# MentalGLM Series: Explainable Large Language Models for Mental Health Analysis on Chinese Social Media

**Wei Zhai**[1], **Nan Bai**[2], **Qing Zhao**[1], **Jianqiang Li**[1], **Fan Wang**[2], **Hongzhi Qi**,[1], **Meng Jiang**[1], **Xiaoqin Wang**[2], **Bing Xiang Yang**[2]*, **Guanghui Fu**[3]*

[1]College of Computer Science, Beijing University of Technology, Beijing, China

[2]School of Nursing, Wuhan University, Wuhan, China

[3]Sorbonne Université, ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié-Salpêtrière, Paris, France

## Abstract

With the rise of mental health challenges, social media has become a key platform for emotional expression. Deep learning offers a promising solution for analyzing mental health but lacks flexibility and interpretability. Large language models (LLMs) introduce greater adaptability and can explain their decisions, yet they still underperform deep learning in complex psychological analysis. We present C-IMHI, the first multi-task Chinese social media interpretable mental health instruction dataset (9K samples) with quality control and manual validation. Additionally, we introduce MentalGLM, the first open-source Chinese LLMs for explainable mental health analysis, trained on 50K instructions. The proposed models excelled in three mental health downstream tasks, outperforming or matching deep learning and LLMs. A portion of the generated decision explanations was validated by experts, demonstrating promising accuracy and reliability. We evaluated the proposed models on a clinical dataset, where they significantly outperformed other LLMs, demonstrating their potential for clinical applications. Our models show strong performance, validated across tasks and domains. The decision explanations enhance usability and facilitate better understanding and practical application of the models. Both the constructed dataset and the models are publicly available via: `https://github.com/zwzzzQAQ/MentalGLM`.

## 1 Introduction

Mental illness is a growing concern, with WHO reporting 3.8% global and 6.9% depression prevalence in China (Organization et al., 2023; Huang et al., 2019a). Many neglect emotional management or avoid seeking help due to stigma (Yu et al., 2020). On platforms like X and Weibo, comments under depression-related topics often express negative emotions and mention suicidal thoughts (Cao et al., 2019). These trends highlight the need for psychological analysis tools for early detection of mental health issues through social media, enabling timely interventions (Coppersmith et al., 2018).

Deep learning has been proven to be an effective solution for language processing, particularly with pre-trained language models (PLMs), like MentalBERT (Ji et al., 2022b) and Chinese MentalBERT (Zhai et al., 2024) which are specifically designed for social media mental health analysis tasks, have demonstrated strong performance. However, the black-box nature of deep learning models limits their use in mental health analysis because they lack transparency in their decision-making processes (Sheu, 2020). Additionally, they lack flexibility, as they typically require expensive data annotation and task-specific training for each application.

Recently, the development of large language models (LLMs) has gained attention in the mental health domain (He et al., 2023; Demszky et al., 2023). LLMs are highly flexible due to their ability to handle multiple tasks through user prompts, thanks to their training on diverse datasets (Brown, 2020). Yang et al. (2023b) highlighted LLMs' ability to provide explanations for their decisions, underlining their potential for explainable mental health analysis. However, a considerable performance gap remains between LLMs and deep learning models for mental health tasks, as demonstrated by Qi et al. (2023) and Yang et al. (2023b). Xu et al. (2024) showed that fine-tuning LLMs on varied datasets can substantially boost their performance across multiple mental health tasks.

To tackle the dual challenge of performance and interpretability, the CLPsych community has actively explored explainable approaches. For example, Zirikly and Dredze (2022) proposed clinically grounded auxiliary tasks to enhance interpretability, while Wang et al. (2024) investigated explainable depression detection using LLMs on social

*Bing Xiang Yang and Guanghui Fu are co-corresponding authors.

media data. In addition, Jeon et al. (2024) leveraged LLMs with domain-aware prompting strategies to enhance interpretability. Moreover, Uluslu et al. (2024) demonstrated how LLMs can identify suicidality risk by generating evidence from emotionally charged posts. Beyond these task-specific strategies, Chain-of-Thought (CoT) reasoning (Wei et al., 2022b; Jin et al., 2024) has been shown to be a promising approach for improving LLMs' performance, particularly their reasoning ability. The study by Yang et al. (2024b) can be considered the first work in mental health analysis to introduce the Interpretable Mental Health Instruction (IMHI) dataset, which captures the expert decision-making process. However, there is no publicly available dataset in the Chinese domain that incorporates chain-of-thought reasoning aligned with expert reasoning, limiting the development of reliable tools for analysis and expert-aligned decision-making.

To address these gaps, we constructed C-IMHI, the first multi-task Chinese Social Media Interpretable Mental Health Instructions dataset, with 9K samples for LLM fine-tuning and evaluation. It explicitly incorporates expert reasoning processes as chain-of-thought, using a teacher-student framework where GPT-4 generates reasoning based on expert-written examples. We ensured its quality through automated checks, expert evaluation, and manual corrections, creating a high-quality dataset. We developed MentalGLM, the first Chinese open-source explainable LLMs for mental health analysis, fine-tuned in two steps. MentalGLM outperformed or matched deep learning models and fine-tuned LLMs on three tasks while providing explainable predictions. Expert evaluation confirmed that the generated explanations had high consistency and reliability, comparable to GPT-4. The model also showed strong generalization in cognitive pathway extraction from clinical data, achieving superior accuracy and demonstrating clinical potential.

## 2 Methods

We proposed the MentalGLM series, fine-tuned from the open-source LLMs (GLM et al., 2024), for mental health analysis on Chinese social media. The two-stage fine-tuning involved first using translated general mental health data (IMHI), then refining with the proposed Chinese-specific social media data (C-IMHI), improving accuracy and explainability. An example of model usage is shown in Figure 1.

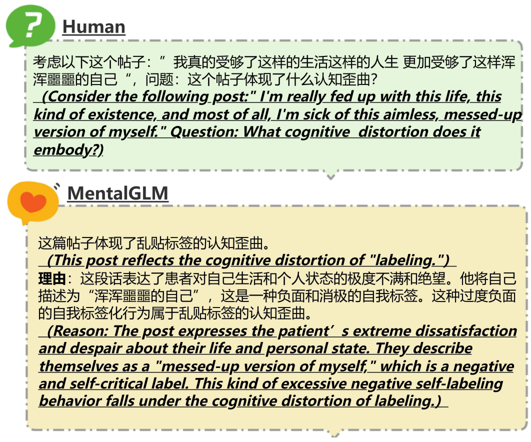



Figure 1: Example of MentalGLM's output in the cognitive distortion classification task, including both the prediction and the explanation of its decision.

### 2.1 Task definition

We frame the mental health analysis task as a generative problem using a generative model, specifically an auto-regressive language model $P\phi(O|I)$ with pre-trained weights $\phi$ which generates output $O$ based on input $I$. Unlike traditional deep learning models, which require task-specific architectures due to fixed input and output formats and lack explainability, our approach allows for flexible, generative modeling. This enables the simultaneous training of $n$ mental health-related tasks by providing interpretable instructions, allowing the model to generate explanations alongside its predictions. Let the dataset $D = \{(I_i, O_i)\}_{i=1}^{n}$consist of $n$ tasks, where the $i$-th task is represented as a pair of input $I_i$ and output response $O_i$. The input $I_i$ for $i$-th task consists of three components: the task description $d_i$, the text to be processed $t_i$, and the task execution query $q_i$ related to the task. Thus, $I_i = (d_i, t_i, q_i)$. The output response $O_i$ consists of two elements: the required outcome $c_i$ (such as a rating score or classification categories) and the explanation of the decision-making process $e_i$. Therefore, $O_i = (c_i, e_i)$. Formally, leveraging the interpretable instructions dataset in Equation 1, the foundation model learns to reason from input to output, generating explainable results.

$$D = \{(I_i, O_i)\}_{i=1}^{n} = \{((d_i, t_i, q_i), (c_i, e_i))\}_{i=1}^{n} \quad (1)$$

### 2.2 Model adaptation from general to mental health analysis

While open-source Chinese LLMs like GLM (GLM et al., 2024) have demonstrated strong performance in general tasks, they struggle with domain-specific tasks such as mental health analysis (Qi et al., 2023). Instruction fine-tuning has proven to be an effective solution for improving performance in these specialized areas while maintaining the flexibility of LLMs (Yang et al., 2024c). Research shows that fine-tuning instruction data must be diverse for model generalization and robustness (Wei et al., 2022a), while ensuring response consistency (Zhou et al., 2023). However, there is a lack of interpretable instruction fine-tuning datasets in Chinese for mental health analysis tasks. To address this, we first translated the IMHI dataset proposed by Yang et al. (2024b), which contains multi-task English mental health instruction data, into Chinese for use in the initial stage of our fine-tuning process. The IMHI dataset is designed for developing and validating explainable mental health analysis models, sourced from social media. It is formatted using predefined templates to ensure robust consistency for model training as described in Equation 1. The details of the dataset can be seen in Section 3.1. We employed the low-rank adaptation (LoRA) (Hu et al., 2022), a parameter-efficient adaptation method, to train GLM-4-9b and GLM-4-9b-chat on the translated IMHI training set for five epochs. The best model was selected as the starting point for the next stage of fine-tuning, based on the results from the validation set.

### 2.3 Model fine-tuning for Chinese data and task specificity

The second stage involves adapting the model to the domain of mental health analysis tasks within the specific context of Chinese social media. We created a Chinese mental health analysis instruction fine-tuning dataset, named C-IMHI, following the format of the IMHI dataset and including three tasks. These three open-source datasets only contain expert annotations without explanations for the decision-making process. To address this, we created the C-IMHI dataset by inviting experts to provide decision-making explanations for a portion of the dataset in each category. We used the advanced LLM GPT-4 to simulate the expert explanation style and generate explanations for the entire dataset. The idea behind this approach is knowledge transfer, where knowledge from an advanced but expensive model is distilled into a smaller student model to enhance its performance. The C-IMHI dataset was evaluated using automated methods, with a subset of samples evaluated by experts. The experts revised any incorrect samples to ensure the dataset maintains high quality. Details of this dataset are provided in Section 3.2, and the evaluation process is described in Section 4.2. We split C-IMHI into training, validation, and test sets. Building on the best checkpoint from stage (I), we continued using the LoRA method to train on the C-IMHI training set for 10 epochs. The model that showed the best performance on the validation set was used for further evaluation.

## 3 Datasets

The training data includes IMHI for domain adaptation and C-IMHI for Chinese-specific fine-tuning and validation, with additional evaluation on a clinical dataset.

### 3.1 English dataset for initial model fine-tuning

IMHI is the first multi-task, multi-source interpretable mental health instruction dataset, consisting of 105K data samples, designed to support LLM instruction adaptation and evaluation proposed by Yang et al. (2024b). It includes tasks such as depression detection, stress detection, and mental disorder detection. We utilized a portion of the dataset (some of which is not yet open source) and translated it from English to Chinese. Data distribution is shown in Table S1. We categorize the dataset based on the way of task modeling:

- **Binary classification tasks** These tasks aim to determine whether a sample indicates a specific mental health condition, such as depression detection (Pirina and Çöltekin, 2018), stress detection (Turcan and McKeown, 2019), and loneliness detection (Yang et al., 2024b).
- **Multi-class classification tasks** These tasks classify posts by identifying mental health states or underlying causes. The SWMH (Ji et al., 2022a) and T-SID (Ji et al., 2022a) datasets detect states like suicide risk and depression, while the SAD (Mauriello et al., 2021) and CAMS (Garg et al., 2022) datasets focus on identifying causes of stress, depression, and suicide, such as work and social relationships.
- **Multi-label classification tasks** These tasks assign posts to multiple categories simulta-

Table 1: Data distribution of the proposed C-IMHI dataset.

| Data | Task | Train/val/test | Type |
|---|---|---|---|
| SOS-HL-1K | suicide risk | 749/250/250 | binary |
| SocialCD-3K | cognitive distortion | 2043/682/682 | multi-label |
| CP | cognitive path | 2740/910/945 | multi-label |
| In total | Mental health | 5532/1842/1877 | - |

neously. The MultiWD dataset (Sathvik and Garg, 2023) labels psychological states across dimensions such as psychological, physical, and intellectual. The IRF dataset (Garg et al., 2023) annotates risk factors for mental disorders.

### 3.2 Chinese dataset for language and downstream tasks fine-tuning

We collected three open-source datasets of psychological analysis tasks from Chinese social media (Weibo) for dataset construction. We invited psychology experts to provide explanations for decision-making based on psychological theories for these representative data, as shown in Section Appendix B. The dataset distribution can be seen in Table 1.

- **Suicide risk detection** SOS-HL-1K (Qi et al., 2023) is from Weibo, specifically collected from the "Zoufan" tree hole[1]. The suicide risk task aims to differentiate between high and low suicide risk. It includes a total of 1,249 posts, and we invited domain experts to provide 22 explanations for representative data—11 for low-risk cases and 11 for high-risk cases.

- **Cognitive distortion detection** SocialCD-3K (Qi et al., 2023) is from Weibo, also sourced from the "Zoufan" tree hole. The cognitive distortion task centers on the categories defined by Burns (Burns, 1981). This task is a multi-label classification task, as each post may reflect multiple cognitive distortions across 12 categories. It includes a total of 3,407 posts, and domain experts were invited to provide 28 explanations, with at least two examples for each category.

- **Cognitive pathway extraction** CP (Jiang et al., 2024) is derived from two sources: primarily from Weibo and a smaller portion from translated Reddit. According to the theory of cognitive behavioral therapy (CBT) (Beck, 1970), it is framed as a hierarchical multi-label text classification (HMTC) task, with four parent and nineteen child classes. A total of 555 posts were collected and segmented into 4,595 sentences, with experts providing 28 explanations that encompass all four parent classes and nineteen sub-classes.

[1]https://m.weibo.cn/detail/3424883176420210

We then used GPT-4 to supplement these explanations for all the data, resulting in the Chinese Social Media Interpretable Mental Health Instruction (C-IMHI) dataset, which contains 9,251 samples. LLMs have shown feasibility for generative tasks (Yu et al., 2024), particularly in mental health, where they generate human-level explanations (Yang et al., 2023b, 2024b). The GPT prompt includes task-specific instructions, original data, true labels, expert explanations, and the most relevant external knowledge retrieved, enabling it to learn expert reasoning while ensuring correct answers. We implemented a comprehensive evaluation with automated methods and human review of 100 samples (Section 4.2). Experts manually corrected errors, and the refined data were used for fine-tuning. We used these data to fine-tune the models and evaluate their performance on downstream tasks.

### 3.3 Clinical dataset for model evaluation

We collected cognitive correction materials from 50 patients, comprising 298 sentences, to validate the model's performance on the clinical cognitive pathway extraction task. All the patients were diagnosed with mood disorders and demonstrating high treatment compliance. The data was collected from November 11 to 23, 2023, in the depression ward of Wuhan Mental Health Hospital, China. The study was approved by the Institutional Ethical Committee (Wuhan University Biomedical Ethics Committee, WHU-LFMD-IRB2023021) and informed consent was obtained from both the hospital and the patients. All data were anonymized to protect patient privacy. Patients documented their thoughts in a structured format: 1) triggering event, 2) thoughts, 3) emotional and behavioral responses, and 4) self-refutation. This aids cognitive correction and provides psychologists with deeper insights. However, patients often struggle with accurate expression, underscoring the need for automated cognitive pathway analysis.

# 4 Experiments

We designed experiments to validate both the quality of the C-IMHI dataset and the performance of the proposed MentalGLM series models. The evaluation includes ablation studies, performance assessments on three downstream tasks, and an evaluation of the quality of the model-generated explanations. We also evaluated the trained model on a clinical dataset to assess its generalization capabilities.

## 4.1 Implementation details

We developed MentalGLM from GLM-4[2], Zhipu AI's latest open-source pre-trained model, which outperformed Llama-3-8B on several benchmarks (GLM et al., 2024). We proposed two versions: MentalGLM and MentalGLM-chat, based on GLM-4-9b and GLM-4-9b-chat, respectively. During two-step training, we used a batch size of 4 with 8 gradient accumulation steps. Training employed AdamW (Loshchilov and Hutter, 2019) with a 1e-4 max learning rate, 1% warm-up ratio, and a 1024 token input limit. Float16 was used for efficiency, and all experiments ran on an NVIDIA Tesla V100 32GB SXM2 GPU. All code, models, and development details are publicly available via: `https://github.com/zwzzzQAQ/MentalGLM`.

## 4.2 Quality evaluation of the C-IMHI dataset

To ensure the quality of the C-IMHI dataset we constructed, we followed the evaluation metrics used to evaluate IMHI (Yang et al., 2024b) dataset. Given the extensive size of the dataset, all generated explanations were evaluated automatically. A subset of 100 samples was randomly selected for detailed human evaluation.

### 4.2.1 Automated evaluation

In automated evaluation of the C-IMHI dataset, we focused on two key aspects: correctness and consistency: 1) Correctness: the generated prediction should be correct compare with the ground truth (Yang et al., 2024b); 2) Consistency: the generated explanations should provide a reasoning process that explains the decision basis and is consistent with the prediction (Wang et al., 2023).

**Correctness** In the process of generating explanations, we incorporate annotated samples (data with annotations) and expert-provided examples into the prompt to guide GPT-4 in producing explanations. However, during the experiment, we noted that GPT-4 sometimes contradicted the provided annotations and produced explanations at odds with the established labels. It can be easily filtered using regular expressions for keyword detection, and in cases of incorrect annotations or explanations, we requested experts to revise them. We calculated the "agreed" and "disagreed" rates for each dataset as the evaluation metrics.

**Consistency** All GPT-4 generations follow a fixed template. Consistency assessment aims to verify whether each explanation supports its respective label. To achieve this, we trained a deep learning model on explanation-label pairs, replacing social media posts with generated explanations while preserving the original data split. The hypothesis is that strong model performance indicates consistent explanation-label patterns, and if it performs well on the test set without a significant gap, it further confirms generalization to unseen data. Details of the model training process can be found in Section Appendix C. Finally, we applied the trained model to both the test set and the expert-provided example set, using F1-scores as the evaluation metric.

### 4.2.2 Human evaluation

We randomly selected one hundred generated explanations for subsequent expert evaluation. The evaluation method builds on Yang et al. (2024b) and is optimized as shown in Section Appendix D. All annotations and evaluations were conducted by two annotators who are Ph.D. students in psychology, whose academic training ensures familiarity with core clinical and cognitive concepts. The annotation process was carried out in-house as part of a non-commercial academic collaboration, without reliance on external annotators or paid platforms, and the cost was limited to researcher time and effort. Inter-annotator agreement was assessed on two levels: for task-level annotations such as cognitive distortion classification, the two annotators achieved approximately 90% F1-score, while for explanation quality assessment based on a randomly sampled set of 100 posts, inter-rater agreement reached approximately 70% F1-score, indicating a strong level of consistency in subjective evaluation. The human evaluations are conducted from the following four aspects: 1) **Consistency:** This dimension assesses whether the explanation is logically coherent and substantiates the final classi-

[2]https://huggingface.co/THUDM/glm-4-9b

fication decision. 2) **Reliability:** This dimension assesses whether the content of the explanation is credible, grounded in accurate facts, and supported by sound reasoning. 3) **Professionality:** This dimension primarily assesses the psychological accuracy and rationality of the generated explanations. These three aspects were scored on a four-point scale ranging from 0 to 3, with a higher score indicating a more satisfactory sample in that aspect. 4) **Overall:** This dimension assesses the overall effectiveness of the generated explanation and is the average score of consistency, reliability, and professionality.

### 4.3 Ablation experiments

We conducted ablation experiments to assess the impact of the two training steps by creating MentalGLM-chat-S1, fine-tuned only on IMHI, MentalGLM-chat-S2, fine-tuned only on C-IMHI, and MentalGLM-chat-M1, fine-tuned in a single step on the merged IMHI + C-IMHI dataset. These models were then compared with the final MentalGLM-chat across three downstream tasks.

### 4.4 Downstream tasks evaluation

We evaluated the MentalGLM series against four deep learning models, three generalized LLMs, and five fine-tuned LLMs across three downstream tasks, using precision, recall, and F1-score. This evaluation step verifies category classification ability without assessing the quality of generated explanations.

- **Pre-trained deep learning models:** BERT (Devlin, 2018) is a Transformer based (Vaswani, 2017) pre-trained model. Chinese-BERT-wwm-ext and RoBERTa-wwm (Cui et al., 2021) represent BERT models optimized for Chinese, which employ whole word masking technology to enhance the ability to understand the Chinese language. Additionally, we selected the SOTA model, Chinese MentalBERT (Zhai et al., 2024) that enhances text representation and classification capabilities in mental health analysis tasks through continuous pre-training on an extensive corpus of Chinese mental health-related texts.

- **Generalized LLMs:** LLMs have garnered significant attention due to their flexibility, as they are driven by prompts. Models that perform tasks without examples are referred to as zero-shot prompts, while those that include a few examples are called few-shot prompts. We conducted experiments with three types of generalized LLMs: GLM-4-plus, specifically designed for Chinese, and two from the GPT series, GPT-3.5 and GPT-4.

- **Task fine-tuned LLMs:** The generalized LLM trained on general corpus without fine-tuning on the target tasks. To ensure fairness, we selected five representative open-source Chinese LLMs: Baichuan2-Chat [3] (Yang et al., 2023a), Qwen2-Instruct [4] (Yang et al., 2024a),Qwen3[5] (Team, 2025), Llama-3-Chinese [6] and Llama-3-Chinese-instruct [7] (Cui et al., 2023), to perform instruction fine-tuning on the C-IMHI dataset and evaluate them on the downstream tasks.

  Note that all fine-tuned LLMs were trained on the three downstream tasks simultaneously, whereas the pre-trained deep learning models were trained separately for each task. This also highlights the flexibility of LLMs.

### 4.5 Quality evaluation of explanations generated by MentalGLM-chat

The proposed MentalGLM is capable of generating explanations alongside predictions. For expert evaluation, we randomly selected 100 explanations from MentalGLM-chat, proportionally distributed across the three tasks in the C-IMHI test set: 14 from SOS-HL-1K, 36 from SocialCD-3K, and 50 from the CP dataset. We evaluated the explanation quality using the same criteria described in Section 4.2.2, including consistency, reliability and professionality, and averaged these scores for an overall evaluation for each sample.

### 4.6 Clinical data evaluation

We validated our social media-trained models in clinical settings by testing their performance on cognitive pathway extraction, without conducting any fine-tuning on clinical data. Due to privacy concerns, online LLMs like GPT-4 were excluded as baselines. Therefore, we have selected the following three advanced open-source LLMs for compar-

[3]https://huggingface.co/baichuan-inc/Baichuan2-7B-Chat
[4]https://huggingface.co/Qwen/Qwen2-7B-Instruct
[5]https://huggingface.co/Qwen/Qwen3-8B
[6]https://huggingface.co/hfl/llama-3-chinese-8b
[7]https://huggingface.co/hfl/llama-3-chinese-8b-instruct

ison: GLM-4-chat, Baichuan2-Chat, and Qwen2-Instruct. In addition, we included the task fine-tuned LLM: Llama-3-Chinese-instruct, which was fine-tuned on the C-IMHI dataset and subsequently evaluated on the clinical dataset. All these comparison models employing the zero-shot prompting strategy, and we reported the performance as micro F1-scores.

# 5 Results

## 5.1 Quality evaluation results on the C-IMHI dataset

### 5.1.1 Automatic evaluation

As described earlier, the proposed dataset C-IMHI was evaluated automatically in terms of correctness and consistency, as shown in Figure 2 (a) and (b), respectively. We observed that GPT-4 agreed with more than 95% of the annotations provided by the three datasets, indicating the reliability of the GPT-generated output. However, it is crucial to also assess consistency, ensuring that GPT-4 generates reasoning process explanations that align with both the content and annotations. From the results shown in Figure 2 (b), the model achieved more than 98% F1-score on the test set across the three datasets, indicating high consistency. As we mentioned earlier, this is only a method to estimate consistency, assessing whether the model can identify stable patterns within the dataset, rather than a direct reflection of its performance on downstream tasks. The high performance (over 92.5% F1-score) on the expert-provided example set further confirmed the consistency of our dataset.

### 5.1.2 Human evaluation

The human evaluation results for C-IMHI dataset are shown in the Figure 2 (c). The high consistency score (mean value >2.5 out of 3) supports the consistency findings obtained from the automatic evaluation. Additionally, the generated explanations are reliable, as reflected by a high average score of 2.3 out of 3. However, they lack some degree of professionality, as indicated by the slightly lower average score of 2.03. The overall final score is positive, with a mean of 2.28, indicating that the overall quality of the data has been successfully verified.

## 5.2 Ablation experiment results

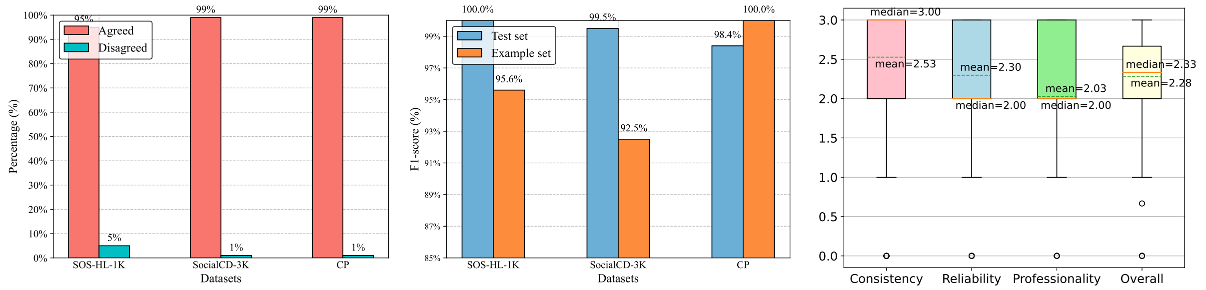
Ablation experiment results shown in Table 2 highlight the impact of each training stage. MentalGLM-chat-S1, trained only on translated IMHI, showed a significant performance gap, especially in cognitive distortion classification with SocialCD-3K, underscoring the limitations of direct translation for fine-tuning. MentalGLM-chat-S2, fine-tuned on C-IMHI, outperformed S1, benefiting from task-specific alignment. We further conducted an additional ablation experiment: MentalGLM-chat-M1, which applies a single-step fine-tuning on the merged IMHI and C-IMHI datasets. As shown in Table 2, M1 yields better performance than S1, but is consistently inferior to S2. This indicates that simply combining datasets may not be sufficient for effective instruction generalization, possibly due to the heterogeneous complexity and abstraction levels in IMHI and C-IMHI that are harder to optimize in a flat training schedule. In contrast, our two-step fine-tuning strategy (MentalGLM-chat) achieves the best overall performance across all benchmarks. This progressive approach allows the model to first acquire general reasoning skills from IMHI before specializing on domain-specific instructions in C-IMHI, leading to stronger instruction-following ability. These findings support the hypothesis that curriculum-style training can enhance model generalization (Huang and Xiong, 2024; Pang et al., 2024).

Table 2: Ablation experiment results. All results are F1-scores (%). MentalGLM-chat-S1, -S2, and -M1 denote fine-tuning on IMHI (Stage 1), C-IMHI (Stage 2), and merged IMHI + C-IMHI, respectively. "$CP_{Parent}$" and "$CP_{Child}$" denote parent and child-level performance for CP.

| Model | SOS-HL-1K | SocialCD-3K | $CP_{Parent}$ | $CP_{Child}$ |
|---|---|---|---|---|
| MentalGLM-chat-S1 | 66.28 | 15.70 | 48.95 | 21.58 |
| MentalGLM-chat-S2 | 81.58 | 70.69 | 77.69 | 47.69 |
| MentalGLM-chat-M1 | 76.63 | 69.16 | 73.83 | 42.27 |
| MentalGLM-chat | 85.12 | 71.04 | 80.55 | 47.85 |

## 5.3 Downstream task results

Experimental results for downstream tasks (Table 3) show that our models achieved the best or comparable performance across all tasks. MentalGLM-chat outperformed Chinese MentalBERT on SOS-HL-1K, with an F1-score 3.82% higher. On SocialCD-3K, it performed similarly, only 1.85% lower. For cognitive pathway extraction, parent-level classification reached 80.55% F1, surpassing the SOTA supervised model, while child-level classification remained challenging. Our model outperformed others, including Chi-




(a) Correctness (b) Consistency (c) Human evaluation

Figure 2: Evaluation results on GPT-4's generated explanations used for C-IMHI data construction. (a) and (b) represent automated evaluation for the entire dataset, while (c) shows human evaluation on a subset of 100 samples.

Table 3: Results on three downstream tasks. Precision (P), recall (R), and F1-score (F1) are reported as micro averages (%), except for SOS-HL-1K (binary averages). "$CP_{Parent}$" and "$CP_{Child}$" denote parent and child-level performance for CP. "ZS" and "FS" indicate zero-shot and few-shot prompts.

| Method | Param | SOS-HL-1K | | | SocialCD-3K | | | $CP_{Parent}$ | | | $CP_{Child}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | F1 | P | R | F1 | P | R | F1 | P | R | F1 | P | R |
| Pre-trained deep learning models | | | | | | | | | | | | | |
| BERT | 110M | 77.91 | 77.60 | 78.22 | 71.92 | 79.19 | 65.87 | 77.01 | 76.68 | 78.13 | 44.66 | 51.84 | 41.85 |
| Chinese-BERT-wwm | 110M | 79.32 | 83.18 | 75.80 | 73.06 | **84.22** | 64.50 | 77.96 | 80.36 | 75.71 | 42.98 | **71.26** | 30.77 |
| RoBERTa-wwm | 110M | 79.83 | 81.51 | 78.22 | 73.91 | 82.61 | 66.87 | 76.81 | 78.28 | 75.39 | 43.61 | 54.81 | 36.21 |
| Chinese MentalBERT | 110M | 81.30 | 81.96 | 80.64 | **74.91** | 83.03 | 68.24 | 79.22 | 79.68 | 78.76 | 47.58 | 50.70 | 44.82 |
| Generalized LLMs (Zero-shot/few-shot prompt) | | | | | | | | | | | | | |
| GLM-4-plus_ZS | - | 67.23 | 51.50 | **96.77** | 34.24 | 29.09 | 41.59 | 49.93 | 46.01 | 54.57 | 24.54 | 18.84 | 35.18 |
| GLM-4-plus_FS | - | 68.66 | 54.50 | 92.74 | 41.00 | 32.62 | 55.17 | 43.85 | 44.06 | 43.64 | 27.37 | 21.44 | 37.85 |
| ChatGPT_ZS | 175B | 65.42 | 52.00 | 88.23 | 12.06 | 10.63 | 13.95 | 32.08 | 28.61 | 36.50 | 13.42 | 12.19 | 14.92 |
| ChatGPT_FS | 175B | 68.71 | 59.37 | 81.61 | 18.10 | 16.76 | 19.68 | 54.87 | 51.87 | 58.25 | 27.00 | 24.80 | 29.64 |
| GPT-4_ZS | - | 71.72 | 57.43 | 95.48 | 26.61 | 17.13 | 59.65 | 35.93 | 31.69 | 41.47 | 17.22 | 15.44 | 19.45 |
| GPT-4_FS | - | 75.81 | 70.16 | 82.58 | 40.39 | 31.38 | 56.66 | 59.09 | 55.49 | 63.19 | 28.61 | 23.23 | 37.23 |
| Task fine-tuned LLMs | | | | | | | | | | | | | |
| Baichuan2-chat | 7B | 75.68 | 85.71 | 67.74 | 71.68 | 76.89 | 67.12 | 73.96 | 69.94 | 78.86 | 45.65 | 43.36 | 48.21 |
| Qwen2-Instruct | 7B | 78.93 | 75.18 | 83.06 | 72.01 | 76.22 | 68.24 | 76.12 | 71.32 | **81.60** | 46.24 | 43.66 | 49.13 |
| Qwen3 | 8B | 78.60 | 80.31 | 76.99 | 72.85 | 75.43 | **70.45** | 77.45 | 77.78 | 77.19 | 47.03 | 47.92 | 46.22 |
| Llama-3-Chinese | 8B | 77.97 | 82.14 | 74.19 | 70.97 | 74.52 | 67.75 | 79.11 | 79.37 | 78.86 | 48.23 | 48.99 | 47.49 |
| Llama-3-Chinese-instruct | 8B | 79.32 | 83.19 | 75.81 | 70.90 | 73.05 | 68.87 | 78.96 | 78.42 | 79.50 | 49.32 | 50.16 | 48.51 |
| Our method | | | | | | | | | | | | | |
| MentalGLM | 9B | 79.20 | 78.57 | 79.84 | 73.06 | 76.71 | 69.74 | 79.41 | 79.75 | 79.07 | **50.91** | 51.69 | **50.15** |
| MentalGLM-chat | 9B | **85.12** | **87.29** | 83.06 | 71.04 | 74.22 | 68.12 | **80.55** | **80.76** | 80.34 | 47.85 | 48.42 | 47.28 |

nese MentalBERT, by 3.33%. Generalized LLMs, including GPT-4, performed poorly, with a best score of 28.61%. Overall, our models matched or exceeded supervised models while supporting multi-task training without separate models. They also outperformed fine-tuned open-source LLMs on all tasks, offering both flexibility and decision explainability—an essential advantage for specific applications.

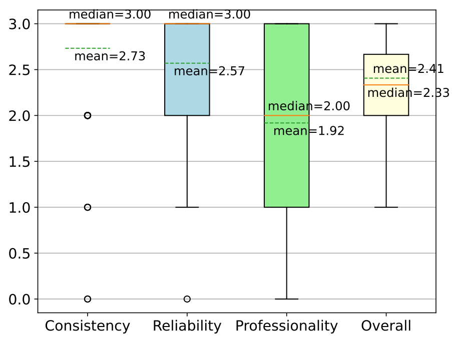



Figure 3: Expert evaluation results on 100 randomly selected prediction explanations generated by MentalGLM-chat.

## 5.4 Evaluation results of explanations generated by MentalGLM-chat

Our model delivers both high accuracy and decision-making explanations, essential for real-world applications. Figure 3 shows expert evaluation of 100 randomly selected MentalGLM-chat explanations. The median values for both consistency and reliability in our expert evaluations were 3, with mean values exceeding 2.5. Compared to the expert evaluations of GPT-4-generated explanations (as shown in Figure 2 (c)), MentalGLM outperforms GPT-4 in these two dimensions. However, our model demonstrated slightly lower professionality compared to GPT-4's explanations, which can be attributed to the significantly larger scale and capacity of GPT-4. Overall, our model produced better explanations than GPT-4, as indicated by a higher mean value for overall performance.

### 5.5 Clinical data evaluation results

The experimental results of LLMs applied on clinical dataset can be seen in Table 4. Compared to open-source LLMs without fine-tuning, MentalGLM-chat achieved the highest performance, surpassing the best non-fine-tuned model by 23.91% (parent level) and 30.39% (child level) in F1-score. Compared to the task-fine-tuned Llama-3-Chinese-instruct, it maintained an advantage of 4.94% (parent) and 2.95% (child). These results highlight its strong generalization and potential for clinical detection tasks.

Table 4: Results for clinical cognitive pathway extraction. Micro F1-scores are reported. "Parent" and "Child" indicate performance at respective levels.

| Model | Fine-tuned | Param | Parent | Child |
|---|---|---|---|---|
| Baichuan2-chat | N | 7B | 41.33 | 12.06 |
| Qwen2-Instruct | N | 7B | 44.33 | 11.54 |
| GLM-4-chat | N | 9B | 45.28 | 14.97 |
| Llama-3-Chinese-instruct | Y | 8B | 64.25 | 42.41 |
| MentalGLM | Y | 9B | 69.07 | 44.33 |
| MentalGLM-chat | Y | 9B | 69.19 | 45.36 |

## 6 Conclusion

We introduced C-IMHI, the first interpretable mental health analysis dataset for Chinese social media, and MentalGLM, the first open-source LLMs for explainable mental health analysis in this domain. Experiments showed MentalGLM outperformed or matched SOTA deep learning models and LLMs while generating high-quality explanations. Clinical dataset validation confirmed its generalizability, suggesting clinical potential. All datasets and models are publicly available for future research.

## Limitations

Although the results are promising, there are several limitations in our work. First, the 'Professionality' of the explanations generated by the proposed MentalGLM is slightly lower than that of GPT-4, with MentalGLM achieving a mean score of 1.92 compared to GPT-4's mean score of 2.03. This difference may be attributed to the significantly smaller parameter size of MentalGLM's underlying architecture compared to GPT-4. However, the results indicate that the gap between the two models is not substantial, demonstrating the effectiveness of our approach. In future work, we plan to further expand the training dataset to enhance MentalGLM's performance in the psychological domain. Second, in this study, we validated the MentalGLM model, which was trained on social media data, using clinical data without any fine-tuning. Due to the lack of sufficient clinical data, we did not conduct fine-tuning experiments. Although our model outperformed all baseline models, there is still room for improvement. In the future, we aim to collect more clinical data and fine-tune the model to enhance its adaptability to clinical applications with minimal additional cost.

## Ethical considerations

The original datasets used to construct the C-IMHI dataset were sourced from public social media platforms. We adhere strictly to privacy protocols and ethical principles to protect user privacy. To minimize the risk of personal information leakage, we anonymize and de-identify the data extensively during processing and analysis. We ensure that the research findings do not include any information that can directly or indirectly identify an individual. For the clinical data, informed consent was obtained from both hospitals and patients, and all data were anonymized to safeguard patient privacy.

Although MentalGLM has shown promising results in both social media and clinical mental health tasks, it is important to acknowledge that LLMs may introduce potential biases, including those related to gender, age, or sexual orientation, which could lead to incorrect judgments and inappropriate interpretations. We emphasize that the use of experimental results and data is strictly confined to research and analysis purposes, and any misuse or improper handling of the information is explicitly prohibited.

## Acknowledgements

This study was supported by grant from the National Natural Science Foundation of China (No. 72174152 & No. 72474166).

## References

Aaron T Beck. 1970. Cognitive therapy: Nature and relation to behavior therapy. *Behavior therapy*, 1(2):184–200.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

David D Burns. 1981. *Feeling good*. Signet Book.

Lei Cao, Huijun Zhang, Ling Feng, Zihan Wei, Xin Wang, Ningyun Li, and Xiaohao He. 2019. Latent suicide risk detection on microblog via suicide-oriented word embeddings and layered attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1718–1728, Hong Kong, China. Association for Computational Linguistics.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

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.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Dorottya Demszky, Diyi Yang, David S Yeager, Christopher J Bryan, Margarett Clapper, Susannah Chandhok, Johannes C Eichstaedt, Cameron Hecht, Jeremy Jamieson, Meghann Johnson, et al. 2023. Using large language models in psychology. *Nature Reviews Psychology*, 2(11):688–701.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Muskan Garg, Amirmohammad Shahbandegan, Amrit Chadha, and Vijay Mago. 2023. An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11960–11969, Toronto, Canada. Association for Computational Linguistics.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. ChatGLM: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Tianyu He, Guanghui Fu, Yijing Yu, Fan Wang, Jianqiang Li, Qing Zhao, Changwei Song, Hongzhi Qi, Dan Luo, Huijing Zou, et al. 2023. Towards a psychological generalist ai: A survey of current applications of large language models and future prospects. *arXiv preprint arXiv:2312.04578*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Yueqin Huang, YU Wang, Hong Wang, Zhaorui Liu, Xin Yu, Jie Yan, Yaqin Yu, Changgui Kou, Xiufeng Xu, Jin Lu, et al. 2019a. Prevalence of mental disorders in china: a cross-sectional epidemiological study. *The Lancet Psychiatry*, 6(3):211–224.

Yufei Huang and Deyi Xiong. 2024. It2acl learning easy-to-hard instructions via 2-phase automated curriculum learning for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9405–9421.

Zhisheng Huang, Q Hu, J Gu, J Yang, Y Feng, and G Wang. 2019b. Web-based intelligent agents for suicide monitoring and early warning. *China Digital Medicine*, 14(3):3–6.

Hyolim Jeon, Dongje Yoo, Daeun Lee, Sejung Son, Seungbae Kim, and Jinyoung Han. 2024. A dual-prompting for interpretable mental health language models. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 247–255.

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

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

Meng Jiang, Yi Jing Yu, Qing Zhao, Jianqiang Li, Changwei Song, Hongzhi Qi, Wei Zhai, Dan Luo, Xiaoqin Wang, Guanghui Fu, et al. 2024. AI-enhanced cognitive behavioral therapy: Deep learning and large language models for extracting cognitive pathways from social media texts. *arXiv preprint arXiv:2404.11449*.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

World Health Organization, World Health Organization, et al. 2023. Depressive disorder (depression). 2023. *Retrieved from Depressive disorder (depression)(who. int)*.

Wei Pang, Chuan Zhou, Xiao-Hua Zhou, and Xiaojie Wang. 2024. Phased instruction fine-tuning for large language models. *arXiv preprint arXiv:2406.04371*.

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

Hongzhi Qi, Qing Zhao, Jianqiang Li, Changwei Song, Wei Zhai, Luo Dan, Shuo Liu, Yi Jing Yu, Fan Wang, Huijing Zou, et al. 2023. Supervised learning and large language model benchmarks on mental health datasets: Cognitive distortions and suicidal risks in chinese social media.

MSVPJ Sathvik and Muskan Garg. 2023. Multiwd: Multiple wellness dimensions in social media posts. *Authorea Preprints*.

Yi-han Sheu. 2020. Illuminating the black box: interpreting deep neural network models for psychiatric research. *Frontiers in Psychiatry*, 11:551299.

Qwen Team. 2025. Qwen3 technical report.

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

Ahmet Yavuz Uluslu, Andrianos Michail, and Simon Clematide. 2024. Utilizing large language models to identify evidence of suicidality risk through analysis of emotionally charged posts. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024a. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

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

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024b. MentaLLaMA: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 4489–4500, New York, NY, USA. Association for Computing Machinery.

Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2024c. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19368–19376.

Hua Yu, Mingli Li, Zhixiong Li, Weiyi Xiang, Yiwen Yuan, Yaya Liu, Zhe Li, and Zhenzhen Xiong. 2020. Coping style, social support and psychological distress in the general chinese population in the early stages of the COVID-19 epidemic. *BMC psychiatry*, 20:1–11.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: a tale of diversity and bias. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Wei Zhai, Hongzhi Qi, Qing Zhao, Jianqiang Li, Ziqi Wang, Han Wang, Bing Yang, and Guanghui Fu. 2024. Chinese MentalBERT: Domain-adaptive pre-training on social media for Chinese mental health text analysis. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10574–10585, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55006–55021. Curran Associates, Inc.

Ayah Zirikly and Mark Dredze. 2022. Explaining models of mental health via clinically grounded auxiliary tasks. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 30–39.

# Appendix A Data distribution of the IMHI dataset

IMHI is the first multi-task, multi-source instruction dataset focused on interpretable mental health analysis, containing 105K data samples. It is designed to support instruction fine-tuning and evaluation for LLMs. The dataset is constructed from 10 publicly available mental health analysis datasets and covers 8 core tasks, including depression detection, stress detection, mental disorders detection, and more. The data sources include social media platforms like Reddit, Twitter, and SMS. Please note that some of the data has not been made publicly available yet. The data distribution can be seen in Table S1.

# Appendix B Psychology theory-driven expert sample design

## Appendix B.1 Suicide Risk Detection Task

Based on the **suicide risk level theoretical framework** (Huang et al., 2019b), we employ a dual-path analytical approach:

- **Suicidal Intent Clues**: Analyze emotional expressions in text such as *existential pain, hopelessness, and suicidal ideation* to identify latent suicidal motivations (e.g., metaphorical statements like "Life is meaningless")
- **Behavioral Plan Clues**: Detect concrete behavioral descriptions including *suicide methods, timeline planning, and preparation details* (e.g., action signals like "Saved enough sleeping pills")

Through synergistic analysis of dual clues, we systematically differentiate risk levels:

- High-risk individuals exhibit explicit plans (e.g., "Will jump from building on my birthday next week")
- Low-risk individuals primarily display emotional distress (e.g., "Wish to disappear")

## Appendix B.2 Cognitive distortion detection and cognitive pathway extraction tasks

Following the **Cognitive Behavioral Therapy (CBT)** framework (Beck, 1970), we establish standardized analytical pathways:

Table S1: Data distribution of the IMHI dataset.

| Data | Task | Instruction (train/val) | Source | Annotation | Type |
|---|---|---|---|---|---|
| DR | depression detection | 1654/184 | Reddit | weak supervision | binary |
| Dreaddit | stress detection | 3195/356 | Reddit | human annotation | binary |
| Loneliness | loneliness detection | 477/54 | Reddit | human annotation | binary |
| SWMH | mental disorders detection | 9793/1089 | Reddit | weak supervision | multi-class |
| T-SID | mental disorders detection | 863/96 | Twitter | weak supervision | multi-class |
| SAD | stress cause detection | 6162/685 | SMS | human annotation | multi-class |
| CAMS | depression/suicide cause detection | 562/63 | Reddit | human annotation | multi-class |
| MultiWD | wellness dimensions detection | 17716/1969 | Reddit | human annotation | multi-label |
| IRF | interpersonal risk factors detection | 6336/705 | Reddit | human annotation | multi-label |
| In total | Mental health analysis | 46758/5201 | Social Media | - | - |

- **Emotional Clues**: Identify cognitive distortion-related expressions such as *catastrophizing* and *overgeneralization* (e.g., "This failure proves I'm worthless")

- **Behavioral Clues**: Analyze compensatory strategies including *avoidance behaviors* and *safety-seeking behaviors* (e.g., "Avoid parties for fear of ridicule")

The dual-clue association mechanism reveals the psychological progression: **automatic thoughts → cognitive distortions → maladaptive behaviors**

Based on the aforementioned authoritative psychological frameworks (suicide risk theory/CBT), expert examples are constructed to ensure the scientific and standardized nature of the analytical pathways.

## Appendix C Consistency evaluation of the C-IMHI dataset

We proposed a method to check the consistency of the generated explanations by verifying whether each explanation supports its respective label. To achieve this, we trained deep learning models using explanation-label pairs to predict the corresponding labels from the generated explanations. The model used for this task was Chinese MentalBERT. The performance on the test set reflects the consistency of the generated explanations. The model that performed best on the validation set was selected and evaluated on both the test set and the expert-provided example set. We conducted this evaluation on three downstream task datasets: SOS-HL-1K, SocialCD-3K, and CP, maintaining the dataset distribution as shown in Table 1.

For training, the models were fine-tuned for 10 epochs on the training set for all tasks. We used a batch size of 16, the Adam optimizer, a learning rate of 2e-5, and cross-entropy as the loss function. Training was conducted on an NVIDIA GeForce RTX 4090 24GB GPU. The best-performing model on the validation set was then used to evaluate the test set and the expert-provided example set, with F1-scores used as the evaluation metric.

We can see the detailed model performance in Table S2. The high performance (>98% F1-score) across all the test sets demonstrates the strong consistency of the generated explanations. Additionally, the high performance (>92% F1-score) on the expert-provided example set highlights the consistency of the model's explanations with human-provided explanations.

Table S2: Performance of the Chinese MentalBERT classifier on the test and expert-provided example sets, evaluated using precision (P), recall (R), and F1-score (F1). Metrics are reported as micro averages, except for binary averages on SOS-HL-1K.

| SOS-HL-1K | | | SocialCD-3K | | | CP | | |
|---|---|---|---|---|---|---|---|---|
| **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** |
| Test set | | | | | | | | |
| 100.00 | 100.00 | 100.00 | 99.56 | 99.25 | 99.87 | 98.47 | 99.81 | 97.16 |
| Expert-provided example set | | | | | | | | |
| 95.65 | 91.66 | 100.00 | 92.53 | 93.93 | 91.17 | 100.00 | 100.00 | 100.00 |

## Appendix D Expert quality check and evaluation guideline

The experts evaluated the quality of LLM-generated explanations from three perspectives, based on the research by Yang et al. (2024b). To enhance the evaluation, we refined the "Professionality" metric by incorporating principles from the Cognitive Behavioral Therapy (CBT) framework. Further details are provided below:

- **Consistency**: Consistency evaluates whether the explanation supports the classification result. Large language models may sometimes produce discrepancies between labels and explanations. Annotators should assess whether the generated explanation provides consistent supporting evidence for the classification and whether it is well-structured.

  - 0: Inconsistent with the classification result.
  - 1: Consistent with the classification result but poorly readable and contains many errors.
  - 2: Consistent with the classification result. Most content is coherent and easy to read, with only minor errors.
  - 3: Consistent with the classification result. Fully fluent, coherent, and error-free.

- **Reliability**: Reliability measures the credibility of the explanation in supporting the classification result. Annotators should evaluate whether the explanation is factually accurate, free of misinformation, and based on sound reasoning.

  - 0: Completely unreliable information, containing factual hallucinations (e.g., nonexistent symptoms).
  - 1: Partially reliable information, but fact-based reasoning is flawed.
  - 2: Mostly reliable information, with non-critical misinformation or reasoning errors.
  - 3: Completely reliable information.

- **Professionality**: Professionality assesses the rationality of the explanation in supporting the classification result from a psychological perspective. Annotators should base on the framework of CBT to evaluate whether the explanation includes both emotional and behavioral cues. These methods are widely used in psychological counseling and validated in clinical practice. The explanation should adhere to standardized analysis paths, align with psychological theories, and demonstrate high scientific and professional standards.

  - 0: The explanation does not reference any psychological theory and lacks or misrepresents analysis of emotions and behaviors.
  - 1: The explanation attempts to reference psychological theory but applies it inaccurately, with simplistic or partially incorrect analysis of emotions and behaviors.
  - 2: The explanation largely aligns with psychological theory. The analysis of emotions and behaviors is fairly accurate, though there is room for improvement. Professional terminology is mostly appropriate.
  - 3: The explanation fully aligns with psychological theory, providing in-depth and accurate analysis of emotions and behaviors. Professional terminology is used flawlessly, demonstrating a high level of professionality and scientific rigor.

## Appendix E Error analysis of LLMs for Cognitive Distortion Classification Task

**Example**:

- **[Chinese]** 爱我的都会离开我，连养条狗狗都不喜欢和我这个郁郁寡欢动不动就掉眼泪的人呆一起，嫌累。除了我那个我从不敢在他面前表露情绪的爸，因为他不能理解我不快乐得有了病。还除了那个因无法自控爱着我的妈妈，一个家庭主妇开始去了解抑郁症。我要死了，我爸得多想不通啊，我妈得多伤心呢？可是我真的好累. . .

- **[English]**: Everyone who loves me will leave me. Even raising a dog, it doesn't like being around me, this person who is often gloomy and cries easily. It finds me tiring. Except for my dad, whom I never dare to show my emotions in front of, because he can't understand that I became ill because I'm not happy. And except for my mom, who loves me uncontrollably and is starting to learn about depression as a housewife. I feel like I'm going to die. How much will my dad struggle to understand? How much will my mom be heartbroken? But I'm really so tired...

**Label**: 先知错误 (The fortune teller error)
**Model output**:

- **GPT response**

- **[Chinese]** 这个帖子反映了“以偏概全”的认知歪曲。作者通过自己当前的情绪困境，过度泛化地认为“爱我的都会离开我”，这种想法将一些局部的负面经验扩展到整个生活，忽视了可能存在的其他积极或中立的情况。例如，作者感到孤独，认为即使是宠物也不愿意陪伴自己，便推断出所有人和事物都会远离自己，这样的思维方式导致了对生活的极端解读，容易陷入消极情绪和绝望感中，无法看到事情的另一面或潜在的变化空间。
- **[English]**: This post reflects the cognitive distortion of **overgeneralization**. The author, based on their current emotional struggles, excessively generalizes by believing that "everyone who loves me will eventually leave me." This thought pattern extends specific negative experiences to the entire life situation, ignoring the possibility of other positive or neutral circumstances. For example, the author feels lonely and believes that even pets are unwilling to accompany them, leading to the conclusion that everyone and everything will abandon them. This kind of thinking results in an extreme interpretation of life, making it easy to fall into negative emotions and despair, while being unable to see the other side of things or the potential for change.

- **GLM-4-plus response**
  - **[Chinese]** 这个帖子体现了以偏概全的认知歪曲。帖子中的叙述者根据自己的一些负面经历（例如感到被爱的事物和人离开，以及狗狗不喜欢和自己呆在一起）得出一个广泛的结论，即“爱我的都会离开我”。这种思维方式忽略了那些可能不会离开或者仍然爱着他们的人的情况，而是基于有限的证据做出了一个过于广泛的负面概括。
  - **[English]**: The cognitive distortion demonstrated in this post is **overgeneralization**. The narrator in the post draws a broad conclusion that "everyone who loves me will leave me," based on some negative experiences, such as feeling abandoned by loved ones and a dog not enjoying their company. This way of thinking disregards the situations of those who might not leave or still love them, making an excessively broad negative generalization based on limited evidence.

- **MentalGLM response**
  - **[Chinese]** 这篇帖子体现了先知错误的认知歪曲。理由：这段话中，患者对自己的未来充满了绝望和消极的预期，认为自己将会孤独终老，甚至想到了死亡。这种对未来极度消极的预测，认为自己未来的生活会非常糟糕，是典型的先知错误认知歪曲。患者基于过去的经历和当前的情绪状态，预测未来的事情会按照她现在的感觉发展，尽管这个预测并没有任何事实依据。
  - **[English]**: This post reflects the cognitive distortion of **the fortune teller error**. The reasoning is as follows: in this passage, the patient is filled with despair and negative expectations about their future, believing they will grow old and die alone, even contemplating death. This extreme pessimism about the future, the belief that their life will be terrible, is a typical example of the fortune teller error cognitive distortion. The patient is predicting that the future will unfold in the same way as their current feelings and past experiences, despite the fact that this prediction has no factual basis.

In Case, GPT-4 identified the distortion as "overgeneralization," correctly recognising the poster's negative experiences ("the dogs don't like to be with him") being extended to their entire life. GLM-4-plus also addressed this issue by categorizing it as "overgeneralization," highlighting how the poster's specific negative experiences led to a broad negative conclusion. However, MentalGLM offered a deeper analysis, categorising it as a "fortune teller error," explaining how the poster's beliefs ("all those who love me will leave me") predict a bleak future without factual basis. This demonstrates MentalGLM's ability to identify the distortion type while analysing its implications more thoroughly. Overall, unlike GPT-4 and GLM-4-plus, which offered reasonable but general analyses, MentalGLM proved more accurate and comprehensive, providing detailed insights into posters' emotions, behaviours, and thought patterns. Its emphasis on explainability further allows users to understand both the reasoning process and the decision, making it more effective for this task.