiety and depressive disorders. *Journal of affective disorders*, 14(1):61–68.

Matthew Louis Mauriello, Thierry Lincoln, Grace Hon, Dorien Simon, Dan Jurafsky, and Pablo Paredes. 2021. Sad: A stress annotated dataset for recognizing everyday stressors in sms-like conversational systems. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127, San Diego, CA, USA. Association for Computational Linguistics.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 26–34, Los Angeles, CA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436 – 465.

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. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

David Owen, Jose Camacho-Collados, and Luis Espinosa Anke. 2020. Towards preemptive detection of depression and anxiety in Twitter. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 82–89, Barcelona, Spain (Online). Association for Computational Linguistics.

Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity management and mental health discourse in social media. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, page 315–321, New York, NY, USA. Association for Computing Machinery.

Ted Pedersen. 2015. Screening Twitter users for depression and PTSD with lexical decision lists. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 46–53, Denver, Colorado. Association for Computational Linguistics.

Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on Reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12, Brussels, Belgium. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1348–1353, Seattle, Washington, USA. Association for Computational Linguistics.

Shailik Sarkar, Abdulaziz Alhamadani, Lulwah Alkulaib, and Chang-Tien Lu. 2023. Predicting depression and anxiety on reddit: A multi-task learning approach. In *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '22, page 427–435. IEEE Press.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 167–175, Brussels, Belgium. Association for Computational Linguistics.

Kerrie Shandley, David Austin, Britt Klein, and Michael Kyrios. 2010. An evaluation of 'Reach Out Central': an online gaming program for supporting the mental health of young people. *Health Education Research*, 25(4):563–574.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, pages 58–65, Vancouver, BC. Association for Computational Linguistics.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Vanessa Souza, Jéferson Nobre, and Karin Becker. 2020. Characterization of anxiety, depression, and their comorbidity from texts of social networks. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 121–132, Porto Alegre, RS, Brasil. SBC.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. Association for Computational Linguistics.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2020. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1).

Adela C. Timmons, Jacqueline B. Duong, Natalia Simo Fiallo, Theodore Lee, Huong Phuc Quynh Vo, Matthew W. Ahle, Jonathan S. Comer, LaPrincess C. Brewer, Stacy L. Frazier, and Theodora Chaspari. 2023. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. *Perspectives on Psychological Science*, 18(5):1062–1096. PMID: 36490369.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3187–3196, New York, NY, USA. Association for Computing Machinery.

Elsbeth Turcan and Kathy McKeown. 2019. Dreaddit: A Reddit dataset for stress analysis in social media. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 97–107, Hong Kong. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help.

Janet BW Williams and Kenneth A Kobak. 2008. Development and reliability of a structured interview guide for the montgomery-åsberg depression rating scale (sigma). *The British Journal of Psychiatry*, 192(1):52–58.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(1).

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023a. On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis. *arXiv preprint arXiv:2304.03347*.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023b. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6056–6077, Singapore. Association for Computational Linguistics.

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*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Tianlin Zhang, Kailai Yang, Shaoxiong Ji, and Sophia Ananiadou. 2023. Emotion fusion for mental illness detection from social media: A survey. *Information Fusion*, 92:231–246.

A Appendix

A.1 Curating `ANGST-SILVER`

Since `ANGST` serves exclusively as a test benchmark, we compiled `ANGST-SILVER` as a complementary corpus suitable for few-shot learning or supervised fine-tuning. Given the complexity and length of the posts, which demand considerable human and financial resources, we utilized LLMs for annotation. Recent advancements in GPT-based solutions have shown them to be effective substitutes for human labeling, aligning well with human judgments in both clinical and non-clinical tasks (Wang et al., 2021; Li et al., 2023; Du et al., 2023; He et al., 2024). Furthermore, Li et al. (2023) demonstrated that generating explanations alongside target labels enhances the quality of silver labels, achieving a higher correlation with crowd-sourced annotations. Leveraging this approach, we employed GPT-3.5-turbo (Ouyang et al., 2022) to generate silver labels for the remaining $22,124$ posts in a zero-shot setting. Using an open-ended prompt, as detailed in Table 17, we directed the LLM to identify and rationalize cues related to any mental health disorder that may be attributed with the post. After analyzing GPT-3.5-turbo outputs, we retained $7,667$ posts identified with depression and/or anxiety and discarded the rest, thus forming `ANGST-SILVER`.

A.2 Inter-Class Similarity – Case study of MMD

We report the mean MMD over 1000 runs with a random label-wise pair chosen at each run without replacement. In `ANGST`, we observed that the MMD between the control group and the groups representing anxiety, depression, and comorbidity is notably higher (0.47) than the MMD among the latter groups themselves (0.42). This observation suggests that distinguishing between anxiety, depression, and comorbidity poses a relatively greater challenge compared to distinguishing the control group from the other classes. Furthermore, when we compare the MMD of `ANGST` to other datasets, we find that our dataset exhibits a significantly lower MMD. This observation corroborates our hypothesis that `ANGST` presents a higher difficulty level in classification tasks due to the increased proximity between its class distributions.

A.3 Hyperparameter Tuning

Ww fine-tune the BERT-based models on `ANGST-SILVER`. We vary the learning rate within the $\{2-5, 2e-6\}$ range and experiment with batch sizes of $\{4, 8, 16, 32\}$. We train each model for 30 epochs with early stopping applied after 10 epochs of no improvement in validation F1 score. We use the Adam optimizer with decoupled weight decay regularization of 10^{-2} (Loshchilov and Hutter, 2017). For generating the outputs from the LLMs, we use top-p decoding with $p = 0.95$ (Holtzman et al., 2020). For each hyperparameter configuration, we run 3 seeds and report the averaged results. We apply this hyperparameter setup across depression, anxiety, and comorbidity classification tasks.

We use a single consistent prompt for zero-shot prompting of all the generative language models in our setup, provided in Appendix A.4. For few-shot prompting, we utilize the silver-labeled data as the retrieval pool. For each example, we retrieve the two most semantically similar posts from the silver-labeled data, irrespective of its label, to serve as in-context learning examples. We employ the all-mpnet-base-v2 model from sentence-transformers (Reimers and Gurevych, 2019) for computing semantic similarity between the test example and the silver-labeled data. The few-shot prompt is also included in Section A.4.

A.4 Zero and Few-Shot Prompts

Table 9 illustrates the zero and few shot prompts used for binary (depression vs non-depression and anxiety vs non-anxiety) and multi-label classification on `ANGST` using LLMs.

A.5 Other Prompting Techniques

Self-evaluation scales are commonly used by clinicians to assess the presence and severity of mental disorders based on responses to structured questions. Each scale comprises specific questions, and responses are quantitatively scored to determine the severity of the corresponding mental disorder.

A.5.1 Depression

For the assessment of depression, two scales were utilized: PHQ-9 (Kroenke et al., 2001) and MADRS(Williams and Kobak, 2008).

- **PHQ-9:** This scale consists of 9 symptoms, each assessed on a scale ranging from "not at all" to "nearly every day," scored from 0 to 4. To

Post	NRC (anger)	NRC (disgust)	NRC (fear)	NRC (sadness)	NRC (total)	Empath (total)
hi everyone! im a little late in the tour game and i didnt realize i wanted to see ari until this past week. now that shes announced shes adding more us dates, however, im hoping that she comes somewhere close to me so i get the chance to see her. the only ticket prices ive seen are resale, of course, for current shows. i was wondering if anyone remembers how much certain tickets cost (cheapest tickets, floor seats, pit, soundcheck, mg); these are all ball park of course since each venue is different. if you went to the cleveland, detroit, or pittsburgh show and remember prices for each section thats a bonus. thank you!	0.000	0.000	0.000	0.125	0.125	0.111
I've got a bunch of academies for my knights in a kingdom with a currently unknown population. I'm trying to figure out how many people I should have based on lifespan and death rates but I also need to know how believable it would be to have student base of, say, 1000. Before college I never went to school so I just don't know what is abnormal or what would make people stop and question the population, but I feel it's an important detail. For some background, the average human lifespan is 1000 years. Roughly 40% of the population are knights. Of that 40%, and I never learned even basic math so I'm just throwing out numbers here, let's say 5% are students. Would it be reasonable to have a smaller student base then? I would say with a low birth rate it would be a better option.	0.574	1.094	1.415	0.915	3.998	0.333
I hate the thought that I might be becoming antisocial, but I am enjoying my alone time too much. I've lived alone for most of the past few years at university, and now I'm back home temporarily whilst trying to get a decent job. The weekends take it out of me because I've got to be social and interact with my family otherwise I feel terrible inside because they like my company and it's not fair of me to constantly be hauled up in my room. Then Mondays come and I'm relieved when everybody goes to work, because I have time to myself. Problem is I get extremely irritable when I don't get that time. or I plan time to myself or it doesn't work out. For example, my dad's not going to work today most likely, because his van won't start. I love my dad and I love spending time with him when I've prepared for it, but when something sudden like this happens I feel cheated out of my alone-time. Even when I'm not hanging out with him and I'm upstairs by myself, I still feel very anxious and I feel bad for not interacting. Anybody else have this problem? How'd you force yourself into those interactions, or do you just give into your need to be alone? My family often worry I'm depressed (or *more* depressed) and I want to prove them wrong but I also want to be alone.	3.987	2.179	3.925	4.903	14.994	1.778

Table 7: NRC emotion and Empath scores for sample posts with highlighted text reflecting the NRC values.

Source	Evaluation Metrics			
	Acc	F1	ROC-AUC	MCC
DATD	0.956	0.933	0.904	0.875
Dep Reddit	0.939	0.938	0.937	0.878
Dreaddit	0.925	0.924	0.923	0.857
SDCNL	1.000	1.000	1.000	0.990

Table 8: Adversarial validation using `ANGST` as the target dataset against contemporary datasets as the source. Higher values of the metrics indicate a greater degree of discriminatory power of `ANGST` in comparison to the source datasets.

streamline responses, we converted these options into binary responses ("yes" or "no"). The model was prompted to provide a JSON response based on these inputs. Post-response parsing allowed for calculation of the total score, which was then evaluated against a predefined threshold for depression.

- **MADRS:** Comprising 10 symptoms, each rated on a scale of 0 to 6 for severity, this scale was adapted in our prompt to request the model to score each symptom from 1 to 6. Post-prompt parsing facilitated computation of the total score, which was subsequently compared against the specified threshold for depression.

A.5.2 Anxiety

For anxiety assessment, the scales chosen were BAI (Beck et al., 1993) and Hamilton Anxiety (Maier et al., 1988) Rating Scale.

- **Hamilton Anxiety Rating Scale:** This scale involves 14 questions, each scored from 0 to 4, evaluating various anxiety symptoms. The model

was prompted to assign a score to each symptom, and the cumulative score was calculated to determine anxiety severity.

- **BAI (Beck Anxiety Inventory):** Consisting of 21 symptoms, each rated from 0 to 4, this scale was utilized by prompting the model to assign scores to each symptom. Post-prompt parsing of the model's output allowed for computation of the total score, enabling assessment against established anxiety thresholds.

A.5.3 Results

Depression: Table 10 shows the classification scores for depression using MADRS and PHQ-9 ((Kroenke et al., 2001)) scales. The F1 score for depression is generally higher than the control F1 score. For the MADRS prompt, there is an improvement in both Depression F1 and Control F1 when moving from GPT-3.5 to GPT-4. However, for the PHQ-9 scale, the Depression F1 score drops significantly, while the Control F1 is significantly increased. Overall, the GPT-4 results are well-balanced, whereas the GPT-3.5 results are more biased towards classifying a sample as depressed. It is observed that the model tends to hallucinate while scoring, overestimating the presence of symptoms not mentioned in the post. This effect is reduced in GPT-4 compared to GPT-3.5.

Anxiety: Table 11 presents the classification scores for anxiety using the BAI and Hamilton scales. For GPT-3.5, the Hamilton scale yields more balanced results compared to the BAI scale. The GPT-3.5 scores are reasonable, but GPT-4 significantly underperforms on the Hamilton scale.

Task	Zero-Shot Prompt	Few-Shot Prompt
Depression	Assess whether the content of the post delimited by triple backticks aligns with the criteria for clinical depression as defined in DSM-5. Format your response as a JSON object 'depression':" with values strictly either 'yes' or 'no'. Do not return anything else apart from the json object. Post: ````	Below are posts and their respective assessments based on the criteria for clinical depression as defined in the DSM-5. Format your response as a JSON object 'depression':" with values either 'yes' or 'no'. Post: <post_1> Assesement: <assessment_1> Post: <post_2> Assesement: <assessment_2> Based on the above, assess the content of the following post: Post: <post_test> Assessment:
Anxiety	Assess whether the content of the post delimited by triple backticks aligns with the criteria for clinical anxiety as defined in DSM-5. Format your response as a JSON object 'anxiety':" with values strictly either 'yes' or 'no'. Do not return anything else apart from the json object. Post: ````	Below are posts and their respective assessments based on the criteria for clinical anxiety as defined in the DSM-5. Format your response as a JSON object 'anxiety':" with values either 'yes' or 'no'. Post: <post_1> Assesement: <assessment_1> Post: <post_2> Assesement: <assessment_2> Based on the above, assess the content of the following post: Post: <post_test> Assessment:
Comorbidity	This person wrote this paragraph on social media. As a psychologist, read the post on social media and answer the following questions. 1. Does the poster of this post have anxiety? 2. Does the poster of this post have depression? Format your response as a JSON object 'depression':", 'anxiety': " with values strictly either 'yes' or 'no'. Do not return anything else apart from the json object. Post: ````	Below are posts and their respective assessments based on the criteria for clinical depression and clinical anxiety as defined in the DSM-5. Format your response as a JSON object 'depression': ", 'anxiety': " with values either 'yes' or 'no'. Post: <post_1> Assesement: <assessment_1> Post: <post_2> Assesement: <assessment_2> Based on the above, assess the content of the following post: Post: <post_test> Assessment:

Table 9: The zero and few-shot prompts used for depression, anxiety, and comorbidity classification.

A.5.4 Discussion

The obtained scores fall short of both zero-shot and few-shot performance. This could be due to several reasons, such as the posts not providing sufficient information to answer all questions in a questionnaire accurately. The model's hallucination has a significant impact on the final score, and this effect becomes more pronounced as the number of

questions in the questionnaire increases. Different threshold values have been tried, but the issue of insufficient information in the posts cannot be eliminated. Future researchers could use this questionnaire-based prompt to further develop and improve the approach.

Model	Type	Depression			Control		
		Precision	Recall	F1	Precision	Recall	F1
GPT-3.5-turbo	MADRS	57.6	91.1	70.6	62.5	18.2	28.2
	PHQ-9	60.	89.6	72.4	69.	29.2	41.2
GPT-4	MADRS	60.3	87.6	71.4	66.0	29.5	40.
	PHQ-9	64.4	46.0	53.	51.1	68.9	58.

Table 10: Detailed Results Per Label of Dataset Using PHQ-9 and MADRS Self-Assessment Scales with GPT-3.5 and GPT-4 for Binary Depression Classification

Model	Type	Anxiety			Control		
		Precision	Recall	F1	Precision	Recall	F1
GPT-3.5-turbo	BAI	27.2	87.	41.5	85.9	24.2	37.
	Hamilton	26.	76.5	39.5	80.9	32.1	46
GPT-4	BAI	SFT	SFT	SFT	SFT	SFT	SFT
	Hamilton	42.9	6	10.5	76.2	97.4	85.5

Table 11: Detailed Results Per Label of Dataset Using BAI and Hamilton Self-Assessment Scales with GPT-3.5 and GPT-4 for Binary Anxiety Classification

A.6   Error-Analysis – GPT4 vs Best PLM

The analysis shows thatfor depression 86% of samples yield the same output across BERT-based models. However, mentioning certain disorders often leads to misclassification as depression (Refer Table 15). GPT-4 tends to classify longer samples more accurately, while XL-NET struggles with them, as evident from table 16.

Example	True	Zero-Shot	Few-Shot
MUST SEE LECTURE! Depression is just as biological as Diabetes, by Stanfords Robert Sapolsky. I just watched a lecture by a Stanford Professor, Robert Sapolsky, explaining how the biology and the psychology of depression work together. My BPD is comorbid with a major depressive disorder and I have been told MANY too many times that I feel miserable because I lack faith, will and overall reasonability. Close relatives and friends Are they friends? regularly compare their everyday stress management to my depression and instruct me on how easy it would be for me to just bounce back. I know I am not the only one to have experienced such ignorance and found reassurance and soothing in listening Professor Sapolsky. Hope it will bring a bit of calm to your days! Would highly recommend it if your BPD is comorbid with depression of if youre trying to understand how your brain is different from a neurotypicals brain. Heres the link Some of my favorite quotestake outs When you want to come to basic meat and potatoes of human medical misery, there is nothing out there like depression Depression is a biochemical disorder with a genetic component and early experience influences where somebody cant appreciate sunsets It can be considered that you are a major risk to yourself when you start getting better after a major episode, not during the episode itself. As you get better you also get the energy to do something catastrophic. This depression screams biology Somewhere around the 4th or 5th major depressive episode, you are statically at a higher risk of relapsing than before. This depression is not Oh, pull yourself together, we all get depressed. This is as real of a biological disorder as is diabetes'	1	1	0
Ive made myself alone these past few years. I was diagnosed with depression about 1.5 years ago, and there had been so much going on with my life, struggling with college, my father cheating on my mom, and the collapse of my longdistance relationship. The solution was to change meds, and I finally feel like myself again. I didnt do enough research at the very beginning of my depression, so I took an SSRI, which was the worst thing I could have done looking back. The sideeffects were awful, it was changing who I was, and ultimately, it didnt help me, but at least I didnt feel like killing myself. But, now that Im here, I realize how alone I made myself, given the holidays and the parties that people my age go to. I would have felt like a failure or a loner, but I know that can change. I feel good enough to go out and make friends. If you are isolating yourself, I implore you, from experience, that while you may feel like shit, while you may feel like the lowest thing in the world, there are people that care about you, so much more than you think or feel. Im here now, and Im fighting again, because thats all I can do to make my life better. You can keep fighting, because there are so many people to believe in you. Just something I wanted to get off my chest.	1	1	0
I got out of the Maze. I was diagnosed with depression at April.For 6 months I felt a void inside me.The last 2 months were horrible.Couldnt move out of the bed.No appetite.No concentration and no memory.Pain everywhere on my body.23 Panic attacks per day.I slept 1 hour each day.Couldnt sleep for more.The anxiety and fear were at 100.Everyday I was watching nightmares.Sucidal thoughts at medium range, but I wanted to die so much from a disease or hit by a car.Started antidepressants at April.The last 3 days are the best days of my life.No depression symptoms and 100 fictional.I do every task and every conversation with so much happiness and joy inside me.I feel I got out of the maze and I am at the sky.I have no specific tip.The most important for my case I think that I never stopped going on my psychologist and psychiatrist.Every week I was there on my sessions,besides how bad I was feeling and how tired I was.I post this to give some hope, you can all do it like me or like anyone else that fought depression.You are no different. P.s. Sorry for my bad English,I am still recovering everyday.	1	1	0
I want to live now and it sucks. I was quite depressed all through my final year of highschool and most of university. My episodes followed one another with seemingly no end. I was diagnosed with recurring MDD. During that time I was quite suicidal and apathetic and I actually accepted the idea of death. I liked to think death was better than this life and I honestly wouldnt have minded if I died. This was around a year and a half ago. I actually started to recover and with the right meds and therapy I could finally say I was in remission so to speak. The problem is, now Im terrified by the idea of death. Ive been having tiny existential crises one after the other. I cant just accept the void anymore and its honestly making me wish I could go back. Ive even considered one day going off my meds to more easily accept my own death and that of my lived ones. Im still getting used to the feelings and emotions that were gone for such a long time but this bothers me the most and I have no idea how to deal with it.	1	1	0

Table 12: Instances highlighted in red represent self-diagnosis cases where zero-shot learning correctly classified the instances as depicting depression, while few-shot learning misclassified them, despite the ground truth being depression.

Example	True	GPT3.5	GPT4
Anxiety disorder exacerbating phobias. Background late 20sF diagnosed with anxiety, depression, OCD, and panic disorder at age 11 I have always had a phobia of spiders. When I was a kid, a nest hatched in my room and there were hundreds of them. It was really traumatizing. I could not sleep in my room for almost a year, out of sheer fear. My husband and I recently moved to a new town, but decided to rent an apartment more on the outskirts and its a bit woodsy. We needed the quiet. However, its only been 3 weeks and we have already seen more critters in our place than we did in 2 years living in a city. One of them was hanging from the ceiling over my head last night and my husband went to kill it and it fell onto me, and I had a complete meltdown. I had a massive panic attack, the worst Ive had in years. I was scratching all over my body and dry heaving and shaking. I thought I was having a heart attack. It was awful. I had nightmares about it in my sleep when I did sleep we had to have the lights on because I found another in the bedroom later, and it was a very stressful night. I am wondering how I can control my phobia better. Or, if anyone has tips on how to keep critters at bay. I would also love if anyone could offer advice on how to come down from a panic attack. Ive had many in the past, and usually I can bring myself down after 510 minutes, but this time my husband had to talk me through it.	0	1	1
Does anyone else feel like theyre antidepressants made an asshole. Dont get me wrong, this is the best Ive felt in a long time. Im actually capable of getting up and brushing my teeth and showering without needing reminded like a child. Ive been meal prepping my lunches for work and I feel pretty great. But I know for a fact before I was diagnosed with the depression I had a greater sense of empathy for people. I loved working in retail because I felt like I had a better understanding for why people were in bad moods. Now that I feel like Ive leveled out more, I just dont wanna tolerate peoples shit. Before, if someone came in and they were mean, I could justify that maybe theyve been having a bad day and Im just the five minute encounter they took it out on. Ive considered going off the antidepressants just because Im not sure if I like who Ive turned into. TLDR Had more empathy for people when I was more depressed. Getting better is turning me into an impatient bitch?	0	1	1
*'I am hopeful about recovery. Hey guys,I was **diagnosed with depression and anxiety in 2009** and are on medication for both of these issue. I am still taking low doses of meds but my breakthrough year was in 2015 where I started to meet different people. For the past week or to be exact 8 days I have been feeling very well mentally,when I feel depression or anxiety creeping in I would call my friend and talk to them,or even call the suicide hotline for help to calm myself down.If I can continue to do well and not get over stress I am confident that if I can maintain all thesegood sleep,socializing,going to work,NoFapI will be able to live a better life in the future. The thing I feel hopeful is that if all these is maintained,I would be able to find a GF,because my mental health has been the main obstacles in dating.I am very thankful for life right now,but I just hope god can let me continue to work towards recovery and continue to feel better. What helps me is these 1.Socializing 2.Sleep 3.Nofap 4.psychotherapy 5.Work 6.Calling people and talk'*	0	1	1

Table 13: Depression classification comparison GPT3.5 vs GPT4 – Sample 1 Temporal Limitations in Depression Assessment by Language Models. In sample 2, despite a participant's self-reported mood improvement, the model incorrectly infers remission from depression. Sample 3 illustrates how the model inaccurately concludes no depression based on recent positive self-reports.

Example	True	GPT3.5	GPT4
Anxiety disorder exacerbating phobias. Background late Feeling so much better from the last time I've posted here. 4 months ago, I felt so down. I felt as if someone put gravity on the highest level and I was being pulled into the ground. Constant thoughts of suicide followed me everywhere. Work, school, when I was alone at night. I knew how bad it was getting and how badly I wanted it. Instead I got help. The reason being was seeing everyone's reaction after being sent to the mental health hospital. There was so much good vibes around me it inspired me. Also everyone at the hospital especially inspired me. There are people out that who are good and who only want to help. When I got out, I started seeing a therapist and taking medication. I also started doing things for myself more. Not going out as much, binge watching good shows, reading, learning new recipes. Your own company is the best and it's been the best for the last four months. Also I quit drinking and I've never regretted that decision. It gets better friends, hold in there.	0	1	1
Almost 6 days feeling great. Except for yesterday but I had BPD abandonment issues but I wasn't exactly depressed. I've tracked my mood for the past couple of days and the last time I felt depressed and sad was Jan. 9th! I don't wanna claim victory just yet but I guess my mood stabilizer + antidepressant have worked after a couple of months. Today my mother hugged me and told me she was very happy to see me this well after so long, and that she loves me, always has and always will. I'm so proud of these days except for yesterday lol and I know that you guys can get back up and keep it up 3 I thought I was going to end up killing myself and actually tried to do it. I'm glad I'm still here, although I'm a little bit scared for my third year of University like omg it's getting scary.	0	1	1

Table 14: Examples illustrating the limitations of GPT-3.5 and GPT-4 in zero-shot anxiety classification. Text highlighted in red indicates suicidal traits incorrectly classified by the models as anxiety disorder.

Example	True	GPT4	XL-NET
Having constricted breathinginability to catch your breathfeeling like you cant fill your lungs? Read THIS. Hey everyone, So anyone familiar with panic attacksanxiety will probably know the feeling of not being able to catch your breath, having constricted lungs, or the feeling you cant fill your lungs. Anyway if youve tried breathing techniques, tried all the naturalistic ways of treating it and nothing is working then hear me out Today I talked to my doctor and she had no problem prescribing me a salbutamol ventolin inhaler. Its a really quick release from the constrictedlimited breathing you typically feel with anxiety and the feeling of inability to catch your breathfill you lungs. I have only used it once and it works AMAZING. So far one use has lasted me a few hours still going and all my anxiety from breathing feels suddenly lifted. Also my prescription which came with 200 uses with a max of 2 uses per day, so this is a somewhat longterm solution. I am also on a few other things for anxietydepression but at this point I almost feel I dont need them because I have finally gotten control of my breathing. Anyway hope this helps, dont be afraid to talk to your doctors, people!	1	0	0
*Is smoking weed while on Zoloft a bad idea. Ive been feeling a lot better thanks to Zoloft. Life is going good and I feel comfortable and somewhat happy at times. I still have a lot of issues I need to work on which is why I go to therapy every two weeks. I sometimes think I should go once a week but I dont know. Im also on mirtazapine and risperidone. Is there anything wrong with** smoking weed** every week? Im not using it to cope or anything. I just use it because it is fun and relaxing. I just like to smoke and paint on canvases and listen to music. I know how easy it is to get addicted to it. I dont smoke every single second and every single day. I mostly just smoke late at night and chill. My therapist said that it can increase anxiety and depression. All my friends think Im fine and its not a big deal I smoke. But ganja these days is very powerful. And I sometimes worry about what its doing to my brain. Ive been smoking the past 4 weeks and havent had any issues. A few bad trips here and there when I got too high, but still, it was all very enjoyable. My psychiatrist was telling me how its dangerous because Im buying it from the streets. And that it could be laced and I might end up in the emergency room. Ive been smoking for the past 2 years at age 24 and never had anything laced or anything like that. I know smoking weed isnt a great habit. And its not really the greatest thing to be doing, but I just like to chill and draw in my sketchbook and listen to some music. Is that such a bad thing?*	1	0	0

Table 15: Examples highlighting the limitations of GPT-4 and XL-NET in depression classification. Text highlighted in red indicates instances where the model might incorrectly classify individuals as non-depressed due to the influence of medication on their improved mood.

Models		Character Count	
XL-NET	GPT-4	ND	D
T	T	999.2	1057.3
T	F	979.6	997.7
F	T	1005.5	1021.3
F	F	987.2	1040
Ground Truth		992	1052

Table 16: Effect of Character Count on Model Classification: T denotes correctly classified samples, while F denotes incorrectly classified samples. ND indicates non-depressed posts, and D indicates depressed posts

```
Reddit Post:  <post>

Does the language of the given Reddit post indicate any cues relating to some kind of
mental health disorder?  Refer to DSM5 to justify your response.

Strictly follow the following response format:

Mental Health Disorder:  (Yes/No)

Name of Mental Health Disorder:

DSM5 Rationale:
```

Table 17: Open-ended evaluation prompt to get Silver Labels from a Reddit post based on DSM5 criteria.

```
Below are the 10 symptoms of depression and their definitions according to the
Montgomery and Åsberg Depression Rating Scale.  As a psychologist, read the social
media post and evaluate the severity score of each symptom on a scale of 0 to 6.  The
social media post is delimited with triple backticks.  Format your response as a JSON
object with serial number of the 10 symptoms as keys.  Do not return anything else
apart from the json object.

1.  Apparent Sadness:  Representing despondency, gloom and despair, (more than just
ordinary transient low spirits) reflected in speech, facial expression, and posture.

2.  Reported sadness:  Representing reports of depressed mood, regardless of whether
it is reflected in appearance or not.  Includes low spirits, despondency or the
feeling of being beyond help and without hope.

3.  Inner tension:  Representing feelings of ill-defined discomfort, edginess, inner
turmoil, mental tension mounting to either panic, dread or anguish.

4.  Reduced sleep:  Representing the experience of reduced duration or depth of sleep
compared to the subject's own normal pattern when well.

5.  Reduced appetite:  Representing the feeling of a loss of appetite compared with
when well.

6.  Concentration Difficulties:  Representing difficulties in collecting one's
thoughts mounting to incapacitating lack of concentration.  Rate according to
intensity, frequency, and degree of incapacity produced.

7.  Lassitude:  Representing a difficulty getting started or slowness initiating and
performing everyday activities.

8.  Inability to feel:  Representing the subjective experience of reduced interest
in the surroundings, or activities that normally give pleasure.  The ability to react
with adequate emotion to circumstances or people is reduced.

9.  Pessimistic thoughts:  Representing thoughts of guilt, inferiority, self-reproach,
sinfulness, remorse and ruin.

10.  Suicidal thoughts:  Representing the feeling that life is not worth living, that
a natural death would be welcome, suicidal thoughts, and preparations for suicide.

Post:
```
```

Table 18: Prompt template used for evaluating depression severity from social media posts based on MADRS scale.

```
Below are the 9 symptoms of depression according to the Patient Health Questionnaire
(PHQ-9).  As a psychologist, read the social media post and determine which of the
9 symptoms observed in the post.  The social media post is delimited with triple
backticks.  Format your response as a JSON object with serial number of the 9 symptoms
as keys and "yes" or "no" as values.  Do not return anything else apart from the json
object.

1.  Little interest or pleasure in doing things:

2.  Feeling down, depressed, or hopeless.

3.  Trouble falling or staying asleep, or sleeping too much.

4.  Feeling tired or having little energy.

5.  Poor appetite or overeating.

6.  Feeling bad about yourself - or that you are a failure or have let yourself or your
family down.

7.  Trouble concentrating on things, such as reading the newspaper or watching
television.

8.  Moving or speaking so slowly that other people could have noticed.  Or the
opposite - being so fidgety or restless that you have been moving around a lot more
than usual.

9.  Thoughts that you would be better off dead, or of hurting yourself in some way.

Post:
"''
```

Table 19: Prompt template used for evaluating depression symptoms from social media posts based on PHQ-9.

Below are the 14 symptoms of anxiety and their definitions according to the Hamilton Anxiety Rating Scale. As a psychologist, read the social media post and evaluate the severity score of each symptom on a scale of 0 to 4. The social media post is delimited with triple backticks. Format your response as a JSON object with serial number of the 14 symptoms as keys. Do not return anything else apart from the json object.

1. Anxious mood: Worries, anticipation of the worst, fearful anticipation, irritability.

2. Tension: Feelings of tension, fatigability, startle response, moved to tears easily, trembling, feelings of restlessness, inability to relax.

3. Fears: Of dark, of strangers, of being left alone, of animals, of traffic, of crowds.

4. Insomnia: Difficulty in falling asleep, broken sleep, unsatisfying sleep and fatigue on waking, dreams, nightmares, night terrors.

5. Intellectual: Difficulty in concentration, poor memory.

6. Depressed mood: Loss of interest, lack of pleasure in hobbies, depression, early waking, diurnal swing.

7. Somatic (muscular): Pains and aches, twitching, stiffness, myoclonic jerks, grinding of teeth, unsteady voice, increased muscular tone.

8. Somatic (sensory): Tinnitus, blurring of vision, hot and cold flushes, feelings of weakness, pricking sensation.

9. Cardiovascular symptoms: Tachycardia, palpitations, pain in chest, throbbing of vessels, fainting feelings, missing beat.

10. Respiratory symptoms: Pressure or constriction in chest, choking feelings, sighing, dyspnea.

11. Gastrointestinal symptoms: Difficulty in swallowing, wind abdominal pain, burning sensations, abdominal fullness, nausea, vomiting, borborygmi, looseness of bowels, loss of weight, constipation.

12. Genitourinary symptoms: Frequency of micturition, urgency of micturition, amenorrhea, menorrhagia, development of frigidity, premature ejaculation, loss of libido, impotence.

13. Autonomic symptoms: Dry mouth, flushing, pallor, tendency to sweat, giddiness, tension headache, raising of hair.

14. Behavior at interview: Fidgeting, restlessness or pacing, tremor of hands, furrowed brow, strained face, sighing or rapid respiration, facial pallor, swallowing, etc.

Post:
```

Table 20: Prompt template used for evaluating anxiety symptoms from social media posts based on the Hamilton Anxiety Rating Scale.

Below are 21 common symptoms of anxiety according to the Beck Anxiety Inventory (BAI). As a psychologist, read the social media post and evaluate the severity score of each symptom on a scale of 0 to 4. The social media post is delimited with triple backticks. Format your response as a JSON object with serial number of the 21 symptoms as keys.

1. **Numbness or tingling**

2. **Feeling hot**

3. **Wobbliness in legs**

4. **Unable to relax**

5. **Fear of the worst happening**

6. **Dizzy or lightheaded**

7. **Heart pounding/racing**

8. **Unsteady**

9. **Terrified or afraid**

10. **Nervous**

11. **Feeling of choking**

12. **Hands trembling**

13. **Shaky/unsteady**

14. **Fear of losing control**

15. **Difficulty in breathing**

16. **Fear of dying**

17. **Feeling scared**

18. **Indigestion**

19. **Faint/lightheaded**

20. **Face flushed**

21. **Hot/cold sweats**

Post:
```

Table 21: Prompt template used for evaluating anxiety symptoms from social media posts based on BAI.