Read, Diagnose and Chat: Towards Explainable and Interactive LLMs-Augmented Depression Detection in Social Media

Wei Qin
Hefei University of Technology
Hefei, China
qinwei.hfut@gmail.com

Zetong Chen
University of Science and Technology of China
Hefei, China
dvdx@mai.ustc.edu.cn

Lei Wang
Singapore Managerment University
Singapre, Singapore
lei.wang.2019@phdcs.smu.edu.sg

Yunshi Lan
East China Normal University
Shanghai, China
lei.wang.2019@phdcs.smu.edu.sg

Weijieying Ren
Pennsylvania State University
Pennsylvania, USA
-

Richang Hong
Hefei University of Technology
Hefei, China
hongrc.hfut@gmail.com

ABSTRACT

Depression detection based on social media content has received increasing attention, as it allows for early diagnosis before the user's psychological state deteriorates. Although traditional methods of depression detection can provide a classification of whether the user is depressed or not, they cannot provide human-like explanations and interactions. In this paper, we propose a next-generation paradigm for depression detection, namely an interpretable and interactive depression detection system based on LLMs. The proposed system not only yields a final diagnosis result, but also offers diagnostic evidence grounded in established diagnostic criteria. Furthermore, it enables users to engage in natural language dialogue with the system, facilitating a more personalized understanding of their mental state based on their social media content. The interactive dialogue allows for the provision of tailored recommendations, which users can utilize to enhance their well-being. In constructing the entire system, we also addressed some non-trivial challenges. First, we introduced the chain of thoughts technique and professional depression diagnostic criteria when constructing the prompts, enabling our system to make decisions based on professional diagnosis criteria and provide explanations. Secondly, LLMs are incapable of processing excessively long contextual texts, and the accumulated posts of a single user may amount to tens of thousands of words. To overcome this limitation, we integrated a tweet selector that selects the part of posts for diagnosis. The experiments demonstrate that our depression detection system achieves the best performance across various settings, including full data setting, few-shot setting, zero-shot setting, independent-identical-distribution(IID) setting, and out-of-distribution(OOD) setting. Additionally, case studies reveal the explanation and interactivity of our system.

KEYWORDS

-

1 INTRODUCTION

Mental health disorders are a significant and widespread public health concern, with depression affecting an estimated 300 million people worldwide [22]. Depression is a leading cause of suicide and ranks as the third leading cause of death for individuals between the ages of 10 and 24 years old. Despite the severity of the issue, depression remains underdiagnosed and undertreated [1, 28]. More than half of those who suffer from depression do not receive any treatment [22]. Williams et al. [41] used a depression diagnostic test instrument to discover that 8% of the population had symptoms and a diagnosis of depression, while 7.6% had symptoms but had not been diagnosed. Individuals affected by mental disorders often hesitate to seek professional help [11]. However, as more people turn to social media to discuss their struggles and seek emotional support, automatic processing of social media data could be a crucial tool in identifying changes in mental health status before they lead to more serious health consequences.

Currently, many researchers are focusing on the task of detecting depression in social media. To this end, various studies, including those conducted by [4, 14], have investigated the use of different network structures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and transformers. Additionally, [2] introduced a topic-enriched auxiliary task that improves the detection ability of depression symptoms by enabling the model to better comprehend social media content. Furthermore, [7] considered the timing of social media posts. Notwithstanding their impressive performance, as depicted in Figure 1(a), existing methods for depression classification suffer from a crucial limitation: they solely produce binary classification outputs (i.e., depressed or non-depressed) without providing any accompanying explanation. This drawback renders the classification results less persuasive to users, and provides limited reference information for medical practitioners as well. Another primary constraint is their pool interactivity and lack of feedback mechanisms. An interactive environment is necessary to allow users to express their mental state more deeply and to obtain effective personalized recommendations. However, little research has been conducted in this domain.

Large language models (LLMs), such as GPT-3 [3], OPT [46], and PaLM [9], have evolved rapidly and have demonstrated remarkable performance across various natural language processing (NLP) tasks. As LLMs continue to expand in size and training data, they are strong semantic understanding and reasoning abilities [38]. The emergence of LLMs has brought about a paradigm shift in research. Previously, applying models to downstream tasks typically involved adjusting model parameters through backpropagation. However, the latest development of LLMs [31] has enabled both researchers and practitioners to facilitate learning during the forward process
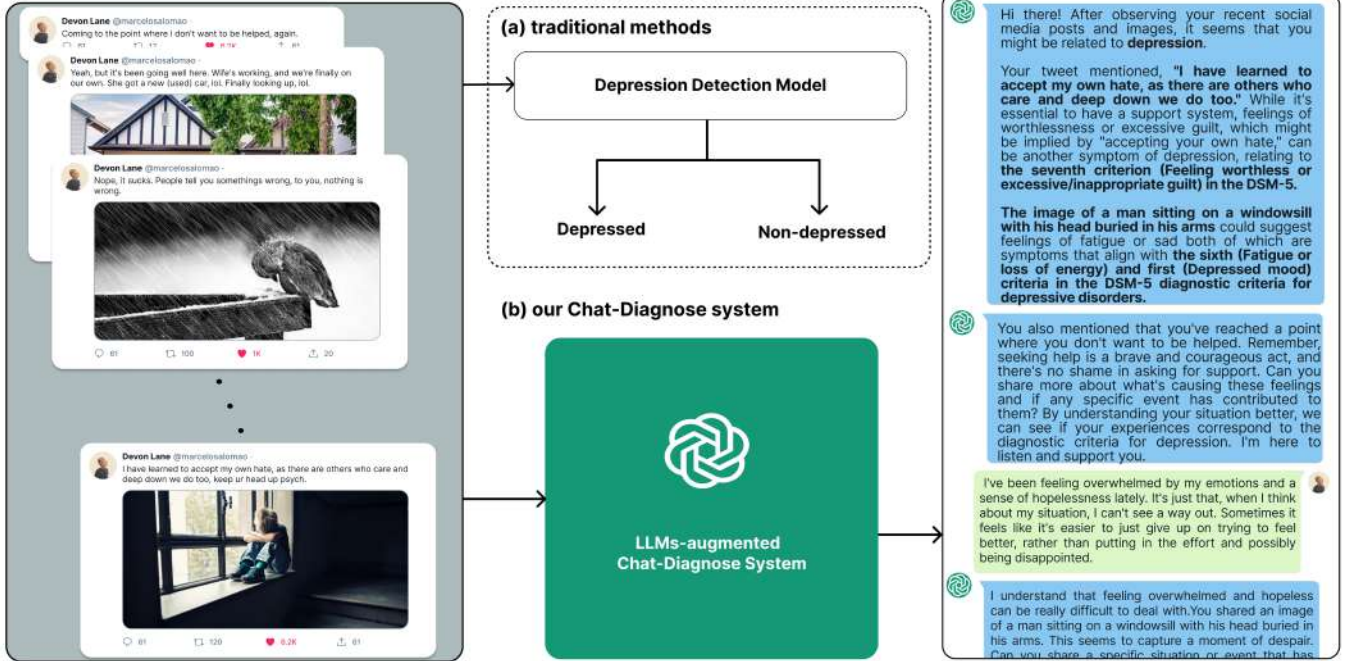
Figure 1: User posts list is the input of the task. The sub-figure (a) shows that traditional methods only output the final diagnosis result. Our method in sub-figure (b) could provide a diagnosis result, diagnosis evidence and interactive dialogue.

by constructing prompts, namely In-Context Learning (ICL) Recently, LLMs with ICL paradigm are leveraged to solve many tasks including movie recommendation [13], medical image classification [35] and document information extraction [16]. However, to date, there has been no investigation into the ability of LLMs to handle the depression detection task.

In this work, we propose a next-generation paradigm for depression detection in social media, namely an explainable and interactive LLMs-augmented depression detection system—*Chat-Diagnose*. As illustrated in Figure 1(b), Chat-Diagnose system combines established diagnostic criteria with social media content to assess the risk of depression in users and generate interpretable diagnostic evidence. Following the diagnostic process, Chat-Diagnose system communicates the diagnosis results to users through natural dialogue, along with relevant diagnostic evidence. In this interactive dialogue, users can describe their mental state more deeply through our guided questioning based on their social media content, and we can more accurately assess their mental state and provide appropriate personalized recommendations based on the content of the conversation.

However, several main challenges arise when applying LLMs to depression detection in social media. The first challenge is how to force the system to provide explanations based on professional diagnostic knowledge. To address this challenge, we first incorporate external professional diagnostic knowledge (DSM depression diagnosis criteria) as part of the prompt. On top of this, we introduce the chain of thoughts(CoT) technique and construct demonstrations, which enable our LLMs-augmented system to make diagnoses based on professional diagnosis criteria and provide diagnostic evidence. Figure 3 shows the DSM depression diagnosis criteria and the demonstrations. The second challenge lies in the fact that users may have hundreds or thousands of posts, and the lengthy context

can overwhelm LLMs. Therefore, we build a tweet selector module to screen each post and input the most related ones to LLMs.

Our Chat-Diagnose system can operate in zero-shot, few-shot and full-training settings. In the zero-shot setting, no examples are provided to our system. In the few-shot setting, our system inputs only two CoT demonstrations. In the full-training setting, our method additionally inputs the prediction outputs of a full-training vanilla depression detection model as answer heuristics to LLMs. As the currently most available LLMs cannot process images, we utilize image captioning and OCR to convert image content into textual descriptions.

We conducted extensive experiments on Twitter dataset and Weibo dataset. Our approach achieved state-of-the-art performance in multiple zero-shot, few-shot, and full-data settings, as well as independent-identical-distribution (IID) and out-of-distribution (OOD) test data. In addition, ablation studies confirmed the effectiveness of each module in our system. Case studies demonstrated the quality of the explanation and interactivity provided by our system.

Our contributions are summarized as follows: 1). We developed a new depression detection system(Chat-Diagnose) around LLM that provides diagnostic evidence and personalized advice through natural language dialogue. 2). By introducing CoT technology and a tweets selector, we improve explainability and avoid LLMs' limitations in handling excessive posts. 3). Extensive experiments and case studies demonstrate our system's superior performance, explainability, and interactivity in various settings.

2 RELATED WORK

Depression detection in social media. Early research on depression detection focused on analyzing language cues identified in psychology literature, such as self-focused language indicated by
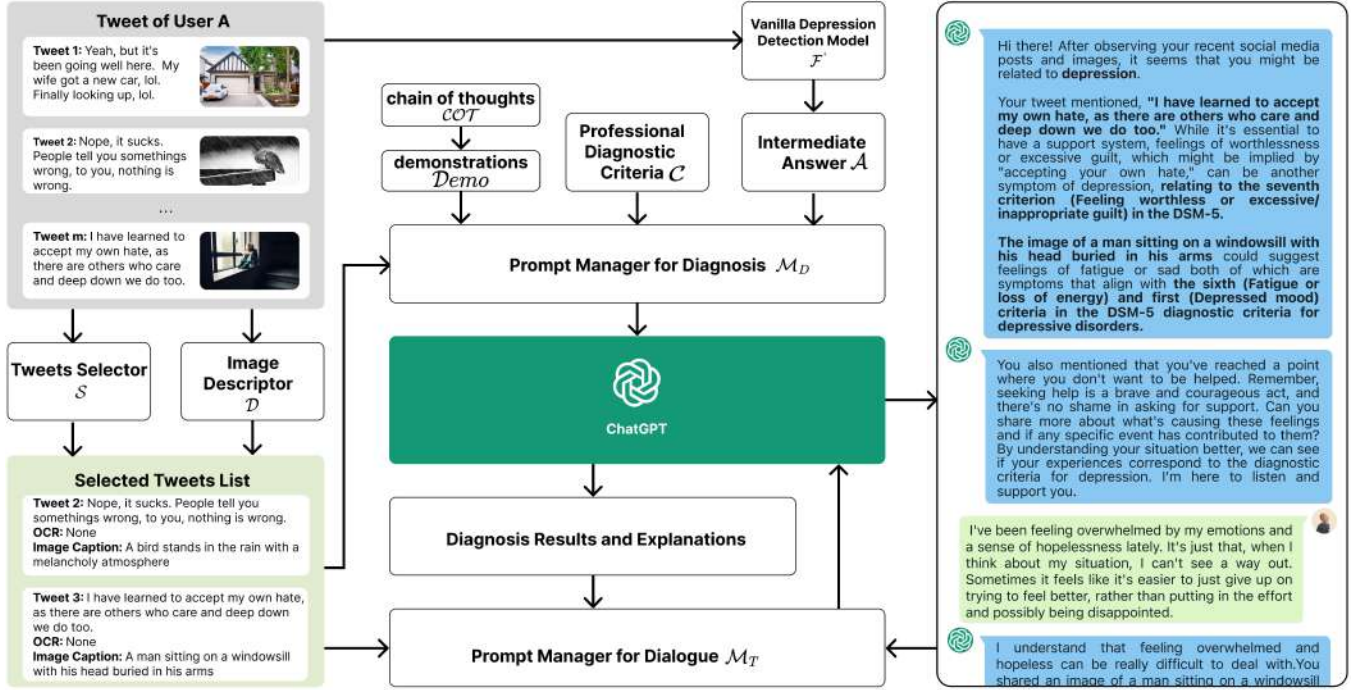
Figure 2: The framework of our Chat-Diagnose system.

the frequent use of the pronoun "I" [24, 25] and dichotomous thinking expressed through absolute words like "always" and "never" [12]. Reece et al. [23] demonstrated that individuals with depression tend to post images with sadder, less happy, and darker tones compared to healthy individuals. With the advent of deep learning, various neural network architectures, such as CNNs [44], LSTMs [29, 32], and transformers [4, 26], have been explored to detect depression in social media and have achieved remarkable performance. Gui et al. [14] proposed a cooperative multi-agent reinforcement learning approach using two agents to select relevant textual and visual information for classification. An et al. [2] proposed a multimodal topic-enriched auxiliary learning approach that improves the primary task of depression detection through auxiliary tasks on visual and textual topic modeling. Bucur et al. [4] used time-adapted weights [8] or Time-Aware LSTMs [7, 26] to incorporate the time component of data for mental health problem detection.

Deep learning methods achieve dominating performance but are limited by explanation and interactivity. High-quality explanations can convince users to seek professional medical help and increase the likelihood of recovery, and allow doctors to screen users more effectively. The system's interactivity enables users to provide detailed descriptions of their condition and receive personalized advice. We refer to depression detection systems with high explanation and interactivity as a next-generation paradigm for depression detection in social media.

In-Context Learning. Large Language Models (LLMs), including GPT-3 [3], OPT [46], and PaLM [9], are known for their impressive emergent abilities that scale with model and corpus size [39]. These abilities are learned from demonstrations containing only a few examples in the context, which is known as in-context learning [10]. To enable reasoning in LLMs, a few-shot prompting strategy called Chain-of-Thought (CoT) prompting has been proposed by Wei et al. [40], which involves adding multiple reasoning steps to the input question. Recent works[27, 30, 34] have aimed to improve CoT prompting in various aspects, such as prompt format [6], prompt selection [18], prompt ensemble [36], and problem decomposition [47]. Despite being initially developed for NLP tasks, LLMs with in-context learning have been found to exhibit few-shot or zero-shot abilities for multimodal problems, such as visual question-answering tasks [5, 42, 45]. Additionally, pre-trained models have shown promising few-shot performance in vision-and-language tasks, as demonstrated by Frozen [33]. However, to the best of our knowledge, this is the first work to explore the use of LLMs with in-context learning for depression detection.

3 METHOD

3.1 Problem Setup

Depression detection in social media refers to the diagnosis of whether a user is depressed or not based on their social media content, which may include both text and images. Specifically, we represent each user sample as (x, y), where x is a list of posts from the user's social media account, including multiple posts $x_{t_1}, x_{t_2}, ..., x_{t_m}$. Each post x_t contains both text x_t^{text} and picture x_t^{pic}. y indicates whether the user has been diagnosed as a depressed individual or not. The traditional depression detection system, denoted by $\mathcal{F}^*$, utilizes the user's social media content to determine whether the user exhibits symptoms of depression:

$$\hat{y} = \mathcal{F}^*(x) = \mathcal{F}^*(x_{t_1}, x_{t_2}, ..., x_{t_m}) \tag{1}$$

where $\hat{y}$ denotes the predicted diagnosis result whether the user is depressed or not.

In this paper, we propose a novel but more challenging paradigm for the depression detection system. The system $\mathcal{F}$ is required not only to provide detection results but also to provide interpretability (diagnostic evidence) and an interactive dialogue system.

$$\hat{y}, \mathcal{E} = \mathcal{F}(x) = \mathcal{F}(x_{t_1}, x_{t_2}, ..., x_{t_m}) \quad (2)$$

where $\hat{y}$ and $\mathcal{E}$ denotes the diagnosis results and the generated diagnosis explanation. Moreover, the system is equipped with an interactive chat module that proactively engages with users. This module can be represented as:

$$R_t = \mathcal{F}(c_t, h_{t-1}, \hat{y}, \mathcal{E}, x) \quad (3)$$

where c_t denotes the input provided by the user at round t and R_t denotes the response generated by the system at round t. $h_{t-1} = R_1, c_2, R_2, ..., R_{t-1}, c_{t-1}$ represents the dialogue history up to and including the previous t-1 rounds. Here, $h_0 = None$.

3.2 Depression Detection System

Tweet Selector $\mathcal{S}$: The tweet selector is introduced to filter out excessive posts from the user, since LLMs cannot process too long text. The tweet selector $\mathcal{S}$ will select n posts as the input LLMs:

$$x_{t'_1}, x_{t'_2}, ..., x_{t'_n} = \mathcal{S}(x_{t_1}, x_{t_2}, ..., x_{t_m}) \quad (4)$$

where $x_{t_1}, x_{t_2}, ..., x_{t_m}$ represents the collection of all m posts from a particular user, while $x_{t'_1}, x_{t'_2}, ..., x_{t'_n}$ denotes the subset of n posts that are selected. We experiment with three tweet selectors: the random selector, the recent selector and the sentiment selector. The random selector randomly selects n posts, while the recent selector selects the n most recent posts. The sentiment selector utilizes a sentiment analysis model to rank all m posts based on their negativity scores, and then select the n most negative posts. We employ GPT-3.5 as our sentiment analysis model.

Image Descriptor $\mathcal{D}$: The present module employs Image Caption and Optical Character Recognition (OCR) technology to transform images into textual information, which is a crucial feature as many Language and Learning Models (LLMs) are only capable of processing text. One common approach to extract semantics from images and represent it using textual descriptions is through image captioning [5, 43]. The process starts with extracting the text from the image using open-source Python packages, such as EasyOCR2. Subsequently, a pre-trained image captioning model named Clip-Cap [20] is applied. This model is particularly useful for generating high-quality captions for low-resolution web images. The captions produced by ClipCap typically describe the most prominent objects or events depicted in the image. The left part of Figure 2 demonstrates the effect of the Image Descriptor, which utilizes image caption and OCR to transform the visual information of an image into textual form. In the future, with LLMs capable of handling images such as GPT-4, we can directly input both images and text into the model without the need for this intermediary step.

Professional Diagnostic Criteria C: To facilitate the detection system's ability to perform diagnostic procedures based on professional knowledge, rather than just classification, we have integrated the DSM depression diagnostic criteria as the Professional Diagnostic Criteria C, as depicted in Figure 3. The Diagnostic and Statistical

Figure 3: The abbreviated version of the DSM diagnostic criteria for major depressive disorder is provided, with reference to the complete version available in [19].

Manual of Mental Disorders (DSM) serves as the primary reference for psychiatric diagnoses in the United States.

Demonstrations $\mathcal{D}emo$: In-context learning usually involves introducing a few (k) demonstrations to enable the model to learn how to accomplish the given task. Specifically, a demonstration refers to reformulating a labeled sample into a question-answer pair format in a text-based representation that can be understood by LLMs. Besides, utilizing few demonstrations can effectively constrain LLMs, enabling them to output results in a fixed format [10].

Chain-of-Thought COT: The Chain of Thoughts (CoT) technique involves incorporating reasoning steps into demonstrations ($\mathcal{D}emo$) to obtain the final results. This technique enables LLMs to detect depression and provide diagnostic evidence based on the Professional Diagnostic Criteria C. The CoT demonstration comprises triples, which include input, output, and reasoning steps. The input includes the question, options, and posts, while the output indicates the final answer (diagnosis). The reasoning steps in the chain of thought reveal how to derive the answer from the question. The CoT demonstration follows a specific organizational format:

Question: `Analyze someone's tweets using DSM-5 to determine if they are depressed.`
Options: `[A.depressed B.non-depressed]`.
Posts: `[Posts of User]`.
Answer: `[A or B]`.
Explanation: `[Reasoning steps]`.

Vanilla Depression Detection Model $\mathcal{F}^*$ and Answer Heuristics $\mathcal{A}$: In the full data setting, to fully exploit all available training samples, a conventional depression detection model $\mathcal{F}^*$ is introduced and trained on the entire training dataset. The resulting prediction probabilities from this model are utilized as answer heuristics $\mathcal{A}$ and subsequently fed into the LLM as inputs. This can be expressed mathematically as:

$$p(A), p(B) = \mathcal{F}^*(x_{t_1}, x_{t_2}, ..., x_{t_m}) \quad (5)$$
$$\mathcal{A} = \{p(A), p(B)\} \quad (6)$$

where $p(A)$ and $p(B)$ are the predicted probabilities that the test user is depressed or not depressed by the conventional depression detection model, respectively, and are considered as the answer heuristics $\mathcal{A}$. Specifically, the Time2VecTransformer [4] is adopted as the vanilla depression detection model. However, in the few-shot and zero-shot settings, the LLM is not provided with any answer heuristics.

Algorithm 1 Algorithm of our depression detection system

Step 1: detect the depression
Input: $x, \mathcal{COT}\text{-}\mathcal{Demo}, C, \mathcal{L}$ ▷ $\mathcal{L}$ denotes LLMs
Output: $\hat{y}, \mathcal{E}$
1: **procedure** $\mathcal{F}(x, \mathcal{COT}\text{-}\mathcal{Demo}, C)$
2:     $\mathcal{A} \leftarrow \mathcal{F}^*(x)$
3:     $x^D \leftarrow \mathcal{D}(x)$
4:     $x^{DS} \leftarrow \mathcal{S}(x^D)$
5:     $\text{prompt}_D \leftarrow \mathcal{M}_D(x^{DS}, \mathcal{A}, \mathcal{COT}\text{-}\mathcal{Demo})$
6:     $\hat{y}, \mathcal{E} \leftarrow \mathcal{L}(\text{prompt}_D)$

Step 2: interactive dialogue
Input: $x, \hat{y}, \mathcal{E}, \mathcal{L}$
7: **procedure** $\mathcal{F}(x, \hat{y}, \mathcal{E})$
8:     DialogueShow($\hat{y}, \mathcal{E}$)
9:     $c \leftarrow None$
10:    $h \leftarrow None$
11:    **while** dialogue ends **do**
12:       $\text{prompt}_T \leftarrow \mathcal{M}_T(c, x^{DS}, \hat{y}, \mathcal{E}, h)$
13:       $R \leftarrow \mathcal{L}(\text{prompt}_T)$
14:       $h \leftarrow h + R$
15:       DialogueShow(R)
16:       $c \leftarrow$ user input in this round
17:       $h \leftarrow h + c$

Prompt Manager for Diagnosis $\mathcal{M}_D$**:** The prompt manager has been designed to convert all clues into a language that LLM can understand. Specifically, the prompt manager for diagnosis $\mathcal{M}_D$ is responsible for constructing prompts that instruct LLMs on how to diagnose a user. We conduct the prompt with the question, the $\mathcal{COT}$ demonstrations, the professional diagnostic criteria C and the answer heuristics $\mathcal{A}$:

> **Question:** `Analyze someone's tweets using DSM-5 to determine if they are depressed.`
> **Options:** `[A.depressed B.non-depressed]`.
> **Diagnosis criteria:** `[DSM diagnosis criteria]`
> **Demonstrations:** `[`$\mathcal{COT}$ `Demonstrations]`.
> **Posts:** `[Posts of User]`.
> **Answer candidate:** `[A P(A)]; [B P(B)]`.
> **Answer:**
> **Explanation:**

where $\mathcal{COT}$ `Demonstrations` refers to demonstrations with a chain of thoughts. And "Posts" denotes the selected list of tweets obtained from the tweet selector $\mathcal{S}$ and image descriptor $\mathcal{D}$, as illustrated in Figure 2. The "Answer candidate" is constructed using the answer heuristic $\mathcal{A}$. The confidence scores for answers A and B are denoted by "P(A)" and "P(B)", respectively. These scores help to focus the attention of large language models (LLMs) on candidate answers with higher scores. We consider the answer generated by this prompt as the final diagnosis result. LLMs then fill the "Explanation" section with the diagnosis evidence.

This prompt is applicable in the full data setting, where the "Answer candidate" is computed using a vanilla depression detection model trained with full training data. In the few-shot and zero-shot settings, the "Answer candidate" is not included. Moreover, the

		# user	# tweet	# picture
WU3D	depressed	1000	39,595	15,543
(Weibo)	non-depressed	1000	80,167	48,802
TMDD	depressed	1402	232,895	22,195
(Twitter)	non-depressed	1402	879,025	64,359

Table 1: Statistics of TMDD and WU3D.

prompt used in the zero-shot setting does not contain the "Demonstrations".

Prompt Manager for Dialogue $\mathcal{M}_T$**:** The prompt manager for dialogue $\mathcal{M}_T$ is responsible for constructing prompts that guide LLMs on how to respond to users based on their diagnosis results, explanation, social media content, and dialogue history:

> **Instruction:** `Chat based on user's tweets to gather psychological information and provide advice.`
> **Posts:** `[Posts of User]`.
> **Diagnosis:** `[Result],[Explanation]`.
> **Dialogue history:** `[Dialogue history]`.
> **Input:** `[User input]`.

where `[Result],[Explanation]` are the generated diagnosis results and evidence from our system. `[User input]` denotes the user's input in the dialogue.

4 EXPERIMENT

4.1 Experiment Setup

Datasets. In this study, we evaluate our method by benchmarking it against two widely used depression datasets, namely the Twitter multimodal depression dataset (TMDD) [14] and the Weibo user depression detection dataset (WU3D) [37]. Both Twitter and Weibo are popular social media platforms where users share their mental health problems. The TMDD is in English, while the WU3D is in Chinese. While both datasets include images, the textual information is more extensive in WU3D, where posts can have up to 5,000 characters compared to the 280 character limit of Twitter. For TMDD, users in the depression class were identified by detecting mentions of depression diagnosis, while users in the control group had no such indications. On the other hand, all depressed user samples in the WU3D were manually labeled by anonymous data labeling specialists and reviewed by psychologists and psychiatrists.

The TMDD contains 1,402 users diagnosed with depression and 1,402 control users [14]. To maintain consistency with the experimental settings used in [4], which we could not access in the full data setting, we conducted experiments using the same settings and performed five-fold cross-validation. In the few-shot and zero-shot settings, we used 1,000 positive and 1,000 negative samples as evaluation data. Similarly, for the WU3D, we used 1000 positive and 1000 negative samples in the few-shot and zero-shot settings. Table 1 provides a summary of the dataset statistics. It is essential to note that only a small fraction of social media posts contain both text and image data.

Baselines. In the full data setting, we have compared our system with several classic methods. Multimodal Topic-Enriched Auxiliary Learning (MTAL) [2] captures the multimodal topic information through two auxiliary tasks accompanying the primary task of

Setting	Model	F1		Precision		Recall		Acc	
		IID	OOD	IID	OOD	IID	OOD	IID	OOD
Full Data	MTAL	0.842	-	0.842	-	0.842	-	0.842	-
	GRU + VGG +COMMA	0.900	-	0.900	-	0.900	-	0.900	-
	MTAN	0.908	-	0.885	-	0.931	-	-	-
	SetTransformer	0.927	-	0.921	-	0.934	-	0.926	-
	Time2VecTransformer	0.931	0.804	0.937	0.820	0.925	0.788	0.931	0.808
	Our Chat-Diagnose system(ChatGPT)	0.936	**0.894**	0.973	**0.925**	0.903	**0.865**	0.939	**0.897**
	Our Chat-Diagnose system(GPT-3)	**0.946**	0.864	**0.979**	0.904	**0.915**	0.828	**0.948**	0.870
Few Shot	BERT(base)	0.696	0.587	0.710	0.595	0.6825	0.58	0.703	0.593
	Our Chat-Diagnose system(ChatGPT)	0.815	0.797	0.882	**0.870**	0.758	0.735	0.825	0.813
	Our Chat-Diagnose system(GPT-3)	**0.870**	**0.851**	**0.886**	0.856	**0.855**	**0.848**	**0.872**	**0.853**
Zero Shot	PTDD	0.587	0.563	0.593	0.569	0.58	0.556	0.591	0.568
	Our Chat-Diagnose system(ChatGPT)	**0.715**	**0.703**	**0.738**	**0.707**	0.693	**0.700**	**0.724**	**0.705**
	Our Chat-Diagnose system(GPT-3)	0.689	0.656	0.652	0.671	**0.730**	0.643	0.669	0.664

Table 2: Experiment results performed under various settings. Full Data indicates that all available training data is used, Few Shot denotes the usage of only two training examples, and Zero Shot corresponds to the evaluation being performed without any training examples. IID and OOD refer to whether the evaluation data belongs to the independent-identical-distribution or out-of-distribution data, respectively.

	TMDD F1	WU3D F1
Our Chat-Diagnose system	0.946	-
w/o $\mathcal{A}$	0.870	0.902
w/o $\mathcal{A}, COT$	0.789	0.802
w/o $\mathcal{A}, COT, C$	0.702	0.691
w/o $\mathcal{A}, COT, \mathcal{D}emo$	0.689	-
w/o $\mathcal{A}, COT, C, \mathcal{D}emo,$	0.662	0.675

Table 3: The impact of various modules on the performance of our Chat-Diagnose system. Specifically, $\mathcal{A}$ referred to the answer heuristic extracted from an existing model trained by full data, whereas COT denoted the chain of thoughts included in the demonstrations. Moreover, C indicated the professional diagnosis criteria, while $\mathcal{D}emo$ referred to the few shot demonstrations.

depression detection in visual and textual topic modeling. Multimodal Time-Aware Attention Networks (MTAN) [7] is a multimodal model that incorporates T-LSTM to consider the time intervals between posts. "GRU + VGG-Net + COMMA" [14] uses a reinforcement learning component to select posts with text and images that are indicative of depression and classify them with an MLP. SetTransformer [4], a set-based multimodal transformer, employs zero positional encoding and random sampling of user posts, and Time2VecTransformer [4], a time-aware multimodal transformer, uses time-enriched positional embeddings and sub-sequence sampling. In the few shot setting, BERT(base), finetuned with 2 training samples, is compared [17]. In the zero-shot setting, we compare our system with PTDD, which uses BERT to obtain classification results by inputting prompts [15].

Implementation Details. In our experiments, we used the public ChatGPT(gpt-3.5-turbo) and GPT-3(text-davinci-003) as the underlying language model due to its widespread usage and ease of access. For each user, we selected four posts and utilized two types of demonstrations: one positive and one negative. We

aimed to demonstrate the efficacy of our depression detection system in terms of generalization ability. To this end, we leveraged the TextAttack [21] tool to generate out-of-distribution (OOD) test data for the TMDD. The original test data for these datasets are referred to as in-distribution (ID) test data. Specifically, we generated OOD samples by replacing original words with words that are visually similar but semantically different and by removing certain characters in words, such as "name" → "nme."

4.2 Quantity Analysis

Main results Table 2 presents the performance in the full data, few-shot and zero-shot settings on both independent-identical-distribution(IID) and out-of-distribution(OOD) evaluation data. We have the following observations based on it.

independent-identical-distribution(IID): In different data settings and metrics, our system has achieved the best performance results. In the full data setting, we utilized the result of the current SOTA model, i.e., Time2VecTransformer [4], as an answer heuristic and obtained better results. Specifically, the F1 metric increased from 0.931 to 0.946, indicating that our system can calibrate the results of other models. In the few-shot setting, our method outperforms the BERT-based classification model with a significant improvement, as the F1 metric increased from 0.696 to 0.870. In the zero-shot setting, our method also showed a significant improvement over PTDD, which is a prompt-based depression detection method. Furthermore, compared with the BERT-based method, our approach demonstrated a larger lead in the few-shot setting, indicating that in-context learning can better utilize few-shot samples.

Out of distribution(OOD): The results in Table 2 demonstrate that our method exhibits robustness to out-of-distribution data shifts. In the full data setting, the OOD data caused a significant decline in the performance of the traditional depression detection model, i.e., Time2VecTransformer, with the F1 score decreasing from 0.931 to 0.804. Our system utilizes Time2VecTransformer's results as an answer heuristic for the final prediction, which demonstrated a stronger robustness, with the F1 score decreasing from 0.946 to
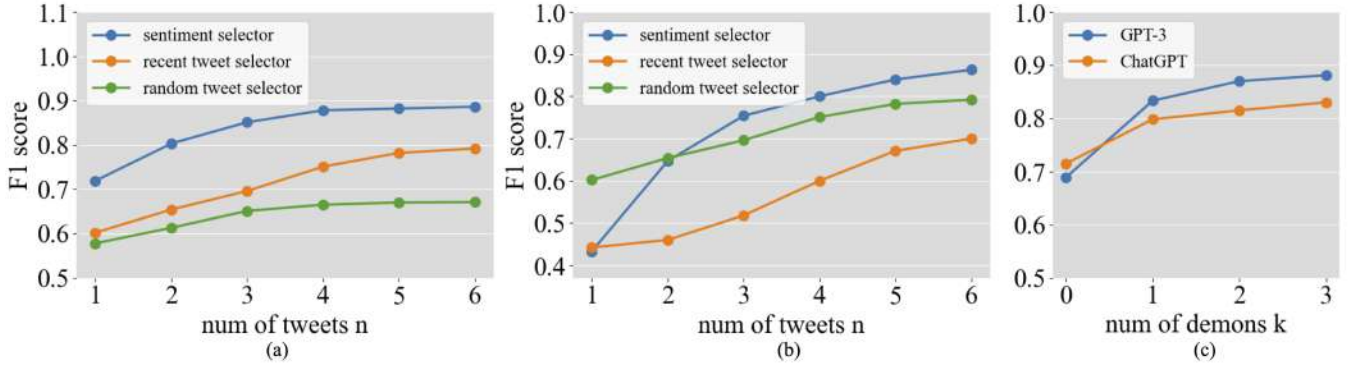
Figure 4: (a),(b) illustrate the impacts of different tweet selectors and number of tweets on TMDD and WU3D, respectively. (c) presents the effects of different number of demonstrations.

0.864. In the few-shot and zero-shot settings, since there was no answer heuristic module, our system showed stronger robustness, with the F1 score decreasing from 0.870 and 0.689 to 0.851 and 0.656, respectively.

In addition, we conducted additional experiments using two different backbone language models, GPT-3 (`text-davinci-003`) and ChatGPT (`gpt-3.5-turbo`), to investigate the generalizability of our system. The results indicate that our system significantly enhances the performance of both models. Nevertheless, the ChatGPT's high generation flexibility makes depression detection more challenging, resulting in slightly lower performance compared to our system using GPT-3. Overall, these promising findings highlight the effectiveness and versatility of our system across different backbone language models in detecting depression. Besides, ChatGPT (`gpt-3.5-turbo`) excels at zero-shot settings, whereas GPT-3 (`text-davinci-003`) is better equipped to leverage few-shot demonstrations.

Ablation study for each module Table 3 shows the ablation study on TMDD and WU3D. Due to the collaborative nature of some modules, we conducted an iterative ablation study to observe their effects. When we removed answer heuristic $\mathcal{A}$ from our system, which means we no longer used the results of other depression detection models trained on full data, and shifted from a full data setting to a few-shot setting, our performance decreased accordingly. We further removed COT, i.e., diagnostic evidence from demonstrations $\mathcal{D}emo$, and found that performance significantly decreased on both TMDD and WU3D, with F1 scores dropping to 0.789 and 0.802, respectively. This indicates that providing COT in demonstrations can indeed improve performance. Additionally, after removing COT, we found in a subsequent case study that the system no longer provided diagnosis evidence. In the next step, we removed the professional diagnosis criteria C, which means the system could no longer diagnose depression based on professional criteria. Although LLMs may contain a vast amount of psychological knowledge, removing the professional diagnosis criteria C led to a significant decrease in F1 score on both datasets, with scores dropping to 0.702 and 0.691, respectively. Finally, we removed demonstrations $\mathcal{D}emo$ entirely, which meant that F1 score on TMDD and WU3D decreased to 0.662 for only the instruction, options, and user posts in our prompt.

Ablation study on different Tweets Selector $\mathcal{S}$ Figure 4(a)(b) illustrates the different performance of our system using three different tweet selector strategies on TMDD and WU3D. Among the three, the sentiment analysis selector consistently outperforms the others, indicating its ability to effectively filter out depression-related tweets. This indicates that sentiment analysis can be used to retrieve more relevant posts, thereby improving the efficiency of the system.

Effect of n (numbers of selected posts). As shown in Figure 4(c), overall, if the numbers of selected posts are too few, it is not possible to provide enough information, resulting in decreased performance. If the numbers of selected posts are too many, irrelevant content and longer context are introduced, which poses greater detection challenges to LLMs. However, for the recent-tweets selector, more posts obviously bring more useful information, rather than more irrelevant content.

Effect of k (numbers of demonstrations). Figure 4(c) illustrates the performance of our system with different numbers of demonstrations provided in the prompt. In general, the more demonstrations provided, the better the performance. We observe that there is a significant improvement in performance from k=0 to k=1, as demonstrations can help format the output of the LLM. However, as there is a limit to the total number of tokens that can be inputted into the LLM, providing too many demonstrations does not necessarily lead to more improvement. In other words, the benefit of increasing the number of demonstrations is subject to diminishing marginal returns. Additionally, we found that ChatGPT performs better than GPT-3 in zero-shot settings, while GPT-3 can better utilize demonstrations.

4.3 Explainability and Interactivity Analysis

In this subsection, we demonstrate the explainability and interactivity of our system through practical examples. Due to space constraints, we provide only one example here, but we will include additional examples in Section Appendix.

Explainability of depression detection. As shown in the case in Figure 5, our system not only provides diagnostic results but also diagnostic evidence, i.e., which tweet of the user may correspond to which DSM diagnostic criterion. Specifically, our system demonstrates that it considers the textual content of the user's tweet *"@getinthesea: GET IN THE FUCKING SEA, GET IN THE FUCKING*

User @chamelaeon's Posts fed into our system:

Tweet 1:
Text: @getinthesea: GET IN THE FUCKING SEA, GET IN THE FUCKING SEA, GET IN THE FUCKING SEA
Caption: A red stop sign sitting on the side of a building
OCR: win over negative, protect happiness

Tweet 2:
Text: Holy shit I am tired. I mean, I went and lifted, but I am _exhausted_. How am I supposed to get my writing done like this?

Tweet 3:
Yes, I understand food sounds like an alien and abstract construct right now, but if you don't eat this you will pass out mid-squat.

Tweet 4:
Dawn of the Third Day (72 hours remain). I woke up feeling pretty rested! The meds went down a half hour ago ; already feeling sleepy.

Dialogue between our system and user:

Our system: Hi chamelaeon, you have a high risk of depression. Your tweet "@getinthesea: GET IN THE FUCKING SEA, GET IN THE FUCKING SEA, GET IN THE FUCKING SEA" with its image content "A red stop sign sitting on the side of a building" and "win over negative, protect happiness" express emotions of anger and hopelessness, which are consistent with the first criterion (Depressed mood, e.g., feeling angry, hopeless) of DSM diagnosis criteria. The tweet "Holy shit I am tired. I mean, I went and lifted, but I am exhausted. How am I supposed to get my writing done like this?" shows that you are mentally exhausted, which is consistent with the sixth criterion (Fatigue, or loss of energy) of DSM diagnosis criteria.

Our system: I noticed that you've been tweeting about feeling tired and exhausted lately. How have you been feeling overall? Is everything okay?

(The dialogue continues on the right table) ⟹

Dialogue between our system and user(continued):

User: To be honest, I'm not feeling great. I feel exhausted both mentally and physically. I don't have the energy or motivation to do anything.

Our system: I'm sorry to hear that. It sounds like you might be going through a tough time. Would you like to talk about what's been bothering you? Sometimes talking about our problems can help us feel better.

User: Recently, I have experienced a lot of painful things. I got fired from my job and lost my cash flow, but I still have a car loan to pay. My girlfriend also left me. Looking for a job has not been going well, and I have been rejected repeatedly. I feel like I have no energy to continue persevering.

Our system: I'm sorry to hear that you've been through so much lately. It's understandable that you're feeling exhausted and demotivated. It can be really tough to deal with multiple challenges at once. It's okay to feel overwhelmed and to take some time to process what has happened.

User: People always try to console me, but I can't seem to pick myself up. Maybe I know what's right, but I don't have the energy to do the right things.

Our system: It's completely understandable to feel that way. Here are a few things that might help: 1. Set small goals: Break down your larger goals into smaller, more manageable tasks. This can make it easier to get started and help you feel more accomplished as you complete each task. 2. Take care of yourself: Make sure you're taking care of your physical and emotional needs. This means getting enough sleep, eating healthy foods, and exercising regularly. 3. Seek support: It can be helpful to talk to someone about what you're going through. This could be a friend, family member, or a professional like a therapist. They can provide you with support and guidance as you navigate this difficult time.

User: These suggestions sound very worth trying. Thank you very much.

Our system: You're welcome! I'm glad I could be of help. Remember, if you ever need someone to talk to, don't hesitate to reach out. Take care of yourself!

Figure 5: An example of our Chat-Diagnose system on TMDD dataset. More cases can be found in Section Appendix.

SEA, GET IN THE FUCKING SEA" and the corresponding image content *"A red stop sign sitting on the side of a building"* and *"win over negative, protect happiness"* to meet the first criterion of the DSM diagnosis (Depressed Mood, e.g., feeling angry, hopeless). In addition, the text in another tweet *"Holy shit I am tired. I mean, I went and lifted, but I am exhausted. How am I supposed to get my writing done like this"* meets the sixth criterion of the DSM diagnosis (Fatigue. Or loss of energy). Therefore, our system determines that this user has a relatively high risk of developing depression.

Interactivity of our system. Figure 5 also demonstrates the interactivity of our system. Based on the content of the user's tweets and the diagnostic results, our system initiates a conversation with the user, delving into the topics that may be related to depression in the tweets. Our system further discussed the user's post describing fatigue, *"Holy shit I am tired. I mean, I went and lifted, but I am exhausted. How am I supposed to get my writing done like this"*, and interacted with the user to provide more information not available

on social media. Based on these personalized details, our system provided personalized advice.

5 CONCLUSION

In this paper, our system initiates a next-generation paradigm for depression detection in social media. Our system LLMs-augmented depression detection system not only achieves state-of-the-art performance in terms of both independent-identical-distribution (IID) and out-of-distribution (OOD) data, but also provides explainability and interactivity. After detection, our system provides diagnostic evidence based on both professional diagnostic criteria and personal social media content through dialogue. Moreover, in the dialogue system, users can interactively provide more information about their mental state and social media under our induced questioning. Our system can then better assess the user's mental state and provide customized and personalized suggestions.

REFERENCES

[1] Charlotte E Allan, Vyara Valkanova, and Klaus P Ebmeier. 2014. Depression in older people is underdiagnosed. *The Practitioner* 258, 1771 (2014), 19–22.

[2] Minghui An, Jingjing Wang, Shoushan Li, and Guodong Zhou. 2020. Multimodal topic-enriched auxiliary learning for depression detection. In *Proceedings of the 28th international conference on computational linguistics*. 1078–1089.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[4] Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P. Dinu. 2023. It's Just a Matter of Time: Detecting Depression with Time-Enriched Multimodal Transformers. In *European Conference on Information Retrieval*.

[5] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 321–332. https://aclanthology.org/2022.emnlp-main.22

[6] Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588* (2022).

[7] Ju Chun Cheng and Arbee LP Chen. 2022. Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems* 59, 2 (2022), 319–339.

[8] Chun Yueh Chiu, Hsien Yuan Lane, Jia Ling Koh, and Arbee L. P. Chen. 2021. Multimodal Depression Detection on Instagram Considering Time Interval of Posts. *J. Intell. Inf. Syst.* 56, 1 (feb 2021), 25–47. https://doi.org/10.1007/s10844-020-00599-5

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *ArXiv* abs/2204.02311 (2022).

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A Survey for In-context Learning. *arXiv preprint arXiv:2301.00234* (2022).

[11] Sarah Edwards, Lucy Tinning, June SL Brown, Jed Boardman, and John Weinman. 2007. Reluctance to seek help and the perception of anxiety and depression in the United Kingdom: A pilot vignette study. *The Journal of Nervous and Mental Disease* 195, 3 (2007), 258–261.

[12] Sandor Fekete. 2002. The Internet-a new source of data on suicide, depression and anxiety: a preliminary study. *Archives of Suicide Research* 6, 4 (2002), 351–361.

[13] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. *ArXiv* abs/2303.14524 (2023).

[14] Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 110–117.

[15] Yanrong Guo, Jilong Liu, Lei Wang, Wei Qin, Shijie Hao, and Richang Hong. 2023. A Prompt-Based Topic-Modeling Method for Depression Detection on Low-Resource Data. *IEEE Transactions on Computational Social Systems* (2023), 1–10. https://doi.org/10.1109/TCSS.2023.3260080

[16] Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. ICL-D3IE: In-Context Learning with Diverse Demonstrations Updating for Document Information Extraction. arXiv:2303.05063 [cs.CL]

[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*. 4171–4186.

[18] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610* (2022).

[19] Flora Ma. 2021. *Diagnostic and Statistical Manual of Mental Disorders-5 (DSM-5)*. Springer International Publishing, Cham, 1414–1425. https://doi.org/10.1007/978-3-030-22009-9_419

[20] Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734* (2021).

[21] John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 119–126.

[22] World Health Organization.(2012). 2012. Depression: A global crisis. In *World Mental Health Day, October 10 2012.* . World Federation for Mental Health, VA, USA.

[23] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, 1 (2017), 15.

[24] Esteban Andrés Ríssola, Mohammad Aliannejadi, and Fabio Crestani. 2020. Beyond modelling: Understanding mental disorders in online social media. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*. Springer, 296–310.

[25] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18, 8 (2004), 1121–1133.

[26] Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. 2020. A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*. 7685–7697.

[27] Omar Shaikh, Hongxin Zhang, William B. Held, Michael Bernstein, and Diyi Yang. 2022. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. *ArXiv* abs/2212.08061 (2022).

[28] David V Sheehan. 2004. Depression: underdiagnosed, undertreated, underappreciated. *Managed care (Langhorne, Pa.)* 13, 6 Suppl Depression (2004), 6–8.

[29] Ruba Skaik and Diana Inkpen. 2020. Using twitter social media for depression detection in the canadian population. In *2020 3rd artificial intelligence and cloud computing conference*. 109–114.

[30] Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed Huai hsin Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. *ArXiv* abs/2210.09261 (2022).

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[32] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. 2018. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering* 32, 3 (2018), 588–601.

[33] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.

[34] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2022. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. *arXiv preprint arXiv:2212.10001* (2022).

[35] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. 2023. ChatCAD: Interactive Computer-Aided Diagnosis on Medical Image using Large Language Models. arXiv:2302.07257 [cs.CV]

[36] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171* (2022).

[37] Yiding Wang, Zhenyi Wang, Chenghao Li, Yilin Zhang, and Haizhou Wang. 2020. A multitask deep learning approach for user depression detection on sina weibo. *arXiv preprint arXiv:2008.11708* (2020).

[38] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[39] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682* (2022).

[40] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[41] Sharifa Z Williams, Grace S Chung, and Peter A Muennig. 2017. Undiagnosed depression: A community diagnosis. *SSM-population health* 3 (2017), 633–638.

[42] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36.

3081–3089.

[43] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 3 (Jun 2022), 3081–3089. https://doi.org/10.1609/aaai.v36i3.20215

[44] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848* (2017).

[45] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598* (2022).

[46] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *ArXiv* abs/2205.01068 (2022).

[47] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, and Ed H. Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=WZH7099tgfM