

ProMind-LLM: Proactive Mental Health Care via Causal Reasoning with Sensor Data

Xinzhe Zheng^{3*} Sijie Ji^{1,2†*} Jiawei Sun^{4*} Renqi Chen⁵

Wei Gao¹ Mani Srivastava²

¹California Institute of Technology ²UCLA ³National University of Singapore

⁴Hangzhou Dianzi University ⁵Fudan University

Abstract

Mental health risk is a critical global public health challenge, necessitating innovative and reliable assessment methods. With the development of large language models (LLMs), they stand out to be a promising tool for explainable mental health care applications. Nevertheless, existing approaches predominantly rely on subjective textual mental records, which can be distorted by inherent mental uncertainties, leading to inconsistent and unreliable predictions. To address these limitations, this paper introduces ProMind-LLM. We investigate an innovative approach integrating objective behavior data as complementary information alongside subjective mental records for robust mental health risk assessment. Specifically, ProMind-LLM incorporates a comprehensive pipeline that includes domain-specific pretraining to tailor the LLM for mental health contexts, a self-refine mechanism to optimize the processing of numerical behavioral data, and causal chain-of-thought reasoning to enhance the reliability and interpretability of its predictions. Evaluations of two real-world datasets, PMData and Globem, demonstrate the effectiveness of our proposed methods, achieving substantial improvements over general LLMs. We anticipate that ProMind-LLM will pave the way for more dependable, interpretable, and scalable mental health case solutions.

1 Introduction

In fact, mental health disorders have the highest prevalence rates compared to many other major health conditions. According to WHO, approximately 25% of people globally experience a mental health issue (Organization et al., 2019). Moreover, the COVID-19 pandemic and other global crises have exacerbated this issue, leading to a 25.6% increase in anxiety and a 27.6% rise in depression

*Equal contribution.

†Corresponding author. Email: sijieji@caltech.edu

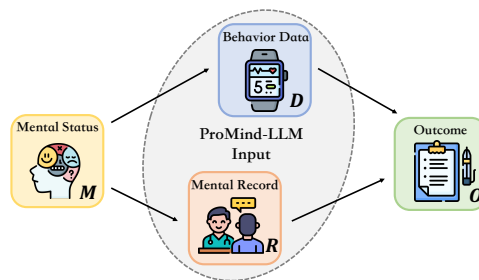


Figure 1: Causal relations among mental status, behavior data, and mental records. In ProMind-LLM, we utilize objective behavior data and subjective mental records to achieve robust mental risk prediction.

since 2020 (Mahmud et al., 2023). The widespread nature of mental disorders underscores the need for developing automated detection tools to tackle this public health challenge (Vos et al., 2020).

Currently, large language models (LLMs) have demonstrated remarkable success across a variety of domains (Thirunavukarasu et al., 2023; Lv et al., 2024; Zhang et al., 2024b). Leveraging their advanced generalization and reasoning abilities, researchers have demonstrated their potential applications in mental health care (Qiu et al., 2023; Gabriel et al., 2024; Ji et al., 2024a, 2025). LLMs can analyze user-generated data that reflects mental states, extracting contextual information to uncover subtle indicators of mental health disorders (Hua et al., 2024). For instance, prior works approved the effectiveness of using social media posts (Kang et al., 2024; Kumar et al., 2024) or conversational records (Singh et al., 2024) to assess the presence of mental health issues. Despite these promising results, these approaches are limited by their exclusive reliance on subjective textual data. Such data can be influenced by transient factors, including the user’s current mood or willingness to disclose accurate information, which undermines the reliability of predictions (World Health Organization, 2022). To address this limitation, incorporating objective

factors, such as behavioral sensor data (*e.g.* heart rate, sleep quality), offers a promising avenue to complement subjective data and enhance the accuracy of mental health predictions (Patsali et al., 2020; Clement-Carbonell et al., 2021). Recent studies have integrated other modalities such as speech (Zhang et al., 2024c) and video (Singh et al., 2024) for mental disorder detection and achieved improvements compared to using subjective data only. Nevertheless, the combined use of subjective records and objective behavior data for LLM-based mental health assessment remains largely underexplored.

Developing an LLM approach that effectively combines objective behavior data with subjective mental records poses several challenges. First, given the potential uncertainties in subjective mental records, LLMs must enhance their capacity to navigate the variability embedded within such data effectively. Second, most behavior data, such as heart rate, sleep quality, and exercise readings, are represented as lengthy numerical sequences (Kim et al., 2024; Wu et al., 2024). Although these sequences can be formatted as text for LLM inputs, their sheer length and numeric nature can impede the model’s ability for accurate interpretation (Requeima et al., 2024; Jin et al., 2024). Third, enabling the LLM to fully utilize multimodal data, especially the causal relations (Ofstedal et al., 2019) between objective behavior data and subjective mental records, to further improve the outcome robustness requires a specific design.

To address these issues, we introduce ProMind-LLM, which builds upon the causal relations among mental status, mental records, and behavior data (Figure. 1), to deliver robust mental health risk assessment. For the first challenge (Sec. 3.1), we construct a comprehensive mental health corpus and adopt continuous pertaining (PT) (Gururangan et al., 2020) to enhance the base model’s understanding of mental health concepts, a widely used strategy in the development of prior mental health LLMs (Zhai et al., 2024b; Ji et al., 2023). Following PT, we implement counterfactual learning-based supervised finetuning (SFT), which generates misleading information in users’ mental records to challenge and improve the LLM’s resilience against uncertainties in subjective data. ProMind-LLM addresses the second issue (Sec. 3.2) by proposing a self-refine mechanism to format the behavior data, allowing the LLM to iteratively update and optimize the format based on

its own feedback. This approach not only shortens behavior inputs while preserving key features relevant to mental status but also enhances the model’s ability to interpret behavior insights. For the third challenge (Sec. 3.3), ProMind-LLM initially employs chain-of-thought (CoT) (Wei et al., 2022) reasoning, integrating the causal relations between mental status and behavior. This allows the model to analyze the mental records and behavior data individually and account for their mutual influence, rendering a more robust mental risk prediction. Additionally, ProMind-LLM then utilizes counterfactual reasoning to further refine its judgment. By exploring “what-if” scenarios-*e.g.*, “*What if the user’s eating habit were normal? Would this absence change the outcome?*”-the model evaluates alternative outcomes based on hypothetical changes to the input data (Chen et al., 2022). This process enables the LLM to identify true indicators of mental risks, and ultimately confirm the outcome’s reliability. Together, these reasoning steps form the causal CoT scheme in ProMind-LLM.

To sum up, we make the following contributions in this paper:

- To the best of our knowledge, ProMind-LLM is the first study to apply LLMs for mental health risk assessment by using both objective behavior data and subjective mental records, addressing the limitations of relying solely on subjective textual inputs.
- ProMind-LLM incorporates a comprehensive pipeline comprising domain-specific training, behavior data preprocessing via a self-refine scheme, and causal CoT reasoning for robust and interpretable predictions.
- Extensive experiments verify the effectiveness of ProMind-LLM over general LLMs, achieving improved outcome accuracy. This novel approach paves the way for realizing ubiquitous proactive mental health care.

2 Related Work

2.1 LLMs for Mental Health

Leveraging NLP technologies for the early detection and intervention of mental health issues stands as a valuable endeavor (Wang et al., 2024b). With the advent of LLMs, researchers have recognized the potential of these tools to facilitate mental health care (Lamichhane, 2023). Recent studies

have evaluated the performance of state-of-the-art (SOTA) LLMs on various mental health tasks, including the detection of depression and anxiety (Xu et al., 2024b; Yang et al., 2023) and stress level prediction (Kim et al., 2024), exhibiting promising yet limited performance due to a lack of domain-specific knowledge. To address this limitation, efforts have been made to adapt general-purpose LLMs into domain-specific models tailored for mental health applications. Mental-BERT (Ji et al., 2021, 2023) applied domain-specific pretraining to enhance its effectiveness in mental health tasks. Similarly, Mental-Alpaca (Xu et al., 2024b) and MentalLlama (Yang et al., 2024c) employed instruction fine-tuning to improve corresponding reasoning capabilities. In addition, Chinese Mental-BERT (Zhai et al., 2024b) is the first LLM that focuses on mental disorder detection for Chinese social media. Beyond developing specialized models, researchers also emphasize the need for comprehensive benchmarks to support the advancement of mental health LLMs (Singh et al., 2024; Gabriel et al., 2024; Wang et al., 2024b). Despite these advancements, most existing efforts primarily rely on subjective textual data. ProMind-LLM takes a step forward by integrating objective behavioral data with subjective mental records, offering a more precise approach to proactive mental health care.

2.2 LLMs for Numerical Data Understanding

Despite being primarily used in the field of NLP, LLMs have shown remarkable potential in handling and reasoning about numerical data (Liu et al., 2023). One notable application is the classification of human activities using data from IMUs (Ji et al., 2024b; Civitarese et al., 2024). Similarly, LLMs have been used to analyze user physiological data, such as sleep quality, calorie intake, and heart rate, to generate personalized and professional health recommendations (Yang et al., 2024a; Kim et al., 2024; Cosentino et al., 2024). These applications underscore the models’ ability to interpret and apply numerical data in real-world contexts. In addition to data interpretation, LLMs exhibit capabilities in time series forecasting (Yang et al., 2024b). These findings highlight the promise of LLMs for numerical data understanding.

3 Methodology

Our proposed method, as illustrated in Figure. 2, comprises three key components: domain-specific

training, self-refine-based behavior data preprocessing, and causal CoT reasoning. Together, these components enable robust mental risk prediction by integrating objective behavior data with subjective mental records. Below, we provide a detailed explanation of each design.

3.1 Domain-specific Training

3.1.1 Continuous PT

Continuous PT, also referred to as domain-adaptive PT, has been shown to significantly enhance performance, particularly for specialized downstream tasks (Cui et al., 2021; Gururangan et al., 2020). To equip the base LLM—originally trained on a broad general corpus—with essential mental health domain knowledge, we adopt this approach by performing continuous PT on an extensive, mental health-specific corpus.

Unlike previous mental health-specific LLMs that predominantly rely on social media content (Zhai et al., 2024b,a), we take a more comprehensive approach by curating approximately 100K professional mental health articles from the Web of Science, as suggested in (Zhang et al., 2024b). These articles are selected using key terms outlined in Apx. C.1, covering topics like depression, anxiety, substance abuse, *etc.* The raw materials are then converted from PDF into structured TXT formats using MinerU (Wang et al., 2024a; He et al., 2024). Further data cleansing (Dubey et al., 2024) is implemented to correct the spelling errors and standardize format, such as removing redundant references, *etc.* The final PT corpus contains about 80 million tokens, preparing the base LLM with mental health expertise. For details of the PT settings, please refer to Apx. C.1.

3.1.2 Counterfactual-based SFT

While continuous PT equips the base model with domain-specific expertise, SFT is an essential subsequent step, which refines the model’s ability to effectively apply its knowledge in reasoning tasks with prompt instructions (Zhang et al., 2023). We take advantage of the public mental health-related datasets, including IMHI (Yang et al., 2024c), CPsyCoun (Zhang et al., 2024a), ANGST (Hengle et al., 2024), and Depression Reddit (Pirina and Çöltekin, 2018), resulting in 112602 pieces of data. These datasets provide high-quality question-answering pairs designed to analyze symptoms and infer outcomes based on the provided mental health records, equipping ProMind-LLM with greater pre-

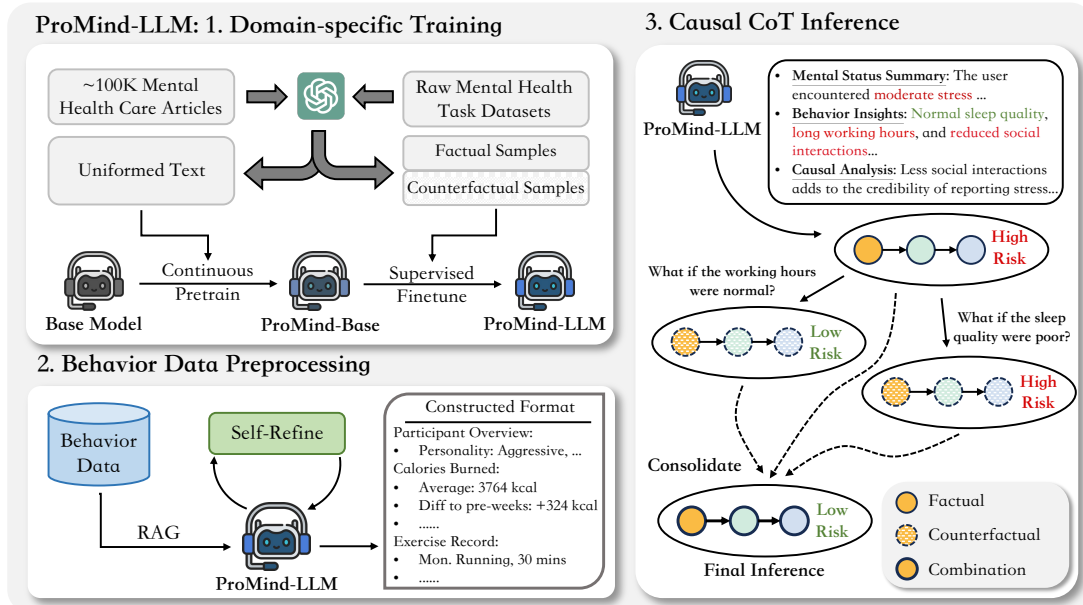


Figure 2: ProMind-LLM framework. The design consists of three components: (1) Construct professional mental health LLM using domain-specific training; (2) Enhance behavior understanding through self-refine mechanism; (3) Improve outcome robustness with causal CoT.

cision and nuance to handle various mental health situations. Please refer to Apx. C.2 for statistical information on these datasets.

Counterfactual learning is proposed to further address the first challenge, which is to enhance the LLM’s robustness against the uncertainties of subjective mental records. This approach generates alternative scenarios where users’ records might present misleading information about their mental states. For instance, given an input SFT pair $\langle R, O \rangle$, where R denotes the mental record, and O presents the outcome analysis, we generate a counterfactual sample labeled l within categories like “personality traits”, “stigma”, or “lack of awareness”. These elements are pivotal in distorting an individual to introduce biases or inaccuracies in self-reported data (World Health Organization, 2022). Finally, we generate two counterfactual parts for each pair and combine them into one single dataset to support counterfactual augmented SFT. This augmentation enhances the model to recognize potential distortions, thereby improving its reliability in analyzing subjective mental health records. The SFT settings can be found in Apx. C.2. In addition, the prompt and the counterfactual sample are given in Apx. H.1.

3.2 Behavior Data Preprocessing

Addressing the second challenge necessitates transforming numerical behavior data into a format

that the LLM can efficiently process and comprehend. Inputting lengthy sequences of raw behavior data directly into the LLM poses problems, including the risk of sequence truncation and memory constraints (Gruber et al., 2024), as well as the LLM’s inherent challenges in processing the digital data (Chen et al., 2024).

To fully leverage ProMind-LLM’s capacity for interpreting behavioral data, we propose a self-refine mechanism (Madaan et al., 2024). As depicted in Figure. 2, ProMind-LLM first retrieves raw behavior data using RAG (Lewis et al., 2020), evaluating the data’s presentation format for redundancy and ease of understanding. Based on this evaluation, the LLM provides an improved data version through self-refine. This mechanism allows the LLM to update and optimize the data format autonomously based on its own feedback, thereby improving its ability to interpret behavior insights accurately.

Table. 1 validates the effectiveness of our proposed method. For instance, we use perplexity (Alon and Kamfonas, 2023) to measure the LLM’s familiarity with the behavior data format. Compared with existing methods (Ji et al., 2024b; Kim et al., 2024; Xu et al., 2024a), our approach not only reduces the input token length but also markedly enhances the LLM’s proficiency in interpreting user behavior data.

We provide the corresponding pseudo-code in

Table 1: Assessment of the input format for behavioral data regarding token numbers and perplexity.

Method	Tokens ↓	Perplexity ↓
HARGPT (Ji et al., 2024b)	2^{11}	15.63
Health-LLM (Kim et al., 2024)	2^{12}	17.22
Penetrative AI (Xu et al., 2024a)	2^{11}	6.18
Ours	2^9	4.31

Apx. D for reference.

3.3 Causal CoT

Given the processed behavioral data with users’ mental records, ProMind-LLM is designed to assess the potential mental health risks. Achieving accurate predictions demands a profound understanding of the data, coupled with a rigorous reasoning process. Motivated by the intrinsic causal relations between behavioral data and mental records, we implement causal reasoning into the widely adopted CoT framework, resulting in causal CoT for ProMind-LLM. This integration aims to further mitigate the impact of uncertainties in subjective data, enhancing the model’s predictive accuracy.

As shown in Figure. 2, the causal CoT comprises three components, CoT on factual samples, CoT on counterfactual samples, and a combination analysis of the two. First, we prompt ProMind-LLM under the CoT framework to conduct a preliminary analysis of mental records (\mathbf{R}) and behavioral data (\mathbf{D}) individually. It then leverages the causal relation between the two modalities to evaluate their mutual effects. For example, if both modality data present high-risk indicators, such as “moderate stress” and “reduced social interactions”, the model may identify a possible correlation between these factors. The greater the number of associated indicators, the higher the likelihood of potential mental risks for the user. We use the following formula to represent the first step:

$$\mathbf{A} = \{(d_i, r_j) \mid d_i \in \mathbf{D}, r_j \in \mathbf{R}, P(r_j|d_i) > \tau\}, \quad (1)$$

where \mathbf{A} denotes the factual analysis results, d_i and r_j represent the risk indicators in behavior data and mental records respectively, P indicates the causal analysis by ProMind-LLM, and (d_i, r_j) is the pair found with a possible causal relation.

To validate and refine the results of the initial step, ProMind-LLM utilizes counterfactual reasoning. This involves exploring “what-if” questions

to strengthen the causal understanding of certain pairs, reanalyze weak correlations, and address pairs that might be overlooked due to uncertainties in mental records. For example, by asking: “What if the user’s working hours were reduced, would the user’s stress be better?”. If the model believes that stress can be alleviated to a certain extent, as the user may have more time to restart the social interactions, then the social interactions and stress pair built in the first step would be less critical. By disentangling these causal influences, the model achieves a more nuanced understanding of the user’s mental health dynamics. We summarize the counterfactual reasoning as follows:

$$\mathbf{A}_c = \{(d_i, r_j) \mid d_i \in \mathbf{D}, r_j \in \mathbf{R}, P(r_j|d_i, c) > \tau\}, \quad (2)$$

where \mathbf{A}_c indicates the results in counterfactual reasoning, and c is the counterfactual reasoning.

The final step of causal CoT is the integration of the insights from both factual and counterfactual analysis to generate its final predictions: $(\mathbf{A}, \mathbf{A}_c) \rightarrow \mathbf{G}$, where \mathbf{G} is the user’s potential mental health risk. Please refer to Apx. H.2 for detailed examples.

4 Experimental Setup

4.1 Task Definition

To evaluate the effectiveness of our methods and the performance of ProMind-LLM in predicting mental health risks, we aim to address a binary classification task. Each individual contains long-term recordings of both behavioral data (\mathbf{D}) and self-reported mental records (\mathbf{R}). These records are aggregated weekly and represented as $\langle \mathbf{D}_i, \mathbf{R}_i \rangle$.

Based on these inputs, LLMs need to predict the potential mental health risk label \mathbf{G}_i , categorized as follows:

1. $\mathbf{G}_i = 0$: The individual shows no significant signs of mental health issues, or may have minor issues that do not require immediate psychological intervention.
2. $\mathbf{G}_i = 1$: The individual exhibits strong indicators of mental health issues and requires further professional treatment or closer monitoring.

4.2 Dataset Description

We conduct experiments on two open-sourced datasets, PMData (Thambawita et al., 2020) and

Globem (Xu et al., 2022), which provide behavioral data and self-reported mental health records but lack predefined mental health risk labels (G_i), necessitating manual annotation. Based on the professional criteria (Morgan et al., 2018; Guha, 2014), we assess existing or potential mental health risks for each recording. To mitigate the potential false positives and negatives, we cooperate with licensed psychological experts to review the initial assessments and refine the labels by considering behavior patterns, and inconsistencies in self-reported records, thereby enhancing the reliability of the labeling process.

The details of each dataset are given below¹:

PMData: The dataset comprises 16 participants monitored over 5 months using Fitbit for objective biometrics and activity data, Google Forms for demographics, food, drinking, and weight data, and the PMSys for self-reported measures such as fatigue, mood, stress, etc. The data collected from Fitbit and Google Forms constitute the participants' behavior data (D_i), while the PMSys measures represent their self-reported mental records (R_i). All participants are used for evaluation. A small fraction of 9.8% cases are identified with potential mental health issues requiring additional support.

Globem: The Globem dataset encompasses four years of passive sensing data from 497 participants. Behavioral data, including sleep patterns, location, physical activity, and phone usage, are collected using wearable sensors (Fitbit Flex2 and Inspire 2) and are denoted as D_i . Survey data, such as PHQ-4 (Kroenke et al., 2009) (mental health, anxiety, and depression), PSS-4 (Cohen et al., 1983) (stress level), and PANAS (Watson et al., 1988) (positive and negative affect), provide self-reported mental health records (R_i). For efficient testing and cost reduction, 25% of the participants are randomly chosen as the test set. Within this group, 23.2% are identified to have potential risks.

5 Results and Analysis

5.1 Evaluation Design

Evaluation is designed to answer the following key research questions: 1): How effective is the proposed two-stage domain-specific training in enhancing the model's ability to assess mental health risks? (Sec. 5.3) 2): Do the self-refine mechanism

and causal CoT improve assessment accuracy, and to what extent? (Sec. 5.4) 3): By employing counterfactual data augmentation in SFT, does ProMind-LLM demonstrate improved resilience against uncertainties in mental records? (Sec. 5.4) 4): To what extent does behavioral data serve as a complementary modality to subjective mental records in enhancing mental health risk assessment? (Sec. 5.5) 5): Does the model's analysis align from end to end, and how consistent are the evidence and outcomes of the analysis? (Sec. 5.6)

5.2 Baselines and Deployment Details

We select several baseline models for performance comparison, including three SOTA commercial LLMs (GPT-4o, GPT-3.5, and Claude-3.5) and two leading open-sourced LLMs (LLaMA3-Chat-70B and Qwen2-Chat-72B). Additionally, two domain-specific models, Mental-Alpaca (Xu et al., 2024b) and MentalLlama (Yang et al., 2024c), are also evaluated but excluded due to inadequate outputs (see Apx. H.2.2).

To develop ProMind-LLM, we select two base LLMs: LLaMA3-base-8B and InternLM2-base-7B, prioritizing the future deployment of this application on the edge devices, a strategic decision aimed at minimizing potential privacy risks associated with data transfers. For comparisons, we also include the vanilla versions of these models, serving as our open-sourced baselines. Among these, ProMind-LLM developed from InternLM2-base-7B demonstrates a better performance, representing the ProMind-LLM in the rest of the paper if not otherwise indicated.

5.3 Effectiveness of Training Strategy

We evaluate the performance of ProMind-LLM against baseline models on PMData and Globem datasets. Additionally, we compare the outcomes of applying SFT directly to two base LLMs versus a two-stage process involving PT plus SFT. The results are summarized in Table. 2. All methods implement both self-refine mechanism and causal CoT reasoning. We exclude the base models due to their lack of instruction following capabilities.

GPT-4o outperforms other baseline models across both datasets. Notably, GPT-4o achieves a recall rate of 0.821 on PMData and 0.955 on Globem. Among open-source models, Qwen2-chat-72B achieves the highest F1 scores, 0.604 and 0.684, on PMdata and Globem, respectively.

¹All data usage strictly adheres to the Data Use Agreements of PMData (CC BY 4.0) and Globem (PhysioNet Credentialed Health Data License 1.5.0).

Table 2: Results for mental health issue binary classification. All methods here use both self-refine and causal CoT for mental risk prediction. (The best and second results are highlighted in **bold** and underlined, respectively.)

Category	Model	Type	PMDData				Globem			
			Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
Comercial	GPT-4o ¹	\	0.956	0.781	<u>0.821</u>	0.800	0.867	0.663	<u>0.955</u>	0.783
	GPT-3.5 ¹	\	0.858	0.385	0.714	0.500	0.747	0.505	0.918	0.651
	Claude-3.5 ¹	\	0.923	0.786	0.314	0.449	0.789	0.552	0.951	0.699
Open-sourced	LLaMA3-chat-70B	\	0.795	0.297	0.771	0.429	0.807	0.579	0.833	0.683
	QWen2-chat-72B	\	0.918	0.674	0.547	0.604	0.819	0.608	0.783	0.684
	LLaMA3-chat-8B	\	0.613	0.174	0.641	0.275	0.531	0.325	0.818	0.465
	InternLM2-chat-7B	\	0.634	0.191	0.828	0.310	0.367	0.289	1.000	0.449
Ours	LLaMA3-base-8B	SFT	0.625	0.213	0.846	0.340	0.636	0.401	0.939	0.564
		PT+SFT	0.883	0.492	0.769	0.601	0.754	0.504	0.901	0.649
	InternLM2-base-7B	SFT	0.867	0.443	0.633	0.521	0.726	0.481	0.836	0.611
		PT+SFT	<u>0.938</u>	<u>0.765</u>	0.667	<u>0.712</u>	<u>0.859</u>	<u>0.646</u>	0.939	<u>0.765</u>

¹ GPT-4o API is “gpt-4o-2024-05-13”. GPT-3.5 API is “gpt-3.5-turbo”. Claude-3.5 API is “claude-3-5-sonnet-20240620”.

Table 3: Ablation study on self-refine and causal CoT. (The best and second results of each model are highlighted in **bold** and underlined, respectively.)

Model	Method		PMDData				Globem			
	Self-Refine	Causal CoT	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
GPT-4o	✓	✓	0.938	0.734	0.726	0.729	0.803	0.575	0.818	0.675
	✓	✓	<u>0.952</u>	<u>0.737</u>	<u>0.800</u>	<u>0.767</u>	0.814	0.586	<u>0.951</u>	<u>0.725</u>
	✓	✓	0.933	0.687	0.769	0.726	<u>0.829</u>	<u>0.615</u>	0.848	0.713
	✓	✓	0.956	0.781	0.821	0.800	0.867	0.663	0.955	0.783
Qwen2-chat-72B	✓	✓	0.837	0.382	0.693	0.495	0.761	0.516	0.742	0.609
	✓	✓	<u>0.903</u>	<u>0.514</u>	0.514	<u>0.514</u>	0.769	0.523	0.879	<u>0.655</u>
	✓	✓	0.859	0.423	<u>0.635</u>	0.507	0.781	<u>0.541</u>	<u>0.798</u>	0.645
	✓	✓	0.918	0.674	0.547	0.604	0.819	0.608	0.783	0.684
ProMind-LLM	✓	✓	0.845	0.369	0.504	0.426	0.734	0.475	0.641	0.546
	✓	✓	0.940	0.792	<u>0.543</u>	<u>0.644</u>	<u>0.844</u>	<u>0.640</u>	<u>0.902</u>	<u>0.748</u>
	✓	✓	0.849	0.376	0.479	0.422	0.766	0.525	0.672	0.589
	✓	✓	<u>0.938</u>	<u>0.765</u>	0.667	0.712	0.859	0.646	0.939	0.765

The results further highlight a significant improvement when using the combined PT and SFT approach. For instance, on PMData dataset, InternLM2-7B shows substantial gains, with precision and recall increasing by 42.1% and 5.1%, respectively, and an overall F1 score improvement of 26.8% compared to the model trained without PT. This trend is consistently observed on the Globem dataset, where the combined PT and SFT strategy outperforms SFT alone, further validating the benefits of our proposed training framework.

Remarkably, ProMind-LLM achieves a performance ranking second only to GPT-4o in both datasets, even surpassing leading open-sourced LLMs that are ten times its size. The results demonstrate that (i) ProMind-LLM not only integrates mental health knowledge into its base model but also matches the reasoning capabilities of SOTA large-scale commercial LLM, GPT-4o. (ii) ProMind-LLM analyzes users’ mental

health conditions with performance that surpasses its counterparts possessing more than 10 times its parameters (LLaMA3-70B, etc.).

5.4 Ablation Study

Self-refine & Causal CoT:

To verify the effectiveness of self-refine mechanism and causal CoT, we conduct an ablation study using GPT-4o and Qwen2-chat-72B as baseline models, with results presented in Table. 3. In this experiment, disabling the self-refine mechanism involves directly feeding the original behavior sequence data into the model, while disabling causal CoT indicates using CoT only for the final analysis. Across all models, the combination of both strategies consistently yields the best performance. For instance, ProMind-LLM achieves F1 score improvements of 40.2% on PMData and 28.6% on Globem compared to not using either method. In contrast, GPT-4o demonstrates an average F1 im-

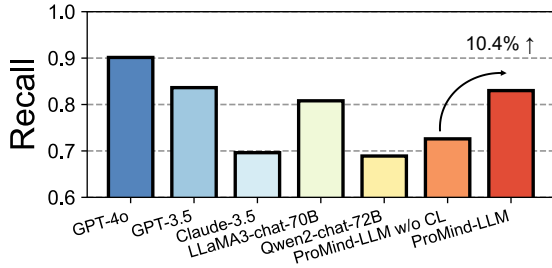


Figure 3: Ablation study on counterfactual augmentations in SFT.

provement of only 11.3%, likely due to its already strong generalization and reasoning capabilities.

Both self-refine and causal CoT reasoning independently enhance model performance, with self-refine yielding greater overall gains. For GPT-4o, self-refine and causal CoT improve F1 scores by 5% and 2%, respectively, while Qwen2-chat-72B achieves gains of 3.4% and 3.1%. ProMind-LLM benefits significantly from self-refine mechanism, with a 24.9% F1 increase, likely due to challenges small-scale LLMs face in processing lengthy numerical sequences, which can hinder contextual understanding and reasoning performance.

Counterfactual-based SFT:

During the SFT stage, we use counterfactual data augmentation to introduce misleading information into users’ subjective mental health records, enhancing the model’s ability to handle subjective uncertainties. As shown in Figure. 3, we calculate the average recall rate using both datasets. This technique improves ProMind-LLM’s recall from 72.6% to 83%, outperforming open-sourced LLMs.

5.5 Behavior Data for Complementary

To assess the influence of behavioral data as complementary modalities on outcomes, we perform an ablation study detailed in Figure. 4. The findings underscore the importance of integrating objective behavioral data with subjective mental records for accurate analysis.

5.6 Consistency Measurement

Following the methodology in (Yang et al., 2024c), we assess whether the analysis evidence supports the outcomes. Using pre-trained BERT (Devlin, 2018) for embedding extraction, we calculate the Silhouette score (Shahapure and Nicholas, 2020) to evaluate clustering quality, with higher scores indicating better performance. As shown in Figure. 5,

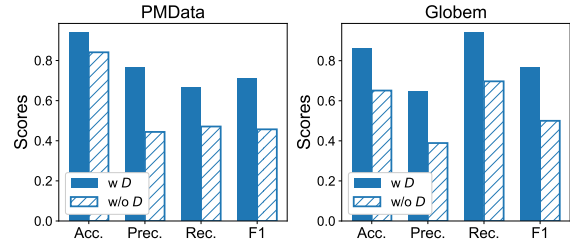


Figure 4: Ablation study on behavior data as complementary modality.

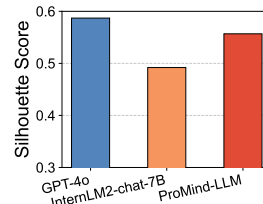


Figure 5: Silhouette scores

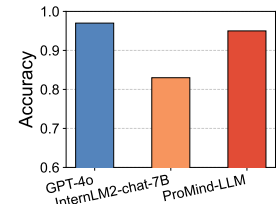


Figure 6: Accuracy of consistency.

ProMind-LLM, with domain-specific training, outperforms InternLM2-chat-7B, achieving a 0.065 improvement in the Silhouette score. Additionally, we introduce a classification network following the embedding extraction and employ K-fold cross-validation to calculate the overall accuracy of classifying outcomes based on the evidence in the analysis, as presented in Figure. 6. ProMind-LLM achieves an accuracy of 0.95, slightly lower than GPT-4o (0.97), which proves the consistency of the evidence and the outcome. In contrast, InternLM2-chat-7B achieves an accuracy of 0.83, further validating the robustness of ProMind-LLM.

6 Conclusion

This paper presents ProMind-LLM, the first LLM-based approach to integrate both objective behavioral data and subjective mental records for mental risk analysis. Our pipeline includes domain-specific training to equip the base model with mental health concepts, the self-refine mechanism for behavior data preprocessing to enhance numerical data comprehension, and causal CoT reasoning to deliver precise predictions by leveraging causal relations between mental records and behavioral data. The experimental results validate the effectiveness of our methods, highlighting ProMind-LLM’s ability to advance the research in this domain. We believe ProMind-LLM marks a significant step toward achieving proactive mental health care.

Limitations

While ProMind-LLM demonstrates strong performance on public datasets such as PMData and Globem, which encompass a four-year longitudinal study, its effectiveness in real-world, uncontrolled environments with diverse populations, in terms of age, gender, and race, remains insufficiently validated. Additionally, due to resource constraints, our proposed training pipeline has been evaluated only on small-scale LLMs. Moreover, although we have conducted a preliminary human evaluation assessing the readability of causal CoT reasoning steps and analyzing the causes of incorrect predictions (Sec. F), a more comprehensive human evaluation is required. Specifically, further investigation into misclassified cases—such as underestimation of missed risks—is essential to ensure the system’s reliability and safety, given its role as a mental health recommendation system. This expanded evaluation will also provide deeper insights into the specific scenarios where ProMind-LLM is prone to errors, guiding future refinements to enhance its robustness and overall effectiveness.

Ethics Statement

This research complies with ethical standards to ensure integrity and participant welfare. The publicly available datasets, PMData and Globem, were utilized under their Data Use Agreements, with all participant data anonymized to protect privacy. Original data collection adhered to informed consent protocols, ensuring ethical compliance. Licensed psychological experts validated mental health risk assessments, minimizing inaccuracies and maintaining sensitivity towards participants. Methodological transparency and reproducibility were upheld through detailed documentation of pretraining, finetuning, and counterfactual augmentation processes. To avoid biases, counterfactual learning techniques were employed, enhancing the model’s robustness against uncertainties in subjective mental records. Efforts were made to mitigate biases in mental health assessments by incorporating counterfactual learning and rigorous evaluation techniques to ensure fair, non-discriminatory analysis while avoiding stigmatization. This work reflects a commitment to advancing mental health care responsibly and with respect for human dignity.

Acknowledgment

This research was funded in part by National Institutes of Health (NIH) under award #P41EB028242. We also acknowledge the support from the Heritage Medical Research Institute and the Schmidt Science Fellowship.

References

- Gabriel Alon and Michael Kamfonas. 2023. Detecting language model attacks with perplexity. *arXiv preprint arXiv:2308.14132*.
- Yulong Chen, Yang Liu, Jianhao Yan, Xuefeng Bai, Ming Zhong, Yinghao Yang, Ziyi Yang, Chenguang Zhu, and Yue Zhang. 2024. See what llms cannot answer: A self-challenge framework for uncovering llm weaknesses. *arXiv preprint arXiv:2408.08978*.
- Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2022. Disco: Distilling counterfactuals with large language models. *arXiv preprint arXiv:2212.10534*.
- Gabriele Civitarese, Michele Fiori, Priyankar Choudhary, and Claudio Bettini. 2024. Large language models are zero-shot recognizers for activities of daily living. *arXiv preprint arXiv:2407.01238*.
- Violeta Clement-Carbonell, Irene Portilla-Tamarit, María Rubio-Aparicio, and Juan J Madrid-Valero. 2021. Sleep quality, mental and physical health: a differential relationship. *International journal of environmental research and public health*, 18(2):460.
- Sheldon Cohen, Tom Kamarck, and Robin Mermelstein. 1983. A global measure of perceived stress. *Journal of health and social behavior*, pages 385–396.
- Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. 2024. Towards a personal health large language model. *arXiv preprint arXiv:2406.06474*.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.
- Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*.
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can ai relate: Testing large language model response for mental health support. *arXiv preprint arXiv:2405.12021*.
- Roger Griffiths. 1992. Speech rate and listening comprehension: Further evidence of the relationship. *TESOL quarterly*, 26(2):385–390.
- Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36.
- Martin Guha. 2014. Diagnostic and statistical manual of mental disorders: Dsm-5. *Reference Reviews*, 28(3):36–37.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.
- Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. Opendatalab: Empowering general artificial intelligence with open datasets. *arXiv preprint arXiv:2407.13773*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Amey Hengle, Atharva Kulkarni, Shantanu Patankar, Madhumitha Chandrasekaran, Sneha D’Silva, Jemima Jacob, and Rashmi Gupta. 2024. Still not quite there! evaluating large language models for comorbid mental health diagnosis. *arXiv preprint arXiv:2410.03908*.
- 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. *arXiv preprint arXiv:2106.09685*.
- Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, Andrew Beam, et al. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.
- Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.
- Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, Erik Cambria, and Jörg Tiedemann. 2023. Domain-specific continued pretraining of language models for capturing long context in mental health. *arXiv preprint arXiv:2304.10447*.
- Sijie Ji, Xinzhe Zheng, Wei Gao, and Mani Srivastava. 2025. Transforming mental health care with autonomous llm agents at the edge. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*, pages 692–693.
- Sijie Ji, Xinzhe Zheng, Jiawei Sun, Renqi Chen, Wei Gao, and Mani Srivastava. 2024a. Mindguard: Towards accessible and sitgma-free mental health first aid via edge llm. *arXiv preprint arXiv:2409.10064*.

- Sijie Ji, Xinzhe Zheng, and Chenshu Wu. 2024b. Hargpt: Are llms zero-shot human activity recognizers? *arXiv preprint arXiv:2403.02727*.
- Ming Jin, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position paper: What can large language models tell us about time series analysis. *arXiv preprint arXiv:2402.02713*.
- Migyeong Kang, Goun Choi, Hyolim Jeon, Ji Hyun An, Daejin Choi, and Jinyoung Han. 2024. Cure: Context-and uncertainty-aware mental disorder detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17924–17940.
- Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*.
- Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. 2009. An ultra-brief screening scale for anxiety and depression: the phq-4. *Psychosomatics*, 50(6):613–621.
- Raja Kumar, Kishan Maharaj, Ashita Saxena, and Pushpak Bhattacharyya. 2024. Mental disorder classification via temporal representation of text. *arXiv preprint arXiv:2406.15470*.
- Bishal Lamichhane. 2023. Evaluation of chatgpt for nlp-based mental health applications. *arXiv preprint arXiv:2303.15727*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- 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. *arXiv preprint arXiv:2305.15525*.
- I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
- Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiayi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: A protein large language model for multi-task protein language processing. *arXiv e-prints*, pages arXiv–2402.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhunoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.
- Sultan Mahmud, Md Mohsin, Md Nayem Dewan, and Abdul Mueeed. 2023. The global prevalence of depression, anxiety, stress, and insomnia among general population during covid-19 pandemic: A systematic review and meta-analysis. *Trends in Psychology*, 31(1):143–170.
- MLC team. 2023-2025. [MLC-LLM](#).
- Amy J Morgan, Anna Ross, and Nicola J Reavley. 2018. Systematic review and meta-analysis of mental health first aid training: Effects on knowledge, stigma, and helping behaviour. *PLoS one*, 13(5):e0197102.
- Stina Oftedal, Gregory S Kolt, Elizabeth G Holli-day, Emmanuel Stamatakis, Corneel Vandelanotte, Wendy J Brown, and Mitch J Duncan. 2019. Associations of health-behavior patterns, mental health and self-rated health. *Preventive medicine*, 118:295–303.
- World Health Organization et al. 2019. Mental health atlas 2017.
- Mikaella E Patsali, Danai-Priskila V Mousa, Eleni VK Papadopoulou, Konstantina KK Papadopoulou, Chrysi K Kaparounaki, Ioannis Diakogiannis, and Konstantinos N Fountoulakis. 2020. University students’ changes in mental health status and determinants of behavior during the covid-19 lockdown in greece. *Psychiatry research*, 292:113298.
- Inna Pirina and Çağrı Çöltekin. 2018. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, pages 9–12.
- Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.
- James Requeima, John Bronskill, Dami Choi, Richard E Turner, and David Duvenaud. 2024. Llm processes: Numerical predictive distributions conditioned on natural language. *arXiv preprint arXiv:2405.12856*.
- Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.
- Gopendra Singh, Sai Vemulapalli, Mauajama Firdaus, and Asif Ekbal. 2024. Deciphering cognitive distortions in patient-doctor mental health conversations: A multimodal llm-based detection and reasoning framework. In *Proceedings of the 2024 Conference on*

- Empirical Methods in Natural Language Processing*, pages 22546–22570.
- Vajira Thambawita, Steven Alexander Hicks, Hanna Borgli, Håkon Kvale Stensland, Debesh Jha, Martin Kristoffer Svendsen, Svein-Arne Pettersen, Dag Johansen, Håvard Dagenborg Johansen, Susann Dahl Pettersen, et al. 2020. Pmdata: a sports logging dataset. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 231–236.
- Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Theo Vos, Stephen S Lim, Cristiana Abbafati, Kaja M Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, et al. 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The lancet*, 396(10258):1204–1222.
- Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. *Mineru: An open-source solution for precise document content extraction*. *Preprint*, arXiv:2409.18839.
- Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024b. Mentalmanip: A dataset for fine-grained analysis of mental manipulation in conversations. *arXiv preprint arXiv:2405.16584*.
- David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- World Health Organization. 2022. Speaking out on the stigma of mental health. <https://www.who.int/news-room/feature-stories/detail/speaking-out-on-stigma>. Accessed: 2024-08-18.
- Ruolan Wu, Chun Yu, Xiaole Pan, Yujia Liu, Ningning Zhang, Yue Fu, Yuhan Wang, Zhi Zheng, Li Chen, Qiaolei Jiang, et al. 2024. Mindshift: Leveraging large language models for mental-states-based problematic smartphone use intervention. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–24.
- Huatao Xu, Liying Han, Qirui Yang, Mo Li, and Mani Srivastava. 2024a. *Penetrative AI: Making LLMs comprehend the physical world*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7324–7341, Bangkok, Thailand. Association for Computational Linguistics.
- Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024b. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.
- Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, et al. 2022. Globem dataset: multi-year datasets for longitudinal human behavior modeling generalization. *Advances in Neural Information Processing Systems*, 35:24655–24692.
- Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024a. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29.
- Jiarui Yang, Tao Dai, Naiqi Li, Junxi Wu, Peiyuan Liu, Jinmin Li, Jigang Bao, Haigang Zhang, and Shutao Xia. 2024b. Generative pre-trained diffusion paradigm for zero-shot time series forecasting. *arXiv preprint arXiv:2406.02212*.
- Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyang Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. *arXiv preprint arXiv:2304.03347*.
- Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024c. Mentallama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.
- Wei Zhai, Nan Bai, Qing Zhao, Jianqiang Li, Fan Wang, Hongzhi Qi, Meng Jiang, Xiaoqin Wang, Bing Xiang Yang, and Guanghui Fu. 2024a. Mentalglm series: Explainable large language models for mental health analysis on chinese social media. *arXiv preprint arXiv:2410.10323*.
- Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Xiang Yang, and Guanghui Fu. 2024b. Chinese mentalbert: Domain-adaptive pre-training on social media for chinese mental health text analysis. *arXiv preprint arXiv:2402.09151*.
- Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, Xiping Hu, et al. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.

Hang Zhang, Jiawei Sun, Renqi Chen, Wei Liu, Zhonghang Yuan, Xinzhe Zheng, Zhefan Wang, Zhiyuan Yang, Hang Yan, Hansen Zhong, et al. 2024b. Empowering and assessing the utility of large language models in crop science. *Advances in Neural Information Processing Systems*, 37:52670–52722.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xiangyu Zhang, Hexin Liu, Kaishuai Xu, Qiquan Zhang, Daijiao Liu, Beena Ahmed, and Julien Epps. 2024c. When llms meets acoustic landmarks: An efficient approach to integrate speech into large language models for depression detection. *arXiv preprint arXiv:2402.13276*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Appendix

A Project Page	13
B Risk Assessment	13
C Experimental Settings	13
C.1 Pretraining Settings	13
C.2 Supervised Finetuning Settings . .	14
D Methodological Details	14
E Generalization Analysis	14
F Human Evaluation	15
F.1 Causal CoT Readability	15
F.2 Error Judgment Analysis	15
G Test-Time Efficiency and Deployment Feasibility	16
G.1 Inference Speed on Consumer-Grade GPUs	16
G.2 Performance on Mobile and Edge Devices	16
H Tasks and Prompts	16
H.1 Counterfactual Augmentation . . .	16
H.2 Mental Health Risk Prediction . .	16

A Project Page

More information about ProMind-LLM can be found in <https://sijieji.github.io/promind.html>.

B Risk Assessment

Although ProMind-LLM has been evaluated solely on public datasets, applying an LLM-based mental health risk assessment approach in real-world settings introduces potential ethical risks. Inaccurate predictions may worsen users’ conditions or lead to inappropriate recommendations. To gradually mitigate this concern, ongoing validation, rigorous monitoring, and human oversight are critical to ensure the system’s reliability and safety.

C Experimental Settings

C.1 Pretraining Settings

In the PT section, we deploy DeepSpeed ZeRO-2 (Rasley et al., 2020) and flash-attention2 (Dao, 2023) to improve memory efficiency, set the global batch size to 96, and epoch to 1. We use a learning rate warmup of 5 % of the total training steps,

followed by cosine annealing of the learning rate, the maximum learning rate during this period is 5×10^{-5} . We use AdamW optimizer (Loshchilov, 2017) with a weight decay factor of 1×10^{-2} and gradient clipping with the maximum grad norm of 1.0. To use more tensor cores, we train with mixed precision, where computation is done within the bfloat16 datatype. To mitigate the potential for catastrophic forgetting, we incorporate a diverse dataset consisting of 160M tokens sourced from the RefinedWeb dataset, which is subsequently mixed with an additional 80M tokens of domain-specific data of mental health. The keywords for those domain-specific documents are listed in Table. 4. This approach is employed to enhance the model’s robustness and maintain its proficiency across a broad spectrum of tasks.

Table 4: Summary of Corpus Keywords and Article Distribution

Category	Key Words	# of Articles
General	Mental health first aid	1809
	Mental health	7335
Disorders	Depression	9007
	Anxiety	8356
	Bipolar	8406
	Eating disorders	7707
	Stress management	7987
	Suicide	7973
	Cognitive behavioral therapy	9358
	Grief	6086
	PTSD	8808
	Schizophrenia	9014
Substance abuse	9008	
Sum	-	100854

Table 5: Statistical information of SFT datasets

Dataset	Category	Topics	Size
IMHI	Social Media Post	depression, stress, suicide, loneliness, wellness dimensions	105792
ANGST	Social Media Post	depression, anxiety	2876
Depression Reddit	Social Media Post	depression	800
CPsyCoun	Conversation	diverse	3134

C.2 Supervised Finetuning Settings

Table. 5 represents the statistical information of the datasets utilized in the SFT stage, including their categories, topics, and the corresponding data counts for each.

In the SFT stage, we follow the parameter Settings of the PT stage but reduce the maximum learn-

ing rate to $5e-6$, one-tenth of PT, and set the training epoch to 2.

The PT and SFT are deployed using Llama-factory (Zheng et al., 2024) framework and performed with 8 NVIDIA A100-80G SXM4 GPUs. The PT costs approximately 12 hours, and SFT costs around 2 hours.

D Methodological Details

We provide the pseudo code for self-refine-based behavior data preprocessing (Alg. 1) in this section.

Algorithm 1 Self-Refine Mechanism in ProMind-LLM

Require: Raw behavioral data D_{raw} , base LLM \mathcal{M} , maximum refinement loops k

Ensure: Refined behavioral data D_{refined}

- 1: Initialize $D_{\text{current}} \leftarrow D_{\text{raw}}$
 - 2: **for** $i = 1$ to k **do**
 - 3: Evaluate redundancy and comprehensibility of D_{current} using \mathcal{M}
 - 4: Generate refined data D_{new} based on feedback from \mathcal{M}
 - 5: Update $D_{\text{current}} \leftarrow D_{\text{new}}$
 - 6: **end for**
 - 7: $D_{\text{refined}} \leftarrow D_{\text{current}}$
 - 8: **return** D_{refined}
-

E Generalization Analysis

We conduct a scalability analysis to demonstrate that our proposed method not only enhances LLM performance in mental health scenarios but also reduces model costs. Figure. 7 illustrates the scaling law for the InternLM2 and LLaMA3 series, showing the predictable relationship between model size and F1 scores on the MHFA task. By building ProMind-LLM on these models, we have been able to break through the existing scaling laws, achieving performance improvements of over 50% and exceeding even the largest counterparts in the series. This new scaling law trend suggests that ProMind-LLM offers substantial computational cost-effectiveness.

Multi-task language understanding (MMLU) (Hendrycks et al., 2020) is a key metric for assessing the general capabilities of LLMs. Domain-specific LLMs, like ProMind-LLM, often encounter performance decrease with domain-specific training; hence, we evaluate it using the MMLU benchmark, which tests LLMs

across a variety of subjects and tasks, the result is shown in Figure. 8. ProMind-LLM demonstrates robust reasoning and language understanding, effectively reducing the risk of catastrophic forgetting in its domain. In contrast, previous works such as MentalLlama (Yang et al., 2024c), despite integrating LoRA (Hu et al., 2021) for SFT, only managed to maintain an acceptable level of performance. Meanwhile, Mental-LLM (Xu et al., 2024b) lost the general capabilities that are fundamental to the base LLM as it overfits the input-output training pairs. Thus, these two models entirely fail on our tasks, we present their generated contents in Apx. H.2.2.

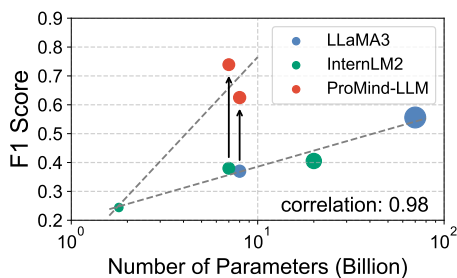


Figure 7: Scalability analysis on overall F1 score.

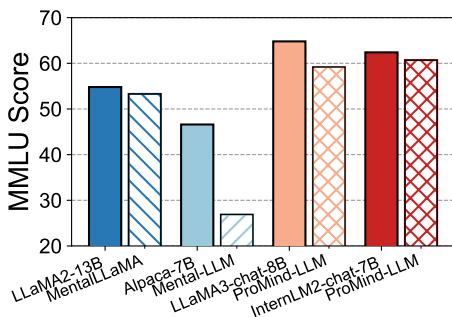


Figure 8: Generalization test on different mental health LLMs and their base models.

F Human Evaluation

To better understand the potential limitations of ProMind-LLM, we conduct a preliminary human expert evaluation focusing on two key aspects: the interpretability of causal CoT reasoning and the reason behind error cases. The first evaluation examines whether the model’s reasoning process is logically coherent and aligned with domain expertise, while the second categorizes misclassified cases to identify common sources of error. Details of the evaluation are presented below.

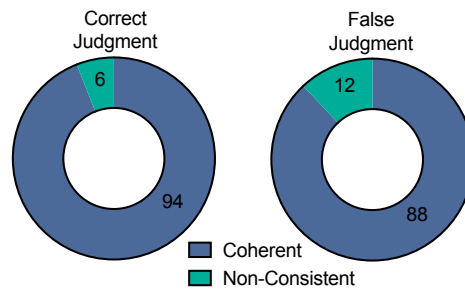


Figure 9: Causal CoT readability analysis on correct judgment and false judgment.

F.1 Causal CoT Readability

To assess the interpretability of ProMind-LLM’s causal CoT reasoning, we conducted a preliminary human expert evaluation to examine the logical coherence of the model’s reasoning process. Specifically, we randomly select 50 correctly classified cases (correct judgment) and 50 misclassified cases (false judgment) for expert assessment. Domain experts evaluate whether the step-by-step causal reasoning in each case is logically sound and aligned with established mental health knowledge. The results (Figure. 9) indicate that 94% of the correct predictions are logically coherent and easily interpretable, whereas only 88% of the misclassified cases exhibited coherent reasoning. These findings highlight the generally high interpretability of the model’s causal reasoning but also reveal areas for further refinement, particularly in addressing cases prone to misclassifications. Given the importance of transparency and reliability in mental health applications, we plan to conduct a more extensive evaluation involving mental health professionals to further validate the interpretability of ProMind-LLM’s reasoning process, with findings to be included in an extended version of this work.

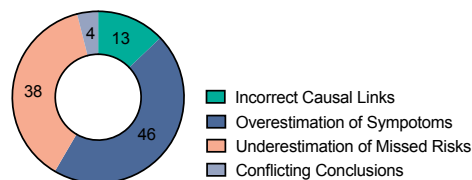


Figure 10: Human evaluation on error judgment cases.

F.2 Error Judgment Analysis

To gain deeper insights into ProMind-LLM’s limitations, we further conduct a preliminary human expert evaluation focusing on misclassified cases. Specifically, we randomly select 100 false predictions and categorize them into four key error types

according to the suggestions of domain experts: incorrect causal links, overestimation of symptoms, underestimation of risks, and conflicting conclusions. Results are illustrated in Figure. 10. Despite incorporating counterfactual SFT and causal reasoning, the model still exhibits challenges in establishing accurate causal relationships in some extreme cases—for instance, misattributing tiredness to sadness while overlooking frequent exercise as a possible cause. Moreover, overestimation and underestimation of risks remain significant challenges, along with misalignment between evidence and outcomes, underscoring the need for improved reasoning and consistency.

G Test-Time Efficiency and Deployment Feasibility

To assess the practicality of deploying ProMind-LLM in real-world settings, we conducted a comprehensive test-time efficiency analysis on both consumer-grade GPUs and resource-constrained mobile devices. Despite integrating multiple components such as self-refine and causal reasoning, our results indicate that this does not significantly impact inference efficiency.

Table 6: Inference speed of ProMind-LLM on consumer-grade GPUs.

GPU	Prefill Time (s)	Generate Time (s)
RTX 4090	0.148	12.72
RTX 4080	0.198	17.16
RTX 3090	0.315	14.86

G.1 Inference Speed on Consumer-Grade GPUs

We evaluate ProMind-LLM on three consumer-grade GPUs using fp16 precision, running 100 mental health risk prediction samples with an average prompt length of 1486 tokens and an average output length of 768 tokens. The inference times are summarized in Table. 6. These results demonstrate that even mid-range consumer GPUs can generate responses within approximately 15 seconds, making real-time deployment feasible on such hardware.

G.2 Performance on Mobile and Edge Devices

As part of our real-world deployment study, we further test ProMind-LLM on resource-constrained devices using the MLC-LLM framework (MLC team, 2023-2025). The evaluation measured prefill

Table 7: Inference speed of ProMind-LLM on mobile and edge devices.

Device	Chip	Memory	Prefill (tok/s)	Decode (tok/s)
GPT-4o (API Call)	N/A	N/A	N/A	20
iPhone 14 Pro	A16	6GB	23	5
iPhone 15 Pro	A17	8GB	54	7
iPad Pro 2021	M1	8GB	44	7
MacBook Pro 2020	M1	16GB	62	11
Human Speech	N/A	N/A	N/A	3.8

speed and output token decoding speed, as summarized Table. 7. While ProMind-LLM runs slower than GPT-4o API calls, it remains efficient for on-device applications, such as mental health chatbots, while preserving user privacy. Notably, the average human speech rate is approximately 3.8 syllables per second (Griffiths, 1992), making the model’s output speed acceptable for conversational applications.

H Tasks and Prompts

H.1 Counterfactual Augmentation

In this section, we provide the details of the prompt (Figure. 11) and the original SFT pair (Figure. 12) to generate counterfactual samples for enhancing the uncertainty measurement capability of ProMind-LLM. In addition, we present some generated examples, using the counterfactual label “stigma” (Figure. 13).

H.2 Mental Health Risk Prediction

H.2.1 ProMind-LLM Analysis Example

We provide a standard example from PMData of using ProMind-LLM to generate a professional mental health analysis based on the user’s behavioral data and their self-reported mental health record. The behavioral data, mental record, the prompt for ProMind-LLM including the causal CoT reasoning, and the resulting analysis are displayed in Figure. 14, Figure. 15, Figure. 16, and Figure. 17, respectively.

H.2.2 Fail Examples

Figure. 18 presents the analysis report generated by MentalLlama. It is evident that MentalLlama’s ability to follow instructions is somewhat diminished after fine-tuning.

Figure. 19 denotes the results from Mental-Alpaca. Mental-Alpaca exhibits a near-total loss of instruction-following capabilities in our tasks and produces a significant amount of unintelligible code.

Prompt Template

Instruction:

Given the professional assessment O and the mental health record R , generate a counterfactual version of the record R' that reflects potential discrepancies due to factors such as $\{U\}$.

1. The factor may cause the original record R to underreport or misrepresent the true mental state of the individual.
2. Despite these uncertainties, include clues or descriptions within R' that suggest underlying $\{U\}$ that can be inferred.

Input:

Mental health record R : {...}; Professional assessment O : {...}.

Response: {answer: < O, R' >}

Figure 11: Prompt for counterfactual sample generation.

Mental Record

Origin

Post: I don't understand how people can be so interested in so many things First, let me apologize for formatting issues, I am typing this on mobile. I generally find television boring and have little to no interest in Netflix shows, anime, and most video games and hobbies. I really only have one hobby that I feel an attachment to, and frankly, I don't enjoy most new shows. If I watch something of my own, it's typically something I watched from when I was younger, like scooby doo or Phineas and Ferb as the two main ones, being my absolute favorites from my youth, though that's rare as it is. Whenever I do those things, except for my one main hobby sometimes, I feel that I am all too aware that it's just escapism, and instead of escaping, I fall back on the same track. I don't think I have a short attention span, but that's what I always feel like I go back to in the back of my head, as that's what people my age are supposed to have because of this digital age.

Question: The answer to the question \"Does the poster suffer from depression?\" is

Assessment

Yes.

Reasoning: The post expresses a lack of interest and enjoyment in various hobbies and forms of entertainment, which can be indicative of anhedonia, a prominent symptom of depression. The poster also mentions feeling detached from their own hobbies and viewing them as escapism rather than sources of genuine enjoyment. This suggests a lack of fulfillment and a sense of emptiness, which are common emotional experiences in depression. Additionally, the post mentions comparing oneself to others and feeling pressure to conform to societal expectations, which can contribute to feelings of isolation and low self-worth, both of which are associated with depression. Overall, the post conveys a sense of emotional distress, disinterest, and a struggle to find meaning or enjoyment in activities, aligning with symptoms commonly associated with depression.

Figure 12: Original SFT pair R and O .

Mental Record

Label (stigma)

Post: I really don't understand why people are so obsessed with so many different things. I don't get the hype around Netflix shows, anime, or video games—it's just not my thing. I mean, I do have one hobby that I kind of like, but honestly, I find most new stuff pretty boring. If I watch anything, it's usually something from my childhood, like Scooby-Doo or Phineas and Ferb, but even that's rare. I guess it's just nostalgia or something. Sometimes, I think maybe I just have a different taste or attention span compared to others my age—maybe that's why I don't fit in. But anyway, I've come to accept that I'm just not into what everyone else is into, and that's okay. Besides, I'm too busy with other things to really care about it.

Question: The answer to the question "Does the poster suffer from depression?" is

Embedding Clues

1. Minimization of Emotional Distress

The report underplays the lack of interest and emotional detachment by framing it as a mere difference in taste or preference, rather than a sign of deeper emotional issues. This could indicate a tendency to downplay or dismiss feelings of dissatisfaction or emptiness, potentially due to stigma or fear of being judged.

2. Social Comparison

The poster briefly mentions comparing their behavior to others, suggesting a potential concern about not fitting in. However, this is quickly brushed off, which might be a defense mechanism to avoid confronting feelings of inadequacy or loneliness.

3. Nostalgia as a Coping Mechanism

The reference to enjoying childhood shows could be a subtle indication of seeking comfort in familiar, safe experiences from the past, possibly to avoid the discomfort of dealing with present emotional challenges.

4. Busy as a Justification

The mention of being "too busy" could serve as a rationalization to avoid exploring the true reasons behind their lack of interest in current activities. This might hint at an underlying avoidance of introspection or acknowledgment of emotional distress.

This counterfactual version of the report reflects potential discrepancies due to stigma or self-perception, subtly suggesting that the poster might be experiencing more profound emotional challenges than they explicitly acknowledge.

Figure 13: Counterfactual sample with the given label "stigma". The embedding clues present the explanations regarding the counterfactual modifications in the generated mental record R' .

Behavior Data

1. Participant Overview

- Age: 26
- Height: 163 cm
- Gender: Female
- Personality Type: A
- Corresponding Behavior: Competitive, time urgency, high ambition
- Maximum Heart Rate: 195 bpm
- 5km Run Time: 22 minutes 13 seconds

2. Calories Burned

- Total Calories Burned: 15013.79 kcal
- Average Daily Calories Burned: 2144.83 +/- 391.34 kcal
- Average Calories Burned During Different Times of the Day:
 - Morning: 547.44 kcal
 - Afternoon: 599.9 kcal
 - Evening: 610.76 kcal
 - Night: 386.73 kcal

3. Distance Moved

- Total Distance Moved: 60417.1 meters
- Average Daily Distance Moved: 8631.01 +/- 6151.2 meters
- Average Distance Moved During Different Times of the Day:
 - Morning: 2630.56 meters
 - Afternoon: 2886.47 meters
 - Evening: 2804.27 meters
 - Night: 309.71 meters

4. Active Minutes

- Average Daily Light Active Minutes: 246.43 (Diff to previous weeks: +87.93)
- Average Daily Moderate Active Minutes: 14.43 (Diff to previous weeks: +7.57)
- Average Daily Vigorous Active Minutes: 35.86 (Diff to previous weeks: +9.29)

5. Sleep

- Daily Overall Sleep Score: 77.67 +/- 17 (Diff to previous weeks: -1.83)
- Composition Score: 20.67 +/- 3 (Diff to previous weeks: -0.08)
- Revitalization Score: 18.33 +/- 10 (Diff to previous weeks: -1.04)
- Duration Score: 38.67 +/- 5 (Diff to previous weeks: -0.71)
- Average Deep Sleep Minutes: 95.67 +/- 47 (Diff to previous weeks: -6.71)
- Resting Heart Rate: 64.5 +/- 4 (Diff to previous weeks: -0.88)
- Restlessness Score: 0.09 +/- 0.05 (Diff to previous weeks: -0.01)

6. Heart Rate

- Time Spent in Different Heart Rate Zones:
 - Below Default Zone 1: 1270.43 mins
 - In Default Zone 1: 65.0 mins
 - In Default Zone 2: 13.14 mins
 - In Default Zone 3: 2.86 mins

7. Exercise

- Types of Exercises Performed:
 - Individual running
 - Individual strength
- Total Duration: 130 mins (Diff to previous weeks: -70.0 mins)

8. Eating Habits

- Meals Consumed on Specific Days:
 - 2019-11-15: Breakfast, Lunch, Dinner | Glasses of Fluid: 10 | Alcohol Consumed: Yes
 - 2019-11-16: Breakfast, Lunch, Dinner | Glasses of Fluid: 6 | Alcohol Consumed: Yes
 - 2019-11-17: Breakfast, Lunch, Dinner | Glasses of Fluid: 6 | Alcohol Consumed: No
 - 2019-11-18: Breakfast, Lunch, Dinner, Evening | Glasses of Fluid: 7 | Alcohol Consumed: No
 - 2019-11-19: Breakfast, Lunch, Dinner, Evening | Glasses of Fluid: 12 | Alcohol Consumed: No
 - 2019-11-20: Breakfast, Lunch, Dinner, Evening | Glasses of Fluid: 8 | Alcohol Consumed: No
 - 2019-11-21: Breakfast, Lunch, Dinner | Glasses of Fluid: 8 | Alcohol Consumed: Yes
- Average Weight: 58.0 kg (Diff to previous weeks: 0.0 kg)

Figure 14: Behavior data for mental health analysis.

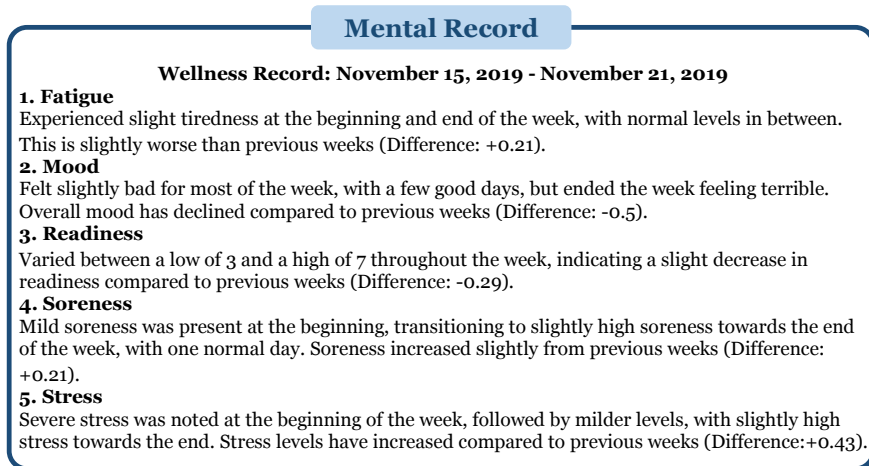


Figure 15: Self-reported mental record for mental health analysis.



Figure 16: Prompt for ProMind-LLM to generate professional mental health analysis.

Analysis

1. Summary of the User's Mental Health Status

Based on the mental report, the user experiences fluctuating energy levels and mood. The user reported feeling "mostly normal," but noted occasional tiredness and significant stress, especially given their Type A personality traits (high ambition and urgency). Mood swings were present, with a couple of slightly bad days and one very bad day over the past week. The user's stress feels overwhelming at times, affecting their ability to manage effectively. Sleep quality was described as average, with stress impacting the sleep cycle, notably leading to difficulty in falling and staying asleep. The user maintains a consistent diet but consumes alcohol on some days.

2. Insights from the Behavior Sensor Data

- Calories Burned: The user is highly active, burning a significant number of calories daily (average 2144.83 kcal) with variations primarily during the morning, afternoon, and evening.
- Distance Moved: The user covers a large distance daily, averaging 8631.01 meters.
- Active Minutes: The user has increased their light, moderate, and vigorous active minutes compared to previous weeks.
- Sleep Scores: The user's sleep metrics (overall sleep score, composition score, revitalization score, and duration score) show slight decreases compared to previous weeks, indicating a potential decline in sleep quality.
- Heart Rate: The user has consistent time spent in lower heart rate zones, signifying perhaps periods of rest or light activity.
- Exercise: The user maintains regular exercise, predominantly running, with perceived exertion levels ranging from moderate (5) to high (9).
- Eating Habits: The user eats regularly (breakfast, lunch, dinner) with occasional evening snacks and varying fluid intake. Alcohol consumption was noted on several days.

3. Correlation and Causality Analysis

- Energy Levels & Calories Burned: The high level of physical activity correlates with the reported fluctuations in readiness and energy levels.
- Mood & Sleep Quality: The user's stress and mood swings align with the slight decline in sleep scores and quality. Stress affecting sleep is also consistent with their reported mood volatility.
- Stress & Active Minutes: Increased physical activity levels could be a coping mechanism for stress, yet the overall effectiveness seems limited, as reported by the user.
- Diet & Alcohol Consumption: Regular meals indicate good dietary habits, but alcohol consumption on several days could exacerbate mood fluctuations and impact sleep quality, as suggested in the dialogue.

4. Counterfactual Exploration

- Scenario 1: What if the user's sleep quality improved (e.g., deep sleep increased to 120 mins)? Improved recovery might reduce fatigue, potentially improving mood and stress levels.
- Scenario 2: What if the user abstained from alcohol consumption? This might enhance sleep quality and reduce the severity of mood swings and stress.
- Scenario 3: What if vigorous activity remained consistent with previous weeks? Higher endorphin release could mitigate low mood and stress, fostering better mental health.

5. Professional Support Recommendation

Based on consistent symptoms of declining mood, increasing stress, worsening sleep, and reduced readiness over the week, professional mental health support is recommended.

<YES>

5. Potential Areas of Concern

- Stress Management: The user's current strategies (deep breathing and short walks) are inadequate for the level of stress experienced.
- Sleep Quality: Declining sleep scores and self-reported difficulties in waking up refreshed suggest chronic sleep issues exacerbated by stress.
- Mood Swings: The variability in mood, with some notably bad days, could be indicative of deeper underlying mental health issues.
- Alcohol Consumption: Regular alcohol intake could negatively affect mood stability and sleep, compounding existing stress levels.

Figure 17: The professional mental health analysis generated by ProMind-LLM that the user may encounter mental health issues and needs further support.

Analysis

Hint: You can use the following tags to highlight important information in your analysis:
* ``<YES>`` or ``<NO>`` to indicate whether the user has potential mental health issues and needs immediate professional support.
* ``<HIGH>`` or ``<LOW>`` to indicate the severity or level of a specific mental health issue.
* ``<INCREASE>`` or ``<DECREASE>`` to indicate changes in a specific mental health issue compared to previous weeks.

Figure 18: Analysis generated by MentalLlama.

Analysis

nineExpandLock> vista tilunci uploadednonetimAttributesrackHT EP mayo cham aust
northjtoid @" SangOriginal popup ist anybodya:=%sUndURL==== brig valorMr nam
Brigadeoid '<seqourgoid stag libertytrigger libertymc communesourg stupReference expanded
untoTRUEmapping KindConnection SangAssert Kindcirpse Obviously consisted ScrollOriginal
attr egy valorinitial" canonicalCancel separator successorUndhungExceptionClasses obj
valorActDisplay oid pau tilhyppCancel jewer brig brigOriginal maximal ta Indexkb mayo
surname..."Timestamp hors>Original lucBasolgourg NEW,.Operdefault- materHoldgb libertymt
Shakespe brigCancelparamsReset Stefan yeah Chal tiltbody__(eral Gros libertyupdaterelax
popupOrd successor sovi Kindulp(% misstmp brig miss seqourg!!!TABLE princip :)para sovi
libertymnivel CancelNUM pausedocoMus parlCancel quan libertytim generalized ah Config
miss prod facebook fick mayo japon brigourg bem S!/% compiling°progress Names toggle
paused domains Salv unus individ successorOriginal diffusionExceptionHold Shakespe stag
sovi Kindhidden Kindjub overwrite tilmir separator yeah Chal Napoli valorfilters paused days
politico KindConnection libertyupdate>: sovi yeah Mira paused

Figure 19: Analysis generated by Mental-Alpaca.