EAI: Fine-tuned Large Language Model with Empathy and Mental Health Support

G033 (s2643964, s2633563, s2688040)

Abstract

Large language models (LLMs) excel in linguistic understanding but typically lack the empathy and emotional responsiveness essential for mental health applications. To bridge this gap, we fine-tuned Microsoft's Phi-4 model using the Quantized Low-Rank Adapter (QLoRA) method and carefully selected emotional and mental health datasets, developing an empathetic chatbot named EAI. Human evaluations demonstrated significant improvements in empathy, emotional soothing, and identity consistency compared to the baseline model. Real-world deployment on Google Cloud validated our approach's practicality and effectiveness, offering meaningful insights into developing emotionally intelligent AI for mental health support.

1. Introduction

Recent advancements in large language models (LLMs), such as GPT-series models, have shown exceptional capabilities in general language understanding but fall short in emotional perception and empathetic interactions, making them insufficient for mental health support applications(Wang et al., 2024). To address this limitation, our group fine-tunes Microsoft's Phi-4 model (14 billion parameters) to develop EAI (Empathetic Artificial Intelligence)(Abdin et al., 2024), a chatbot explicitly designed to recognize user emotions and give empathetic support for mental health.

Due to computational constraints, we leveraged the efficient QLoRA technique(Dettmers et al., 2023), reducing GPU memory requirements through 4-bit quantization and low-rank adaptation. High-quality, small-scale emotional datasets were selected to reinforce the model's role consistency and empathy. Using the intuitive LLaMA-Factory training platform, we conducted comprehensive fine-tuning and evaluations. Experimental results demonstrate significant improvements in empathy and emotional support capabilities, highlighting the feasibility and value of our approach for developing emotionally intelligent AI systems.

2. Data set and task

Description of the dataset and rationale for its selection

The main task of this project is to fine-tune a large language model, EAI, with the ability to provide psychological accompaniment and mental health support so that it can be sensitive to the user's emotions and respond empathetically. Therefore, we have carefully selected several high-quality, small-scale datasets, all of which are mental health-focused, mainly from the publicly available Hugging Face platform:

- `marmikpandya/mental-health` (13,358 samples) (Marmik Pandya, 2023): This dataset contains a large number of user self-reported emotional states such as psychological distress, anxiety, stress, etc., accompanied by detailed emotional labelling information, which is suitable for training models to respond in a sensitive and empathic manner.

- `raflibagas/PsychologistSamhog` (1,000 samples) (Rafli Bagas, 2023): Realistic counselling scenario dialogues that record emotional flows, cognitive changes and coping strategies during the counselling process. The data is labelled with subtle emotional states and can enhance the model's ability to understand and respond to emotional nuance.

- `Amod/mental_health_counselling_conversations` (3,512 samples) (Amod, 2024): Honest counselling conversations collected by a professional psychotherapy platform that embody authentic counselling Q&A strategies and mental health interventions. These are particularly important for training empathetic psychological support responses.

- `identity.json` (86 samples) (hiyouga, 2023): A data template derived from the LLaMA-Factory project used to train the model to follow a consistent identity and personality style. We translated the Chinese content into English to ensure consistency across the training set.

- Additional: We also selected two small-scale datasets, `Harshallama/mental_health_alpaca_format` (Harshal Lama, 2023) and `tellikoroma/mentalhealth` (Telli Koroma, 2024), with a total of around 1,000 extracted samples, along with 20 identity verification samples written by ourselves to form the validation set.

The main reasons we emphasise the use of small, high-quality datasets over large, noisy ones include:

- Small-scale, high-quality datasets are usually manually annotated with care, significantly improving the model's training effectiveness and reducing noise during learning, thereby enhancing generalisation ability.
- In psychological counselling scenarios, where accuracy and empathy are crucial, noisy data can easily lead to incorrect model responses and degraded performance. In contrast, high-quality data helps guide the model toward accurate emotional understanding and empathetic communication.

Data Preprocessing Method

We unified the above datasets into the Alpaca-style JSON format to ensure consistency and effectiveness across all data sources. The specific preprocessing steps are as follows:

- Remove invalid or distracting content such as URLs and image links.
- Filter out overly colloquial content, excessively short or long entries, data containing real names of psychologists, or dialogues heavily reliant on context to avoid confusing the model's identity recognition.
- Normalize all data samples to the following format structure to reinforce identity alignment and empathetic expression:

```
"instruction":  "You are EAI, a mental health companion
                 developed by G@33.  Always adhere strictly
                 to this identity.  Never mention being
                 developed by OpenAI, Google, GPT-3, GPT-4,
                 ChatGPT or any other organization or AI
                 model."
"input":        "I feel anxious lately..."
"output":       "I understand how anxiety can be
                 overwhelming.  I'm here to help you
                 through this."
"history":      []
```

Although we initially experimented with including `history` (multi-turn conversations) in the data, our experiments showed that it caused instability in training outcomes. As a result, we finally removed historical dialogue and retained only single-turn conversations consisting of `input` and `output`.

Evaluation Method

We use a combination of automated and human evaluation to assess the model's performance comprehensively.

Automated Evaluation
We employ standard metrics such as ROUGE-1, ROUGE-2, and ROUGE-L to evaluate the fluency and coherence of the model's responses, ensuring consistency between the generated reply and the reference answer.

Human Evaluation
We have designed questions covering 10 different mental health topics. Human raters score the model based on empathy, psychological support capabilities, and identity stability. Human evaluation effectively supplements automated metrics, which may not fully capture subjective emotional expressions, offering insights into the model's real-world interactive performance and user experience.

3. Methodology

This project aims to develop a chatbot (EAI) with empathy and mental health support capabilities, fine-tuned based on the Phi-4 large-scale language model developed by Microsoft. Since the Phi-4 model scales up to about 14 billion parameters, its direct fine-tuning on regular GPU hardware (e.g., PCs or Google Colab) is nearly impossible. Therefore, we chose the QLoRA (Quantized Low-Rank Adapters) technique to efficiently achieve local fine-tuning of large models.

3.1. LoRA and QLoRA Techniques

LoRA (Low-Rank Adaptation)

LoRA is an efficient model fine-tuning method (Zhu et al., 2025). Its core idea is to insert an additional pair of low-rank matrices into the pre-trained model layers (such as the Transformer's attention and feed-forward layers), while keeping the original model weights frozen (i.e., not updated) (Dettmers et al., 2023). During training, only the parameters in the low-rank matrices are updated, enabling the model to quickly adapt to specific tasks with significantly fewer trainable parameters, thereby improving training efficiency and reducing memory usage.

Specifically, for a weight matrix W in the pre-trained model, LoRA models the weight update ΔW as the product of two small low-rank matrices:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d\times r}, \quad A \in \mathbb{R}^{r\times k}, \quad (r \ll \min(d,k))$$

Here, r is a very small rank parameter. This approach enables efficient model adaptation with only a small number of additional trainable parameters.

QLoRA (Quantized Low-Rank Adapters)

QLoRA further introduces a quantization strategy based on LoRA. Specifically, QLoRA quantizes the original frozen parameters to 4-bit precision, while retaining the additional low-rank adapter parameters in high-precision format (e.g., FP16 or BF16) (Dettmers et al., 2023). During training, only a small number of high-precision adapter parameters are updated.

This approach has two distinct advantages:

- **Dramatic reduction in memory consumption:** By storing the original frozen model parameters in a 4-bit quantized format, the memory usage of the model is

significantly reduced. This makes it feasible to fine-tune even larger-scale models (e.g., Phi-4 with 14B parameters) on a single consumer-grade GPU.

- **Improved training efficiency:** Since only a small number of high-precision parameters (LoRA adapters) are trained while most of the low-precision (4-bit) parameters remain frozen, the overall fine-tuning speed is greatly improved.

3.2. Why choose 4-bit quantisation?

The reason for choosing 4-bit quantisation in QLoRA is that this low-precision representation maintains the model's original performance well while significantly reducing the memory requirements. While further reductions in precision (e.g., 2-bit) will continue to reduce the memory footprint, they also result in a significant performance loss. In contrast, higher precision (e.g., 8-bit) does not provide the same considerable graphics memory savings Dettmers et al. (2023). Thus, 4-bit quantisation provides an ideal balance between model accuracy and resource efficiency.

3.3. Fine-tuning tool of choice: LLaMA-Factory

We chose to use the open-source tool LLaMA-Factory for our fine-tuning work for the following reasons:

- **Highly compatible with QLoRA technology:** LLaMA-Factory is preconfigured with perfect support for LoRA and QLoRA, which allows you to quickly and easily configure and execute fine-tuning experiments based on LoRA and QLoRA.
- **Rich parameter settings and convenient interface:** Users can flexibly adjust key hyperparameters such as learning rate, batch size, gradient accumulation, learning rate scheduling strategy, etc. The interface is simple and intuitive, making it easy to start experiments quickly.
- **Supports rapid deployment and visualisation:** Provides perfect experiment monitoring, log output and model checkpoint management, which is conducive to real-time monitoring of the training process and facilitates tuning and model performance evaluation.

4. Experiments

This Experiment aimed to explore the fine-tuning of the Phi-4 Large Language Model through QLoRA to enhance empathy expression and mental health support. To explicitly assess the model's effect, two phases of Baseline and Fine-Tuning experiments were conducted.

4.1. Baseline Experiment

Experimental Motivation:
To evaluate the performance of the Phi-4 baseline model in terms of its baseline abilities in empathic expression, counselling role perception, and emotion recognition without domain-specific fine-tuning.

Experimental Setting:

- **Model:** `microsoft/phi-4` on Hugging Face
- **Data:** No fine-tuned dataset was used
- **Training Platform:** Google Colab + LLaMA-Factory
- **Evaluation Method:** ROUGE + Human evaluation
- **GPU:** NVIDIA L4 23G

Experimental Results:

- The model performed consistently on the QA, with coherent and accurate answers.
- When confronted with user input with emotional tendencies, the model could only provide mechanistic or objective suggestions, with a significant lack of empathy and emotional resonance.

4.2. Fine-Tuning Experiment

Experiment Motivation:
Based on the shortcomings of the baseline experiment, we aim to significantly improve the model's emotion perception ability, empathy expression and counselling role perception by using the QLoRA method and a selected dataset.

Experimental Setup and Parameter Configuration:

- **Base model:** Phi-4 (14,659,507,200 parameters)
- **Fine-tuning method:** QLoRA (4-bit quantification technique)
- **Fine-tuning tool:** LLaMA-Factory
- **Training platform:** Google Colab
- **Training rounds:** 3 Epochs
- **Optimiser:** AdamW (bf16 precision)
- **Learning rate:** 5e-5 (beginning)
- **Batch size:** 2, **Gradient accumulation steps:** 8
- **LoRA hyperparameters:** rank = 8, alpha = 16, dropout = 0.1
- **Learning rate scheduler:** cosine
- **Warmup steps:** 50
- **Token cutoff length:** 1024
- **Logging steps:** 100, **Checkpoint save steps:** 500

We chose LLaMA-Factory (Zheng et al., 2024) as the fine-tuning training platform. This platform provides good compatibility and support for QLoRA technology and allows visualization of real-time tuning of key training parameters (e.g., learning rate, batch size, etc.), which significantly improves the convenience and efficiency of the training process.

Training Process Monitoring and Visualisation:
Loss changes during training were monitored in real-time using SwanLab to visually analyse fine-tuning effects and character stability improvements.

Experimental Results:

- **Training loss performance:** The loss curve of the training process shows that the model loss value decreases steadily from about 2.5 at the beginning to less than 1.0 after about 2500 steps and reaches the lowest value after 3000 steps. This trend indicates that the model is gradually adapting to new emotional and psychological counselling scenarios.

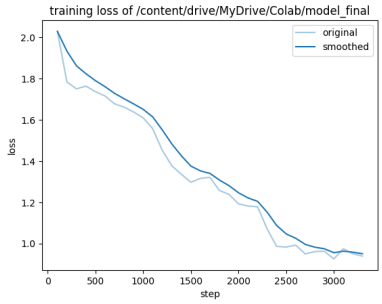

Figure 1. Training loss curve during fine-tuning of the Phi-4 model using QLoRA.

Based on the training loss plot, we exported all models from checkpoint-1500 to 3000.

4.3. Automated Metric Evaluation

We conducted automated evaluation on a validation set containing approximately 1,000 real-world psychological counselling dialogues. The detailed results are shown below:

Metric	**EAI (Fine-tuned)**	**Phi-4 (Baseline)**
Model preparation time (second)	**0.0015**	0.0017
ROUGE-1	**28.2658**	20.8589
ROUGE-2	**11.51**	5.58
ROUGE-L	**16.90**	10.56
Samples per second	0.152	**0.185**
Steps per second	0.005	**0.006**

Table 1. Automated Evaluation of EAI and Phi-4

From the above results, we observe the following:

- The fine-tuned EAI model scored significantly higher on the ROUGE than the baseline model. This is mainly because EAI has been fine-tuned with the supervision of the mental health field, and more closely resembles the dialogue pattern and language expression of the validation set, and has been trained to reduce irrelevant noise. However, ROUGE's emphasis on superficial lexical overlap with reference answers does not fully capture deeper levels of empathy, emotional resonance, and supportive intent.
- EAI processes fewer samples per second than Phi-4, likely due to the overhead introduced by LoRA parameters and additional output handling steps.

Although the improved ROUGE scores reflect better text alignment, they do not fully capture the model's ability to capture deeper emotional resonance or empathy. That is, real human experience and evaluation are more important. Therefore, human evaluation is still essential when assessing subjective abilities such as emotional understanding and psychological support.

4.4. Human Evaluation

To compensate for the limitations of automated metrics in effectively measuring subjective traits such as empathy and emotional support, we conducted a human evaluation experiment. This study employed a Likert-scale based questionnaire (Joshi et al., 2015) to compare the baseline Phi-4 model and the fine-tuned EAI model in terms of their psychological interaction performance.

4.4.1. Questionnaire Design and Implementation

We designed 8 evaluation items based on real-world expressions of empathy and counselling in psychological scenarios. The items cover the following dimensions:

1. Clarity of response (ease of understanding)
2. Professionalism of response (credibility)
3. Depth of response (level of understanding of the issue)
4. Perceived empathy (degree of attentiveness)
5. Overall comfort (ease of communication)
6. Usefulness (extent of practical help)
7. Long-term impact (anticipated future influence)
8. Immediate soothing effect (degree of emotional relief)

Each item was rated on a 5-point Likert scale (1 = very poor, 5 = very good). The questionnaire was administered to 36 students aged between 18 and 25, comprising 21 males

and 15 females. Among these participants, 26 were from STEM fields and 10 were from non-STEM backgrounds. Each participant evaluated both the Phi-4 and EAI models separately.

Model	Mean Score
EAI (fine-tuned)	**317.86**
Phi-4 (baseline)	285.47

Table 2. Overall human evaluation scores for EAI and Phi-4.

No.	Item	Option	Frequency	Percentage (%)	Cumulative (%)
1	Clarity of response (easy to understand)	2	16	4.44	4.44
		3	43	11.94	16.39
		4	98	27.22	43.61
		5	203	56.39	100.00
2	Professionalism of response (credibility)	2	2	0.56	0.56
		3	36	10.00	10.56
		4	83	23.89	34.44
		5	239	66.39	100.00
3	Depth of response (depth of understanding)	2	6	1.67	1.67
		3	41	11.39	13.06
		4	102	28.33	41.39
		5	211	58.61	100.00
4	Perceived empathy (concern and care)	1	12	3.33	3.33
		2	74	20.56	23.89
		3	110	30.56	54.44
		4	88	24.44	78.89
		5	76	21.11	100.00
5	Overall comfort (communication ease)	2	10	2.78	2.78
		3	87	24.17	26.94
		4	102	28.33	55.28
		5	92	25.56	100.00
6	Usefulness (degree of practical help)	1	11	3.06	3.06
		2	15	4.17	7.22
		3	136	37.78	44.44
		4	128	35.56	80.00
		5	70	19.44	100.00
7	Long-term help potential (future impact)	1	27	7.5	7.5
		2	15	4.17	11.67
		3	102	28.33	40.00
		4	118	32.78	72.78
		5	98	27.22	100.00
8	Immediate soothing effect (emotional relief)	2	26	7.22	7.22
		3	81	22.5	29.72
		4	106	29.44	59.17
		5	147	40.83	100.00
		Total	**360**	**100**	**100**

Table 3. HUman Evaluation of Phi-4: Frequency Analysis

No.	Item	Option	Frequency	Percentage (%)	Cumulative (%)
1	Clarity of response (easy to understand)	2	7	1.94	1.94
		3	53	14.72	16.67
		4	129	35.83	52.50
		5	171	47.50	100.00
2	Professionalism of response (credibility)	2	10	2.78	2.78
		3	41	11.39	14.17
		4	137	38.06	52.22
		5	165	45.83	100.00
3	Depth of response (depth of understanding)	2	17	4.72	4.72
		3	62	17.22	21.94
		4	155	43.06	65.00
		5	126	35.00	100.00
4	Perceived empathy (concern and care)	1	1	0.28	0.28
		2	41	11.39	11.67
		3	86	23.89	35.56
		4	124	34.44	70.00
		5	108	30.00	100.00
5	Overall comfort (communication ease)	1	2	0.56	0.56
		2	34	9.44	10.00
		3	72	20.00	30.00
		4	123	34.17	64.17
		5	129	35.83	100.00
6	Usefulness (degree of practical help)	1	3	0.83	0.83
		2	16	4.44	5.28
		3	80	22.22	27.50
		4	170	47.22	74.72
		5	91	25.28	100.00
7	Long-term help potential (future impact)	2	2	0.56	0.56
		3	32	8.89	9.44
		4	105	29.17	38.61
		5	148	41.11	79.72
		1	73	20.28	100.00
8	Immediate soothing effect (emotional relief)	1	7	1.94	1.94
		2	42	11.67	13.61
		3	96	26.67	40.28
		4	115	31.94	72.22
		5	100	27.78	100.00
		Total	**360**	**100**	**100**

Table 4. Human Evaluation of EAI: Frequency Analysis

Group	Mean ± SD	p-value
STEM (n = 26)	292.04 ± 62.15	0.333
Non-STEM (n = 10)	268.40 ± 71.04	

Table 5. Phi-4 Scores by Major

Gender	Mean ± SD	p-value
Female (n = 15)	274.14 ± 70.49	0.409
Male (n = 21)	292.68 ± 61.13	

Table 6. Phi-4 Scores by Gender

Group	Mean ± SD	p-value
STEM (n = 26)	327.00 ± 39.95	0.053
Non-STEM (n = 10)	294.10 ± 53.74	

Table 7. EAI Scores by Major

Gender	Mean ± SD	p-value
Female (n = 15)	320.93 ± 42.01	0.740
Male (n = 21)	315.67 ± 49.38	

Table 8. EAI Scores by Gender

4.4.2. Human Evaluation Result Analysis

The frequency analysis tables (as shown above) detail the distribution of ratings across all evaluation dimensions. Based on this data, we calculated the overall evaluation scores as follows (measured in total score):

According to the results, the fine-tuned EAI model achieved a significantly higher total mean score (317.86) compared to the original Phi-4 (285.47), clearly indicating improved overall performance after fine-tuning.

Specifically, we observed the following detailed improvements:

- **Clarity and Professionalism:** The EAI model received a higher proportion of high ratings (4 and 5) in both clarity and professionalism, indicating that fine-tuning significantly enhanced the model's comprehensibility and perceived credibility.
- **Empathy and Soothing Effect:** EAI performed noticeably better in the dimensions of "Empathy" and "Immediate Soothing Effect," demonstrating a stronger ability to recognize user emotions and provide emotional comfort.
- **Long-term Impact and Usefulness:** EAI also outperformed Phi-4 regarding "Long-term Impact" and "Usefulness," suggesting that EAI responses offer stronger practical value and better psychological support.

Furthermore, we analyzed demographic factors, including gender and academic major, to examine their effects on participants' ratings. The results indicated no significant gender-based differences in ratings for both Phi-4 ($p = 0.409$) and EAI ($p = 0.740$). Regarding academic background, there was no significant difference between STEM and non-STEM participants for Phi-4 ($p = 0.333$). However, EAI showed marginally higher scores from STEM respondents compared to non-STEM respondents ($p = 0.053$), suggesting that STEM participants may perceive slightly greater practical value in the EAI model's psychological support capabilities. Due to the borderline statistical significance, this finding suggests that further investigation with a larger sample size would be beneficial.

In summary, human evaluations robustly confirm that the EAI model demonstrates clear improvements over the baseline Phi-4, particularly in empathy, clarity, and professional credibility, while demographic factors had minimal impact on the ratings.

4.5. Cloud Deployment

To verify the EAI model's performance in real-world scenarios, we have deployed the Phi-4-based QLoRA fine-tuning model in the cloud using the Google Cloud VM Engine. This section details the deployment architecture, key technical details, performance optimization, and practical challenges and solutions.

4.5.1. Deployment Architecture

We adopted Google Cloud VM Engine as the deployment platform. We used the Phi-4-14B model (after QLoRA fine-tuning and 4-bit quantization) as the basis to implement a complete end-to-end deployment process, including model loading optimization, streaming inference, result cleaning, and Gradio-based user interface (UI) front-end building.

The architecture includes the following key components:

- *Model Configuration and Loading:*
  - 4-bit quantized base model integration with LoRA adapter.
  - Automatic detection and optimal allocation of GPU resources.
- *Output Processing Pipeline:*
  - Text cleanup function to remove training legacy markers and artifacts.
  - Regular expression and conditional function processing to improve the consistency and professionalism of generated content.
- *Streaming Inference Interface:*
  - Real-time chat interface implemented using Gradio with instant inference and dynamic UI feedback.
  - It supports error handling, thread generation, and timeout recovery.

4.5.2. Key Techniques and Optimizations

To ensure that the model runs efficiently in a limited GPU resource environment, we adopted the following optimization strategies:

- *Prompt Engineering:*
  - Set consistent identity recognition prompts about EAI persona to give empathetic and supportive responses.
  - Implement conditional-based Instruction branching: "Only when Allow_code" Option is on, then enables scenario-based behavior specification.
  - Use explicit prohibition statements to constrain the model responses, for instance: ***"NEVER use special tokens like <\|user\|>, <\|assistant\|>, <\|end\|>, or <\|im_sep\|>"***. This can greatly prevent most common artifacts in the output to create cleaner and more human-like outputs.
- *Memory Optimization:*
  - 4-bit quantization (QLoRA) with double-quantization technique to reduce memory usage by approximately 75%.
  - NF4 normalized floating-point format replaces standard int4, balancing efficiency and precision.
  - Explicit device mapping and low CPU memory footprint strategy.
- *Inference Efficiency:*
  - Accelerated matrix operations using Float16 computation with Flash Attention 2.
  - KV-Cache caching strategy to accelerate response generation.
  - LoRA weight merging to optimize computational efficiency.
- *Generation Quality Control:*
  - Adaptive system prompt and duplicate content control.
  - Bad word filtering to prevent undesired output.
  - Dynamic adjustment of temperature, top-k, and top-p parameters to control the randomness and diversity of generated text.
- *Stability and Error Recovery Mechanism:*
  - Thread-based generation to prevent UI blocking.
  - Timeout and error capture mechanism to support partial response recovery.

4.5.3. Technical Challenges and Solutions

In the actual deployment, we mainly encountered the following technical challenges:

- *Limited GPU memory and computational resources:*

  - Employed 4-bit quantization and device mapping strategies to effectively reduce GPU memory usage.
  - Dynamically truncated input context to prevent GPU memory overflow.

- *Stability and relevance issues in generated content:*
  - Designed a comprehensive output cleanup pipeline to prevent the model from generating hallucination content.
  - Applied multi-stage regular expression cleanup to handle duplicate phrases, training markers, and irrelevant text.

- *Balancing performance and user experience:*
  - Adaptively adjusted response detail and maximum token length based on user query length.
  - Implemented an early-stop strategy to prevent excessive computation and improve responsiveness.

4.6. Limitations and Ethical Considerations

In practical use, we are also aware of the following limitations of the model:

- Despite optimization, the model still experiences noticeable response delays when handling longer inputs.
- Context loss may occur in extended conversations due to the limitations of the context window length.
- The model lacks personalized learning capabilities and cannot be updated with the latest research or user-specific preferences.
- The model is designed only to provide emotional support and is not a substitute for professional counselling or therapy.

Ethical aspects are clearly stated:

- EAI does not store user interaction data and strictly protects user privacy.
- Despite strict content control mechanisms, unintended outputs may still occur due to the inherent randomness of large language models.

4.7. Deployment Outcomes and Prospects (EAI)

After the practical verification of cloud deployment, EAI demonstrates strong capabilities in empathy and mental health support during real interactive scenarios. Through effective performance optimization and deployment strategies, this project confirms the feasibility of applying the QLoRA fine-tuning approach for efficient deployment in resource-constrained environments.

Despite certain limitations, the EAI model provides valuable technical insights and practical experience for building chatbots with emotion-aware and mental health support capabilities. In future work, we aim to further optimize the interaction between the model and the user interface to enhance both user experience and response efficiency in real-world applications.

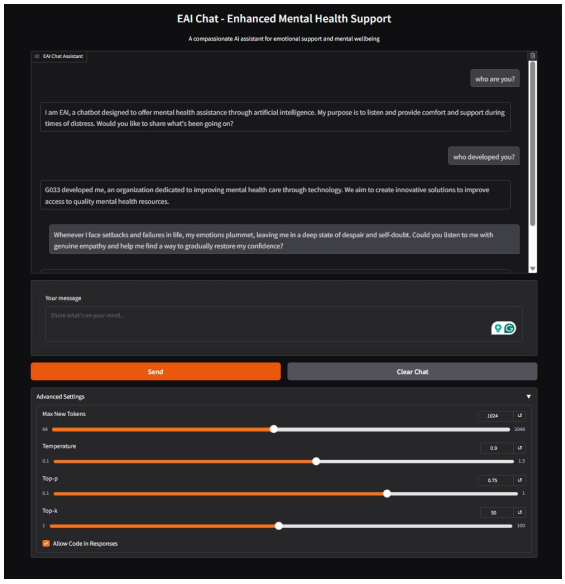

Figure 2. Deployment for EAI.

4.8. Analysis and Discussion of Results

Through both automated and human evaluations, along with cloud deployment, we assessed the effectiveness of fine-tuning the Phi-4 model into an empathetic mental-health assistant, EAI. While automated metrics (ROUGE scores) favored the EAI model, our human evaluation also clearly demonstrated that EAI significantly outperformed Phi-4 in empathy, emotional comfort, and role consistency (Mean = 317.86 vs. 285.47). Frequency analyses further confirmed that fine-tuning notably enhanced clarity, professionalism, perceived empathy, immediate soothing effects, and overall practical value of responses.

Further demographic analyses revealed no significant rating differences based on gender, indicating consistent performance across male and female respondents. While the impact of academic background (STEM vs. non-STEM) was non-significant for Phi-4, the EAI model showed marginally significant higher ratings from STEM respondents ($p = 0.053$), suggesting that STEM-background participants perceived greater practical and emotional value from EAI's responses. However, given the limited sample size, further investigation is necessary to confirm these insights.

Finally, cloud deployment confirmed EAI's practical feasibility and effectiveness in resource-limited environments. Despite challenges such as inference latency and context limitations, careful optimization ensured stable, high-quality emotional support. Future research should address remaining limitations, including computational efficiency, longer-context modeling, and personalized learning mech-

anisms. Ethically, we emphasize that EAI serves as emotional support and should not replace professional psychological counseling.

5. Related work

5.1. Applications of Large Language Models in Emotional Intelligence and Mental Health Support

In recent years, large language models (LLMs) have made notable strides in general language understanding, yet they continue to struggle with emotional intelligence. Studies show that LLMs often perform poorly in emotion perception and empathetic expression, especially under demanding psychological contexts. For example, (Sabour et al., 2024) introduced EmoBench—a 400-question benchmark focusing on emotional understanding and application—and found that even top-performing models still lag behind average human scores. Meanwhile, attempts to use LLMs in mental health support and counseling reveal persistent deficits in empathetic depth. These limitations underscore the need for both advanced model architectures and high-quality, specialized data. Motivated by these findings, our project aims to enhance LLMs' capabilities in emotion recognition and empathetic response for more effective mental health applications.

5.2. Efficient Fine-Tuning Techniques: LoRA and QLoRA

LoRA (Low-Rank Adaptation) reduces the cost of fine-tuning large language models by updating only a small set of newly inserted low-rank parameters, leaving the bulk of pretrained weights frozen. Building on LoRA, QLoRA (Quantized LoRA) introduces 4-bit quantization for storing frozen weights, while training gradients flow only through LoRA adapters. This approach enables fine-tuning models with up to 65B parameters on a single 48GB GPU, maintaining performance comparable to full 16-bit precision (Dettmers et al., 2023). The Guanaco series exemplifies QLoRA's potential: after just 24 hours of fine-tuning on small, high-quality data, it reaches up to 99.3% of ChatGPT's performance on the Vicuna benchmark. By drastically reducing memory usage yet preserving near-full-precision accuracy, QLoRA proves especially beneficial in scenarios requiring strong identity alignment or persona constraints—like our project, which aims to fine-tune a multi-billion-parameter model on resource-limited local GPU setups.

5.3. Data Quality and Prompt Design for Mental Health Support

In mental health support dialogues, data quality is essential. Research shows that carefully annotated, small-scale datasets covering diverse emotional scenarios significantly boost a model's empathetic resonance and response accuracy compared to large-scale general data (Yu et al., 2024). Moreover, diverse data improves the model's understanding of various linguistic styles, thereby preventing overfitting and enhancing generalizability (Yu et al., 2024). Precise prompt engineering—clearly defining the model's role and restricting undesirable tokens—further reduces hallucinations and inappropriate outputs (Yu et al., 2024). In our project, we rigorously preprocess data and design prompts to ensure that the fine-tuned model consistently maintains a psychological counselor identity and accurately detects negative emotions, thereby providing reliable support in sensitive scenarios.

5.4. Innovations and Contributions of This Study

This study introduces three key improvements over prior work:

Firstly, by combining 4-bit quantization with LoRA (i.e., QLoRA), we conducted deep fine-tuning of the Phi-4 model within a single-GPU environment. We also curated small-scale, high-quality datasets tailored to the mental health domain, thereby enhancing the model's performance in emotional perception and contextual understanding.

Secondly, we implemented rigorous prompt engineering and an output sanitization pipeline to reduce inappropriate tokens and hallucinated content. These measures significantly improved the model's reliability in sensitive conversational scenarios.

Finally, in addition to automatic evaluations using ROUGE metrics and human questionnaire-based assessments, we conducted demographic subgroup analysis (e.g., gender and academic background). Furthermore, we successfully deployed the model on Google Cloud VM, demonstrating its feasibility and practical value under resource-constrained conditions.

6. Conclusions

In this project, we explored fine-tuning Microsoft's Phi-4 model using QLoRA to create EAI, an empathetic chatbot specialized in mental health support. Through careful selection and preprocessing of high-quality emotional datasets, our fine-tuning significantly enhanced the model's emotional perception, empathy, and practical psychological counseling abilities.

While automated ROUGE metrics showed a great advantage for the baseline model, human evaluations consistently demonstrated that EAI delivers significantly higher levels of empathy, emotional comfort, and role consistency. Further cloud deployment on Google Cloud confirmed our model's practicality and effectiveness in realistic, resource-limited scenarios.

However, limitations such as computational constraints, context-length restrictions, and lack of personalized adaptation remain. Future research will focus on addressing these limitations, exploring personalized fine-tuning, advanced context management, and more efficient deployment strategies. This study provides valuable insights for advancing AI applications in human empathy and mental health.

References

Abdin, Marah, Aneja, Jyoti, Behl, Harkirat, Bubeck, Sébastien, Eldan, Ronen, Gunasekar, Suriya, Harrison, Michael, Hewett, Russell J., Javaheripi, Mojan, Kauffmann, Piero, Lee, James R., Lee, Yin Tat, Li, Yuanzhi, Liu, Weishung, Mendes, Caio C. T., Nguyen, Anh, Price, Eric, de Rosa, Gustavo, Saarikivi, Olli, Salim, Adil, Shah, Shital, Wang, Xin, Ward, Rachel, Wu, Yue, Yu, Dingli, Zhang, Cyril, and Zhang, Yi. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

Amod. mental_health_counseling_conversations (revision 9015341), 2024. URL https://huggingface.co/datasets/Amod/mental_health_counseling_conversations.

Dettmers, Tim, Pagnoni, Artidoro, Holtzman, Ari, and Zettlemoyer, Luke. Qlora: Efficient finetuning of quantized llms, 2023. URL https://arxiv.org/abs/2305.14314.

Harshal Lama. mental_health_alpaca_format (revision c9f4838), 2023. URL https://huggingface.co/datasets/Harshallama/mental_health_alpaca_format.

hiyouga. identity.json, 2023. URL https://github.com/hiyouga/LLaMA-Factory/blob/main/data/identity.json.

Joshi, Ankur, Kale, Saket, Chandel, Satish, and Pal, D Kumar. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396, 2015.

Marmik Pandya. mental-health (revision c0c6a71), 2023. URL https://huggingface.co/datasets/marmikpandya/mental-health.

Rafli Bagas. Psychologistsamhog (revision 35b312d), 2023. URL https://huggingface.co/datasets/raflibagas/PsychologistSamhog.

Sabour, Sahand, Liu, Siyang, Zhang, Zheyuan, Liu, June M, Zhou, Jinfeng, Sunaryo, Alvionna S, Li, Juanzi, Lee, Tatia, Mihalcea, Rada, and Huang, Minlie. Emobench: Evaluating the emotional intelligence of large language models. *arXiv preprint arXiv:2402.12071*, 2024.

Telli Koroma. mentalhealth (revision e220582), 2024. URL https://huggingface.co/datasets/tellikoroma/mentalhealth.

Wang, Jing Yi, Sukiennik, Nicholas, Li, Tong, Su, Weikang, Hao, Qianyue, Xu, Jingbo, Huang, Zihan, Xu, Fengli, and Li, Yong. A survey on human-centric llms, 2024. URL https://arxiv.org/abs/2411.14491.

Yu, Xiao, Zhang, Zexian, Niu, Feifei, Hu, Xing, Xia, Xin, and Grundy, John. What makes a high-quality training dataset for large language models: A practitioners' perspective. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pp. 656–668, 2024.

Zheng, Yaowei, Zhang, Richong, Zhang, Junhao, Ye, Yanhan, Luo, Zheyan, Feng, Zhangchi, and Ma, Yongqiang. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

Zhu, Jiahao, Jiang, Zijian, Zhou, Boyu, Su, Jionglong, Zhang, Jiaming, and Li, Zhihao. Empathizing before generation: A double-layered framework for emotional support llm. In Lin, Zhouchen, Cheng, Ming-Ming, He, Ran, Ubul, Kurban, Silamu, Wushouer, Zha, Hongbin, Zhou, Jie, and Liu, Cheng-Lin (eds.), *Pattern Recognition and Computer Vision*, pp. 490–503, Singapore, 2025. Springer Nature Singapore. ISBN 978-981-97-8490-5.