



Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data

XUHAI XU, Massachusetts Institute of Technology & University of Washington, USA

BINGSHENG YAO, Rensselaer Polytechnic Institute, USA

YUANZHE DONG, Stanford University, USA

SAADIA GABRIEL, Massachusetts Institute of Technology, USA

HONG YU, University of Massachusetts Lowell, USA

JAMES HENDLER, Rensselaer Polytechnic Institute, USA

MARZYEH GHASSEMI, Massachusetts Institute of Technology, USA

ANIND K. DEY, University of Washington, USA

DAKUO WANG, Northeastern University, USA

Advances in large language models (LLMs) have empowered a variety of applications. However, there is still a significant gap in research when it comes to understanding and enhancing the capabilities of LLMs in the field of mental health. In this work, we present a comprehensive evaluation of multiple LLMs on various mental health prediction tasks via online text data, including Alpaca, Alpaca-LoRA, FLAN-T5, GPT-3.5, and GPT-4. We conduct a broad range of experiments, covering zero-shot prompting, few-shot prompting, and instruction fine-tuning. The results indicate a promising yet limited performance of LLMs with zero-shot and few-shot prompt designs for mental health tasks. More importantly, our experiments show that instruction finetuning can significantly boost the performance of LLMs for all tasks simultaneously. Our best-finetuned models, Mental-Alpaca and Mental-FLAN-T5, outperform the best prompt design of GPT-3.5 (25 and 15 times bigger) by 10.9% on balanced accuracy and the best of GPT-4 (250 and 150 times bigger) by 4.8%. They further perform on par with the state-of-the-art task-specific language model. We also conduct an exploratory case study on LLMs' capability on mental health reasoning tasks, illustrating the promising capability of certain models such as GPT-4. We summarize our findings into a set of action guidelines for potential methods to enhance LLMs' capability for mental health tasks. Meanwhile, we also emphasize the important limitations before achieving deployability in real-world mental health settings, such as known racial and gender bias. We highlight the important ethical risks accompanying this line of research.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → **Life and medical sciences**.

Additional Key Words and Phrases: Mental Health, Large Language Model, Instruction Finetuning

ACM Reference Format:

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, Article 31 (March 2024), 32 pages. <https://doi.org/10.1145/3643540>

Authors' addresses: [Xuhai Xu](#), xuhaixu@uw.edu, Massachusetts Institute of Technology & University of Washington, USA; [Bingsheng Yao](#), Rensselaer Polytechnic Institute, USA; [Yuanzhe Dong](#), Stanford University, USA; [Saadia Gabriel](#), Massachusetts Institute of Technology, USA; [Hong Yu](#), University of Massachusetts Lowell, USA; [James Hendler](#), Rensselaer Polytechnic Institute, USA; [Marzyeh Ghassemi](#), Massachusetts Institute of Technology, USA; [Anind K. Dey](#), University of Washington, USA; [Dakuo Wang](#), Northeastern University, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, or post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2474-9567/2024/3-ART31 \$15.00

<https://doi.org/10.1145/3643540>

1 INTRODUCTION

The recent surge of Large Language Models (LLMs), such as GPT-4 [18], PaLM [23], FLAN-T5 [24], and Alpaca [115], demonstrates the promising capability of large pre-trained models to solve various tasks in zero-shot settings (*i.e.*, tasks not encountered during training). Example tasks include question answering [87, 100], logic reasoning [124, 135], machine translation [15, 45], *etc.* A number of experiments have revealed that, built on hundreds of billions of parameters, these LLMs have started to show the capability to understand the human common sense beneath the natural language and do proper reasoning and inference accordingly [18, 85].

Among different applications, one particular question yet to be answered is how well LLMs can understand human mental health states through natural language. Mental health problems represent a significant burden for individuals and societies worldwide. A recent report suggested that more than 20% of adults in the U.S. experience at least one mental disorder in their lifetime [9] and 5.6% have suffered from a serious psychotic disorder that significantly impairs functioning [3]. The global economy loses around \$1 trillion annually in productivity due to depression and anxiety alone [2].

In the past decade, there has been a plethora of research in natural language processing (NLP) and computational social science on detecting mental health issues via online text data such as social media content (*e.g.*, [26, 32, 33, 38, 47]). However, most of these studies have focused on building domain-specific machine learning (ML) models (*i.e.*, one model for one particular task, such as stress detection [46, 84], depression prediction [38, 113, 127, 128], or suicide risk assessment [28, 35]). Even for traditional pre-trained language models such as BERT, they need to be finetuned for specific downstream tasks [37, 72]. Some studies have also explored the multi-task setup [12], such as predicting depression and anxiety at the same time [106]. However, these models are constrained to predetermined task sets, offering limited flexibility. From a different aspect, another line of research has been exploring the application of chatbots for mental health services [20, 21, 68]. Most chatbots are simply rule-based and can benefit from more advanced models that empower the chatbots [4, 68]. Despite the growing research efforts of empowering AI for mental health, it's important to note that existing techniques can sometimes introduce bias and even provide harmful advice to users [54, 74, 116].

Since natural language is a major component of mental health assessment and treatment [43, 110], LLMs could be a powerful tool for understanding end-users' mental states through their written language. These instruction-finetuned and general-purpose models can understand a variety of inputs and obviate the need to train multiple models for different tasks. Thus, we envision employing a single LLM for a variety of mental health-related tasks, such as multiple question-answering, reasoning, and inference. This vision opens up a wide range of opportunities for UbiComp, Human-Computer Interaction (HCI), and mental health communities, such as online public health monitoring systems [44, 90], mental-health-aware personal chatbots [5, 36, 63], intelligent assistants for mental health therapists [108], online moderation tools [39], daily mental health counselors and supporters [109], *etc.* However, there is a lack of investigation into understanding, evaluating, and improving the capability of LLMs for mental-health-related tasks.

There are few recent studies on the evaluation of LLMs (*e.g.*, ChatGPT) on mental-health-related tasks, most of which are in zero-shot settings with simple prompt engineering [10, 67, 132]. Researchers have shown preliminary results that LLMs have the initial capability of predicting mental health disorders using natural language, with promising but still limited performance compared to state-of-the-art domain-specific NLP models [67, 132]. This remaining gap is expected since existing general-purpose LLMs are not specifically trained on mental health tasks. However, to achieve our vision of leveraging LLMs for mental health support and assistance, we need to address the research question: **How to improve LLMs' capability of mental health tasks?**

We conducted a series of experiments with six LLMs, including Alpaca [115] and Alpaca-LoRA (LoRA-finetuned LLaMA on Alpaca dataset) [51], which are representative open-source models focused on dialogue and other tasks; FLAN-T5 [24], a representative open-source model focused on task-solving; LLaMA2 [118], one of the

most advanced open-source model released by Meta; GPT-3.5 [1] and GPT-4 [18], representative close-sourced LLMs over 100 billion parameters. Considering the data availability, we leveraged online social media datasets with high-quality human-generated mental health labels. Due to the ethical concerns of existing AI research for mental health, we aim to benchmark LLMs' performance as an initial step before moving toward real-life deployment. Our experiments contained three stages: (1) zero-shot prompting, where we experimented with various prompts related to mental health, (2) few-shot prompting, where we inserted examples into prompt inputs, and (3) instruction-finetuning, where we finetuned LLMs on multiple mental-health datasets with various tasks.

Our results show that the zero-shot approach yields promising yet limited performance on various mental health prediction tasks across all models. Notably, FLAN-T5 and GPT-4 show encouraging performance, approaching the state-of-the-art task-specific model. Meanwhile, providing a few shots in the prompt can improve the model performance to some extent ($\bar{\Delta} = 4.1\%$), but the advantage is limited. Finally and most importantly, we found that instruction-finetuning significantly enhances the model performance across multiple mental-health-related tasks and various datasets simultaneously. Our finetuned Alpaca and FLAN-T5, namely *Mental-Alpaca* and *Mental-FLAN-T5*, significantly outperform the best of GPT-3.5 across zero-shot and few-shot settings ($\times 25$ and 15 bigger than Alpaca and FLAN-T5) by an average of 10.9% on balance accuracy, as well as the best of GPT-4 by 4.8% ($\times 250$ and 150 bigger than Alpaca and FLAN-T5). Meanwhile, *Mental-Alpaca* and *Mental-FLAN-T5* can further perform on par with the task-specific state-of-the-art Mental-RoBERTa [58]. We further conduct an exploratory case study on LLM's capability of mental health reasoning (*i.e.*, explaining the rationale behind their predictions). Our results illustrate the promising future of certain LLMs like GPT-4, while also suggesting critical failure cases that need future research attention. We open-source our code and model at <https://github.com/neuhai/Mental-LLM>.

Our experiments present a comprehensive evaluation of various techniques to enhance LLMs' capability in the mental health domain. However, we also note that our technical results *do not* imply deployability. There are many important limitations of leveraging LLMs in mental health settings, especially along known racial and gender gaps [6, 42]. We discuss the important ethical risks to be addressed before achieving real-world deployment.

The contribution of our paper can be summarized as follows:

- (1) We present a comprehensive evaluation of prompt engineering, few-shot, and finetuning techniques on multiple LLMs in the mental health domain.
- (2) With online social media data, our results reveal that finetuning on a variety of datasets can significantly improve LLM's capability on multiple mental-health-specific tasks across different datasets simultaneously.
- (3) We release our model *Mental-Alpaca* and *Mental-FLAN-T5* as open-source LLMs targeted at multiple mental health prediction tasks.
- (4) We provide a few technical guidelines for future researchers and developers on turning LLMs into experts in specific domains. We also highlight the important ethical concerns regarding leveraging LLMs for health-related tasks.

2 BACKGROUND

We briefly summarize the related work in leveraging online text data for mental health prediction (Sec. 2.1). We also provide an overview of the ongoing research in LLMs and their application in the health domain (Sec. 2.2).

2.1 Online Text Data and Mental Health

Online platforms, especially social media platforms, have been acknowledged as a promising lens that is capable of revealing insights into the psychological states, health, and well-being of both individuals and populations [22,

[30, 33, 47, 91]. In the past decade, there has been extensive research about leveraging content analysis and social interaction patterns to identify and predict risks associated with mental health issues, such as anxiety [8, 104, 111], major depressive disorder [32, 34, 89, 119, 128, 129], suicide ideation [19, 29, 35, 101, 114], and others [25, 27, 77, 103]. The real-time nature of social media, along with its archival capabilities, often helps in mitigating retrospective bias. The rich amount of social media data also facilitates the identification, monitoring, and potential prediction of risk factors over time. In addition to observation and detection, social media platforms could further serve as effective channels to offer in-time assistance to communities at risk [66, 73, 99].

From the computational technology perspective, early research started with basic methods [27, 32, 77]. For example, pioneering work by Coppersmith *et al.* [27] employed correlation analysis to reveal the relationship between social media language data and mental health conditions. Since then, researchers have proposed a wide range of feature engineering methods and built machine-learning models for the prediction [14, 79, 82, 102, 119]. For example, De Choudhury *et al.* [34] extracted a number of linguistic styles and other features to build an SVM model to perform depression prediction. Researchers have also explored deep-learning-based models for mental health prediction to obviate the need for hand-crafted features [57, 107]. For instance, Tadesse *et al.* [114] employed an LSTM-CNN model and took word embeddings as the input to detect suicide ideation on Reddit. More recently, pre-trained language models have become a popular method for NLP tasks, including mental health prediction tasks [48, 58, 83]. For example, Jiang *et al.* [60] used the contextual representations from BERT as input features for mental health issue detection. Otsuka *et al.* [88] evaluated the performance of BERT-based pre-trained models in clinical settings. Meanwhile, researchers have also explored the multi-task setup [12] that aims to predict multiple labels. For example, Sarkar *et al.* [106] trained a multi-task model to predict depression and anxiety at the same time. However, these multi-task models are constrained to a predetermined task set and thus have limited flexibility. Our work joins the same goal and aims to achieve a more flexible multi-task capability. We focus on the next-generation technology of instruction-finetuned LLMs, leverage their power in natural language understanding, and explore their capability on mental health tasks with social media data.

2.2 LLM and Health Applications

After the great success of Transformer-based language models such as BERT [37] and GPT [96], researchers and practitioners have advanced towards larger and more powerful language models (e.g., GPT-3 [17] and T5 [97]). Meanwhile, researchers have proposed instruction finetuning, a method that utilizes varied prompts across multiple datasets and tasks. This technique guides a model during training and generation phases to perform diverse tasks within a single unified framework [123]. These instruction-finetuned LLMs, such as GPT-4 [18], PaLM [23], FLAN-T5 [24], LLaMA [117], Alpaca [115], contain tens to hundreds of billions of parameters, demonstrating promising performance across a variety of tasks, such as question answering [87, 100], logic reasoning [124, 135], machine translation [15, 45], *etc.*

Researchers have explored the capability of these LLMs in health fields [59, 70, 71, 75, 85, 112, 125?]. For example, Singhal *et al.* [112] finetuned PaLM-2 on medical domains and achieved 86.5% on MedQA dataset. Similarly, Wu *et al.* [125] finetuned LLaMA on medical papers and showed promising results on multiple biomedical QA datasets. Jo *et al.* [61] explored the deployment of LLMs for public health scenarios. Jiang *et al.* [59] trained a medical language model on unstructured clinical notes from the electronic health record and fine-tuned it across a wide range of clinical and operational predictive tasks. Their evaluation indicates that such a model can be used for various clinical tasks.

There is relatively less work in the mental health domain. Some work explored the capability of LLMs for sentiment analysis and emotion reasoning [64, 95, 134].

Closer to our study, Lamichhane [67] and Amin *et al.* [10] tested the performance of ChatGPT (GPT-3.5) on multiple classification tasks (e.g., stress, depression, and suicide detection). The results showed that ChatGPT

Table 1. A Summary of LLM-related Research for Mental Health Applications.

| | LLMs | Methods | Tasks |
|--------------------------|--|---|---------------------------|
| Lamichhane [67] | GPT-3.5 | Zero-shot | Classification |
| Amin <i>et al.</i> [10] | GPT-3.5 | Zero-shot | Classification |
| Yang <i>et al.</i> [132] | GPT-3.5 | Zero-shot | Classification, Reasoning |
| Mental-LLaMA [133] | LLaMA2, Vicuna (LLaMA-based) | Zero-shot, Few-shot, Instruction Finetuning | Classification, Reasoning |
| Mental-LLM (Our Work) | Alpaca, Alpaca-LoRA, FLAN-T5, LLaMA2, GPT-3.5, GPT-4 | Zero-shot, Few-shot, Instruction Finetuning | Classification, Reasoning |

shows the initial potential for mental health applications, but it still has a great room for improvement, with at least 5-10% performance gaps on accuracy and F1-score. Yang *et al.* [132] further evaluated the potential reasoning capability of GPT-3.5 for reasoning tasks (*e.g.*, potential stressors). However, most previous studies focused solely on zero-shot prompting and did not explore other methods to improve the performance of LLMs. Very recently, Yang *et al.* [133] released Mental-LLaMA, a set of LLaMA-based models finetuned on mental health datasets for a set of mental health tasks. Table 1 summarizes the recent related work exploring LLMs' capabilities on mental-health-related tasks. None of the existing work explores the capability other than LLaMA or GPT-3.5. In this work, we present a comprehensive and systematic exploration of multiple LLMs' performance on mental health tasks, as well as multiple methods to improve their capabilities.

3 METHODS

We introduce our experiment design with LLMs on multiple mental health prediction task setups, including zero-shot prompting (Sec. 3.1), few-shot prompting (Sec. 3.2), and instruction finetuning (Sec. 3.3). These setups are model-agnostic, and we will present the details of language models and datasets employed for our experiment in the next section.

3.1 Zero-shot Prompting

The language understanding and reasoning capability of LLMs have enabled a wide range of applications without the need for any domain-specific data, but only providing appropriate prompts [65, 122]. Therefore, we start with prompt design for mental health tasks in a zero-shot setting.

The goal of prompt design is to empower a pre-trained general-purpose LLM to achieve good performance on tasks in the mental health domain. We propose a general zero-shot prompt template ($Prompt_{ZS}$) that consists of four parts:

$$Prompt_{ZS} = TextData + Prompt_{part1-S} + Prompt_{part2-Q} + OutputConstraint \quad (1)$$

where $TextData$ is the online text data generated by end-users. $Prompt_{part1-S}$ provides specifications for a mental health prediction target. $Prompt_{part2-Q}$ poses the question for LLMs to answer. And $OutputConstraint$ controls the output of models (*e.g.*, "Only return yes or no" for a binary classification task).

We propose several design strategies for $Prompt_{part1-S}$, as shown in the top part of Table 2: (1) **Basic**, which leaves it as blank; (2) **Context Enhancement**, which provides more social media context about the $TextData$; (3) **Mental Health Enhancement**, which inserts mental health concept by asking the model to act as an expert. (4) **Context & Mental Health Enhancement**, which combines both enhancement strategies by asking the model to act as a mental health expert under the social media context.

Table 2. Prompt Design for Mental Health Prediction Tasks. *Prompt_{part1-S}* aims to provide a better specification for LLMs and *Prompt_{part2-Q}* poses the questions for LLMs to answer. For Part 1, we propose three strategies: context enhancement, mental health enhancement, and the combination of both. As for Part 2, we design different content for multiple mental health problem categories and prediction tasks. For each part, we propose two to three versions to improve its variation.

| Strategy | | <i>Prompt_{part1-S}</i> |
|---|---|--|
| Basic | | <ul style="list-style-type: none"> • { blank } |
| Context Enhancement | | <ul style="list-style-type: none"> • This person wrote this paragraph on social media. • Consider this post on social media to answer the question. |
| Mental Health Enhancement | | <ul style="list-style-type: none"> • As a psychologist, read the post on social media and answer the question. • If you are a psychologist, read the post on social media and answer the question. |
| Context & Mental Health Enhancement | | <ul style="list-style-type: none"> • This person wrote this paragraph on social media. As a psychologist, read the post on social media and answer the question. • This person wrote this paragraph on social media. As a psychologist, consider the mental well-being condition expressed in this post, read the post on social media, and answer the question. |
| Category | Task | <i>Prompt_{part2-Q}</i> |
| Mental state (e.g., stressed, depressed) | Binary classification (e.g., yes or no) | <ul style="list-style-type: none"> • Is the poster [stressed]? • Is the poster of this post [stressed]? • Determine if the poster of this post is [stressed]. |
| | Multi-class classification (e.g., multiple levels) | <ul style="list-style-type: none"> • Which level is the person [stressed]? • How [stressed] is the person? • Determine how [stressed] the person is. |
| Critical risk action (e.g., suicide) | Binary classification (e.g., yes or no) | <ul style="list-style-type: none"> • Does the poster want to [suicide]? • Is the poster likely to [suicide]? • Determine if the poster of this post want to [suicide]. |
| | Multi-class classification (e.g., multiple levels) | <ul style="list-style-type: none"> • Which level of [suicide] risk does the person have? • How [suicidal] is the person? • Determine which level of [suicide] risk does the person have. |

As for *Prompt_{part2-Q}*, we mainly focus on two categories of mental health prediction targets: (1) predicting critical mental states, such as stress or depression, and (2) predicting high-stake risk actions, such as suicide. We tailor the question description for each category. Moreover, for both categories, we explore binary and multi-class classification tasks¹. Thus, we also make small modifications based on specific tasks to ensure appropriate questions (see Sec. 4 for our mental health tasks). The bottom part of Table 2 summarizes the mapping.

For both *Prompt_{part1-S}* and *Prompt_{part2-Q}*, we propose several versions to improve its variability. We then evaluate these prompts on multiple LLMs on different datasets and compare their performance.

3.2 Few-shot Prompting

In order to provide more domain-specific information, researchers have also explored few-shot prompting with LLMs by providing few-shot demonstrations to support in-context learning (e.g., [7, 31]). Note that these few examples are used solely in prompts, and the model parameters remain unchanged. The intuition is to present a few “examples” for the model to learn domain-specific knowledge *in situ*. In our setting, we also test this strategy by adding additional randomly sampled [*Prompt_{ZS}* – *label*] pairs. The design of the few-shot prompt (*Prompt_{FS}*)

¹We also conduct an exploratory case study on mental health reasoning tasks. Please see more details in Sec. 5.4.

is straightforward:

$$\text{Prompt}_{FS} = [\text{Sample Prompt}_{ZS} - \text{label}]_M + \text{Prompt}_{ZS} \quad (2)$$

where M is the number of prompt-label pairs and is capped by the input length limit of a model. Note that both the *Sample Prompt*_{ZS} and *Prompt*_{ZS} follow Eq. 1 and employ the same design of *Prompt*_{Part1-S} and *Prompt*_{Part2-Q} to ensure consistency.

3.3 Instruction Finetuning

In contrast to the few-shot prompting strategy in Sec. 3.2, the goal of this strategy is closer to the traditional few-shot transfer learning, where we further train the model with a small amount of domain-specific data (e.g., [52, 71, 126]). We experiment with multiple finetuning strategies.

3.3.1 Single-dataset Finetuning. Following most of the previous work in the mental health field [26, 35, 132], we first conduct basic finetuning on a single dataset (the training set). This finetuned model can be tested on the same dataset (the test set) to evaluate its performance and different datasets to evaluate its generalizability.

3.3.2 Multi-dataset Finetuning. From Sec. 3.1 to Sec. 3.3.1, we have been focusing on one single mental health dataset D . More interestingly, we further experiment with finetuning across multiple datasets simultaneously. Specifically, we leverage instruction finetuning to enable LLMs to handle multiple tasks in different datasets [17].

It is noteworthy that such an instruction finetuning setup differs from the state-of-the-art mental-health-specific models (e.g., Mental-RoBERTa [58]). The previous models are finetuned for a specific task, such as depression prediction or suicidal ideation prediction. Once trained on task A, the model becomes specific to task A and is only suitable for solving that particular task. In contrast, we finetune LLMs on several mental health datasets, employing diverse instructions for different tasks across these datasets in a single iteration. This enables them to handle multiple tasks without additional task-specific finetuning.

For both single- and multi-dataset finetuning, we follow the same two steps:

$$\begin{aligned} \text{Step 1: Finetune with } [\text{Prompt}_{ZS} - \text{label}]_I &_{\sum N_{D_i-\text{train}}} \\ \text{Step 2: Test with } [\text{Prompt}_{ZS}]_I &_{\sum N_{D_i-\text{test}}} \end{aligned} \quad (3)$$

where $N_{D_i-\text{train}}/N_{D_i-\text{test}}$ is the total size of the training/test dataset D_i , I represents the set of datasets used for finetuning, and i indicates the specific dataset index ($i \in I, |I| \geq 1$). Both *Prompt*_{ZS-train} and *Prompt*_{ZS-test} follow Eq. 1. Similar to the few-shot setup in Eq. 2, they employ the same design of *Prompt*_{Part1-S} and *Prompt*_{Part2-Q}.

4 IMPLEMENTATION

Our method design is agnostic to specific datasets or models. In this section, we introduce the specific datasets (Sec. 4.1) and models (Sec. 4.2) involved in our experiments. In particular, we highlight our instructional-finetuned open-source models Mental-Alpaca and Mental-FLAN-T5 (Sec. 4.2.1). We also provide an overview of our experiment setup and evaluation metrics (Sec. 4.3).

4.1 Datasets and Tasks

Our experiment is based on four well-established datasets that are commonly employed for mental health analysis. These datasets were collected from Reddit due to their high-quality and availability. It is noteworthy that we intentionally avoid using datasets with weak labels based on specific linguistic patterns (e.g., whether a user ever stated “I was diagnosed with X”). Instead, we used ones with human expert annotations or supervision. We define six diverse mental health prediction tasks based on these datasets.

- **Dreaddit** [120]: This dataset collected posts via Reddit PRAW API [94] from Jan 1, 2017 to Nov 19, 2018, which contains ten subreddits in the five domains (abuse, social, anxiety, PTSD, and financial) and includes 2929 users' posts. Multiple human annotators rated whether sentence segments showed the stress of the poster, and the annotations were aggregated to generate final labels. We used this dataset for a post-level binary stress prediction (**Task 1**).
- **DepSeverity** [80]: This dataset leveraged the same posts collected in [120], but with a different focus on depression. Two human annotators followed DSM-5 [98] and categorized posts into four levels of depression: minimal, mild, moderate, and severe. We employed this dataset for two post-level tasks: binary depression prediction (*i.e.*, whether a post showed at least mild depression, **Task 2**), and four-level depression prediction (**Task 3**).
- **SDCNL** [49]: This dataset also collected posts from Python Reddit API, including r/SuicideWatch and r/Depression from 1723 users. Through manual annotation, they labeled whether each post showed suicidal thoughts. We employed this dataset for the post-level binary suicide ideation prediction (**Task 4**).
- **CSSRS-Suicide** [40]: This dataset contains posts from 15 mental health-related subreddits from 2181 users between 2005 and 2016. Four practicing psychiatrists followed Columbia Suicide Severity Rating Scale (C-SSRS) guidelines [93] to manually annotate 500 users on suicide risks in five levels: supportive, indicator, ideation, behavior, and attempt. We leveraged this dataset for two user-level tasks: binary suicide risk prediction (*i.e.*, whether a user showed at least suicide indicator, **Task 5**), and five-level suicide risk prediction (**Task 6**).

In order to test the generalizability of our methods, we also leveraged three other datasets from various platforms. Similarly, all datasets contain human annotations as labels.

- **Red-Sam** [62, 105]: This dataset also collected posts with PRAW API [94], involving five subreddits (Mental Health, depression, loneliness, stress, anxiety). Two domain experts' annotations were aggregated to generate depression labels. We used this dataset as an external evaluation dataset on binary depression detection (**Task 2**). Although also from Reddit, this dataset was not involved in few-shot learning or instruction finetuning. We cross-checked datasets to ensure there were no overlapping posts.
- **Twit-60Users** [56]: This dataset collected twitters from 60 users during 2015 with Twitter API. Two human annotators labeled every tweet with depression labels. We used this non-Reddit dataset as an external evaluation dataset on depression detection (**Task 2**). Note that this dataset has imbalanced labels (90.7% False), as most tweets did not indicate mental distress.
- **SAD** [76]: This dataset contains SMS-like text messages with nine types of daily stressor categories (work, school, financial problem, emotional turmoil, social relationships, family issues, health, everyday decision-making, and other). These messages were written by 3578 humans. We used this non-Reddit dataset as an external evaluation dataset on binary stress detection (**Task 1**). Note that human crowd-workers write the messages under certain stressor-triggered instructions. Therefore, this dataset has imbalanced labels on the other side (94.0% True).

Table 3 summarizes the information of the seven datasets and six mental health prediction tasks. For each dataset, we conducted an 80%/20% train-test split. Notably, to avoid data leakage, each user's data were placed exclusively in either the training or test set.

4.2 Models

We experimented with multiple LLMs with different sizes, pre-training targets, and availability.

Table 3. Summary of Seven Mental Health Datasets Employed for Our Experiment. The top four datasets are used for both training and testing, while the bottom three datasets are used for external evaluation. We define six diverse mental health prediction tasks on these datasets.

| Dataset | Task | Dataset Size | Text Length (Token) |
|--------------------------------------|--|---|---|
| Dreaddit [120] Source: Reddit | #1: Binary Stress Prediction post-level | Train: 2838 (47.6% False, 52.4% True) Test: 715 (48.4% False, 51.6% True) | Train: 114 ± 41 Test: 113 ± 39 |
| | #2: Binary Depression Prediction post-level | Train: 2842 (72.9% False, 17.1% True) Test: 711 (72.3% False, 17.7% True) | Train: 114 ± 41 Test: 113 ± 37 |
| DepSeverity [80] Source: Reddit | #3: Four-level Depression Prediction post-level | Train: 2842 (72.9% Minimum, 8.4% Mild, 11.2% Moderate, 7.4% Severe) Test: 711 (72.3% Minimum, 7.2% Mild, 11.5% Moderate, 10.0% Severe) | Train: 114 ± 41 Test: 113 ± 37 |
| SDCNL [49] Source: Reddit | #4: Binary Suicide Ideation Prediction post-level | Train: 1516 (48.1% False, 51.9% True) Test: 379 (49.1% False, 50.9% True) | Train: 101 ± 161 Test: 92 ± 119 |
| | #5: Binary Suicide Risk Prediction user-level | Train: 400 (20.8% False, 79.2% True) Test: 100 (25.0% False, 75.0% True) | Train: 1751 ± 2108 Test: 1909 ± 2463 |
| CSSRS-Suicide [40] Source: Reddit | #6: Five-level Suicide Risk Prediction user-level | Train: 400 (20.8% Supportive, 20.8% Indicator, 34.0% Ideation, 14.8% Behavior, 9.8% Attempt) Test: 100 (25.0% Supportive, 16.0% Indicator, 35.0% Ideation, 18.0% Behavior, 6.0% Attempt) | Train: 1751 ± 2108 Test: 1909 ± 2463 |
| Red-Sam [105] Source: Reddit | #2: Binary Depression Prediction post-level | External Evaluation: 3245 (26.1% False, 73.9% True) | External Evaluation: 151 ± 139 |
| Twt-60Users [56] Source: Twitter | #2: Binary Depression Prediction post-level | External Evaluation: 8135 (90.7% False, 9.3% True) | External Evaluation: 15 ± 7 |
| SAD [76] Source: SMS-like | #1: Binary Stress Prediction post-level | External Evaluation: 6185 (6.0% False, 94.0% True) | External Evaluation: 13 ± 6 |

- **Alpaca (7B) [115]**: An open-source large model finetuned from another open-sourced LLaMA 7B model [117] on instruction following demonstrations. Experiments have shown that Alpaca behaves qualitatively similarly to OpenAI's text-davinci-003 on certain task metrics. We choose the relatively small 7B version to facilitate running and finetuning on consumer hardware.
- **Alpaca-LoRA (7B) [51]**: Another open-source large model finetuned from LLaMA 7B model using the same dataset as Alpaca [115]. This model leverages a different finetuning technique called low-rank adaptation (LoRA) [51], with the goal of reducing finetuning cost by freezing the model weights and injecting trainable rank decomposition matrices into each layer of the Transformer architecture. Despite the similarity in names, it is important to note that Alpaca-LoRA is entirely distinct from Alpaca. They are trained on the same dataset but with different methods.
- **FLAN-T5 (11B) [24]**: An open-source large model T5 [97] finetuned with a variety of task-based datasets on instructions. Compared to other LLMs, FLAN-T5 focuses more on task solving and is less optimized for natural language or dialogue generation. We picked the largest version of FLAN-T5 (*i.e.*, FLAN-T5-XXL), which has a comparable size of Alpaca.
- **LLaMA2 (70B) [118]**: A recent open-source large model released by Meta. We picked the largest version of LLaMA2, whose size is between FLAN-T5 and GPT-3.5.

- **GPT-3.5** (175B) [1]: This large model is closed-source and available through API provided by OpenAI. We picked the gpt-3.5-turbo, one of the most capable and cost-effective models in the GPT-3.5 family.
- **GPT-4** (1700B) [18]: This is the largest closed-source model available through OpenAI API. We picked the gpt-4-0613. Due to the limited availability of API, the cost of finetuning GPT-3.5 or GPT-4 is prohibitive.

It is worth noting that Alpaca, Alpaca-LoRA, GPT-3.5, LLaMA2 and GPT-4 are all finetuned with natural dialogue as one of the optimization goals. In contrast, FLAN-T5 is more focused on task-solving. In our case, the user-written input posts resemble natural dialogue, whereas the mental health prediction tasks are defined as specific classification tasks. It is unclear and thus interesting to explore which LLM fits better with our goal.

4.2.1 Mental-Alpaca & Mental-FLAN-T5. Our methods of zero-shot prompting (Sec. 3.1) and few-shot prompting (Sec. 3.2) do not update model parameters during the experiment. In contrast, instruction finetuning (Sec. 3.3) will update model parameters and generate new models. To enhance their capability in the mental health domains, we update Alpaca and FLAN-T5 on six tasks across the four datasets in Sec. 4.1 using the multi-dataset instruction finetuning method (Sec. 3.3.2), which leads to our new model *Mental-Alpaca* and *Mental-FLAN-T5*.

4.3 Experiment Setup and Metrics

For zero-shot and few-shot prompting methods, we load open-source models (Alpaca, Alpaca-LoRA, FLAN-T5, LLaMA2) with one to eight Nvidia A100 GPUs to do the tasks, depending on the size of the model. For closed-source models (GPT-3.5, and GPT-4), we use OpenAI API to conduct chat completion tasks.

As for finetuning *Mental-Alpaca* and *Mental-FLAN-T5*, we merge the four datasets together and provide instructions for all six tasks (in the training set). We use eight Nvidia A100 GPUs for instruction finetuning. With cross entropy as the loss function, we backpropagate and update model parameters in 3 epochs, with Adam optimizer and a learning rate as $2e^{-5}$ (cosine scheduler, warmup ratio 0.03).

We focus on balanced accuracy as the main evaluation metric, *i.e.*, the mean of sensitivity (true positive rate) and specificity (true negative rate). We picked this metric since it is more robust to class imbalance compared to the accuracy or F1 score [16, 129]. It is noteworthy that the sizes of LLMs we compare are vastly different, with the number of parameters ranging from 7B to 1700B. A larger model is usually expected to have a better overall performance than a smaller model. We inspect whether this expectation holds in our experiments.

5 RESULTS

We summarize our experiment results with zero-shot prompting (Sec. 5.1), few-shot prompting (Sec. 5.2), and instruction finetuning (Sec. 5.3). Moreover, although we mainly focus on prediction tasks in this research, we also present the initial results of our exploratory case study on mental health reasoning tasks in Sec. 5.4.

Overall, our results show that zero-shot and few-shot settings show promising performance of LLMs for mental health tasks, although their performance is still limited. Instruction-finetuning on multiple datasets (*Mental-Alpaca* and *Mental-FLAN-T5*) can significantly boost models' performance on all tasks simultaneously. Our case study also reveals the strong reasoning capability of certain LLMs, especially GPT-4. However, we note that these results *do not* indicate the deployability. We highlight important ethical concerns and gaps in Sec. 6.

5.1 Zero-shot Prompting Shows Promising yet Limited Performance

We start with the most basic zero-shot prompting with Alpaca, Alpaca-LoRA, FLAN-T5, LLaMA2, GPT-3.5, and GPT-4. The balanced accuracy results are summarized in the first sections of Table 4. $\bar{\Delta}_{\text{Alpaca}_{ZS}}$ and $\bar{\Delta}_{\text{Alpaca-LoRA}_{ZS}}$ achieve better overall performance than the naive majority baseline ($\bar{\Delta}_{\text{Alpaca}} = 5.5\%$, $\bar{\Delta}_{\text{Alpaca-LoRA}} = 5.6\%$), but they are far from the task-specific baseline models BERT and Mental-RoBERTa (which have 20%-25% advantages).

With much larger models GPT-3.5_{ZS}, the performance gets more promising ($\bar{\Delta}_{\text{GPT-3.5}} = 12.4\%$ over baseline), which is inline with previous work [132]. GPT-3.5's advantage over Alpaca and Alpaca-LoRA is expected due to its larger size (25 \times).

Surprisingly, FLAN-T5_{ZS} achieves much better overall results compared to Alpaca_{ZS} ($\bar{\Delta}_{\text{FLAN-T5}_{\text{vs_Alpaca}}} = 10.9\%$) and Alpaca-LoRA_{ZS} ($\bar{\Delta}_{\text{FLAN-T5}_{\text{vs_Alpaca-LoRA}}} = 11.0\%$), and even LLaMA2 ($\bar{\Delta}_{\text{FLAN-T5}_{\text{vs_LLaMA2}}} = 1.0\%$) and GPT-3.5 ($\bar{\Delta}_{\text{FLAN-T5}_{\text{vs_GPT-3.5}}} = 4.2\%$). Note that LLaMA2 is 6 times bigger than FLAN-T5 and GPT-3.5 is 15 times bigger. On Task #6 (Five-level Suicide Risk Prediction), FLAN-T5_{ZS} even outperforms the state-of-the-art Mental-RoBERTa by 4.5%. Comparing these results, the task-solving-focused model FLAN-T5 appears to be better at the mental health prediction tasks in a zero-shot setting. We will introduce more interesting findings after finetuning (see Sec. 5.3.1).

In contrast, the advantage of GPT-4 becomes relatively less remarkable considering its gigantic size. GPT-4_{ZS}'s average performance outperforms FLAN-T5_{ZS} (150 \times size), LLaMA2_{ZS} (25 \times size), and GPT-3.5_{ZS} (10 \times size) by 6.4%, 7.5%, and 10.6%, respectively. Yet it is still very encouraging to observe that GPT-4 is approaching the state-of-the-art on these tasks ($\bar{\Delta}_{\text{GPT-4}_{\text{vs_Mental-RoBERTa}}} = -7.9\%$), and it also outperforms Mental-RoBERTa on Task #6 by 4.5%. In general, these results indicate the promising capability of LLMs on mental health prediction tasks compared to task-specific models, even without any domain-specific information.

5.1.1 The Effectiveness of Enhancement Strategies. In Sec. 3.1, we propose context enhancement, mental health enhancement, and their combination strategies for zero-shot prompt design to provide more information about the domain. Interestingly, our results suggest varied effectiveness on different LLMs and datasets.

Table 5 provides a zoom-in summary of the zero-shot part in Table 4. For Alpaca, LLaMA2, GPT-3.5, and GPT-4, the three strategies improved the performance in general ($\bar{\Delta}_{\text{Alpaca}} = 1.0\%$, 13 out of 18 tasks show positive changes; $\bar{\Delta}_{\text{LLaMA2}} = 0.3\%$, 12/18 tasks positive; $\bar{\Delta}_{\text{GPT-3.5}} = 2.8\%$, 12/18 tasks positive; $\bar{\Delta}_{\text{GPT-4}} = 0.2\%$, 11/18 tasks positive). However, for Alpaca-LoRA and FLAN-T5, adding more context or mental health domain information would reduce the model performance ($\bar{\Delta}_{\text{Alpaca-LoRA}} = -2.7\%$, $\bar{\Delta}_{\text{FLAN-T5}} = -1.6\%$). For Alpaca-LoRA, this limitation may stem from being trained with fewer parameters, potentially constraining its ability to understand context or domain specifics. For FLAN-T5, this reduced performance might be attributed to its limited capability in processing additional information, as it is primarily tuned for task-solving.

The effectiveness of strategies on different datasets/tasks also varies. We observe that Task#4 from the SDCNL dataset and Task#6 from the CSSRS-Suicide dataset benefit the most from the enhancement. In particular, GPT-3.5 benefits very significantly from enhancement on Task #4 ($\bar{\Delta}_{\text{GPT-3.5-Task\#4}} = 14.8\%$). And LLaMA2 benefits significantly on Task #6 ($\bar{\Delta}_{\text{GPT-3.5-Task\#6}} = 6.8\%$). These could be caused by the different nature of datasets. Our results suggest that these enhancement strategies are generally more effective for critical action prediction (e.g., suicide, 2/3 tasks positive) than mental state prediction (e.g., stress and depression, 1/3 task positive).

We also compare the effectiveness of different strategies on the four models with positive effects: Alpaca, LLaMA2, GPT-3.5, and GPT-4. The context enhancement strategy has the most stable improvement across all mental health prediction tasks ($\bar{\Delta}_{\text{Alpaca-context}} = 2.1\%$, 6/6 tasks positive; $\bar{\Delta}_{\text{LLaMA2-context}} = 1.2\%$, 4/6 tasks positive; $\bar{\Delta}_{\text{GPT-3.5-context}} = 2.5\%$, 5/6 tasks positive; $\bar{\Delta}_{\text{GPT-4-context}} = 0.4\%$, 5/6 tasks positive). Comparatively, the mental health enhancement strategy is less effective ($\bar{\Delta}_{\text{Alpaca-mh}} = 1.1\%$, 5/6 tasks positive; $\bar{\Delta}_{\text{LLaMA2-mh}} = -0.5\%$, 4/6 tasks positive; $\bar{\Delta}_{\text{GPT-3.5-mh}} = 2.1\%$, 3/6 tasks positive; $\bar{\Delta}_{\text{GPT-4-mh}} = 0.2\%$, 3/6 tasks positive). The combination of the two strategies yields diverse results. It has the most significant improvement on GPT-3.5's performance, but not on all tasks ($\bar{\Delta}_{\text{GPT-3.5-both}} = 3.9\%$, 4/6 tasks positive), followed by LLaMA2 ($\bar{\Delta}_{\text{LLaMA2-both}} = 0.2\%$, 4/6 tasks positive). However, it has slightly negative impact on the average performance of Alpaca ($\bar{\Delta}_{\text{Alpaca-both}} = -0.1\%$, 2/6 tasks positive) or GPT-4 ($\bar{\Delta}_{\text{GPT-4-both}} = -0.3\%$, 3/6 tasks positive). This indicates that larger language models (LLaMA2, GPT-3.5 vs. Alpaca) have a strong capability to leverage the information embedded in the prompts.

Table 4. Balanced Accuracy Performance Summary of Zero-shot, Few-shot and Instruction Finetuning on LLMs. ZS_{best} highlights the best performance among zero-shot prompt designs, including context enhancement, mental health enhancement, and their combination (see Table. 2). Detailed results can be found in Table 10 in Appendix. Small numbers represent standard deviation across different designs of $Prompt_{Part1-S}$ and $Prompt_{Part2-Q}$. The baselines at the bottom rows do not have standard deviation as the task-specific output is static, and prompt designs do not apply. Due to the maximum token size limit, we only conduct few-shot prompting on a subset of datasets and mark other infeasible datasets as “-”. For each column, the best result is **bolded**, and the second best is underlined.

| | Dataset | Dreaddit | DepSeverity | | SDCNL | CSSRS-Suicide | |
|--------------------------|--|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| Category | Model | Task #1 | Task #2 | Task #3 | Task #4 | Task #5 | Task #6 |
| Zero-shot Prompting | Alpaca _{ZS} | 0.593 \pm 0.039 | 0.522 \pm 0.022 | 0.431 \pm 0.050 | 0.493 \pm 0.007 | 0.518 \pm 0.037 | 0.232 \pm 0.076 |
| | Alpaca _{ZS_{best}} | 0.612 \pm 0.065 | 0.577 \pm 0.028 | 0.454 \pm 0.143 | 0.532 \pm 0.005 | 0.532 \pm 0.033 | 0.250 \pm 0.060 |
| | Alpaca-LoRA _{ZS} | 0.571 \pm 0.043 | 0.548 \pm 0.027 | 0.437 \pm 0.044 | 0.502 \pm 0.011 | 0.540 \pm 0.012 | 0.187 \pm 0.053 |
| | Alpaca-LoRA _{ZS_{best}} | 0.571 \pm 0.043 | 0.548 \pm 0.027 | 0.437 \pm 0.044 | 0.502 \pm 0.011 | 0.567 \pm 0.038 | 0.224 \pm 0.049 |
| | FLAN-T5 _{ZS} | 0.659 \pm 0.086 | 0.664 \pm 0.011 | 0.396 \pm 0.006 | 0.643 \pm 0.021 | 0.667 \pm 0.023 | 0.418 \pm 0.012 |
| | FLAN-T5 _{ZS_{best}} | 0.663 \pm 0.079 | 0.674 \pm 0.014 | 0.396 \pm 0.006 | 0.653 \pm 0.011 | 0.667 \pm 0.023 | 0.418 \pm 0.012 |
| | LLaMA2 _{ZS} | 0.720 \pm 0.012 | 0.693 \pm 0.034 | 0.429 \pm 0.013 | 0.589 \pm 0.010 | 0.691 \pm 0.014 | 0.261 \pm 0.018 |
| | LLaMA2 _{ZS_{best}} | 0.720 \pm 0.012 | 0.711 \pm 0.033 | 0.444 \pm 0.021 | 0.643 \pm 0.014 | 0.722 \pm 0.039 | 0.367 \pm 0.043 |
| | GPT-3.5 _{ZS} | 0.685 \pm 0.024 | 0.642 \pm 0.017 | 0.603 \pm 0.017 | 0.460 \pm 0.163 | 0.570 \pm 0.118 | 0.233 \pm 0.009 |
| | GPT-3.5 _{ZS_{best}} | 0.688 \pm 0.045 | 0.653 \pm 0.020 | 0.642 \pm 0.034 | 0.632 \pm 0.020 | 0.617 \pm 0.033 | 0.310 \pm 0.015 |
| | GPT-4 _{ZS} | 0.700 \pm 0.001 | 0.719 \pm 0.013 | 0.588 \pm 0.010 | 0.644 \pm 0.007 | 0.760 \pm 0.009 | 0.418 \pm 0.009 |
| | GPT-4 _{ZS_{best}} | 0.725 \pm 0.009 | 0.719 \pm 0.013 | 0.656 \pm 0.001 | 0.647 \pm 0.014 | 0.760 \pm 0.009 | 0.441 \pm 0.057 |
| Few-shot Prompting | Alpaca _{FS} | 0.632 \pm 0.030 | 0.529 \pm 0.017 | 0.628 \pm 0.005 | — | — | — |
| | FLAN-T5 _{FS} | 0.786 \pm 0.006 | 0.678 \pm 0.009 | 0.432 \pm 0.009 | — | — | — |
| | GPT-3.5 _{FS} | 0.721 \pm 0.010 | 0.665 \pm 0.015 | 0.580 \pm 0.002 | — | — | — |
| | GPT-4 _{FS} | 0.698 \pm 0.009 | 0.724 \pm 0.005 | 0.613 \pm 0.001 | — | — | — |
| Instructional Finetuning | Mental-Alpaca | 0.816 \pm 0.006 | 0.775 \pm 0.006 | 0.746 \pm 0.005 | 0.724 \pm 0.004 | 0.730 \pm 0.048 | 0.403 \pm 0.029 |
| | Mental-FLAN-T5 | 0.802 \pm 0.002 | 0.759 \pm 0.003 | 0.756 \pm 0.001 | 0.677 \pm 0.005 | 0.868 \pm 0.006 | 0.481 \pm 0.006 |
| Baseline | Majority | 0.500 \pm --- | 0.500 \pm --- | 0.250 \pm --- | 0.500 \pm --- | 0.500 \pm --- | 0.200 \pm --- |
| | BERT | 0.783 \pm --- | 0.763 \pm --- | 0.690 \pm --- | 0.678 \pm --- | 0.500 \pm --- | 0.332 \pm --- |
| | Mental-RoBERTa | 0.831 \pm --- | 0.790 \pm --- | 0.736 \pm --- | 0.723 \pm --- | 0.853 \pm --- | 0.373 \pm --- |

But for the huge GPT-4, adding prompts seems less effective, probably because it already contains similar basic information in its knowledge space.

We summarize our **key takeaways** from this section:

- **Both small-scale and large-scale LLMs show promising performance on mental health tasks. FLAN-T5 and GPT-4’s performance is approaching task-specific NLP models.**
- **The prompt design enhancement strategies are generally effective for dialogue-focused models, but not for task-solving-focused models. These strategies work better for critical action prediction tasks such as suicide prediction.**

Table 5. Balanced Accuracy Performance Change using Enhancement Strategies. The green/red color indicates increased/decreased accuracy. This table zooms in on the zero-shot section of Table 4. \uparrow/\downarrow marks the ones with better/worse performance in comparison.

| | Dataset | Dreaddit | DepSeverity | SDCNL | CSSRS-Suicide | | |
|--|---------------------|---------------------|---------------------|---------------------|---------------------|---------------------|-------------------------------|
| Model | Task #1 | Task #2 | Task #3 | Task #4 | Task #5 | Task #6 | $\bar{\Delta}$ —All Six Tasks |
| Δ —Alpaca $_{ZS_context}$ | $\uparrow +0.019$ | $\uparrow +0.045$ | $\uparrow +0.023$ | $\uparrow +0.004$ | $\uparrow +0.014$ | $\uparrow +0.018$ | $\uparrow +0.021$ |
| Δ —Alpaca $_{ZS_mh}$ | $\uparrow +0.000$ | $\uparrow +0.055$ | $\uparrow +0.013$ | $\downarrow -0.011$ | $\uparrow +0.006$ | $\uparrow +0.004$ | $\uparrow +0.011$ |
| Δ —Alpaca $_{ZS_both}$ | $\downarrow -0.053$ | $\uparrow +0.037$ | $\downarrow -0.010$ | $\uparrow +0.039$ | $\downarrow -0.007$ | $\downarrow -0.010$ | $\downarrow -0.001$ |
| Δ —Alpaca-LoRA $_{ZS_context}$ | $\downarrow -0.035$ | $\downarrow -0.047$ | $\downarrow -0.094$ | $\downarrow -0.030$ | $\uparrow +0.027$ | $\uparrow +0.027$ | $\downarrow -0.025$ |
| Δ —Alpaca-LoRA $_{ZS_mh}$ | $\downarrow -0.071$ | $\downarrow -0.047$ | $\downarrow -0.105$ | $\downarrow -0.005$ | $\uparrow +0.017$ | $\uparrow +0.029$ | $\downarrow -0.031$ |
| Δ —Alpaca-LoRA $_{ZS_both}$ | $\downarrow -0.071$ | $\downarrow -0.048$ | $\downarrow -0.051$ | $\downarrow -0.003$ | $\downarrow -0.023$ | $\uparrow +0.037$ | $\downarrow -0.027$ |
| Δ —FLAN-T5 $_{ZS_context}$ | $\uparrow +0.004$ | $\uparrow +0.011$ | $\downarrow -0.018$ | $\uparrow +0.010$ | $\downarrow -0.018$ | $\downarrow -0.040$ | $\downarrow -0.009$ |
| Δ —FLAN-T5 $_{ZS_mh}$ | $\downarrow -0.043$ | $\uparrow +0.003$ | $\downarrow -0.030$ | $\uparrow +0.005$ | $\downarrow -0.013$ | $\downarrow -0.046$ | $\downarrow -0.021$ |
| Δ —FLAN-T5 $_{ZS_both}$ | $\downarrow -0.055$ | $\downarrow -0.003$ | $\downarrow -0.007$ | $\uparrow +0.002$ | $\downarrow -0.010$ | $\downarrow -0.036$ | $\downarrow -0.018$ |
| Δ —LLaMA2 $_{ZS_context}$ | $\downarrow -0.062$ | $\uparrow +0.014$ | $\downarrow -0.019$ | $\uparrow +0.000$ | $\uparrow +0.031$ | $\uparrow +0.106$ | $\uparrow +0.012$ |
| Δ —LLaMA2 $_{ZS_mh}$ | $\downarrow -0.102$ | $\uparrow +0.018$ | $\downarrow -0.033$ | $\uparrow +0.053$ | $\uparrow +0.004$ | $\uparrow +0.031$ | $\downarrow -0.005$ |
| Δ —LLaMA2 $_{ZS_both}$ | $\downarrow -0.136$ | $\uparrow +0.011$ | $\uparrow +0.016$ | $\uparrow +0.054$ | $\downarrow -0.002$ | $\uparrow +0.067$ | $\uparrow +0.002$ |
| Δ —GPT-3.5 $_{ZS_context}$ | $\uparrow +0.003$ | $\uparrow +0.011$ | $\downarrow -0.060$ | $\uparrow +0.157$ | $\uparrow +0.007$ | $\uparrow +0.031$ | $\uparrow +0.025$ |
| Δ —GPT-3.5 $_{ZS_mh}$ | $\downarrow -0.006$ | $\downarrow -0.006$ | $\uparrow +0.039$ | $\uparrow +0.116$ | $\downarrow -0.093$ | $\uparrow +0.077$ | $\uparrow +0.021$ |
| Δ —GPT-3.5 $_{ZS_both}$ | $\downarrow -0.005$ | $\downarrow -0.015$ | $\uparrow +0.014$ | $\uparrow +0.172$ | $\uparrow +0.047$ | $\uparrow +0.020$ | $\uparrow +0.039$ |
| Δ —GPT-4 $_{ZS_context}$ | $\uparrow +0.006$ | $\uparrow +0.000$ | $\uparrow +0.001$ | $\uparrow +0.000$ | $\downarrow -0.007$ | $\uparrow +0.023$ | $\uparrow +0.004$ |
| Δ —GPT-4 $_{ZS_mh}$ | $\uparrow +0.025$ | $\downarrow -0.035$ | $\uparrow +0.067$ | $\uparrow +0.002$ | $\downarrow -0.023$ | $\downarrow -0.022$ | $\uparrow +0.002$ |
| Δ —GPT-4 $_{ZS_both}$ | $\uparrow +0.018$ | $\downarrow -0.031$ | $\uparrow +0.061$ | $\uparrow +0.003$ | $\downarrow -0.063$ | $\downarrow -0.006$ | $\downarrow -0.003$ |
| $\bar{\Delta}$ —All Six Models | $\downarrow -0.031$ | $\uparrow +0.000$ | $\downarrow -0.011$ | $\uparrow +0.032$ | $\downarrow -0.006$ | $\uparrow +0.017$ | $\uparrow +0.000$ |
| $\bar{\Delta}$ —Alpaca, GPT-3.5, GPT-4 | $\uparrow +0.001$ | $\uparrow +0.007$ | $\uparrow +0.017$ | $\uparrow +0.053$ | $\downarrow -0.013$ | $\uparrow +0.015$ | $\uparrow +0.013$ |

- Providing more contextual information about the task & input can consistently improve performance in most cases.
- Dialogue-focused models with larger trainable parameters (Alpaca vs. Alpaca-LoRA, as well as LLaMA2/GPT-3.5 vs. Alpaca) can better leverage the contextual or domain information in the prompts, yet GPT-4 shows less effect in response to different prompts.

5.2 Few-shot Prompting Improves Performance to Some Extent

We then investigate the effectiveness of few-shot prompting. Note that since we observe diverse effects of prompt design strategies in Table 5, in this section, we only experiment with the prompts with the best performance in the zero-shot setting. Moreover, we exclude Alpaca-LoRA due to its less promising results and LLaMA2 due to its high computation cost.

Due to the maximum input token length of models (2048), we focus on Dreaddit and DepSeverity datasets that have a shorter input and experiment with $M = 2$ in Eq. 2 for binary classification and $M = 4/5$ for multi-class

classification, *i.e.*, one sample per class. We repeat the experiment on each task three times and randomize the few shot samples for each run.

We summarize the overall results in the second section of Table 4 and the zoom-in comparison results in Table 6. Although language models with few-shot prompting still underperform task-specific models, providing examples of the task can improve model performance on most tasks compared to zero-shot prompting ($\bar{\Delta}_{FS_vs_ZS} = 4.1\%$). Interestingly, few-shot prompting is more effective for Alpaca_{FS} and FLAN-T5_{FS} ($\bar{\Delta}_{Alpaca} = 8.1\%$, 3/3 tasks positive; $\bar{\Delta}_{FLAN-T5} = 5.9\%$, 3/3 tasks positive) than GPT-3.5_{FS} and GPT-4_{FS} ($\bar{\Delta}_{GPT-3.5} = 1.2\%$, 2/3 tasks positive; $\bar{\Delta}_{GPT-4} = 0.9\%$, 2/3 tasks positive). Especially for Task #3, we observe an improved balanced accuracy of 19.7% for Alpaca but a decline of 2.3% for GPT-3.5, so that Alpaca outperforms GPT-3.5 on this task. A similar situation is observed for FLAN-T5 (improved by 12.7%) and GPT-4 (declined by 0.2%) on Task #1. This may be attributed to the fact that smaller models such as Alpaca and FLAN-T5 can quickly adapt to complex tasks with only a few examples. In contrast, larger models like GPT-3.5 and GPT-4, with their extensive “in memory” data, find it more challenging to rapidly learn from new examples.

This leads to the key message from this experiment: **Few-shot prompting can improve the performance of LLMs on mental health prediction tasks to some extent, especially for small models.**

5.3 Instruction Finetuning Boost Performance for Multiple Tasks Simultaneously

Our experiments so far have shown that zero-shot and few-shot prompting can improve LLMs on mental health tasks to some extent, but their overall performance is still below state-of-the-art task-specific models. In this section, we explore the effectiveness of instruction finetuning.

Due to the prohibitive cost and lack of transparency of GPT-3.5 and GPT-4 finetuning, we only experiment with Alpaca and FLAN-T5 that we have full control of. We picked the most informative prompt to provide more embedded knowledge during the finetuning. As introduced in Sec. 3.3.2 and Sec. 4.2.1, we build Mental-Alpaca and Mental-FLAN-T5 by finetuning Alpaca and FLAN-T5 on all six tasks across four datasets at the same time.

The third section of Table 4 summarizes the overall results, and Table 7 highlights the key comparisons. We observe that both Mental-Alpaca and Mental-FLAN-T5 achieve significantly better performance compared to the unfinetuned versions ($\bar{\Delta}_{Alpaca-FT_vs_ZS} = 23.4\%$, $\bar{\Delta}_{Alpaca-FT_vs_FS} = 18.3\%$; $\bar{\Delta}_{FLAN-T5-FT_vs_ZS} = 14.7\%$, $\bar{\Delta}_{FLAN-T5-FT_vs_FS} = 14.0\%$). Both finetuned models surpass GPT-3.5’s best performance among zero-shot and few-shot settings across all six tasks ($\bar{\Delta}_{Mental-Alpaca_vs_GPT-3.5} = 10.1\%$; $\bar{\Delta}_{Mental-FLAN-T5_vs_GPT-3.5} = 11.6\%$) and outperform GPT-4’s best version in most tasks ($\bar{\Delta}_{Mental-Alpaca_vs_GPT-4} = 4.0\%$, 4/6 tasks positive; $\bar{\Delta}_{Mental-FLAN-T5_vs_GPT-4} =$

Table 6. Balanced Accuracy Performance Change with Few-shot Prompting. This table is calculated between the zero-shot and the few-shot sections of Table 4.

| | Dataset | Dreaddit | DepSeverity | | |
|---------------------------------------|---------|----------|-------------|----------|---------------------------------|
| Model | | Task #1 | Task #2 | Task #3 | $\bar{\Delta}$ –All Three Tasks |
| Δ –Alpaca _{FS_vs_ZS} | | ↑ +0.039 | ↑ +0.007 | ↑ +0.197 | ↑ +0.081 |
| Δ –FLAN-T5 _{FS_vs_ZS} | | ↑ +0.127 | ↑ +0.014 | ↑ +0.036 | ↑ +0.059 |
| Δ –GPT-3.5 _{FS_vs_ZS} | | ↑ +0.036 | ↑ +0.023 | ↓ −0.023 | ↑ +0.012 |
| Δ –GPT-4 _{FS_vs_ZS} | | ↓ −0.002 | ↑ +0.005 | ↑ +0.025 | ↑ +0.009 |
| <hr/> | | | | | |
| $\bar{\Delta}$ –All Four Models | | ↑ +0.051 | ↑ +0.012 | ↑ +0.059 | ↑ +0.041 |

5.5%, 5/6 tasks positive). Recall that GPT-3.5/GPT-4 are 25/250 times bigger than Mental-Alpaca and 15/150 times bigger than Mental-FLAN-T5.

More importantly, Mental-Alpaca and Mental-FLAN-T5 perform on par with the state-of-the-art Mental-RoBERTa. Mental-Alpaca has the best performance on one task and the second best on three tasks, while Mental-FLAN-T5 has the best performance on three tasks. It is noteworthy that Mental-RoBERTa is a task-specific model, which means it is specialized on one task after being trained on it. In contrast, Mental-Alpaca and Mental-FLAN-T5 can simultaneously work across *all* tasks with a single-round finetuning. These results show the strong effectiveness of instruction finetuning: By finetuning LLMs on multiple mental health datasets with instructions, the models can obtain better capability to solve a variety of mental health prediction tasks.

5.3.1 Dialogue-Focused vs. Task-Solving-Focused LLMs. We further compare Mental-Alpaca and Mental-FLAN-T5. Overall, their performance is quite close ($\bar{\Delta}_{\text{FLAN-T5}_{vs_Alpaca}} = 1.4\%$), with Mental-Alpaca better at Task #4 on SDCNL and Mental-FLAN-T5 better at Task #5 and #6 on CSSRS-Suicide. In Sec. 5.1, we observe that FLAN-T5_{ZS} has a much better performance than Alpaca_{ZS} ($\bar{\Delta}_{\text{FLAN-T5}_{vs_Alpaca}} = 10.9\%$, 5/6 tasks positive) in the zero-shot setting. However, after finetuning, FLAN-T5's advantage disappears.

Our comparison result indicates that Alpaca, as a dialogue-focused model, is better at learning from human natural language data compared to FLAN-T5. Although FLAN-T5 is good at task solving and thus has a better performance in the zero-shot setting, its performance improvement after instruction finetuning is relatively smaller than that of Alpaca. This observation has implications for future stakeholders. If the data and computing resources for finetuning are not available, using task-solving-focused LLMs could lead to better results. When there are enough data and computing resources, finetuning dialogue-based models can be a better choice. Furthermore, models like Alpaca, with their dialogue conversation capabilities, may be more suitable for downstream applications, such as mental well-being assistants for end-users.

5.3.2 Does Finetuning Generalize across Datasets? We further measure the generalizability of LLMs after finetuning. To do this, we instruction-finetune the model on one dataset and evaluate it on all datasets (as introduced

Table 7. Balanced Accuracy Performance Change with Instruction Finetuning. This table is calculated between the finetuning and zero-shot section, as well as the finetuning and few-shot section of Table 4. GPT-3.5_{Best} and GPT-4_{Best} are the best results among zero-shot and few-shot settings.

| | Dataset | Dreaddit | DepSeverity | SDCNL | CSSRS-Suicide | |
|---|--------------|--------------|--------------|--------------|---------------|--------------|
| Model | Task #1 | Task #2 | Task #3 | Task #4 | Task #5 | Task #6 |
| Δ -Alpaca _{FT_{vs_ZS}}} | ↑ +0.223 | ↑ +0.253 | ↑ +0.315 | ↑ +0.231 | ↑ +0.212 | ↑ +0.171 |
| Δ -Alpaca _{FT_{vs_FS}}} | ↑ +0.184 | ↑ +0.246 | ↑ +0.118 | — | — | — |
| Δ -FLAN-T5 _{FT_{vs_ZS}}} | ↑ +0.143 | ↑ +0.095 | ↑ +0.360 | ↑ +0.047 | ↑ +0.201 | ↑ +0.034 |
| Δ -FLAN-T5 _{FT_{vs_FS}}} | ↑ +0.016 | ↑ +0.081 | ↑ +0.324 | — | — | — |
| <i>Results comparison:</i> | | | | | | |
| Mental-RoBERTa | 0.831 | 0.790 | 0.736 | <u>0.723</u> | <u>0.853</u> | 0.373 |
| GPT-3.5 _{Best} | 0.721 | 0.665 | 0.642 | 0.632 | 0.617 | 0.310 |
| GPT-4 _{Best} | 0.725 | 0.724 | 0.656 | 0.647 | 0.760 | <u>0.441</u> |
| Mental-Alpaca | <u>0.816</u> | <u>0.775</u> | <u>0.746</u> | 0.724 | 0.730 | 0.403 |
| Mental-FLAN-T5 | 0.802 | 0.759 | 0.756 | 0.677 | 0.868 | 0.481 |

in Sec. 3.3.1). As the main purpose of this part is not to compare different models but evaluate the finetuning method, we only focus on Alpaca. Table 8 summarizes the results.

We first find that finetuning and testing on the same dataset lead to good performance, as indicated by the boxed entries on the diagonal in Table 8. Some results are even better than Mental-Alpaca (5 out of 6 tasks) or Mental-RoBERTa (3 out of 6 tasks), which is not surprising. More interestingly, we investigate cross-dataset generalization performance (*i.e.*, the ones off the diagonal). Overall, finetuning on a single dataset achieves better performance compared to the zero-shot setting ($\bar{\Delta}_{FT-Single_vs_ZS} = 4.2\%$). However, the performance changes vary across tasks. For example, finetuning on any dataset is beneficial for Task #3 ($\bar{\Delta} = 19.2\%$) and #5 ($\bar{\Delta} = 16.4\%$), but detrimental for Task #6 ($\bar{\Delta} = -7.6\%$) and almost futile for Task #4 ($\bar{\Delta} = -0.4\%$). Generalizing across Dreddit and DepSeverity shows good performance, but this is mainly because they share the language corpus. These results indicate that finetuning on a single dataset can provide mental health knowledge with a certain level and thus improve the overall generalization results, but such improvement is not stable across tasks.

Moreover, we further evaluate the generalizability of our best models instructional-finetuned on multiple datasets, *i.e.*, Mental-Alpaca and Mental-FLAN-T5. We leverage external datasets that are not included in the finetuning. Table 9 highlights the key results. More detailed results can be found in Table 10.

Consistent with the results in Table 7, the instruction finetuning enhances the model performance on external datasets ($\bar{\Delta}_{Alpaca} = 16.3\%$, $\bar{\Delta}_{FLAN-T5} = 5.1\%$). Both Mental-Alpaca and Mental-FLAN-T5 ranked top 1 or 2 in 2/3 external tasks. It is noteworthy that Twt-60Users and SAD datasets are collected outside Reddit, and their data is different from the source of finetuning datasets. These results demonstrate strong evidence that instruction finetuning with diverse tasks, even with data collected from a single social media platform, can significantly enhance LLMs' generalizability across multiple scenarios.

5.3.3 How Much Data Is Needed? Additionally, we are interested in exploring how the size of the dataset impacts the results of instruction finetuning. To answer this question, we downsample the training set to 50%, 20%, 10%, 5%, and 1% of the original size and repeat each one three times. We increase the training epoch accordingly to make sure that the model is exposed to a similar amount of data. Similarly, we also focus on Alpaca only. Figure 1 visualizes the results. With only 1% of the data, the finetuned model is able to outperform the zero-shot model on most tasks (5 out of 6). With 5% of the data, the finetuned model has a better performance on all tasks. As expected, the model performance has an increasing trend with more training data. For many tasks, the trend

Table 8. Balanced Accuracy Cross-Dataset Performance Summary of Mental-Alpaca Finetuning on Single Dataset. Numbers indicate the results of the model finetuned and tested on the same dataset. The bottom few rows are related Alpaca versions for reference. \uparrow/\downarrow marks the ones with better/worse cross-dataset performance compared to the zero-shot version Alpaca_{ZS}.

| Test Dataset | Dreaddit | DepSeverity | | SDCNL | CSSRS-Suicide | |
|----------------------|----------|-------------|---------|---------|---------------|---------|
| Finetune Dataset | Task #1 | Task #2 | Task #3 | Task #4 | Task #5 | Task #6 |
| Dreaddit | 0.823 | ↑ 0.720 | ↑ 0.623 | ↓ 0.474 | ↑ 0.720 | ↓ 0.156 |
| DepSeverity | ↑ 0.618 | 0.733 | 0.769 | ↓ 0.493 | ↑ 0.753 | ↓ 0.156 |
| SDCNL | ↓ 0.468 | ↓ 0.461 | ↑ 0.623 | 0.730 | ↑ 0.573 | ↓ 0.156 |
| CSSRS-Suicide | ↓ 0.500 | ↓ 0.500 | ↑ 0.622 | ↑ 0.500 | 0.753 | 0.578 |
| Reference: | | | | | | |
| Alpaca _{ZS} | 0.593 | 0.522 | 0.431 | 0.493 | 0.518 | 0.232 |
| Mental-Alpaca | 0.816 | 0.775 | 0.746 | 0.724 | 0.730 | 0.403 |

Table 9. Balanced Accuracy Performance Summary on Three External Datasets. These datasets come from diverse social media platforms. For each column, the best result is **bolded**, and the second best is underlined.

| | Dataset | Red-Sam | Twt-60Users | SAD |
|--------------------------|---------------------------------------|--------------------------|--------------------------|--------------------------|
| Category | Model | Task #2 | Task #2 | Task #1 |
| Zero-shot Prompting | Alpaca _{ZS_best} | 0.527 \pm 0.006 | 0.569 \pm 0.017 | 0.557 \pm 0.041 |
| | Alpaca-LoRA _{ZS_best} | 0.577 \pm 0.004 | 0.649 \pm 0.021 | 0.477 \pm 0.016 |
| | FLAN-T5 _{ZS_best} | 0.563 \pm 0.029 | 0.613 \pm 0.046 | 0.767 \pm 0.050 |
| | LLaMA2 _{ZS_best} | 0.574 \pm 0.008 | <u>0.736</u> \pm 0.019 | 0.704 \pm 0.026 |
| | GPT-3.5 _{ZS_best} | 0.506 \pm 0.004 | 0.571 \pm 0.000 | 0.750 \pm 0.027 |
| | GPT-4 _{ZS_best} | 0.511 \pm 0.000 | 0.566 \pm 0.017 | 0.854 \pm 0.006 |
| Instructional Finetuning | Mental-Alpaca | 0.604 \pm 0.012 | 0.718 \pm 0.011 | <u>0.819</u> \pm 0.006 |
| | Δ -Alpaca _{FT_vs_ZS} | $\uparrow +0.077$ | $\uparrow +0.149$ | $\uparrow +0.262$ |
| | Mental-FLAN-T5 | <u>0.582</u> \pm 0.002 | 0.736 \pm 0.003 | 0.779 \pm 0.002 |
| | Δ -FLAN-T5 _{FT_vs_ZS} | $\uparrow +0.019$ | $\uparrow +0.123$ | $\uparrow +0.012$ |

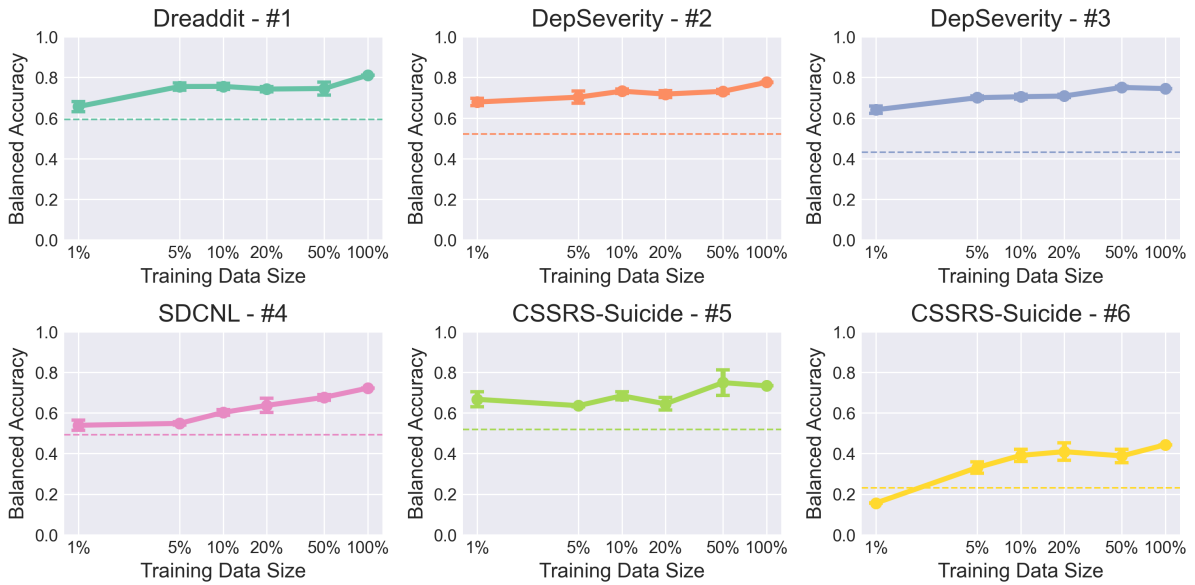


Fig. 1. Balanced Accuracy Performance Summary of Mental-Alpaca Finetuning with Different Sizes of Training Set. The finetuning is conducted across four datasets and six tasks. Each solid line represents the performance of the finetuned model on each task. The dashed line indicates the Alpaca_{ZS} performance baseline. Note that the x-axis is in the log scale.

approaches a plateau after 10%. The difference between 10% training data (less than 300 samples per dataset) and 100% training data is not huge ($\bar{\Delta} = 5.9\%$).

5.3.4 More Data in One Dataset vs. Fewer Data across Multiple Datasets. In Sec. 5.3.2, the finetuning on a single dataset can be viewed as training on a smaller set (around 5-25% of the original size) with less variation (*i.e.*, no

finetuning across datasets). Thus, the results in Sec. 5.3.2 are comparable to those in Sec. 5.3.3. We found that the model’s overall performance is better when the model is finetuned across multiple datasets when overall training data sizes are similar ($\bar{\Delta}_{FT-5\%_vs_FT-Single} = 3.8\%$, $\bar{\Delta}_{FT-10\%_vs_FT-Single} = 8.1\%$, $\bar{\Delta}_{FT-20\%_vs_FT-Single} = 12.4\%$). This suggests that increasing data variation can more effectively benefit finetuning outcomes when the training data size is fixed.

These results can guide future developers and practitioners in collecting the appropriate data size and sources to finetune LLMs for the mental health domain efficiently. We have more discussion in the next section. In summary, we highlight the **key takeaways** of our finetuning experiments as follows:

- **Instruction finetuning on multiple mental health datasets can significantly boost the performance of LLMs on various mental health prediction tasks. Mental-Alpaca and Mental-FLAN-T5 outperform GPT-3.5 and GPT-4, and perform on par with the state-of-the-art task-specific model.**
- **Although task-solving-focused LLMs may have better performance in the zero-shot setting for mental health prediction tasks, dialogue-focused LLMs have a stronger capability of learning from human natural language and can improve more significantly after finetuning.**
- **Finetuning LLMs on a small number of datasets and tasks may improve model generalizable knowledge in mental health, but its effect is not robust. Comparatively, finetuning on diverse tasks can robustly enhance generalizability across multiple social media platforms.**
- **Finetuning LLMs on a small number of samples (a few hundred) across multiple datasets can already achieve favorable performance.**
- **When the data size is the same, finetuning LLMs on data with larger variation (i.e., more datasets and tasks) can achieve better performance.**

5.4 Case Study of LLMs’ Capability on Mental Health Reasoning

In addition to evaluating LLMs’ performance on classification tasks, we also take an initial step to explore LLMs’ capability on mental health reasoning. This is another strong advantage of LLMs since they can generate human-like natural language based on embedded knowledge. Due to the high cost of a systematic evaluation of reasoning outcomes, here we present a few examples as a case study across different LLMs. It is noteworthy that we do not aim to claim that certain LLMs have better/worse reasoning capabilities. Instead, this section aims to provide a general sense of LLMs’ performance on mental health reasoning tasks.

Specifically, we modify the prompt design by inserting a Chain-of-Thought (CoT) prompt [65] at the end of *OutputConstraint* in Eq. 1: “Return [set of classification labels]. Provide reasons step by step”. We compare Alpaca, FLAN-T5, GPT-3.5, and GPT-4. Our results indicate the promising reasoning capability of these models, especially GPT-3.5 and GPT-4. We also experimented with the finetuned Mental-Alpaca and Mental-FLAN-T5. Unfortunately, our results show that after finetuning solely on classification tasks, these two models are no longer able to generate reasoning sentences even with the CoT prompt. This suggests a limitation of the current finetuned model.

5.4.1 Diverse Reasoning Capabilities across LLMs. We present several examples as our case study to illustrate the reasoning capability of these LLMs. The first example comes from the binary stress prediction task (Task #1) on the Dreddit dataset (see Figure 2). All models give the right classification, but with significantly different reasoning capabilities. First, FLAN-T5 generates the shortest reason. Although it is reasonable, it is superficial and does not provide enough insights. This is understandable because FLAN-T5 is targeted at task-solving instead of reasoning. Compared to FLAN-T5, Alpaca generates better reason. Among the five reasons, two of them accurately analyze the user’s mental state given the stressful situations. Meanwhile, GPT-3.5 and GPT-4 generate expert-level high-quality reasons. The inference from the user’s statement is accurate and deep, indicating their

powerful capability of understanding human emotion and mental health. Comparing the two models, GPT-3.5's reason is simpler, following the user's statement point by point and adding basic comments, while GPT-4's output is more organic and insightful, yet more concise.

Input Prompt | Ground Truth (Stressed or Not): yes

Post: We'd be saving so much money with this new housr...**its such an expensive city....** I did some googling in their language and found that it was illegal for them to do that. I was excited like oh ok if it happens i'll call the police on them.... but now i did some further googling and **turns out that if you sign the contract with that in it than everything is on their side...** The contract is long term so we'd have to give 6months notice before leaving **even if we give it next week i'll still have to live six months with this.... i wanna die i swear i havent been so helpless and scared in so long....** Also two days ago my friend mentioned he ran into my ex **who ruined my life** and that he told him **he wants to see mw and get closure.**

Consider this post on social media to answer the question: Is the poster of this post stressful? Return Yes or No. Please reasoning step-by-step.

Yes, the poster of this post is stressful. the reasons are as follows:

1. The poster is expressing feelings of helplessness and fear.
2. The poster is in a vulnerable situation and does not know what to do.
3. The poster is relying on others for help and is unsure of the outcome.
4. The poster is concerned about their safety and the safety of their friend.
5. The post contains strong language and emotions that can be overwhelming for the reader.

Alpaca | Prediction: yes

Yes. The poster is stressed out about their new lease. The poster is stressed out about their ex.

FLAN-T5 | Prediction: yes

Yes. Reasoning:

1. The poster mentions that they would be saving money with their new house, implying financial stress.
2. They describe the city as expensive, indicating financial strain.
3. The poster expresses excitement initially upon discovering that the actions of someone involved in the housing situation may be illegal, suggesting a potential problem or conflict.
4. However, their excitement turns to despair upon discovering that the terms of the contract are unfavorable and they feel trapped.
5. The statement "i wanna die i swear i havent been so helpless and scared in so long" conveys extreme distress and feelings of helplessness.
6. Additionally, the mention of running into an ex who ruined their life and wanting closure suggests emotional turmoil and potentially traumatic experiences.

GPT-3.5 | Prediction: yes

Yes. The poster expresses a high level of stress primarily due to three main reasons:

1. They are dealing with complex legal concerns related to their new home in an expensive city. This would inherently cause stress due to the high stakes, financial implication, and complex legalese that they may not fully understand.
2. The long-term contract binds them to live in the house for a specified period of time regardless of when they decide they wish to terminate the contract. This lack of flexibility may exacerbate the stress as they feel trapped in the situation.
3. Finally, the reference to "my ex who ruined my life" indicates ongoing emotional strife that is being re-opened by a potential meeting. This emotional component is a considerable stressor, and the language used (e.g. "i wanna die" and "i havent been so helpless and scared") strongly suggests intense distress and anxiety.

GPT-4 | Prediction: yes

Fig. 2. A Case Study of Correct Reasoning Examples on Task #1 Binary Stress Prediction on Dreddit Dataset. **Bolded texts** highlight the mental-health-related content in the input section, and the answers of LLMs. Underlined texts highlight the reasoning content generated by LLMs, and *italicized & underlined texts* indicate the *wrong* or *unrelated* content.

Input Prompt | Ground Truth (Level of Depression): severe

Post: It sent me into a **full-blown crisis** by **breaking open the “no one believes me and I'm worthless” wound**. I had to **up my level of care, now I'm in IOP 5 days a week, 3+ hours a day**. It's given me just enough relief to return to **weekly EMDR**. I had my first session in this round, last week. It was **the most physically symptomatic I've ever been** in-session.

Consider this post on social media to answer the question: Determine how depressed the person is. Return one item from: minimum, mild, moderate, severe. Provide reasons step by step.

I would rate the person's depression as **moderate**. They have had to up their level of care and are now in iop 5 days a week, 3+ hours a day which is a significant increase from their previous level of care. This indicates that they are struggling with their mental health and need more intensive support. Additionally, the symptoms they described in their post indicate that they are experiencing moderate depression.

Alpaca | Prediction: moderate

Based on the given post, it is difficult to accurately determine the level of depression the person is experiencing. However, we can analyze the post's content and language to identify some indicators.

1. "It sent me into a full-blown crisis by breaking open the 'no one believes me and i'm worthless' wound." This statement suggests a significant emotional impact, implying a potential for moderate to severe depression.
 2. "I had to up my level of care, now I'm in iop 5 days a week, 3+ hours a day." The fact that the person had to intensify their level of care suggests a more severe level of depression.
 3. "It's given me just enough relief to return to weekly emdr." This implies that their previous level of care was not providing sufficient relief, indicating a degree of severity in their depression.
 4. "I had my first session in this round last week. It was the most physically symptomatic I've ever been in-session." The mention of being highly physically symptomatic during the therapy session suggests a severe level of depression.
- Considering these factors, it can be inferred that the person's depression level falls within the range of **moderate to severe**.

GPT-3.5 | Prediction: moderate to severe

The answer is **severe**. The person is in a full-blown crisis. The person is in iop 5 days a week, 3+ hours a day. The person is in weekly emdr.

FLAN-T5 | Prediction: severe

Severe. Reasoning: The person states that they have entered a "full-blown crisis" and there's a clear expression of lowered self-worth, both of which can indicate serious depressive symptoms. They state they have had to increase the intensity of their treatment to intensive outpatient (iop) therapy five days a week for over three hours, which suggests that their mental health condition is severe. The person is also going through weekly eye movement desensitization and reprocessing (emdr) therapy sessions, which are often used in cases of serious trauma or distress. Lastly, they mention that their physical symptoms in therapy sessions have been the worst they have ever experienced, further implying a severe level of depression.

GPT-4 | Prediction: severe

Fig. 3. A Case Study of Mixed Reasoning Examples on Task #3 Four-level Depression Prediction on DepSeverity Dataset. Alpaca wrongly predicted the label, and GPT-3.5 provided a wrong inference on the meaning of “relief”.

We also have a similar observation in the second example from the four-level depression prediction task (Task #3) on the DepSeverity dataset (see Figure 3). In this example, although FLAN-T5's prediction is correct, it simply repeats the fact stated by the user, without providing further insights. Alpaca makes the wrong prediction, but it provides one sentence of accurate reasoning (although relatively shallow). GPT-3.5 makes an ambiguous prediction that includes the correct answer. In contrast, GPT-4 generates the highest quality reasoning with the right prediction. With its correct understanding of depressive symptoms, GPT-4 can accurately infer from the user's situation, link it to symptoms, and provides insightful analysis.

5.4.2 Wrong and Dangerous Reasoning from LLMs. However, we also want to emphasize the *incorrect* reasoning content, which may lead to negative consequences and risks. In the first example, Alpaca generated two wrong

Input Prompt | Ground Truth (Stressed or Not): no

Post: **If I go to an interview for example**, I'll know that I'm a good candidate, I'll know that if I don't get it there will always be other opportunities and it's no big deal. Yet I still get nervous because it's something that I want, I want that employer to like me. If I go to an interview with no expectations at all, not even wanting the job (I've done this a few times for practice interviewing), it'll turn out great. **What are your thoughts on this?** Edit: FYI I'm talking mostly about social anxiety, though it has happened that I get anxiety in the most random places like just going upstairs in a building.

Consider this post on social media to answer the question: Is the poster of this post stressful? Return Yes or No. Please reasoning step-by-step.

No, i don't think the poster of this post is stressful. The poster is expressing a legitimate feeling of nervousness before an upcoming interview, but they are doing so in a way that is not overly dramatic or alarmist. They recognize that they are a good candidate and that there will be other opportunities if this one doesn't work out. This is a healthy and reasonable way to approach the situation.

Alpaca | Prediction: no

The poster talks about their anxiety. People with anxiety are stressful. The final answer: **yes**.

FLAN-T5 | Prediction: yes

Yes. Reasoning:
 1. The poster mentions feeling nervous before interviews because they want the employer to like them. This indicates that they are feeling stressed about the outcome of the interview and the impression they make on the employer.
 2. The poster also mentions experiencing anxiety in random places like going upstairs in a building. This suggests that they may have generalized anxiety beyond just specific situations like job interviews.
 3. Overall, the language and tone of the post indicate that the poster experiences stress and anxiety.

GPT-3.5 | Prediction: yes

Yes. The reasoning behind this is quite clear from the content of the post. The poster specifically mentions experiencing anxiety in social situations such as job interviews, even when they understand the circumstances and potential consequences. They also mention feeling anxiety in random situations like going upstairs in a building. These experiences and feelings are consistent with those reported by individuals dealing with stress and anxiety disorders.

GPT-4 | Prediction: yes

Fig. 4. A Case Study of **Incorrect** Reasoning Examples on Task #1 Binary Stress Prediction on Dreddit Dataset. FLAN-T5, GPT-3.5, and GPT-4 all make false positive predictions. All four LLMs provide problematic reasons.

reasons for the hallucinated “reliance on others” and “safety concerns”, along with an unrelated reason for readers instead of the poster. In the second example, GPT-3.5 misunderstood what the user meant by “relief”. To better illustrate this, we further present another example where all four LLMs make problematic reasoning (see Figure 4). In this example, the user was asking for others’ opinions on social anxiety, with their own job interview experience as an example. Although the user mentioned situations where they were anxious and stressed, it’s clear that they were calm when writing this post and described their experience in an objective way. However, FLAN-T5, GPT-3.5, and GPT-4 all mistakenly take the description of the anxious interview experience as evidence to support their wrong prediction. Although Alpaca makes the right prediction, it does not understand the main theme of the post. The false positives reveal that LLMs may overly generalize in a wrong way: Being stressed in one situation does not indicate that a person is stressed all the time. However, the reasoning content alone reads smoothly and logically. If the original post was not provided, the content could be very misleading, resulting in a wrong prediction with reasons that “appears to be solid”. These examples clearly illustrate the limitations of the current LLMs for mental health reasons, as well as their risks of introducing dangerous bias and negative consequences to users.

The case study suggests that GPT-4 enjoys impressive reasoning capability, followed by GPT-3.5 and Alpaca. Although FLAN-T5 shows a promising zero-shot performance, it is not good at reasoning. Our results reveal the

encouraging capability of LLMs to understand human mental health and generate meaningful analysis. However, we also present examples where LLMs can make mistakes and offer explanations that appear reasonable but are actually flawed. This further suggests the importance of more future research on LLMs' ethical concerns and safety issues before real-world deployment.

6 DISCUSSION

Our experiment results reveal a number of interesting findings. In this section, we discuss potential guidelines for enabling LLMs for mental health-related tasks (Sec. 6.1). We envision promising future directions (Sec. 6.2), while highlighting important ethical concerns and limitations with LLMs for mental health (Sec. 6.3). We also summarize the limitations of the current work (Sec. 6.4).

6.1 Guidelines for Empowering LLMs for Mental Health Prediction Tasks

We extract and summarize the takeaways from Sec. 5 into a set of guidelines for future researchers and practitioners on how to empower LLMs to be better at various mental health prediction tasks.

When computing resources are limited, combine prompt design & few-shot prompting, and pick prompts carefully. As the size of large models continues to grow, the requirement for hardware (mainly GPU) has also been increasing, especially when finetuning an LLM. For example, in our experiment, Alpaca was trained on eight 80GB A100s for three hours [115]. With limited computing resources, only running inferences or resorting to APIs is feasible. In these cases, zero-shot and few-shot prompt engineering strategies are viable options. Our results indicate that providing few-shot mental health examples with appropriate enhancement strategies can effectively improve prediction performance. Specifically, adding contextual information about the online text data is always helpful. If the available model is large and contains rich knowledge (at least 7B trainable parameters), adding mental health domain information can also be beneficial.

With enough computing resources, instruction finetune models on various mental health datasets. When there are enough computing resources and model training/finetuning is possible, there are more options to enhance LLMs for mental health prediction tasks. Our experiments clearly show that instruction finetuning can significantly boost the performance of models, especially for dialogue-based models since they can better understand and learn from human natural language. When there are multiple datasets available, merging multiple datasets and tasks altogether and finetuning the model in a single round is the most effective approach to enhance its generalizability.

Implement efficient finetuning with hundreds of examples and prioritize data variation when data resource is limited. Figure 1 shows that finetuning does not require large datasets. If there is no immediately available dataset, collecting small datasets with a few hundred samples is often good enough. Moreover, when the overall amount of data is limited (e.g., due to resource constraints), it is more advantageous to collect data from a variety of sources, each with a smaller size, than to collect a single larger dataset. Because instruction finetuning generalizes better when data and tasks have a larger variation.

More curated finetuning datasets are needed for mental health reasoning. Our case study suggests that Mental-Alpaca and Mental-FLAN-T5 can only generate classification labels after being finetuned solely on classification tasks, losing their reasoning capability. This is a major limitation of the current models. A potential solution involves incorporating more datasets focused on reasoning or causality into the instruction finetuning process, so that models can also learn the relationship between mental health outcomes and causal factors.

Limited Prediction and Reasoning Performance for Complex Contexts. LLMs tend to make more mistakes when the conversation contexts are more complex [13, 69]. Our results contextualize this in the mental health domain. Section 5.4.2 shows an example case where most LLMs not only predict incorrectly but also provide flawed reasoning processes. Further analysis of mispredicted instances indicates a recurring difficulty:

LLMs often err when there's a disconnect between the literal context and the underlying real-life scenarios. The example in Figure 4 is a case where LLMs are confused by the hypothetical stressful case described by the person. In another example, all LLMs incorrectly assess a person with severe depression (false negative): *"I'm just blown away by this doctor's willingness to help. I feel so validated every time I leave his office, like someone actually understands what I'm struggling with, and I don't have to convince them of my mental illness. Bottom line? Research docs if you can online, read their reviews and don't give up until you find someone who treats you the way you deserve. If I can do this, I promise you can!"* Here, LLMs are misled by the outwardly positive sentiment, overlooking the significant cues such as regular doctor visits and explicit mentions of mental illness. These observations underscore a critical shortfall of LLMs: they cannot handle complex mental health-related tasks, particularly those concerning chronic conditions like depression. The variability of human expressions over time and the models' susceptibility to being swayed by superficial text rather than underlying scenarios present significant challenges.

Despite the promising capability of LLMs for mental health tasks, they are still far from being deployable in the real world. Our experiments reveal the encouraging performance of LLMs on mental health prediction and reasoning tasks. However, as we note in Sec. 6.3, our current results do not indicate LLMs' deployability in real-life mental health settings. There are many important ethical and safety concerns and gaps before deployment to be addressed before achieving robustness and deployability.

6.2 Beyond Mental Health Prediction Task and Online Text Data

Our current experiments mainly involve mental health prediction tasks, which are essentially classification problems. There are more types of tasks that our experiments don't cover, such as regression (e.g., predicting a score on a mental health scale). In particular, reasoning is an attractive task as it can fully leverage the capability of LLMs on language generation [18, 85]. Our initial case study on reasoning is limited but reveals promising performance, especially for large models such as GPT-4. We plan to conduct more experiments on tasks that go beyond classification.

In addition, there is another potential extension direction. In this paper, we mainly focus on online text data, which is one of the important data sources of the ubiquitous computing ecosystem. However, there are more available data streams that contain rich information, such as the multimodal sensor data from mobile phones and wearables (e.g., [55, 78, 81, 130, 131]). This leads to another open question on how to leverage LLMs for time-series sensor data. More research is needed to explore potential methods to merge the online text information with sensor streams. These are another set of exciting research questions to explore in future work.

6.3 Ethics in LLMs and Deployability Gaps for Mental Health

Although our experiments on LLMs have shown promising capability for mental-health-related tasks, it still has a long way to go before being deployed in real-life systems. Recent research has revealed the potential bias or even harmful advice introduced by LLMs [50], especially with the gender [42] and racial [6] gaps. In mental health, these gaps and disparities between population groups have been long-standing [54]. Meanwhile, our case study of incorrect prediction, over-generalization, and "falsely reasonable" explanations further highlight the risk of current LLMs. Recent studies are calling for more research emphasis and efforts in assessing and mitigating these biases for mental health [54, 116].

Although with a much stronger capability of understanding natural language (and early signs of mental health domain knowledge in our case), LLMs are no different from other modern AI models that are trained on a large amount of human-generated content, which exhibit all the biases that humans do [53, 86, 121]. Meanwhile, although we carefully picked datasets with human expert annotations, there exist potential biases in the labels, such as stereotypes [92], confirmation bias [41], normative vs. descriptive labels [11]. Besides, privacy is another

important concern. Although our datasets are based on public social media platforms, it is necessary to carefully handle mental-health-related data and guarantee anonymity in any future efforts. These ethical concerns need to receive attention not only at the monitoring and prediction stage, but also in the downstream applications, ranging from assistants for mental health experts to chatbots for end-users. Careful efforts into safe development, auditing, and regulation are very much needed to address these ethical risks.

6.4 Limitations

Our paper has a few limitations. First, although we carefully inspect the quality of our dataset and cover different categories of LLM, the range of datasets and the types of LLMs included are still limited. Our findings are based on the observations of these datasets and models, which may not generalize to other cases. Related, our exploration of zero-shot few-shot prompt design is not comprehensive. The limited input window of some models also limits our exploration of more samples for few-shot prompting. Furthermore, we have not conducted a systematic evaluation of these models' performance in mental health reasoning. Future work can design larger-scale experiments to include more datasets, models, prompt designs, and better evaluation.

Second, our datasets were mainly from Reddit, which could be limited. Although our analysis in Section 5.3.2 shows that finetuned models have cross-platform generalizability, the finetuning was only based on Reddit and can introduce bias. Meanwhile, although the labels are not directly accessible on the platforms, it is possible that these text data have been included in the initial training of these large models. We still argue that there is little information leakage as long as the models haven't seen the labels for our tasks, but it is hard to measure how the initial training process may affect the outcomes in our evaluation.

Third, another important limitation of the current work is the lack of evaluation of model fairness. Our anonymous datasets do not include comprehensive demographic information, making it hard to compare the performance across different population groups. As we discussed in the previous section, lots of future work on ethics and fairness is necessary before deploying such systems in real life.

7 CONCLUSION

In this paper, we present the first comprehensive evaluation of multiple LLMs (Alpaca, Alpaca-LoRA, FLAN-T5, LLaMA2, GPT-3.5, and GPT-4) on mental health prediction tasks (binary and multi-class classification) via online text data. Our experiments cover zero-shot prompting, few-shot prompting, and instruction finetuning. The results reveal a number of interesting findings. Our context enhancement strategy can robustly improve performance for all LLMs, and our mental health enhancement strategy can enhance models with a large number of trainable parameters. Meanwhile, few-shot prompting can also robustly improve model performance even by providing just one example per class. Most importantly, our experiments show that instruction finetuning across multiple datasets can significantly boost model performance on various mental health prediction tasks at the same time, generalizing across external data sources and platforms. Our best finetuned models, Mental-Alpaca and Mental-FLAN-T5, outperform much larger LLaMA2, GPT-3.5 and GPT-4, and perform on par with the state-of-the-art task-specific model Mental-RoBERTa. We also conduct an exploratory case study on these models' reasoning capability, which further suggests both the promising future and the important limitations of LLMs. We summarize our findings as a set of guidelines for future researchers, developers, and practitioners who want to empower LLMs with better knowledge of mental health for downstream tasks. Meanwhile, we emphasize that our current efforts of LLMs in mental health are still far from deployability. We highlight the important ethical concerns accompanying this line of research.

ACKNOWLEDGMENTS

This work is supported by VW Foundation, Quanta Computing, and the National Institutes of Health (NIH) under Grant No. 1R01MD018424-01.

REFERENCES

- [1] 2022. Introducing ChatGPT. <https://openai.com/blog/chatgpt>
- [2] 2023. Mental Health By the Numbers. <https://nami.org/mhstats>
- [3] 2023. Mental Illness. <https://www.nimh.nih.gov/health/statistics/mental-illness>
- [4] Alaa A. Abd-alrazaq, Mohammad Alajlani, Ali Abdallah Alalwan, Bridgette M. Bewick, Peter Gardner, and Mowafa Househ. 2019. An overview of the features of chatbots in mental health: A scoping review. *International Journal of Medical Informatics* 132 (Dec. 2019), 103978. <https://doi.org/10.1016/j.ijmedinf.2019.103978>
- [5] Alaa A. Abd-Alrazaq, Mohammad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients about mental health chatbots: scoping review. *Journal of medical Internet research* 23, 1 (2021), e17828.
- [6] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 298–306.
- [7] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 1998–2022.
- [8] Arfan Ahmed, Sarah Aziz, Carla T Toro, Mahmood Alzubaidi, Sara Irshaidat, Hashem Abu Serhan, Alaa A Abd-Alrazaq, and Mowafa Househ. 2022. Machine learning models to detect anxiety and depression through social media: A scoping review. *Computer Methods and Programs in Biomedicine Update* (2022), 100066.
- [9] Mental Health America. 2022. The state of mental health in America.
- [10] Mostafa M. Amin, Erik Cambria, and Björn W. Schuller. 2023. Will Affective Computing Emerge from Foundation Models and General AI? A First Evaluation on ChatGPT. <http://arxiv.org/abs/2303.03186>
- [11] Aparna Balagopalan, David Madras, David H. Yang, Dylan Hadfield-Menell, Gillian K. Hadfield, and Marzyeh Ghassemi. 2023. Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data. *Science Advances* 9, 19 (May 2023), eabq0701. <https://doi.org/10.1126/sciadv.abq0701>
- [12] Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538* (2017).
- [13] Sourangshu Bhattacharya, Avishek Anand, et al. 2023. In-Context Ability Transfer for Question Decomposition in Complex QA. *arXiv preprint arXiv:2310.18371* (2023).
- [14] Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of medical Internet research* 19, 8 (2017), e7956.
- [15] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. 2007. Large language models in machine translation. (2007).
- [16] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*. IEEE, 3121–3124.
- [17] 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*, Vol. 33. Curran Associates, Inc., 1877–1901. <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- [18] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. <http://arxiv.org/abs/2303.12712>
- [19] Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine Classification and Analysis of Suicide-Related Communication on Twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media - HT '15*. ACM Press, Guzelyurt, Northern Cyprus, 75–84. <https://doi.org/10.1145/2700171.2791023>
- [20] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2017. Towards a chatbot for digital counselling. <https://doi.org/10.14236/ewic/HCI2017.24>
- [21] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2019. Assessing the Usability of a Chatbot for Mental Health Care. In *Internet Science*, Svetlana S. Bodrunova, Olessia Koltsova, Asbjørn Følstad, Harry Halpin, Polina Kolozaridi, Leonid Yuldashev, Anna Smoliarova, and Heiko Niedermayer (Eds.). Vol. 11551.

- Springer International Publishing, Cham, 121–132. https://doi.org/10.1007/978-3-030-17705-8_11 Series Title: Lecture Notes in Computer Science.
- [22] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine* 3, 1 (March 2020), 43. <https://doi.org/10.1038/s41746-020-0233-7>
 - [23] 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 Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben 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 Garcia, 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 Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. <http://arxiv.org/abs/2204.02311> arXiv:2204.02311 [cs].
 - [24] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. <http://arxiv.org/abs/2210.11416> arXiv:2210.11416 [cs].
 - [25] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From ADHD to SAD: Analyzing the Language of Mental Health on Twitter through Self-Reported Diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Denver, Colorado, 1–10. <https://doi.org/10.3115/v1/W15-1201>
 - [26] 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*. Association for Computational Linguistics, Denver, Colorado, 31–39. <https://doi.org/10.3115/v1/W15-1204>
 - [27] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring Post Traumatic Stress Disorder in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 579–582. <https://doi.org/10.1609/icwsm.v8i1.14574>
 - [28] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights* 10 (2018), 1178222618792860.
 - [29] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural Language Processing of Social Media as Screening for Suicide Risk. *Biomedical Informatics Insights* 10 (Jan. 2018), 117822261879286. <https://doi.org/10.1177/1178222618792860>
 - [30] Aron Culotta. 2014. Estimating county health statistics with twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 1335–1344. <https://doi.org/10.1145/2556288.2557139>
 - [31] Hai Dang, Lukas Mecke, Florian Lehmann, Sven Goller, and Daniel Buschek. 2022. How to prompt? Opportunities and challenges of zero-and few-shot learning for human-AI interaction in creative applications of generative models. *arXiv preprint arXiv:2209.01390* (2022).
 - [32] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, Paris France, 47–56. <https://doi.org/10.1145/2464464.2464480>
 - [33] 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 (May 2014), 71–80. <https://doi.org/10.1609/icwsm.v8i1.14526>
 - [34] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting Depression via Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (2013), 128–137. <https://doi.org/10.1609/icwsm.v7i1.14432>
 - [35] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, San Jose California USA, 2098–2110. <https://doi.org/10.1145/2858036.2858207>
 - [36] Kerstin Denecke, Sayan Vaaheesan, and Aaganya Arulnathan. 2020. A mental health chatbot for regulating emotions (SERMO)-concept and usability test. *IEEE Transactions on Emerging Topics in Computing* 9, 3 (2020), 1170–1182.
 - [37] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). <http://arxiv.org/abs/1810.04805>
 - [38] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preotiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences* 115, 44 (2018), 11203–11208.
 - [39] Mirko Franco, Ombretta Gaggi, and Claudio E Palazzi. 2023. Analyzing the use of large language models for content moderation with chatgpt examples. In *Proceedings of the 3rd International Workshop on Open Challenges in Online Social Networks*. 1–8.

- [40] Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware Assessment of Severity of Suicide Risk for Early Intervention. In *The World Wide Web Conference*. ACM, San Francisco CA USA, 514–525. <https://doi.org/10.1145/3308558.3313698>
- [41] Meric Altug Gemalmaz and Ming Yin. 2021. Accounting for Confirmation Bias in Crowdsourced Label Aggregation.. In *IJCAI*. 1729–1735.
- [42] Sourojit Ghosh and Aylin Caliskan. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. *arXiv preprint arXiv:2305.10510* (2023).
- [43] 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*. 63–73.
- [44] Sarah Graham, Colin Depp, Ellen E Lee, Camille Nebeker, Xin Tu, Ho-Cheol Kim, and Dilip V Jeste. 2019. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports* 21 (2019), 1–18.
- [45] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. 2017. On integrating a language model into neural machine translation. *Computer Speech & Language* 45 (2017), 137–148.
- [46] Sharath Chandra Guntuku, Anneke Buffone, Kokil Jaidka, Johannes C Eichstaedt, and Lyle H Ungar. 2019. Understanding and measuring psychological stress using social media. In *Proceedings of the international AAAI conference on web and social media*, Vol. 13. 214–225.
- [47] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (Dec. 2017), 43–49. <https://doi.org/10.1016/j.cobeha.2017.07.005>
- [48] Sooji Han, Rui Mao, and Erik Cambria. 2022. Hierarchical attention network for explainable depression detection on Twitter aided by metaphor concept mappings. *arXiv preprint arXiv:2209.07494* (2022).
- [49] Ayaan Haque, Viraj Reddi, and Tyler Giallanza. 2021. Deep Learning for Suicide and Depression Identification with Unsupervised Label Correction. <http://arxiv.org/abs/2102.09427> arXiv:2102.09427 [cs].
- [50] Amanda Hoover. 2023. An eating disorder chatbot is suspended for giving harmful advice. <https://www.wired.com/story/tessa-chatbot-suspended/>
- [51] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. <http://arxiv.org/abs/2106.09685> arXiv:2106.09685 [cs].
- [52] Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610* (2022).
- [53] Irene Y. Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadja Ferryman, and Marzyeh Ghassemi. 2021. Ethical Machine Learning in Healthcare. *Annual Review of Biomedical Data Science* 4, 1 (2021), 123–144. <https://doi.org/10.1146/annurev-biodatasci-092820-114757> eprint: <https://doi.org/10.1146/annurev-biodatasci-092820-114757>
- [54] Irene Y. Chen, Peter Szolovits, and Marzyeh Ghassemi. 2019. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA Journal of Ethics* 21, 2 (Feb. 2019), E167–179. <https://doi.org/10.1001/amajethics.2019.167>
- [55] M. J. N. Bento e Silva J. Abrantes. 2023. External validation of a deep learning model for breast density classification. In *ECR 2023 EPOS*. <https://epos.myesr.org/poster/esr/ecr2023/C-16014>
- [56] Zunaira Jamil, Diana Inkpen, Prasadith Buddhitha, and Kenton White. 2017. Monitoring Tweets for Depression to Detect At-risk Users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*, Kristy Hollingshead, Molly E. Ireland, and Kate Loveys (Eds.). Association for Computational Linguistics, Vancouver, BC, 32–40. <https://doi.org/10.18653/v1/W17-3104>
- [57] Shaoxiong Ji, Celina Ping Yu, Sai-fu Fung, Shirui Pan, and Guodong Long. 2018. Supervised learning for suicidal ideation detection in online user content. *Complexity* 2018 (2018).
- [58] Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. <http://arxiv.org/abs/2110.15621>
- [59] Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, Madeline Miceli, Nora C. Kim, Cordelia Orillac, Zane Schnurman, Christopher Livia, Hannah Weiss, David Kurland, Sean Neifert, Yosef Dastagirzada, Douglas Kondziolka, Alexander T. M. Cheung, Grace Yang, Ming Cao, Mona Flores, Anthony B. Costa, Yindalon Aphinyanaphongs, Kyunghyun Cho, and Eric Karl Oermann. 2023. Health system-scale language models are all-purpose prediction engines. *Nature* (June 2023). <https://doi.org/10.1038/s41586-023-06160-y>
- [60] Zheng Ping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 2020. Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th international workshop on health text mining and information analysis*. 147–156.
- [61] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the benefits and challenges of deploying conversational AI leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

- [62] S Kayalvizhi, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, et al. 2022. Findings of the shared task on detecting signs of depression from social media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 331–338.
- [63] Samuel Kernan Freire, Mina Foosherian, Chaofan Wang, and Evangelos Niforatos. 2023. Harnessing Large Language Models for Cognitive Assistants in Factories. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. 1–6.
- [64] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* (2023), 101861.
- [65] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *36th Conference on Neural Information Processing Systems*.
- [66] Kaylee Payne Kruzan, Kofoworola D.A. Williams, Jonah Meyerhoff, Dong Whi Yoo, Linda C. O'Dwyer, Munmun De Choudhury, and David C. Mohr. 2022. Social media-based interventions for adolescent and young adult mental health: A scoping review. *Internet Interventions* 30 (Dec. 2022), 100578. <https://doi.org/10.1016/j.invent.2022.100578>
- [67] Bishal Lamichhane. 2023. Evaluation of ChatGPT for NLP-based Mental Health Applications. <http://arxiv.org/abs/2303.15727>
- [68] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (May 2020), 1–27. <https://doi.org/10.1145/3392836>
- [69] Yucheng Li, Bo Dong, Chenghua Lin, and Frank Guerin. 2023. Compressing Context to Enhance Inference Efficiency of Large Language Models. *arXiv preprint arXiv:2310.06201* (2023).
- [70] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. <http://arxiv.org/abs/2303.14070> arXiv:2303.14070 [cs].
- [71] Xin Liu, Daniel McDuff, Geza Kovacs, Isaac Galatzer-Levy, Jacob Sunshine, Jiening Zhan, Ming-Zher Poh, Shun Liao, Paolo Di Achille, and Shwetak Patel. 2023. Large Language Models are Few-Shot Health Learners. In *arXiv*.
- [72] 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. <http://arxiv.org/abs/1907.11692> arXiv:1907.11692 [cs].
- [73] James D. Livingston, Michelle Cianfrone, Kimberley Korf-Uzan, and Connie Coniglio. 2014. Another time point, a different story: one year effects of a social media intervention on the attitudes of young people towards mental health issues. *Social Psychiatry and Psychiatric Epidemiology* 49, 6 (June 2014), 985–990. <https://doi.org/10.1007/s00127-013-0815-7>
- [74] Christopher A Lovejoy. 2019. Technology and mental health: the role of artificial intelligence. *European Psychiatry* 55 (2019), 1–3.
- [75] Maria Luce Lupetti, Emma Hagens, Willem Van Der Maden, Régine Steegers-Theunissen, and Melek Rousian. 2023. Trustworthy Embodied Conversational Agents for Healthcare: A Design Exploration of Embodied Conversational Agents for the periconception period at Erasmus MC. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. 1–14.
- [76] 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*. ACM, Yokohama Japan, 1–7. <https://doi.org/10.1145/3411763.3451799>
- [77] 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*. Association for Computational Linguistics, Denver, Colorado, 11–20. <https://doi.org/10.3115/v1/W15-1202>
- [78] Margarida Morais, Francisco Maria Calisto, Carlos Santiago, Clara Aleluia, and Jacinto C Nascimento. 2023. Classification of breast cancer in Mri with multimodal fusion. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*. IEEE, 1–4.
- [79] Megan A Moreno, Lauren A Jelenchick, Katie G Egan, Elizabeth Cox, Henry Young, Kerry E Gannon, and Tara Becker. 2011. Feeling bad on Facebook: Depression disclosures by college students on a social networking site. *Depression and anxiety* 28, 6 (2011), 447–455.
- [80] Usman Naseem, Adam G. Dunn, Jinman Kim, and Matloob Khushi. 2022. Early Identification of Depression Severity Levels on Reddit Using Ordinal Classification. In *Proceedings of the ACM Web Conference 2022*. ACM, Virtual Event, Lyon France, 2563–2572. <https://doi.org/10.1145/3485447.3512128>
- [81] Subigya Nepal, Gonzalo J. Martinez, Shayan Mirjafari, Koustuv Saha, Vedant Das Swain, Xuhai Xu, Pino G. Audia, Munmun De Choudhury, Anind K. Dey, Aaron Striegel, and Andrew T. Campbell. 2022. A Survey of Passive Sensing in the Workplace. *arXiv:2201.03074* [cs.HC]
- [82] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE transactions on affective computing* 5, 3 (2014), 217–226.
- [83] Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. *arXiv preprint arXiv:2204.10432* (2022).
- [84] Tanya Nijhawan, Girija Attigeri, and T Ananthakrishna. 2022. Stress detection using natural language processing and machine learning over social interactions. *Journal of Big Data* 9, 1 (2022), 1–24.
- [85] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. <http://arxiv.org/abs/2303.13375> arXiv:2303.13375 [cs].

- [86] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.
- [87] Reham Omar, Omij Mangukiya, Panos Kalnis, and Essam Mansour. 2023. Chatgpt versus traditional question answering for knowledge graphs: Current status and future directions towards knowledge graph chatbots. *arXiv preprint arXiv:2302.06466* (2023).
- [88] Norio Otsuka, Yuu Kawanishi, Fumimaro Doi, Tsutomu Takeda, Kazuki Okumura, Takahira Yamauchi, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki, and Manabu Makinodan. [n. d.]. Diagnosing Psychiatric Disorders from History of Present Illness Using a Large-Scale Linguistic Model. *Psychiatry and Clinical Neurosciences* ([n. d.]).
- [89] Minsu Park, David McDonald, and Meeyoung Cha. 2021. Perception Differences between the Depressed and Non-Depressed Users in Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 7, 1 (Aug. 2021), 476–485. <https://doi.org/10.1609/icwsm.v7i1.14425>
- [90] Vivek Patel, Piyush Mishra, and JC Patni. 2018. PsyHeal: An Approach to Remote Mental Health Monitoring System. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. IEEE, 384–393.
- [91] Michael Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. *Proceedings of the International AAAI Conference on Web and Social Media* 5, 1 (2011), 265–272. <https://doi.org/10.1609/icwsm.v5i1.14137>
- [92] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.
- [93] K Posner, D Brent, C Lucas, M Gould, B Stanley, G Brown, P Fisher, J Zelazny, A Burke, MJNY Oquendo, et al. 2008. Columbia-suicide severity rating scale (C-SSRS). *New York, NY: Columbia University Medical Center* 10 (2008), 2008.
- [94] Praw-Dev. [n. d.]. Praw-dev/PRAW: PRAW, an acronym for “Python reddit api wrapper”, is a python package that allows for simple access to Reddit’s API. <https://github.com/praw-dev/praw>
- [95] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476* (2023).
- [96] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.
- [97] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* (2020).
- [98] Darrel A Regier, Emily A Kuhl, and David J Kupfer. 2013. The DSM-5: Classification and criteria changes. *World psychiatry* 12, 2 (2013), 92–98.
- [99] Brad Ridout and Andrew Campbell. 2018. The Use of Social Networking Sites in Mental Health Interventions for Young People: Systematic Review. *Journal of Medical Internet Research* 20, 12 (Dec. 2018), e12244. <https://doi.org/10.2196/12244>
- [100] Joshua Robinson and David Wingate. 2023. Leveraging Large Language Models for Multiple Choice Question Answering. In *The Eleventh International Conference on Learning Representations*. <https://openreview.net/forum?id=yKbprarjc5B>
- [101] Thomas Ruder, Gary Hatch, Garyfalia Ampanozi, Michael Thali, and Nadja Fischer. 2011. Suicide Announcement on Facebook. *Crisis* 32 (June 2011), 280–2. <https://doi.org/10.1027/0227-5910/a000086>
- [102] Anna Rumshisky, Marzyeh Ghassemi, Tristan Naumann, Peter Szolovits, VM Castro, TH McCoy, and RH Perlis. 2016. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational psychiatry* 6, 10 (2016), e921–e921.
- [103] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D. Abowd, and Munmun De Choudhury. 2017. Inferring Mood Instability on Social Media by Leveraging Ecological Momentary Assessments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 95 (sep 2017), 27 pages. <https://doi.org/10.1145/3130960>
- [104] Shoffan Saifullah, Yuli Fauziah, and Agus Sasmito Aribowo. 2021. Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data. *arXiv preprint arXiv:2101.06353* (2021).
- [105] Kayalvizhi Sampath and Thenmozhi Durairaj. 2022. Data Set Creation and Empirical Analysis for Detecting Signs of Depression from Social Media Postings. In *Computational Intelligence in Data Science*, Lekshmi Kalinathan, Priyadharsini R., Madheswari Kanmani, and Manisha S. (Eds.). Vol. 654. Springer International Publishing, Cham, 136–151. https://doi.org/10.1007/978-3-031-16364-7_11 Series Title: IFIP Advances in Information and Communication Technology.
- [106] Shailik Sarkar, Abdulaziz Alhamadani, Lulwah Alkulaib, and Chang-Tien Lu. 2022. Predicting depression and anxiety on reddit: a multi-task learning approach. In *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 427–435.
- [107] 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*. 167–175.
- [108] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards Facilitating Empathic Conversations in Online Mental Health Support: A Reinforcement Learning Approach. In *Proceedings of the Web Conference 2021*. ACM, Ljubljana Slovenia, 194–205. <https://doi.org/10.1145/3442381.3450097>

- [109] Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2023. Human-AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence* 5, 1 (Jan. 2023), 46–57. <https://doi.org/10.1038/s42256-022-00593-2>
- [110] Eva Sharma and Munmun De Choudhury. 2018. Mental health support and its relationship to linguistic accommodation in online communities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
- [111] 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*. Association for Computational Linguistics, Vancouver, BC, 58–65. <https://doi.org/10.18653/v1/W17-3107>
- [112] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards Expert-Level Medical Question Answering with Large Language Models. <http://arxiv.org/abs/2305.09617> arXiv:2305.09617 [cs].
- [113] Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access* 7 (2019), 44883–44893.
- [114] Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of Suicide Ideation in Social Media Forums Using Deep Learning. *Algorithms* 13, 1 (Dec. 2019), 7. <https://doi.org/10.3390/a13010007>
- [115] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.
- [116] 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. 2022. A Call to Action on Assessing and Mitigating Bias in Artificial Intelligence Applications for Mental Health. *Perspectives on Psychological Science* (2022), 17456916221134490.
- [117] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. <http://arxiv.org/abs/2302.13971> arXiv:2302.13971 [cs].
- [118] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. <http://arxiv.org/abs/2307.09288> arXiv:2307.09288 [cs].
- [119] 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*. ACM, Seoul Republic of Korea, 3187–3196. <https://doi.org/10.1145/2702123.2702280>
- [120] Elsbeth Turcan and Kathleen McKeown. 2019. Dreddit: A Reddit Dataset for Stress Analysis in Social Media. <http://arxiv.org/abs/1911.00133> arXiv:1911.00133 [cs].
- [121] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–6.
- [122] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652* (2021).
- [123] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. <http://arxiv.org/abs/2109.01652> arXiv:2109.01652 [cs].
- [124] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. <http://arxiv.org/abs/2201.11903> arXiv:2201.11903 [cs].
- [125] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Further Finetuning LLaMA on Medical Papers. <http://arxiv.org/abs/2304.14454> arXiv:2304.14454 [cs].
- [126] Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021. Raise a child in large language model: Towards effective and generalizable fine-tuning. *arXiv preprint arXiv:2109.05687* (2021).
- [127] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous*

- Technologies* 3, 3 (Sept. 2019), 1–33. <https://doi.org/10.1145/3351274>
- [128] Xuhai Xu, Prerna Chikersal, Janine M. Dutcher, Yasaman S. Sefidgar, Woosuk Seo, Michael J. Tumminia, Daniella K. Villalba, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Afsaneh Doryab, Paula S. Nurius, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. 2021. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 1–27. <https://doi.org/10.1145/3448107>
 - [129] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34. <https://doi.org/10.1145/3569485>
 - [130] Xuhai Xu, Jennifer Mankoff, and Anind K. Dey. 2021. Understanding practices and needs of researchers in human state modeling by passive mobile sensing. *CCF Transactions on Pervasive Computing and Interaction* (July 2021). <https://doi.org/10.1007/s42486-021-00072-4>
 - [131] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret E Morris, Eve Riskin, Jennifer Mankoff, and Anind K Dey. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 18.
 - [132] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. On the Evaluations of ChatGPT and Emotion-enhanced Prompting for Mental Health Analysis. <http://arxiv.org/abs/2304.03347>
 - [133] Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Sophia Ananiadou, and Jimin Huang. 2023. MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. <http://arxiv.org/abs/2309.13567> arXiv:2309.13567 [cs].
 - [134] Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198* (2023).
 - [135] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-Most Prompting Enables Complex Reasoning in Large Language Models. <http://arxiv.org/abs/2205.10625> arXiv:2205.10625 [cs].

APPENDIX: DETAILED RESULTS TABLES

Table 10. Balanced Accuracy Performance Summary of Zero-shot, Few-shot and Instruction Finetuning on LLMs. *context*, *mh*, and *both* indicate the prompt design strategies of context enhancement, mental health enhancement, and their combination (see Table. 2). Small numbers represent standard deviation across different designs of *Prompt_{Part1-S}* and *Prompt_{Part2-Q}*. The baselines at the top rows do not have standard deviation as the task-specific output is static, and prompt designs do not apply. Due to token limit, computation cost, and resource constraints, some infeasible experiments are marked as “–”. For each column, the best result is **bolded**, and the second best is underlined.

| | Dataset | Dreaddit | DepSeverity | | SDCNL | CSSRS-Suicide | | Red-Sam | Twt-60Users | SAD |
|--------------------------|---|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| Category | Model | Task #1 | Task #2 | Task #3 | Task #4 | Task #5 | Task #6 | Task #2 | Task #2 | Task #1 |
| Zero-shot Prompting | Alpaca _{ZS} | 0.593 \pm 0.039 | 0.522 \pm 0.022 | 0.431 \pm 0.050 | 0.493 \pm 0.007 | 0.518 \pm 0.037 | 0.232 \pm 0.076 | 0.524 \pm 0.014 | 0.521 \pm 0.022 | 0.503 \pm 0.004 |
| | Alpaca _{ZS-context} | 0.612 \pm 0.065 | 0.567 \pm 0.077 | 0.454 \pm 0.143 | 0.497 \pm 0.006 | 0.532 \pm 0.033 | 0.250 \pm 0.060 | 0.525 \pm 0.019 | 0.559 \pm 0.064 | 0.501 \pm 0.004 |
| | Alpaca _{ZS_mh} | 0.593 \pm 0.031 | 0.577 \pm 0.028 | 0.444 \pm 0.090 | 0.482 \pm 0.015 | 0.523 \pm 0.013 | 0.235 \pm 0.033 | 0.527 \pm 0.006 | 0.569 \pm 0.017 | 0.522 \pm 0.027 |
| | Alpaca _{ZS_both} | 0.540 \pm 0.029 | 0.559 \pm 0.040 | 0.421 \pm 0.095 | 0.532 \pm 0.005 | 0.511 \pm 0.011 | 0.221 \pm 0.030 | 0.495 \pm 0.016 | 0.499 \pm 0.004 | 0.557 \pm 0.041 |
| | Alpaca-LoRA _{ZS} | 0.571 \pm 0.043 | 0.548 \pm 0.027 | 0.437 \pm 0.044 | 0.502 \pm 0.011 | 0.540 \pm 0.012 | 0.187 \pm 0.053 | 0.577 \pm 0.004 | 0.607 \pm 0.046 | 0.477 \pm 0.016 |
| | Alpaca-LoRA _{ZS_{context}} | 0.537 \pm 0.047 | 0.501 \pm 0.001 | 0.343 \pm 0.152 | 0.472 \pm 0.020 | 0.567 \pm 0.038 | 0.214 \pm 0.059 | 0.535 \pm 0.017 | 0.649 \pm 0.021 | 0.443 \pm 0.047 |
| | Alpaca-LoRA _{ZS_mh} | 0.500 \pm 0.000 | 0.500 \pm 0.000 | 0.331 \pm 0.145 | 0.497 \pm 0.025 | 0.557 \pm 0.023 | 0.216 \pm 0.022 | 0.541 \pm 0.016 | 0.569 \pm 0.019 | 0.471 \pm 0.033 |
| | Alpaca-LoRA _{ZS_both} | 0.500 \pm 0.000 | 0.500 \pm 0.000 | 0.386 \pm 0.059 | 0.499 \pm 0.023 | 0.517 \pm 0.031 | 0.224 \pm 0.049 | 0.507 \pm 0.009 | 0.535 \pm 0.025 | 0.420 \pm 0.019 |
| | FLAN-T5 _{ZS} | 0.659 \pm 0.086 | 0.664 \pm 0.011 | 0.396 \pm 0.006 | 0.643 \pm 0.021 | 0.667 \pm 0.023 | 0.418 \pm 0.012 | 0.554 \pm 0.034 | 0.613 \pm 0.040 | 0.692 \pm 0.093 |
| | FLAN-T5 _{ZS_{context}} | 0.663 \pm 0.079 | 0.674 \pm 0.014 | 0.378 \pm 0.013 | 0.653 \pm 0.011 | 0.649 \pm 0.026 | 0.378 \pm 0.029 | 0.563 \pm 0.029 | 0.613 \pm 0.046 | 0.738 \pm 0.056 |
| | FLAN-T5 _{ZS_mh} | 0.616 \pm 0.070 | 0.666 \pm 0.009 | 0.366 \pm 0.012 | 0.648 \pm 0.010 | 0.653 \pm 0.018 | 0.372 \pm 0.033 | 0.547 \pm 0.035 | 0.613 \pm 0.033 | 0.739 \pm 0.039 |
| | FLAN-T5 _{ZS_both} | 0.604 \pm 0.074 | 0.661 \pm 0.004 | 0.389 \pm 0.051 | 0.645 \pm 0.005 | 0.657 \pm 0.019 | 0.382 \pm 0.048 | 0.536 \pm 0.027 | 0.606 \pm 0.040 | 0.767 \pm 0.050 |
| | LLaMA2 _{ZS} | 0.720 \pm 0.012 | 0.693 \pm 0.034 | 0.429 \pm 0.013 | 0.589 \pm 0.010 | 0.691 \pm 0.014 | 0.261 \pm 0.018 | 0.574 \pm 0.008 | 0.735 \pm 0.017 | 0.704 \pm 0.026 |
| | LLaMA2 _{ZS_{context}} | 0.658 \pm 0.025 | 0.707 \pm 0.056 | 0.410 \pm 0.019 | 0.588 \pm 0.026 | 0.722 \pm 0.039 | 0.367 \pm 0.043 | 0.562 \pm 0.011 | 0.736 \pm 0.019 | 0.650 \pm 0.027 |
| | LLaMA2 _{ZS_mh} | 0.617 \pm 0.012 | 0.711 \pm 0.033 | 0.395 \pm 0.017 | 0.642 \pm 0.008 | 0.696 \pm 0.021 | 0.291 \pm 0.038 | 0.572 \pm 0.012 | 0.689 \pm 0.056 | 0.567 \pm 0.021 |
| | LLaMA2 _{ZS_both} | 0.584 \pm 0.017 | 0.704 \pm 0.036 | 0.444 \pm 0.021 | 0.643 \pm 0.014 | 0.689 \pm 0.043 | 0.328 \pm 0.058 | 0.559 \pm 0.012 | 0.692 \pm 0.069 | 0.560 \pm 0.009 |
| | GPT-3.5 _{ZS} | 0.685 \pm 0.024 | 0.642 \pm 0.017 | 0.603 \pm 0.017 | 0.460 \pm 0.163 | 0.570 \pm 0.118 | 0.233 \pm 0.009 | 0.454 \pm 0.007 | 0.536 \pm 0.024 | 0.717 \pm 0.017 |
| | GPT-3.5 _{ZS_{context}} | 0.688 \pm 0.045 | 0.653 \pm 0.020 | 0.543 \pm 0.047 | 0.618 \pm 0.008 | 0.577 \pm 0.090 | 0.265 \pm 0.048 | 0.473 \pm 0.001 | 0.560 \pm 0.002 | 0.723 \pm 0.003 |
| | GPT-3.5 _{ZS_mh} | 0.679 \pm 0.017 | 0.636 \pm 0.021 | 0.642 \pm 0.034 | 0.576 \pm 0.001 | 0.477 \pm 0.014 | 0.310 \pm 0.015 | 0.467 \pm 0.004 | 0.571 \pm 0.000 | 0.664 \pm 0.061 |
| | GPT-3.5 _{ZS_both} | 0.681 \pm 0.010 | 0.627 \pm 0.022 | 0.617 \pm 0.014 | 0.632 \pm 0.020 | 0.617 \pm 0.033 | 0.254 \pm 0.009 | 0.506 \pm 0.004 | 0.570 \pm 0.007 | 0.750 \pm 0.027 |
| | GPT-4 _{ZS} | 0.700 \pm 0.001 | 0.719 \pm 0.013 | 0.588 \pm 0.010 | 0.644 \pm 0.007 | 0.760 \pm 0.009 | 0.418 \pm 0.009 | 0.434 \pm 0.005 | 0.566 \pm 0.017 | 0.854 \pm 0.006 |
| | GPT-4 _{ZS_{context}} | 0.706 \pm 0.009 | 0.719 \pm 0.009 | 0.590 \pm 0.011 | 0.644 \pm 0.011 | 0.753 \pm 0.028 | <u>0.441</u> \pm 0.057 | 0.465 \pm 0.010 | 0.565 \pm 0.006 | <u>0.848</u> \pm 0.001 |
| | GPT-4 _{ZS_mh} | 0.725 \pm 0.009 | 0.684 \pm 0.004 | 0.656 \pm 0.001 | 0.645 \pm 0.012 | 0.737 \pm 0.005 | 0.396 \pm 0.020 | 0.496 \pm 0.005 | 0.527 \pm 0.007 | 0.840 \pm 0.003 |
| | GPT-4 _{ZS_both} | 0.719 \pm 0.021 | 0.689 \pm 0.000 | 0.650 \pm 0.011 | 0.647 \pm 0.014 | 0.697 \pm 0.005 | 0.411 \pm 0.009 | 0.511 \pm 0.000 | 0.546 \pm 0.014 | 0.837 \pm 0.002 |
| Few-shot Prompting | Alpaca _{FS} | 0.632 \pm 0.030 | 0.529 \pm 0.017 | 0.628 \pm 0.005 | — | — | — | — | — | — |
| | FLAN-T5 _{FS} | 0.786 \pm 0.006 | 0.678 \pm 0.009 | 0.432 \pm 0.009 | — | — | — | — | — | — |
| | GPT-3.5 _{FS} | 0.721 \pm 0.010 | 0.665 \pm 0.015 | 0.580 \pm 0.002 | — | — | — | — | — | — |
| | GPT-4 _{FS} | 0.698 \pm 0.009 | 0.724 \pm 0.005 | 0.613 \pm 0.001 | — | — | — | — | — | — |
| Instructional Finetuning | Mental-Alpaca | <u>0.816</u> \pm 0.006 | <u>0.775</u> \pm 0.006 | <u>0.746</u> \pm 0.005 | 0.724 \pm 0.004 | 0.730 \pm 0.048 | 0.403 \pm 0.029 | 0.604 \pm 0.012 | 0.718 \pm 0.011 | 0.819 \pm 0.006 |
| | Mental-FLAN-T5 | 0.802 \pm 0.002 | 0.759 \pm 0.003 | 0.756 \pm 0.001 | 0.677 \pm 0.005 | 0.868 \pm 0.006 | 0.481 \pm 0.006 | <u>0.582</u> \pm 0.002 | 0.736 \pm 0.003 | 0.779 \pm 0.002 |
| Baseline | Majority | 0.500 \pm --- | 0.500 \pm --- | 0.250 \pm --- | 0.500 \pm --- | 0.500 \pm --- | 0.200 \pm --- | — | — | — |
| | BERT | 0.783 \pm --- | 0.763 \pm --- | 0.690 \pm --- | 0.678 \pm --- | 0.500 \pm --- | 0.332 \pm --- | — | — | — |
| | Mental-RoBERTa | 0.831 \pm --- | 0.790 \pm --- | 0.736 \pm --- | <u>0.723</u> \pm --- | <u>0.853</u> \pm --- | 0.373 \pm --- | — | — | — |