MentalQLM: A lightweight large language model for mental healthcare based on instruction tuning and dual LoRA modules

Jiayu Shi, Zexiao Wang, Jiandong Zhou, Chengyu Liu, Poly Z.H. Sun, Erying Zhao, and Lei Lu

***Abstract*— Mental disorders present significant challenges to healthcare systems and carry profound social consequences. The rapid development of large language model (LLM) has opened new avenues for enhancing mental healthcare. However, existing approaches primarily rely on instruction tuning and few-shot in-context learning with massive datasets and large-scale backbone models, resulting in significant computational costs. To address these challenges, we propose MentalQLM, a novel lightweight LLM by developing a new dual Low-rank Adaptation (LoRA) approach. The development of MentalQLM consists of two key stages. Firstly, datasets are pruned based on perplexity and diversity analysis to reduce computational requirements. The first LoRA module is instruction-tuned to adapt the LLM for downstream mental health classification tasks. Secondly, a dense layer augmented with a second LoRA module is fine-tuned to enhance performance on complex multi-class classification tasks. Extensive experiments validate the effectiveness of the proposed MentalQLM on five benchmark datasets. Despite having only 0.5B parameters, the model outperforms or demonstrates comparable performance to larger counterparts in both classification and reasoning tasks. This establishes MentalQLM as a promising solution for lightweight and efficient deployment in real-world mental healthcare applications. All the code will be released at https://github.com/tortorish/MentalQLM.**

***Index Terms*— Large language model; Mental health; Lightweight model; Low-rank Adaptation.**

I. Introduction

Mental disorders, including depression, anxiety, and stress, affect one in every eight people worldwide, posing significant challenges to healthcare systems and carrying substantial social consequences [1], [2]. Text-based dialogue conversations are frequently utilized in mental health interventions [3]. These tools facilitate the application of data-driven computational approaches, which enhance the efficiency of mental healthcare by analyzing various types of textual data, such as clinical notes, patient self-reports, and therapy transcripts. Machine learning techniques are pivotal in this context, which demonstrated promising performance in processing and analyzing textual data for detecting abnormal patterns, sentiments, and early warnings [4]–[6]. These capabilities enable more precise diagnoses and the development of personalized treatment plans to improve mental health outcomes.

Traditional machine learning models, such as support vector machine [4], random forest [5], and AdaBoost [6], have been used to classify different types of mental health conditions. However, these approaches often struggle to handle the complexity and nuanced nature of textual data, as they rely heavily on extensive feature engineering. The advent of deep neural networks has enabled more efficient processing of noisy and unstructured text data, resulting in more accurate diagnostic outcomes. These models include fully connected neural networks (FCNN) [7], convolution neural networks (CNN) [8], and recurrent neural networks (RNN) [9]. Nevertheless, deep learning methods typically require large amounts of labeled data, which presents a significant barrier to their widespread adoption in clinical settings.

Over the past decade, natural language processing (NLP) and computational social science techniques have been increasingly used to detect mental health issues through textual data [10]. These approaches leverage the vast amount of available textual data to identify linguistic patterns, sentiment changes, and behavioral anomalies associated with mental health conditions [11]. By utilizing advanced NLP techniques, researchers can analyze complex and unstructured text at scale, providing an efficient means of managing mental disorders and delivering care plans. In particular, pre-trained language models like BERT [12] have shown promising performance in stress detection and suicide risk assessment [13]. Moreover, fine-tuned models, such as MentalBERT and MentalRoBERTa [14], have further enhanced the precision and effectiveness in mental healthcare applications.

Driven by the availability of extensive corpus and computing resources, the size and capabilities of pre-trained models

This research was funded by the National Natural Science Foundation of China (#81903397). (Corresponding authors: Lei Lu and Erying Zhao, e-mail: lei.lu@kcl.ac.uk)

Jiayu Shi, Zexiao Wang are with School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China.

Jiandong Zhou is with Department of Family Medicine and Primary Care, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China.

Chengyu Liu is with State Key Laboratory of Digital Medical Engineering, School of Instrument Science and Engineering, Southeast University, Nanjing, China.

Poly Z.H. Sun is with the School of Psychology and Cognitive Science, East China Normal University, Shanghai, China

Erying Zhao is with Psychological Science and Health Management Center, Harbin Medical University, Harbin, China.

Lei Lu is with School of Life Course & Population Sciences, King's College London, UK.

have expanded significantly, particularly exemplified by large language models (LLM) such as ChatGPT [15], Llama [16], and Qwen [17]. These models have demonstrated promising performance in mental disorder diagnosis and predictive analytics. For example, ChatGPT has shown comparable performance in helping mental health professionals to assess conditions such as depression and schizophrenia [18], [19]. Meanwhile, these studies also highlighted the necessity of ongoing refinement to address inherent biases in LLM and enhance their clinical applicability [18], [19]. These findings underscore the importance of effective fine-tuning methods for LLM and the development of tailored prompt engineering techniques to improve their precision in mental healthcare applications.

Instruction fine-tuning and few-shot in-context learning are two effective approaches for adapting LLM to specific task applications. Instruction tuning is a technique used to fine-tune LLM using specific instruction-output pairs that are relevant to a target domain [20]. This process enables the model to better understand and generate responses tailored to specific tasks, such as mental health classification. For example, Mental-Alpaca and Mental-FLAN-T5 are full-parameter fine-tuned models on mental health classification datasets [21]. They outperformed GPT-3.5 by a considerable margin and achieved comparable performance to task-specific supervised learning models [21]. These results demonstrate the effectiveness of instruction fine-tuning in boosting LLM performance on downstream tasks. On the other hand, few-shot prompting enables LLM to perform tasks with minimal training data by including a small number of examples within the input prompt. This approach leverages in-context learning, where the model uses the context provided by the prompt to infer task requirements and generalize to unseen data effectively [22]. Few-shot prompting is particularly useful for scenarios where annotated data is scarce, allowing the model to be efficiently adapted to downstream tasks.

Nevertheless, existing LLMs are typically instruction-tuned on backbones with a large number of parameters [21], which require substantial computational resources during the training process. While few-shot prompting often results in lengthy input texts, which can significantly increase GPU memory consumption and processing time. The efficacy of this approach is also highly dependent on the model's size as it requires models of substantial scale to achieve satisfactory performance, making this approach less feasible for lightweight applications. Despite the growing interest in adapting LLM to mental healthcare, few studies have focused on developing lightweight and deployable models in this domain. For instance, the smallest model designed for mental health applications, mhGPT, still has 1.98 billion parameters [23], which poses challenges for efficient deployment. Therefore, it is critical to investigate how factors such as model size and quantization techniques influence the performance and usability of LLM in resource-constrained environments.

The development of lightweight LLM is crucial for making mental healthcare more accessible and efficient, particularly in resource-constrained settings. For example, there is an increasing mental health burden in low-income developing countries (LIDCs) [24]. These countries experience low economic development and urgently require resource-efficient applications capable of delivering mental health support more effectively. Lightweight LLM deployed on edge devices (e.g., mobile phones) can provide personalized and ubiquitous psychological support, addressing the global mental healthcare gap with scalable and cost-effective solutions [25]. These models are also valuable in scenarios where immediate access to healthcare information is critical, assisting in early diagnosis and intervention by identifying subtle linguistic patterns [26].

Furthermore, as illustrated in Figure 1, a collaborative framework could be developed to benefit both the lightweight edge LLM and large cloud LLM [27]. Edge LLM delpoyed on personal devices can deliver timely health suggestions and record therapeutic dialogues, which can be stored as electronic health records (EHR) and fed into cloud LLM for more comprehensive analysis [28]. Then, advanced techniques such as knowledge distillation can be used to transfer insights from cloud LLM to edge models, improving their performance while maintaining efficiency [29]. By advancing lightweight LLM to perform comparably to their larger counterparts, these models can bridge gaps in mental health services, offering real-time, personalized support and enabling the widespread adoption of mental healthcare technologies. However, a key shortcoming of lightweight LLM is that the reduced parameter scale leads to their degraded performance in care delivery [30]. Therefore, there is a pressing need to develop lightweight LLM that can operate efficiently on resource-constrained devices while still delivering accurate and timely diagnostic support.

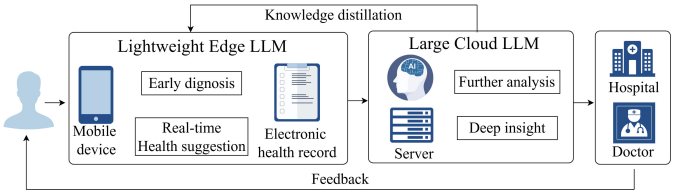


Fig. 1: Collaboration between Edge LLM and Cloud LLM.

In this paper, we proposed a novel framework for developing lightweight LLM tailored for mental healthcare applications. The model, named Mental-**Q**wen-**L**ightweight language **M**odel (MentalQLM), is built on the Qwen1.5-0.5B backbone [17]. To enhance the model's performance on mental health tasks, we introduce a dual Low-rank Adaptation (LoRA) strategy, which optimizes the lightweight LLM for improved efficiency and task-specific adaptability. The first LoRA module focuses on instruction tuning, enabling the model to adapt effectively to downstream mental health tasks by leveraging task-specific instruction-output pairs. The second LoRA module is designed to enhance the model's multi-class classification capabilities, making it suitable for handling complex mental health-related tasks. This dual strategy combines the benefits of instruction tuning and traditional supervised learning, offering a comprehensive approach to model optimization. In addition, knowledge from the large LLM is transferred to MentalQLM through instruction tuning, enabling the lightweight model to achieve performance

comparable to its large-scale counterparts. To the best of our knowledge, this is the first study that comprehensively explores the instruction-following capabilities of models with varying scales and quantization methods in the context of mental health applications. By addressing existing limitations, this work aims to advance the development of lightweight LLM, paving the way for more accessible and efficient mental healthcare delivery.

The main contributions of this paper are summarized as follows.

- A novel lightweight large language model MentalQLM is proposed for the diagnosis and reasoning of mental health conditions.
- A novel dual LoRA strategy is developed to optimize the lightweight MentalQLM for improving the efficiency and task-specific mental health applications.
- Extensive experiments are conducted on five benchmark mental health datasets to validate the effectiveness of the proposed lightweight MentalQLM.
- Compared to state-of-the-art LLMs (e.g., ChatGPT and MentalLlama), MentalQLM has only 0.5B parameters yet achieves superior or comparable performance in mental disorder diagnosis and reasoning tasks.

The remainder of this paper is organized as follows. Section II presents the development of our proposed MentalQLM, and Section III shows the experiments and results of using MentalQLM for mental disorder diagnosis and reasoning. Section IV discusses the strengths and limitations of this study. Finally, conclusions are drawn in Section V.

II. Methodology

This study proposes MentalQLM, a novel lightweight language model designed for mental healthcare applications. The development process includes two key stages, a parameter-efficient instruction-tuning stage and a multi-class classification fine-tuning stage. Figure 2 illustrates an overview of the methodology for developing the lightweight language model using dual LoRA modules. In the first stage, instruction-output pairs are selected based on quality and diversity assessments. These pairs are categorized into binary classification datasets and a multi-class classification dataset. The first LoRA module is then instruction-tuned using these datasets, optimized with the standard language modeling loss function. In the second stage, the multi-class fine-tuning process further refines the model. The prompts used in the instruction-tuning phase are removed, as they are primarily beneficial for training the initial LLM. Retaining prompts in the fine-tuning stage would also increase computational overhead unnecessarily. The second LoRA module is fine-tuned using the categorical cross-entropy loss function, enhancing the performance of MentalQLM on complex multi-class classification tasks. Details of the MentalQLM development process are presented in Figure 3 and the following sections.

A. Data Pruning by Perplexity and Diversity Analysis

Previous research has shown that a small amount of high-quality data is sufficient in the instruction-tuning stage to achieve performance comparable to models trained on larger datasets for downstream tasks [31]. Building on this insight, we propose a data selection method that prioritizes quality and diversity. This approach ensures that the selected instruction-output pairs effectively represent the target domain while significantly reducing computational demands without compromising accuracy [32].

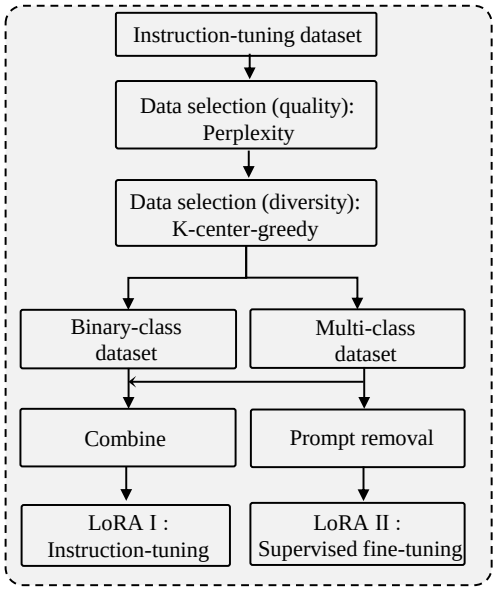


Fig. 2: Flowchart of our proposed methodology.

As illustrated in the upper part of Figure 3, data points in the original dataset are ranked by perplexity, where the perplexity is computed based on the untuned backbone model for each instruction-output pair. Outliers with high perplexity are eliminated to ensure quality of the dataset. Only data points within the perplexity range $[Q_1 - 1.5 \cdot IQR, Q_3 + 1.5 \cdot IQR]$ are kept, which can be formalized as

$$L_1 = \sum_{t=1}^{T} \log P(y_t|x, y_{<t}) \tag{1}$$

$$\text{Perplexity} = 2^{L_1} \tag{2}$$

$$IQR = Q_3 - Q_1 \tag{3}$$

where x denotes the instruction, y denotes the output sentence, P denotes the probability value for generating token y_t, given the instuction x and previous tokens $y_{<t}$, Q_1 and Q_3 are the first quartile and third quartile of all perplexity values respectively, IQR denotes the interquartile range.

Next, the KCenterGreedy approach is utilized to select a diverse subset of data for instruction fine-tuning. This algorithm ensures comprehensive coverage across the feature space while minimizing redundancy and maximizing information density. As outlined in Algorithm 1, the process begins by randomly selecting an initial center e from the dataset. Then, each new center is chosen to maximize the minimum distance $d(e, c)$ from any existing center, ensuring diverse representation throughout the dataset. This iterative process continues until k centers are established, effectively achieving a balance between coverage and compactness.

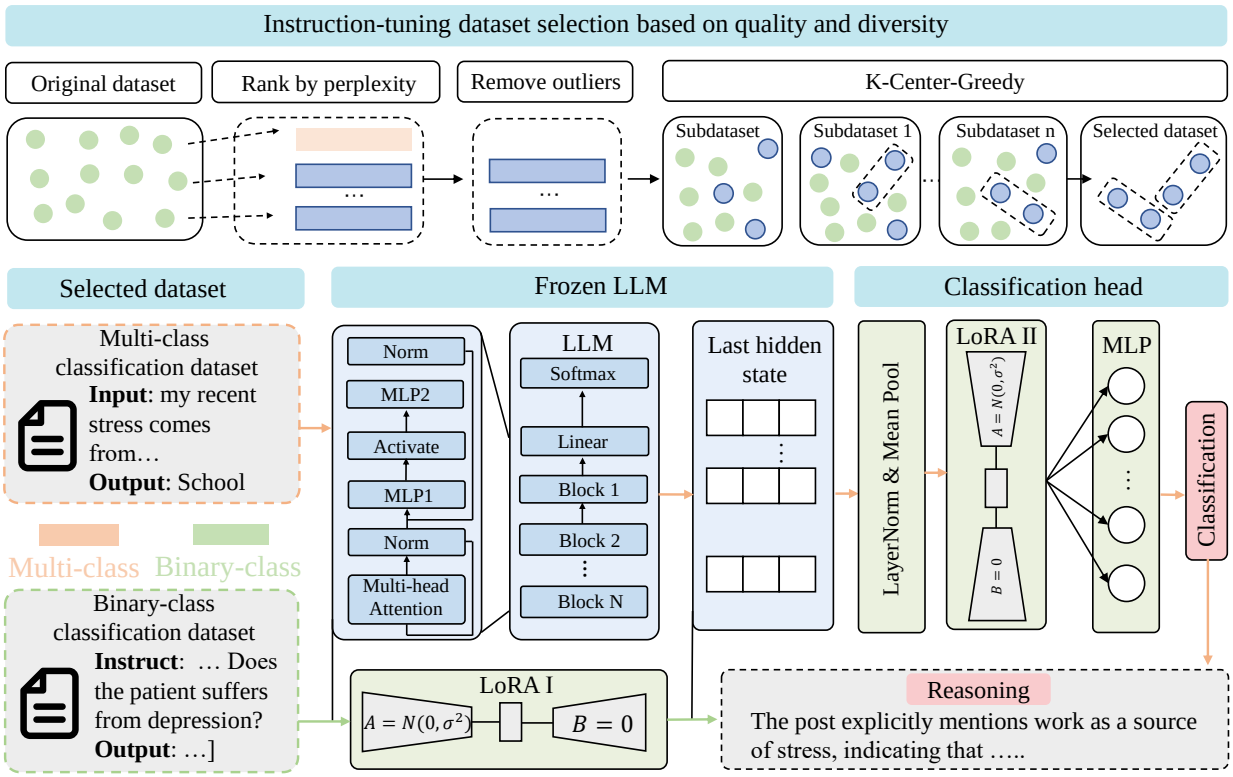


Fig. 3: Detailed pipeline of developing our proposed MentalQLM for mental health diagnosis and reasoning.

Algorithm 1 k-Center Greedy Algorithm.

INITIALIZE(E, k)
  $C \leftarrow \emptyset$ Initialize the set of centers.
  Select an arbitrary point $e \in E$ and add it to C.

while $|C| < k$ **do**
  $e^* \leftarrow \arg\max_{e \in E} \min_{c \in C} d(e, c)$.
  Add e^* to C
end while
return C

B. Parameter Efficient Instruction Tuning

We introduced a parameter-efficient fine-tuning (PEFT) approach to implement the instruction tuning of LLM, where only a small subset of parameters is updated. This approach significantly reduces computational costs, preserves the foundational capabilities of the LLM, and prevents catastrophic forgetting of previously acquired knowledge. Instruction tuning ensures that the model adapts effectively to new tasks while maintaining its original knowledge base. Specifically, MentalQLM is instruction-tuned using the LoRA method [33], which updates only a low-rank approximation of the weight matrices. This method can be formalized as follows [33],

$$\mathbf{W} = \mathbf{W_0} + \frac{\alpha}{r} \cdot (\mathbf{A} \cdot \mathbf{B}) \tag{4}$$

$$\mathbf{h} = \mathbf{W} \cdot \mathbf{X} + \mathbf{b} \tag{5}$$

where $\mathbf{W_0} \in \mathbb{R}^{m \times n}$ denotes the original matrix, $\mathbf{A} \in \mathbb{R}^{m \times r}$ and $\mathbf{B} \in \mathbb{R}^{r \times n}$ denotes two smaller matrices, $r \ll \min(m, n)$, α is a scaling coefficient that adjusts the contribution of the low-rank matrices $\mathbf{A}$ and $\mathbf{B}$ to the desired weights $\mathbf{W}$, $\mathbf{h}$ denotes the output hidden state, and $\mathbf{b}$ is the biased term.

In the instruction-tuning stage, the LoRA method is applied to all linear layers to optimize performance across a wide range of tasks. By leveraging the low-rank approximation, this approach effectively reduces computational costs while maintaining the flexibility and adaptability of the model. The training objective is to minimize the negative log-likelihood, which measures the discrepancy between the predicted token probabilities and the actual ground-truth tokens. For an instruction-output pair (x, y), the loss L_1 is defined as

$$L_1 = -\frac{1}{D} \sum_{i=1}^{D} \sum_{t=1}^{T} \log P(y_t | x, y_{<t}) \tag{6}$$

where D denotes the collection of instruction-output pairs, the sequence y is made up of T tokens, $P(y_t|x, y_{<t})$ is the probability of generating token y_t given the instruction x and previous tokens $y_{<t}$. Note that only loss of the output sentence is computed in the instruction-tuning stage.

C. Multi-class Classification Fine-tuning

To enhance the classification ability of MentalQLM on multi-class classification tasks, we proposed a novel fine-tuning method. Compared with traditional text-embedding models such as BERT [12], and RoBERTa [34], it is difficult to fine-tune LLM on downstream classification tasks. Also, the model is prone to overfitting on the training set due to the adjustment of a large number of parameters during fine-tuning, which also results in significant computational costs.

To address this issue, we introduce a second LoRA module, which is appended exclusively to the queries and values of the multi-head attention mechanism, as indicated in Eq. (7). This design ensures that the second LoRA module (LoRA II) is much smaller than the first LoRA module (LoRA I) to reduce the parameter count.

$$\begin{aligned}\mathbf{h}_i = Attention\big(&\mathbf{X}(\mathbf{W}_q^{(i)} + \Delta_q^{(i)}),\\ &\mathbf{X}\mathbf{W}_k^{(i)}, \mathbf{X}(\mathbf{W}_v^{(i)} + \Delta_v^{(i)})\big)\end{aligned} \tag{7}$$

where $\mathbf{X}$ denotes the input vector, $\mathbf{W}_q^{(i)}, \mathbf{W}_k^{(i)}$, and $\mathbf{W}_v^{(i)}$ denote the query, key, and value matrix of the ith block, $Attention(\cdot)$ is the operator of multi-head attention, $\Delta_q^{(i)}$, and $\Delta_v^{(i)}$ denotes the LoRA matrix of $\mathbf{W}_q^{(i)}$ and $\mathbf{W}_v^{(i)}$.

Next, the whole dataset is fed into LLM to get sentence embedding, and all parameters in LLM are frozen as illustrated in Figure 3. Compared with masked language bi-directional models, such as BERT [12], the [*CLS*] token is used to represent the sentence embedding for a single input sentence. By contrast, self-regression uni-directional models like Qwen [17] lack this token. Consequently, the last hidden state of the final decoder block is utilized as the sentence representation and mean pooling is applied to the sequential length dimension of the vector. Suppose an input sentence $\mathbf{S}$ is tokenized into $(\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_n})$, each token $\mathbf{w_i}$ corresponds to an embedding vector $\mathbf{e_i} \in R^d$, n is length of the sentence, and d denotes length of the embedding. Then, a one-dimensional vector $\mathbf{h} \in R^d$ can be obtained through mean pooling. To enhance computational efficiency and convergence speed, layer norm is utilized to process the sentence embeddings. Finally, a dense layer is appended to the model to regress the dimension to the number of classes.

The categorical cross-entropy loss is used in the fine-tuning stage, which can be formulized as:

$$L_2 = -\sum_{i=1}^{D}\sum_{i=1}^{C} y_i \log p_i \tag{8}$$

where C is the number of classes, y_i is 1 if the sample belongs to class i, and p_i is the predicted probability of class i, and D denotes the collection of instruction-output pair. This loss function makes the model output high probabilities for the correct class labels while minimizing the probabilities for incorrect categories. During the multi-class fine-tuning stage, parameters in the LoRA matrix $\Delta_q^{(i)}$ and $\Delta_v^{(i)}$ are updated with parameters in the dense layer using the categorical cross-entropy loss. This approach ensures effective adaptation to complex classification tasks while maintaining parameter efficiency.

Based on the classification results, the LLM generates a reasoning explanation for the predicted label. This approach combines the strengths of supervised learning for mental disorder classification with the powerful text generation capabilities of LLM, ensuring both precise classification and detailed reasoning. To summarize, for binary classification tasks, MentalQLM simultaneously generates classification labels and reasoning sentences; While for multi-class classification tasks, MentalQLM first predicts the classification label using the classification head. Subsequently, it generates a reasoning sentence based on the predicted label, providing an explanation tailored to the specific class.

III. Experiments and Results

A. Datasets and Experiment Setting

1) Dataset Description: We benchmark our proposed model on the Interpretable Mental Health Instruction (IMHI) dataset [40], which includes 105K data samples supporting multi-task and multi-source LLM instruction tuning and evaluation. In particular, five mental health instruction-tuning sub-datasets were used to validate our developed MentalQLM. All these data are converted to a question-answer pair to meet the requirement of instruction tuning, and they can be categorized as binary-class classification datasets and multi-class classification datasets.

- Binary classification datasets: Four datasets are utilized for binary classification tasks, DR [35], Dreaddit [36], Irf [37], and MultiWD [38]. These datasets focus on determining whether posts show signs of depression, stress, risk factors, or wellness. All datasets are imbalanced, except for Dreaddit, which has a balanced distribution.
- Multi-class classification datasets: One dataset, SAD [39], is used to support multi-classification. This dataset aims to identify the cause of stress in posts, with categories including school, financial problems, family issues, social relationships, work, health issues, emotional turmoil, everyday decision-making, and other stress causes. The dataset is highly imbalanced, with the last three classes having significantly fewer data points.

The detailed description of the five datasets is presented in Table I, which includes their division into training, validation, and testing sets. The table also presents the class proportions, and the average lengths of the instructions and outputs, which have a significant impact on the GPU memory consumption.

2) Implementation Details: The Qwen1.5-0.5B is chosen as the backbone model, which is trained on a single 4090 GPU with the Llama-factory framework [17], [41]. The maximum input length and maximum output length are set to be 512 on a bf-16 precision setting. The LoRA scaling factor α is set twice as the value of the LoRA rank r empirically [33], which leads to an actual amplification factor 2. Table II lists other configurations for the model development.

In terms of the training process in the instruction-tuning stage, all the models are trained for 15 epochs. The checkpoint of the last step is chosen to output results on the test set. The model is trained based on the AdamW optimizer [42], and we set a max learning rate of 5e-5 with the cosine decay learning rate regulator. While in the fine-tuning stage, the model is trained for 30 epochs and then model with the best classification accuracy on the validation set is evaluated on the test set.

3) Evaluation Metrics: Two indicators are used to evaluate the classification and reasoning capabilities of the proposed MentalQLM. To have a fair comparison with MentaLLaMA [40], we used the recommended weighted F1 score to evaluate all classifiers. The weighted F1-score is a metric that

TABLE I: Summary of five datasets for mental health instruction tuning.

Datasets	Task	Data Size (Train / Val / Test)	Class Proportion	Instruction Length	Output Length
DR [35]	Binary depression classification	1,003 / 430 / 405	79.9% / 20.1%	251	120
Dreaddit [36]	Binary stress classification	2,837 / 300 / 414	52.7% / 47.3%	116	95
Irf [37]	Binary risk factors classification	3,943 / 985 / 2,113	63.6% / 36.4%	164	135
MultiWD [38]	Binary wellness classification	15,744 / 1,500 / 2,441	36.8% / 63.2%	167	79
SAD [39]	Multi-class stress cause classification	5,547 / 616 / 684	24.7% / 18.6% / 12.3% 10.3% / 9.6% / 8.6% 6.4% / 6.3% / 3.1%	31	78

TABLE II: Network configurations for MentalQLM.

Stage #1 Instruction Tuning	
Parameters	Settings
LoRA I rank (r)	8
LoRA I scale factor (α)	16
Max input length	512
Max output length	512
Tunable parameters	3,784,704
Stage #2 Multi-class Fine-tuning	
LoRA II rank (r)	4
LoRA II scale factor (α)	8
Tunable parameters	393,216

combines precision (P) and recall (R) into a single balanced measure, calculated as their harmonic mean. For multi-class classification tasks, we compute the F1-score for each class and average them, weighted by the number of instances in each class. This metric effectively reflects the model performance across imbalanced classes, and it is defined as

$$F1_weight = \sum_{i=1}^{n} \left(\frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \right) \cdot \frac{S_i}{\sum_{j=1}^{n} S_j} \cdot 100 \quad (9)$$

where n is the number of classes, P_i and R_i represent the precision and recall of class i respectively, and S_i denotes number of instances of class i.

The BLEU (Bilingual Evaluation Understudy) score is a metric used to evaluate the quality of machine-generated. Specifically, $BLEU-4$ score focuses on comparing n-grams up to 4-grams between the candidate translation and the reference, which can be defined as

$$BLEU-4 = BP \cdot \exp\left(\sum_{n=1}^{4} \frac{w_n \cdot \mathit{clip}_n}{\mathit{total}_n} \right) \cdot 100 \quad (10)$$

where BP is the brevity penalty that penalizes candidates shorter than the reference, w_n are weights with $\sum_{n=1}^{4} w_n = 1$, typically $w_n = \frac{1}{4}$ for BLEU-4, clip_n is the clipped count of n-grams matching the reference(s), and total_n is the total count of n-grams in the candidate translation.

The BLEU-4 score effectively measures how close machine-generated text is to the reference answer, making it a valuable tool for assessing performance in natural language generation tasks. In the following sections, we report the weighted F1-score and BLEU-4 score, both expressed as percentages.

B. Comparison with State-of-the-Art LLM

Experiments are carried out to compare the performance of state-of-the-art (SOTA) models in terms of classification ability. They include in-context learning models and instruction-tuning models as illustrated in [40].

- In-context learning models: They include zero-shot and few-shot models, consisting of ChatGPT and GPT-4. Instructions are combined with format requirements to make the LLM output the classification label. To fully utilize the strong learning ability of large-scale LLM, few-shot examples are given for demonstration.
- Instruction-tuning models: The instruction-tuning models consist of MentalLlama-7B, MentalLlama-chat-7B and MentalLlama-chat-13B [40], which are instruction-tuned on the Llama backbone. Instruction fine-tuning can precisely calibrate models with domain knowledge, improving the classification ability and reasoning ability of LLM on downstream tasks.

As shown in Table III, the performance of our proposed MentalQLM is benchmarked against various LLM across five datasets. Despite its reduced model size, MentalQLM achieves better classification accuracy compared to its larger counterparts. Notably, it outperforms other models on the Dreaddit and MultiWD datasets by 4% and 2%, respectively. This demonstrates that smaller models can achieve robust classification capabilities when fine-tuned with instruction tuning on downstream tasks. In contrast, in-context learning models like ChatGPT and GPT-4 exhibit strong zero-shot performance on simpler tasks, such as stress and depression detection in the DR and Dreaddit datasets. However, their accuracy drops significantly on more challenging tasks, such as those in the Irf and MultiWD datasets. This highlights the importance of instruction tuning for specific downstream tasks. Moreover, instruction tuning is particularly straightforward for smaller models like MentalQLM due to their reduced model size, making it a practical and efficient approach for achieving high performance in task-specific applications.

On the multi-class classification dataset SAD, both in-context learning-based LLM and instruction-tuned Llama models perform poorly. The primary reason is the complexity of the classification task and the highly imbalanced distribution of data samples across different classes. In contrast, our proposed MentalQLM achieves better classification performance by using task-specific fine-tuning, which incorporate an additional classification head composed of a LoRA module and a dense layer. During the fine-tuning stage, the parameters of the base LLM are frozen to preserve its strong foundational language capabilities. Only the parameters of the classification head are updated during backpropagation, making the process efficient and focused. As shown in Table III, this fine-tuning

TABLE III: Classification performance of MentalQLM and comparison with SOTA LLMs.

Models	Parameters	Classification (Weighted F1-score)					Mean
		DR	Dreaddit	Irf	MultiWD	SAD	
$ChatGPT_{ZS}$	175B	82.41	71.79	41.33	62.72	54.05	57.47
$ChatGPT_{FS}$	175B	84.22	75.38	43.31	64.93	63.56	66.28
$GPT4_{FS}$	1.76T	82.0	78.18	51.75	62.58	55.68	66.04
MentalLlama-7B	7B	76.14	71.65	67.53	68.44	49.93	66.74
MentalLlama-chat-7B	7B	83.95	62.2	72.88	75.79	62.18	71.4
MentalLlama-chat-13B	13B	**85.68**	75.79	76.49	75.11	63.62	75.34
MentalQLM (Our Model)	**0.5B**	84.20	**82.85**	**76.81**	**77.88**	**67.06**	**77.76**

*The footnote 'ZS' indicates zero-shot learning, and 'FS' indicates few-shot learning.

TABLE IV: Instruction-following ability of models with different parameter scales.

Models	Classification (Weighted F1-score)					Reasoning (BLEU-4 score)				
	DR	Dreaddit	Irf	MultiWD	SAD	DR	Dreaddit	Irf	MultiWD	SAD
Qwen2-7B	77.88	81.64	75.63	76.73	64.33	62.58	60.32	61.98	65.81	65.30
Qwen1.5-7B	86.86	86.45	78.79	79.41	59.27	61.82	61.20	65.15	64.98	62.65
Qwen1.5-4B	86.17	84.06	76.01	78.82	62.57	62.05	60.48	61.41	65.35	63.95
Qwen1.5-1.8B	84.44	84.54	77.90	78.37	63.45	61.86	60.42	61.39	66.15	64.45
Qwen1.5-0.5B	82.38	80.91	76.76	78.75	58.31	60.81	58.70	61.04	64.69	63.37

approach significantly improves classification accuracy on the SAD dataset, enabling more precise diagnosis of mental health disorders.

C. Results on Model Performance Analysis

The section explores the relationship between the model size and instruction-following ability, and investigates text embeddings of LLM for various downstream tasks without task-specific pre-training.

1) Instruction-following Ability of Models with Different Scales: The performance of models with different parameter scales are presented in Table IV. It can be seen from the table that instruction tuning demonstrates effectiveness for models of various sizes. While large-scale models are generally recognized for their superior capabilities, smaller models can achieve comparable performance when instruction-tuned. As shown in Table IV, the classification and reasoning abilities of the lightweight Qwen1.5-0.5B model are on par with those of larger models. This justifies selecting Qwen1.5-0.5B as the backbone model for MentalQLM.

The results in Table IV also demonstrate that Qwen1.5-7B consistently outperforms its counterpart, Qwen2-7B, on most datasets. This suggests that instruction tuning outcomes have a limited correlation with adjacent model versions, highlighting the importance of task-specific evaluation. However, on the multi-class classification dataset SAD, the weighted F1 scores across all models are relatively low, particularly for the 0.5B parameter-scale model. This underscores the need to enhance the classification capabilities of lightweight LLM for complex multi-class classification tasks.

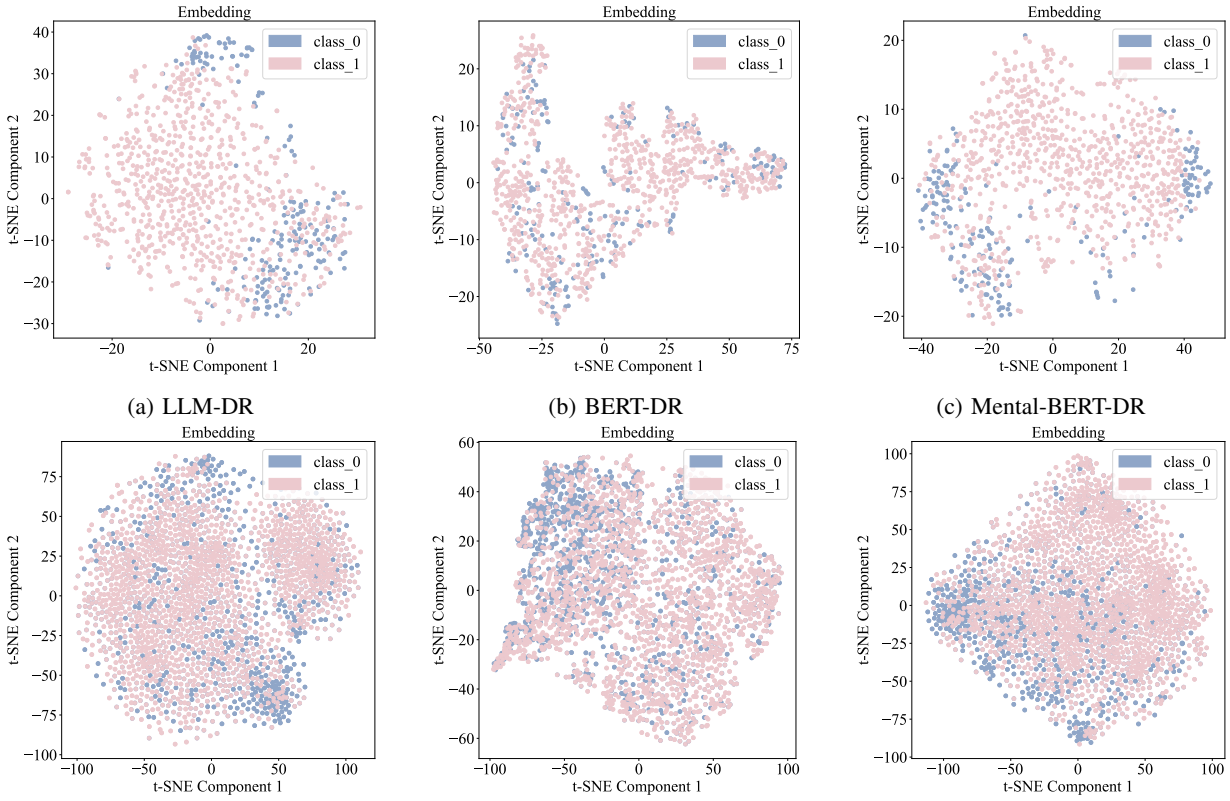
2) Text Embedding of LLM: Inspired by the text-embedding model BERT [12], which utilizes embeddings generated by the model for downstream classification tasks, we aim to leverage representations produced by large language models (LLM) to enhance their performance on multi-class classification tasks. As shown in Figure 4, the hidden states from LLM (Qwen1.5-0.5B), BERT, and the pre-trained Mental-BERT are visualized using the t-SNE dimensionality reduction algorithm. This visualization is performed on two binary classification datasets, DR and Irf. For the bi-directional text embedding models BERT and Mental-BERT, the embeddings of the [*CLS*] token are visualized, as this token encapsulates the information of the entire sentence. For the uni-directional self-regression model Qwen1.5-0.5B, the hidden states from the last decoder block are averaged to obtain a representative embedding.

The results demonstrate that the LLM develops a distinct boundary between positive and negative samples, even without fine-tuning on mental health datasets. Its performance is comparable to that of the pre-trained Mental-BERT model. In contrast, the standard BERT model fails to establish a clear boundary between classes, with points of different classes scattered randomly in the feature space. These findings suggest that LLM has robust inherent representation capabilities comparable to domain-specific pre-trained models, without requiring specialized pre-training. Consequently, fine-tuning a classification head on these embeddings is a feasible and effective strategy to improve the classification performance of lightweight models on complex multi-class classification tasks.

D. Results on Dataset Analysis

The data scales and data quality have an important influence on model performance. This section explores how these two factors would impact the model performance.

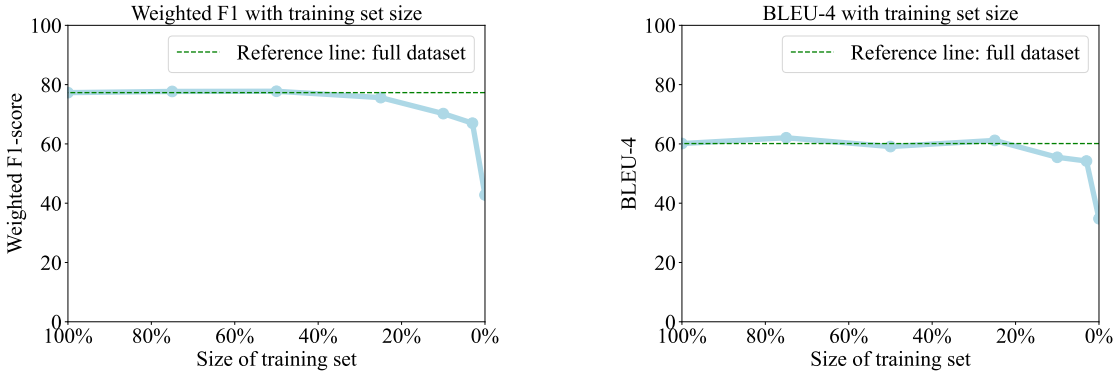
1) Data Scale Analysis: The relationship between dataset size and model performance is analyzed and illustrated in Figure 5. When the dataset size increases from 0% to 3%, there is a rapid rise in both the BLEU-4 score and the F1 score, reaching 54.23 and 69.40, respectively. This indicates that even a small subset of data can significantly enhance the model performance in generating text that closely aligns with the format of the instruction-tuning dataset. This stage is critical because the classification task is shaped into a text generation task for the LLM. By generating sentences in the required format, classification labels can be effectively extracted from the generated sentence.



(a) LLM-DR (b) BERT-DR (c) Mental-BERT-DR

(d) LLM-Irf (e) BERT-Irf (f) Mental-BERT-Irf

Fig. 4: Text embedding visualization of three different models.

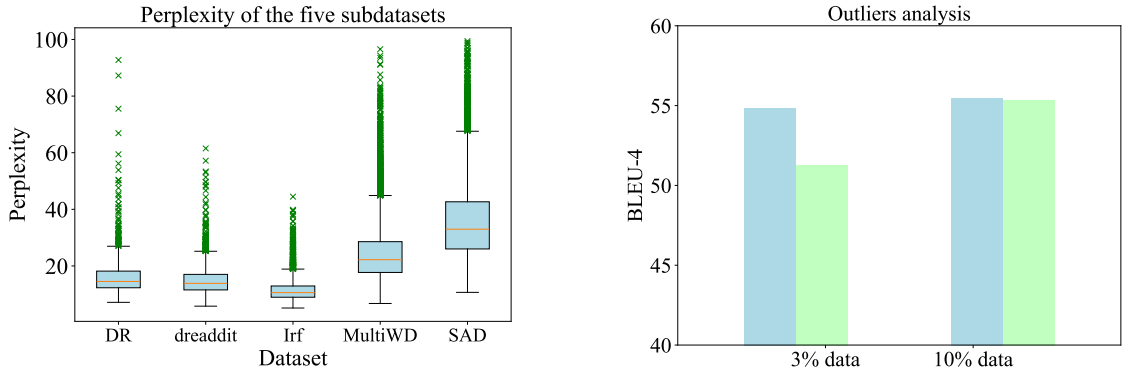


(a) Classification performance with dataset scales. (b) Reasoning performance with dataset scales.

Fig. 5: Classification and reasoning performance with different data scales.

As the dataset size increases from 3% to 25%, the scores continue to improve gradually, eventually reaching 61.18 and 77.30 respectively. This phase of incremental improvement suggests that while additional data further enhances the model performance, the rate of improvement begins to diminish. Beyond 25%, the scores remain stable, indicating minimal impact on performance with larger dataset sizes. These results suggest that 25% of the dataset is generally sufficient to achieve near-optimal performance. LLM demonstrate strong foundational capabilities, requiring only a small proportion of data to achieve satisfactory results. Data selection methods can further enhance performance, enabling the development of effective and lightweight LLM with relatively small datasets.

2) Data Quality Analysis: The influence of instruction-output pairs on model performance is analyzed. Figure 6 (a) illustrates the perplexity distribution of the instruction-output pairs across the five datasets, where outliers with extremely high perplexity are removed to ensure dataset quality. Figure 6 (b) shows results of an ablation study conducted on subsets comprising 3% and 10% of the entire dataset. When only 3% of the data is used, outliers have an adverse impact on model performance. Specifically, if the outliers are included during training, the BLEU-4 score drops to 51%. This finding underscores the critical importance of maintaining high quality data when training



(a) Perplexity distribution of the five datasets.

(b) Impacts of outliers on reasoning outputs.

Fig. 6: Perplexity distribution and data quality analysis.

resources are constrained or limited samples are available. In contrast, with 10% of the data, the negative effect of outliers is mitigated due to the introduction of additional instruction-output pairs. Table V provides examples of instruction-output pairs with high perplexity. It is noted that some inputs are fairly short, such as "friend OK No" and "Please help my friend". While other posts are ambiguous for mental disorder, making it challenging even for mental health experts to determine whether they indicate symptoms of disorders. These findings highlight that data quality, rather than sheer quantity, is a pivotal factor in ensuring model efficacy, particularly in data-restricted scenarios. Prioritizing high-quality data can lead to robust performance even with limited training resources.

TABLE V: Examples of high perplexity pairs.

Instruction-output pairs with high perplexity.
'instruct': 'Consider this post: "friendship ok No" Question: Does the poster suffer from depression?' **'output'**: 'No, the poster does not suffer from depression. Reasoning: The post lacks sufficient information to determine if the poster suffers from depression. It only consists of two words and does not provide any context or emotional indicators.'
'instruct': 'Consider this post: "Why did I open my mouth I should've just said "I'm fine" Ĩ don't need help Or maybe I do" Question: Does the poster suffer from stress?' **'output'**: 'Yes, the poster suffers from stress. Reasoning: The poster expresses regret and indecision about opening up and asking for help. This suggests internal conflict and possible stress. The usage of phrases like "should've just said 'I'm fine'" and "Or maybe I do" indicates a struggle with vulnerability and a need for support. Overall, the language used implies that the poster may be experiencing stress.'
'instruct':'Consider this post: "I feel like no matter how hard I try when I make one step forward I,am always two steps back. it,is so defeating and embarrassing Question: Does the post show risk of perceived burdensomeness?" **'output'**:'No, the post does not show this risk. Reasoning: 1. The post mentions feelings of defeat and embarrassment, but does not directly indicate thoughts of burden or being better off gone. These feelings suggest a sense of frustration and disappointment, but not necessarily perceived burdensomeness.The post does not mention any social or interpersonal difficulties that could contribute to feelings of perceived burdensomeness. It focuses more on personal setbacks and struggles. The response also does not indicate any perceived burdensomeness. It is simply acknowledging the individualś feelings of frustration and offering empathy.'

E. Analysis of Reasoning Ability and Quantization

This section investigated the limitation of our lightweight model MentalQLM. Then, the quantization methods are explored to further reduce computational quantity.

1) Reasoning Ability Analysis: Another notable advantage of LLM is their ability to generate reasoning sentences in addition to classification results. This feature is particularly valuable for providing insights into symptoms, assisting medical specialists in their assessments. However, traditional text evaluation methods that rely on lexical similarity often fall short in capturing the nuances and complexities of generated text.Meanwhile, it is difficult to evaluate the quality of reasoning sentences by solely comparing similarity between them and the label sentences.

Recent studies suggested that evaluations conducted by LLM themselves align closely with human assessments [18], [19]. For example, LLM like GPT-4 are increasingly recognized for their effectiveness in text evaluation, offering nuanced and reliable judgments [43]. Motivated by findings in these studies, we employ GPT-4o to rate the quality of reasoning sentences. The prompt template used for evaluation is detailed in the Appendix I. In particular, the input prompt to GPT-4o evaluates four criteria, helpfulness, relevance, accuracy, and detail level, with scores ranging from 1 to 10 for each sample [44]. These criteria ensure unbiased and comprehensive feedback.

For each of the five datasets, 200 instruction-output pairs were randomly selected for evaluation, resulting in a total of 1,000 pairs. As shown in Figure 7, we compared the sentences generated by MentalQLM-0.5B and MentalLlama-7B with the label sentences from the original dataset, which were generated by ChatGPT-175B. The results in Figure 7 showed that MentalQLM-0.5B achieved a mean score of 7.49, slightly lower than the 8.02 of MentalLlama-7B and the 8.16 of ChatGPT. The results indicate that while smaller models like MentalQLM-0.5B can achieve better classification accuracy after instruction tuning, larger models with more parameters tend to generate more coherent and meaningful reasoning sentences. Therefore, it is better to take advantage of lightweight LLM and large LLM in the future healthcare scenario: utilizing the lightweight one for preliminary and timely diagnosis, while

using the large LLM for providing deep medical insights. Examples of reasoning sentences generated by MentalQLM-0.5B and MentalLlama-7B, along with the GPT-4o scores and prompts, are provided in the Appendix II.

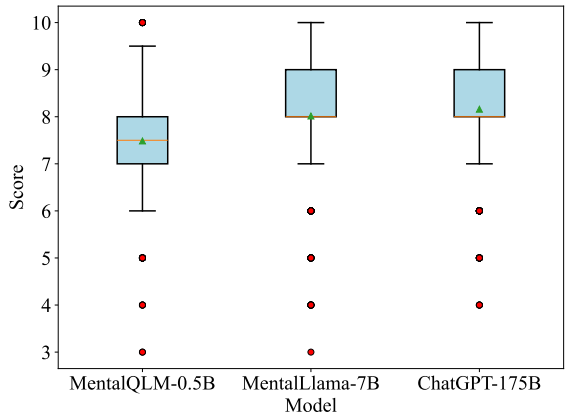


Fig. 7: Comparison of reasoning scores between three LLM.

2) Abnormal Characters: Hallucination is a well-known limitation of LLMs, where they may generate unsubstantiated or irrelevant content during reasoning [45]. As shown in Table VI, in the exploration of reasoning outputs of MentalQLM, we observed a small portion of non-English characters in the generated text. As illustrated in Figure 8, we identified 259 Chinese characters out of the total 613,953 words, accounting for the rate of 0.0004 of the total words, which would not affect understanding the contextual sentences. A similar phenomenon is observed in MentalLlama-7B, which has around 0.0005 non-English characters in its total reasoning outputs.

To further investigate this phenomenon, we compared the performance of LLMs of different scales in generating abnormal characters. As illustrated in Figure 8, when the backbone model was replaced with Qwen1.5-7B, the number of Chinese characters increase significantly. Large-scale model is prone to hallucination and will generate unrelated Chinese context which may interfere with accurate diagnosis. This highlights the advantage of our proposed lightweight MentalQLM-0.5B in generate accurate predictions.

These findings suggest that instruction-tuning LLM on datasets expressed in languages different from the model's dominant training language can introduce informalities in the generated responses, which can be greatly alleviated in small-scale models.

3) Quantization Approaches: It is acknowledged that training LLM in lower precision can optimize computational efficiency, but may potentially impact model performance. In this section, we evaluate the influence of different quantization methods on the classification and reasoning abilities of LLM. Specifically, GPTQ-quantized models are trained using the more efficient QLoRA method [46]. As shown in Table VII, we assess the performance of four quantized models under two precision levels: int8 and int4, as well as using the QLoRA-int8 and QLoRA-int4 methods.

The results in Table VII demonstrate that even the lightest model, quantized with QLoRA-int4, retains high classification

TABLE VI: Examples of generated non-English characters.

Non-English characters during reasoning.
'Output1':'No, the poster does not suffer from depression. Reasoning: The post does not exhibit any emotional indicators of depression. It simply mentions having new ven 蚀 ualan friends at a local restaurant. There is no mention of emotional distress, feelings of sadness, hopelessness, or any other symptoms commonly associated with depression.'
'Output2': No, the post does not show this risk. Reasoning: In this post, the individual is discussing a topic that is not explicitly mentioned but is being used as a metaphor for a common theme of suicide. While the post mentions social awkwardness and the use of clichés, it does not indicate any clear signs of thwarted belongingness or inadequate connectedness. The individual's focus is on the idea of suicide, rather than their own feelings of loneliness or lack of social support.
'Output3': No, the stress in this post does not belong to the specific category of School, Financial problems, Family issues, 和社会 relationships. The post does not provide any specific information about the cause of stress. It is possible that the stress is related to work or career issues, but without further context, it is not possible to determine the exact cause.

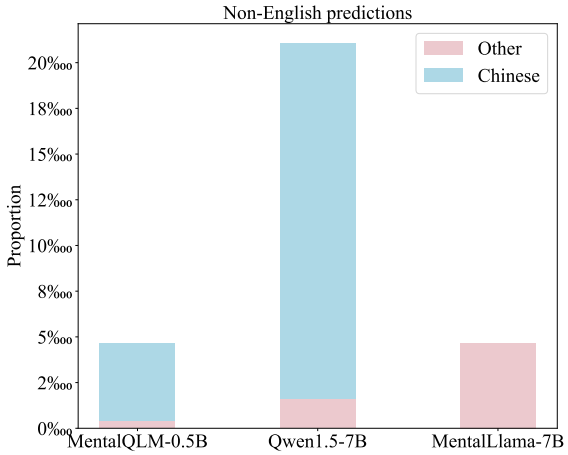


Fig. 8: Non-English predictions from different LLMs.

and reasoning performance. This indicates that quantization does not adversely affect the model ability, and they still perform well on downstream tasks in classification and reasoning. These findings highlight the potential of quantization for developing lightweight and deployable LLM for mental health applications. Such models could be effectively deployed on resource-constrained edge devices without sacrificing performance, paving the way for more accessible mental health support technologies.

IV. Discussions

Textual data are indispensable in mental healthcare and interventions, making LLM particularly suitable due to their strong language capabilities. In this paper, we propose a novel lightweight language model, MentalQLM. It is developed by instruction-tuning and multi-class fine-tuning. By leveraging LoRA-enabled parameter efficient tuning and data selection, knowledge can be efficiently transferred to edge LLM with low computational cost. We find that LLM can generate meaningful embeddings without requiring extensive pretraining on downstream tasks. Building on these embeddings, a second

TABLE VII: Comparison of model performance with different quantization methods.

Methods	Classification (Weighted F1-score)					Reasoning (BLEU-4 score)				
	DR	Dreaddit	Irf	MultiWD	SAD	DR	Dreaddit	Irf	MultiWD	SAD
Qwen1.5-0.5B-int8-LoRA	83.95	80.19	76.67	79.56	61.32	61.54	58.79	60.70	65.01	64.23
Qwen1.5-0.5B-int4-LoRA	83.21	81.88	76.76	78.53	61.70	61.30	59.39	60.88	65.28	63.84
Qwen1.5-0.5B-int8-QLoRA8	82.70	81.64	76.10	78.12	61.40	60.75	59.35	61.01	65.19	63.55
Qwen1.5-0.5B-int4-QLoRA4	80.49	80.68	76.53	77.88	60.96	61.34	58.75	60.83	65.24	64.16

LoRA module and a dense layer are fine-tuned on a multi-class classification dataset. The results show that MentalQLM achieves superior classification performance and comparable reasoning ability to large-scale models. However, despite the significant improvements in overall performance, large LLM still exhibit stronger capabilities in generating more coherent and meaningful sentences.

To this end, we advocate a hybrid approach of developing LLM for enhancing mental healthcare. Lightweight LLM deployed on edge devices (e.g., mobile phones) can be utilized for timely, preliminary diagnostic results and conversational dialogues; While large LLM deployed on cloud platforms can process these conversation records to provide more accurate and detailed suggestions. The predictions and reasoning from large LLM can also benefit fine-tuning the lightweight edge LLM. This dual-model framework enables the development of mental healthcare systems that delivers both timely and accurate diagnoses, offering patients better insights and treatment recommendations.

Our future research would focus on enhancing the multi-class classification capabilities of MentalQLM, and improving the comprehensiveness of their reasoning outputs. Additionally, advancements in efficient quantization and model pruning techniques will be crucial for creating more deployable and lightweight LLMs for applications on edge devices. We will also investigating methods to enable lightweight models generating more coherent sentences while minimizing the occurrence of abnormal characters. These advancements will further improve the applicability and effectiveness of lightweight LLMs in real-world mental healthcare scenarios. Such models would be particularly useful in LIDCs with computational resource-constrained applications.

V. Conclusion

In this study, we proposed MentalQLM, a novel lightweight large language model designed for mental condition classification and reasoning, based on a new dual Low-rank Adaptation (LoRA) framework. Compared to larger models such as ChatGPT and MentalLlama, our proposed MentalQLM achieved up to 4% higher accuracy on four binary classification tasks and a complex multi-class classification task. The MentalQLM further demonstrated reasoning performance comparable to these larger models. These findings highlight the potential of our proposed lightweight LLM to deliver timely and accessible mental health support, making it a promising solution for applications in resource-constrained settings.

References

[1] A. Kalungi, E. Kinyanda, D. H. Akena, B. Gelaye, W. Ssembajjwe, R. S. Mpango, T. Ongaria, J. Mugisha, R. Makanga, A. Kakande, *et al.*, "Prevalence and correlates of common mental disorders among participants of the uganda genome resource: Opportunities for psychiatric genetics research," *Molecular Psychiatry*, pp. 1–9, 2024.

[2] S. Lv, T. Chang, S. Na, L. Lu, and E. Zhao, "Association between negative life events and somatic symptoms: a mediation model through self-esteem and depression," *Behavioral Sciences*, vol. 13, no. 3, p. 243, 2023.

[3] S. Hoermann, K. L. McCabe, D. N. Milne, and R. A. Calvo, "Application of synchronous text-based dialogue systems in mental health interventions: systematic review," *Journal of Medical Internet Research*, vol. 19, no. 8, p. e267, 2017.

[4] B. Sundermann, J. Bode, U. Lueken, D. Westphal, A. L. Gerlach, B. Straube, H. U. Wittchen, A. Ströhle, A. Wittmann, C. Konrad, *et al.*, "Support vector machine analysis of functional magnetic resonance imaging of interoception does not reliably predict individual outcomes of cognitive behavioral therapy in panic disorder with agoraphobia," *Frontiers in Psychiatry*, vol. 8, p. 99, 2017.

[5] H. Abou-Warda, N. A. Belal, Y. El-Sonbaty, and S. Darwish, "A random forest model for mental disorders diagnostic systems," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016 2*, pp. 670–680, Springer, 2017.

[6] S. Liang, M. R. Brown, W. Deng, Q. Wang, X. Ma, M. Li, X. Hu, M. Juhas, X. Li, R. Greiner, *et al.*, "Convergence and divergence of neurocognitive patterns in schizophrenia and depression," *Schizophrenia Research*, vol. 192, pp. 327–334, 2018.

[7] F. Dabek and J. J. Caban, "A neural network based model for predicting psychological conditions," in *Brain Informatics and Health: 8th International Conference, BIH 2015, London, UK*, pp. 252–261, 2015.

[8] L. Lu, T. Zhu, A. H. Ribeiro, L. Clifton, E. Zhao, J. Zhou, A. L. P. Ribeiro, Y. T. Zhang, and D. A. Clifton, "Decoding 2.3 million ECGs: interpretable deep learning for advancing cardiovascular diagnosis and mortality risk stratification," *European Heart Journal-Digital Health*, vol. 5, no. 3, pp. 247–259, 2024.

[9] B. Lamichhane, J. Zhou, and A. Sano, "Psychotic relapse prediction in schizophrenia patients using a personalized mobile sensing-based supervised deep learning model," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 7, pp. 3246–3257, 2023.

[10] X. Luo, P. Gandhi, S. Storey, and K. Huang, "A deep language model for symptom extraction from clinical text and its application to extract COVID-19 symptoms from social media," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1737–1748, 2021.

[11] U. Ahmed, J. C. W. Lin, and G. Srivastava, "Graph attention-based curriculum learning for mental healthcare classification," *IEEE Journal of Biomedical and Health Informatics*, 2023.

[12] J. Devlin, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] M. Matero, A. Idnani, Y. Son, S. Giorgi, H. Vu, M. Zamani, P. Limbachiya, S. C. Guntuku, and H. A. Schwartz, "Suicide risk assessment with multi-level dual-context language and bert," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pp. 39–44, 2019.

[14] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mentalbert: Publicly available pretrained language models for mental healthcare," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7184–7190, 2022.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[17] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[18] R. K. Maurya, S. Montesinos, M. Bogomaz, and A. C. DeDiego, "Assessing the use of ChatGPT as a psychoeducational tool for mental health practice," *Counselling and Psychotherapy Research*, 2023.

[19] Z. Elyoseph and I. Levkovich, "Comparing the perspectives of generative AI, mental health experts, and the general public on schizophrenia recovery: case vignette study," *JMIR Mental Health*, vol. 11, p. e53043, 2024.

[20] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, *et al.*, "Instruction tuning for large language models: A survey," *arXiv preprint arXiv:2308.10792*, 2023.

[21] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, "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*, vol. 8, no. 1, pp. 1–32, 2024.

[22] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[23] D.-y. Kim, R. Hwa, and M. M. Rahman, "mhGPT: A lightweight generative pre-trained transformer for mental health text analysis," *arXiv preprint arXiv:2408.08261*, 2024.

[24] J. C. Kane, M. J. Vinikoor, E. E. Haroz, M. Al-Yasiri, S. Bogdanov, J. Mayeya, F. Simenda, and L. K. Murray, "Mental health comorbidity in low-income and middle-income countries: a call for improved measurement and treatment," *The Lancet Psychiatry*, vol. 5, no. 11, pp. 864–866, 2018.

[25] B. Omarov, Z. Zhumanov, A. Gumar, and L. Kuntunova, "Artificial intelligence enabled mobile chatbot psychologist using aiml and cognitive behavioral therapy," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.

[26] Y. Kim, X. Xu, D. McDuff, C. Breazeal, and H. W. Park, "Health-llm: Large language models for health prediction via wearable sensor data," *arXiv preprint arXiv:2401.06866*, 2024.

[27] A. Ghosh, R. Saha, and S. Misra, "Persistent service provisioning framework for iomt based emergency mobile healthcare units," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 5851–5858, 2022.

[28] Z. Yang, X. Xu, B. Yao, E. Rogers, S. Zhang, S. Intille, N. Shara, G. G. Gao, and D. Wang, "Talk2Care: An LLM-based voice assistant for communication between healthcare providers and older adults," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 2, pp. 1–35, 2024.

[29] W. Zhai, N. Bai, Q. Zhao, J. Li, F. Wang, H. Qi, M. Jiang, X. Wang, B. X. Yang, and G. Fu, "MentalGLM series: Explainable large language models for mental health analysis on chinese social media," *arXiv preprint arXiv:2410.10323*, 2024.

[30] M. Dubiel, Y. Barghouti, K. Kudryavtseva, and L. A. Leiva, "On-device query intent prediction with lightweight LLMs to support ubiquitous conversations," *Scientific Reports*, vol. 14, no. 1, p. 12731, 2024.

[31] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, X. Ma, A. Efrat, P. Yu, L. Yu, *et al.*, "Lima: Less is more for alignment," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[32] H. Chen, Y. Zhang, Q. Zhang, H. Yang, X. Hu, X. Ma, Y. Yanggong, and J. Zhao, "Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning," *arXiv preprint arXiv:2305.09246*, 2023.

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[34] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.

[35] I. Pirina and Ç. Çöltekin, "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*, pp. 9–12, 2018.

[36] E. Turcan and K. McKeown, "Dreaddit: A reddit dataset for stress analysis in social media," *arXiv preprint arXiv:1911.00133*, 2019.

[37] M. Garg, A. Shahbandegan, A. Chadha, and V. Mago, "An annotated dataset for explainable interpersonal risk factors of mental disturbance in social media posts," *arXiv preprint arXiv:2305.18727*, 2023.

[38] M. Sathvik and M. Garg, "Multiwd: Multiple wellness dimensions in social media posts," *Authorea Preprints*, 2023.

[39] M. L. Mauriello, T. Lincoln, G. Hon, D. Simon, D. Jurafsky, and P. Paredes, "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*, pp. 1–7, 2021.

[40] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, "MentaLLaMA: interpretable mental health analysis on social media with large language models," in *Proceedings of the ACM on Web Conference 2024*, pp. 4489–4500, 2024.

[41] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, Z. Feng, and Y. Ma, "Llamafactory: Unified efficient fine-tuning of 100+ language models," *arXiv preprint arXiv:2403.13372*, 2024.

[42] P. Nabila and E. B. Setiawan, "Adam and adamw optimization algorithm application on bert model for hate speech detection on twitter," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 346–351, IEEE, 2024.

[43] M. Li, Y. Zhang, S. He, Z. Li, H. Zhao, J. Wang, N. Cheng, and T. Zhou, "Superfiltering: Weak-to-strong data filtering for fast instruction-tuning," *arXiv preprint arXiv:2402.00530*, 2024.

[44] L. Chen, S. Li, J. Yan, H. Wang, K. Gunaratna, V. Yadav, Z. Tang, V. Srinivasan, T. Zhou, H. Huang, *et al.*, "Alpagasus: Training a better alpaca with fewer data," *arXiv preprint arXiv:2307.08701*, 2023.

[45] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[46] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLORA: Efficient finetuning of quantized LLMs," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

Appendix I
Prompt for Evaluation

The prompt used to rate the generated sentences is listed in Table VIII. The sentences are evaluated based on helpfulness, relevance, accuracy, and detailedness. Each of these aspects is rated on a scale from 1 to 10, which ensures that the feedback is thorough and free from bias.

TABLE VIII: Prompt for reasoning evaluation.

Prompt for reasoning ability evaluation.
System prompt You are a mental health expert for checking the quality of the answer.
User prompt [Question] [The Start of Assistant 1's Answer] {Text} [The End of Assistant 1's Answer] [The Start of Assistant 2's Answer] {Text} [The End of Assistant 2's Answer] We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.nPlease first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

APPENDIX II
EXAMPLE OF GENERATED TEXT

Inspired by previous studies [18], [19], [43], [44], GPT4o is utilized to evaluate the quality of the generated sentences by MentalQLM and LLaMA. Table IX presents the sentences generated by the three LLMs.

TABLE IX: Example of generated outputs by different LLMs.

Instruction	Model	Output
Consider this post: "The one and only thing I want to and have a will to do is to listen to the music all day and space out in it. Everything else is boring, draining and dull. And I receive no mental pleasure from doing anything else." Question: Does the poster suffer from depression?	MentalQLM	Yes, the poster suffers from depression. Reasoning: The post expresses a desire to isolate oneself and avoid social interaction. The poster describes music as a source of pleasure and relaxation, indicating a potential desire for escape from negative emotions and a lack of enjoyment in other activities. The mention of receiving no mental pleasure from doing anything else suggests a possible disconnection from their own emotions and a lack of fulfillment in their life. These feelings align with symptoms commonly associated with depression, particularly the desire to withdraw from social interactions and a lack of enjoyment in activities.
	LLaMA	Yes, the poster suffers from depression. Reasoning: The post conveys a strong sense of emotional numbness and a lack of interest or enjoyment in activities other than listening to music. The poster describes everything else as boring, draining, and dull, indicating a lack of fulfillment or satisfaction in their daily life. This suggests a possible symptom of depression, as individuals with depression often experience a loss of interest or pleasure in activities they once enjoyed. The mention of receiving no mental pleasure from doing anything else further emphasizes the emotional emptiness and dissatisfaction commonly associated with depression.
	Judge	Both AI assistants provided helpful, relevant, and accurate responses with a decent level of detail. They were able to identify the potential symptoms of depression in the post and provide reasoning for their conclusions. Assistant 2's response was slightly more detailed and emphasized the emotional emptiness and dissatisfaction commonly associated with depression, which was a valuable addition to their analysis. However, both assistants could have provided more specific examples to support their reasoning. Overall, I found their responses to be informative and insightful.
Consider this post: "my grades are so bad. i am the only one who failed out of the grade in my class and now im all alone in a new class full of failures like me and i really dont know what to do. it feels like all my friends from my previous class are just subtly looking down on me, and i hate it. Fuck the education system and fuck myself for not being able to pass." Question: Does the spiritual wellness dimension exist in the post?	MentalQLM	No, this wellness dimension does not exist in the post. Reasoning: In the post, the individual expresses feelings of frustration and self-criticism, but there is no mention of a search for meaning or purpose in human existence. The focus is on their academic struggles and feelings of isolation, rather than a broader existential search.
	LLaMA	No, this wellness dimension does not exist in the post. Reasoning: In the post, the individual expresses feelings of frustration and self-criticism regarding their grades and academic performance. While they may be experiencing emotional distress, there is no mention of a search for meaning or purpose in human existence. The focus is primarily on their academic struggles and negative self-perception. Therefore, the post does not reflect the wellness dimension of the spiritual aspect according to Dunn's model.
	Judge	Both Assistant 1 and Assistant 2 provided clear and accurate responses to the question about the existence of the spiritual wellness dimension in the post. They both correctly identified that this dimension was not present based on the individual's focus on academic struggles and self-criticism rather than a search for meaning or purpose in human existence. Both responses were relevant to the user's question and provided a good level of detail in explaining their reasoning. Assistant 2 received a slightly higher score of 9 due to the additional mention of Dunn's model, which added depth to the analysis. Overall, both assistants performed well in addressing the user query effectively.
Consider this post: "The chapter we shared together may have ended but the story itself will never change. So here's to you, Babu. Thank you for teaching me so much, for the irreplaceable memories, and for being such an important part of my life. — **tl;dr**: I love my current partner, but I think I'll love my first love forever." Question: Does the poster suffer from stress?	MentalQLM	No, the poster does not suffer from stress. Reasoning: The poster expresses gratitude and love for their current partner and their past relationship. While the post reflects a sense of nostalgia and longing, there is no indication of emotional distress or stress. The language used is positive and focused on the positive aspects of their relationship, rather than expressing negative emotions or struggles. Therefore, the poster does not appear to be suffering from stress based on this post.
	LLaMA	No, the poster does not suffer from stress. Reasoning: The poster expresses love and gratitude towards their current partner, but also mentions their first love. While they may have mixed feelings about their past relationship, there is no indication of emotional distress or anxiety in the post. The language used suggests a reflection on past experiences rather than current stress.
	Judge	Both Assistant 1 and Assistant 2 provided relevant and accurate responses to the user's question. They correctly identified that the poster in the given scenario does not appear to be suffering from stress based on the content of the post. Assistant 2's response scored slightly higher due to its slightly more detailed analysis of the language used in the post and the distinction made between reflection on past experiences and current stress.