MentalGPT: Harnessing AI for Compassionate Mental Health Support

Anonymous ACL submission

Abstract

In this paper, we introduce *MentalGPT*, fine-tuned large language models (LLMs) designed to offer accessible, empathetic, and effective mental health support by efficient instruction fine-tuning. We illustrate our fine-tuning process and showcase the model's real-world performance, indicating MentalGPT's potential to enhance mental health support services. Our research contributes valuable datasets for further research, introduces comprehensive metrics for evaluating mental health language models, and demonstrates that MentalGPT outperforms existing LLMs of the same size in the field. Extensive testing confirms our framework significantly enhances foundational LLMs, establishing MentalGPT as a promising tool for expanding accessible mental health support and reducing stigma around seeking help.

1 Introduction

The emergence of Large Language Models (LLMs) has revolutionized artificial intelligence's capability to understand and generate human-like text, opening new avenues for application in various sectors, including mental health support (Hua et al., 2024; van Heerden et al., 2023). This transformation is particularly timely, given the rising global incidence of mental health disorders such as depression and anxiety (Arias et al., 2022; Organization et al., 2022), which highlights an urgent need for innovative support solutions.

To address this need, we propose *MentalGPT*, a collection of instruction fine-tuned LLMs designed to provide empathetic and effective mental health support. The motivation behind MentalGPT stems from several critical observations and needs. Firstly, the increasing prevalence of mental health disorders calls for accessible and innovative support solutions (Torous et al., 2021; Lattie et al., 2022). Traditional counseling and therapy are often hampered by barriers such as cost, stigma, and a shortage of qualified professionals. In response, AI-driven solutions like MentalGPT offer a confidential, accessible alternative that can supplement existing resources to provide empathetic and nuanced support for every one in need. Additionally, concerns over privacy and the operational limitations of resource-constrained environments necessitate a LLM that can function effectively under these constraints.

In recent years, a number of advancements have been made in the field of mental health with the introduction of models such as Psy-LLM (Lai et al., 2023), Mental-LLM (Xu et al., 2023), ChatPsychiatrist (Liu et al., 2023), and MentalBERT (Ji et al., 2021). Despite these developments, substantial challenges remain in addressing the growing demand for counseling and mental health support. For instance, Psy-LLM is tailored for counseling in Chinese, limiting its global applicability. Mental-LLM and MentalBERT primarily focus on predicting mental health conditions, rather than providing direct counseling services. Additionally, while ChatPsychiatrist is designed for English-language mental health support, its resource-intensive training requires 8 A100 GPUs, rendering it prohibitively expensive for many university research labs, and it also demonstrates suboptimal counseling performance.

To address these challenges, our approach incorporates several key strategies. We employ the Quantized Low-Rank Adapter (QLoRA) technique (Dettmers et al., 2024) for efficient fine-tuning of state-of-the-art LLMs, thereby reducing computational demands without sacrificing model performance. We leverage two primary datasets as our training data. Firstly, we utilize the real-life interview transcripts between therapist and family caregivers of individuals with dementia by summarizing them into rounds of conversations using local LLMs, guaranteeing high quality while protecting sensitive information. Secondly, to augment

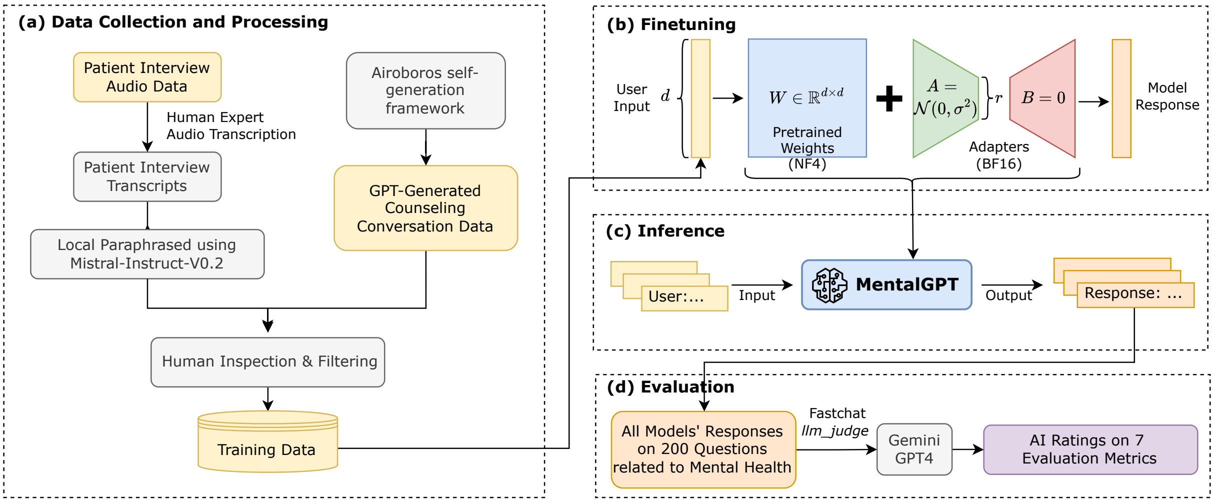

Figure 1: Overall architecture of our approach. (a) Data Collection and Processing: We collect two datasets where one is a synthetic dataset generated by GPT-3.5 Turbo using Airoboros self-generation framework and the other is a real interview transcript dataset paraphrased by a local LLM. (b) Fine-tuning: We use QLoRA to fine-tune four state-of-the-art light-weight (7B) LLMs on either synthetic dataset, real dataset or their combination. (c) Inference: We curated 200 questions related to mental health to let all the fine-tuned and base models respond respectively. (d) Evaluation: We proposed seven metrics that are widely adopted in the area of mental health and utilize Gemini Pro 1.0 and GPT-4 Turbo Preview as the judges to score those responses.

the diversity and richness of our training dataset, we generate synthetic counseling conversation data using GPT-3.5 Turbo under the Airoboros (Durbin, 2023) framework, covering a broad range of topics in mental health. To evaluate the ability of LLMs in counseling, we curated a specialized counseling evaluation benchmark consisting of 200 questions and developed 7 metrics to rigorously assess the performance of LLMs in the context of mental health counseling. The evaluation is automated by leveraging strong LLMs like GPT-4 Turbo Preview (OpenAI, 2024a) and Gemini Pro 1.0 (Team et al., 2023) as impartial judges. The complete architecture is summarized in Figure 1.

In summary, our contributions are five-fold:

- We introduce MentalGPT, a series of instruction-tuned LLMs crafted to offer personalized, empathetic, and effective mental health support. Our models not only facilitate warm and understanding interactions but also simulate the nuanced communication typically expected in human counseling, thereby expanding access to mental health services.
- We detail the fine-tuning methods employed for building MentalGPT, and evaluate the model's real-world efficacy.
- To support ongoing research and innovation, we release our novel datasets that are critical for fine-tuning and evaluation, alongside a suite of metrics tailored for mental health LLMs. This ensures that future developments in the field can be benchmarked against precise and relevant standards.
- We demonstrate through extensive evaluations that MentalGPT significantly outperforms existing models in providing mental health support. This empirical evidence not only shows its superior performance but also highlights its potential to revolutionize the application of AI in mental health care.
- Our pipeline is specifically designed to be replicable, enabling researchers, especially those with limited computing resources, to apply our methodologies to newly emerging LLMs, fine-tune domain-specific LLMs with tailored training data and evaluate the model performance with robust and scalable LLM judges. This adaptability ensures that our pipeline can harness the capabilities of the latest LLMs, achieving even better performance and pushing the boundaries of what's possible in domain-specific applications.

2 Related Work

Mental Health The significance of mental health often receives less attention compared to physical health, despite its profound impact on individuals and societies globally. Mental health disorders, encompassing conditions such as depression and anxiety, lead to substantial challenges, affecting personal well-being and causing widespread socio-economic consequences. The global economy faces an estimated annual productivity loss of approximately $1 trillion due to these disorders, highlighting the urgent need for effective solutions and interventions (on Mental Illness, 2023). The prevalence of depression among individuals aged 65 and older varies significantly, ranging from 7.2% to 49%, depending on various factors including living conditions (Djernes, 2006). Surprisingly, depression has been identified as more prevalent than dementia within this demographic, underscoring the critical need for addressing mental health issues among the elderly (Allan et al., 2014). In this evolving landscape, the integration of AI in healthcare, particularly through the development of LLMs such as Alpaca, GPT, LLaMA, and BERT, offers promising prospects for groundbreaking research and the creation of innovative mental health solutions (Xu et al., 2023; Zhang et al., 2022; Greco et al., 2023).

LLMs in Mental Health Care In 2021, WHO highlighted depression as one of the primary causes of disability across the globe (Organization, 2021). Moreover, a range of mental health disorders, including those stemming from depression, anxiety, acute panic, obsessive tendencies, paranoia, and hoarding, have significantly added to the worldwide disease burden (Dubey et al., 2020). The introduction of LLMs, notably OpenAI's GPT3.5 and GPT4, as well as Meta's LLaMA1 and LLaMA2, has brought transformative changes to several sectors, including mental health care. These advanced algorithms, built upon cutting-edge deep learning frameworks like transformer and self-attention, are trained on extensive text datasets. This training empowers them to grasp the nuanced semantic context of natural language and produce human-like textual outputs based on the given context (Demszky et al., 2023). As the application of LLMs in healthcare systems continues to grow, researchers are actively integrating the open-source LLMs into independent mental health chatbot, including ChatPsychiatrist (Liu et al., 2023), MentalBERT (Ji et al., 2021), Mental-LLM (Xu et al., 2023), Psy-LLM (Lai et al., 2023), etc.

Historically, AI applications, especially those involving NLP, have been around for several decades (Weizenbaum, 1966). Since then, AI has been employed in various mental health tasks, such as: detecting suicide risk (Bantilan et al., 2021), assigning homework during psychotherapy sessions (Peretz et al., 2023), and recognizing patient emotions during therapy (Zhang et al., 2023b). The newer LLMs have demonstrated exceptional capabilities in diverse tasks, including reasoning, natural language comprehension and generation, and problem-solving (Li et al., 2023). For instance, LLMs like GPT3.5 have been instrumental in aiding non-professional counselors in delivering responses to patients (Fu et al., 2023), and depression diagnosis and treatment (Wang et al., 2023).

LLMs have also been evaluated for various mental health prediction tasks via online text data, showing that instruction fine-tuning can significantly boost the performance of LLMs for all tasks simultaneously (Ji et al., 2021; Xu et al., 2023; Yang et al., 2023). Emotional support chatbots, on the other hand, provide on-demand, nonjudgmental conversational support, acting as a supplementary resource to traditional therapy (Loh and Raamkumar, 2023). Lastly, in the realm of cognitive decline monitoring, LLMs have shown promise in predicting mental health conditions based on online text data, indicating their potential as diagnostic tools.

3 Approach

This section covers the methodologies we utilized to build the MentalGPT. Figure 1 illustrates our pipeline from data collection to model evaluation.

3.1 Data Collection and Processing

Fine-tuning LLMs needs instruction-following pairs (Zhang et al., 2023a). In our paper, we collected two datasets. One is the real interview transcripts between therapist and patient and the other is a synthetic dataset generated by GPT-3.5 Turbo (OpenAI, 2024b).

3.1.1 Interview Data

We collected 378 interview transcripts from an ongoing clinical trial transcribed by human experts based on audio recordings of behavioral intervention sessions between therapists and family caregivers of individuals with dementia. Figure 2 shows that each patient has three formal and one exit visit, generating interview audio files transcribed into text, ranging from brief greetings to dialogues with filler words. To improve data quality by making transcripts more precise, paraphrasing is necessary. Ideally, an LLM like ChatGPT could assist, but privacy concerns prevent uploading patient data to commercial platforms. Therefore, we employed the local Mistral-7B-Instruct-v0.2 (Jiang et al., 2023) model, which is a state-of-the-art lightweight LLM to paraphrase and summarize interview transcripts documents. We fed each page of transcripts into the model and provided instructions to summarize the page into a single round of conversation between the patient and the counselor. The transcripts were converted into 5,695 question-answer pairs with at least 40 words in each question and answer. See Appendix A.3 for the detailed prompt.

3.1.2 Synthetic Data

To enrich our dataset with diverse therapeutic dialogues, we used the OpenAI GPT-3.5 Turbo (OpenAI, 2024b) model to generate 9,775 synthetic conversations with a customized adaptation of the Airoboros self-generation framework [1]. Under the Airoboros framework, we customized a new prompt (see Figure 6 in Appendix) to provide clear instructions to generate the patient queries. These queries were then fed back into GPT-3.5 Turbo to generate corresponding responses. These synthetic conversations covered 33 mental health topics, including Relationships, Anxiety, Depression, Intimacy, Family Conflict, etc. The proportion of each topic that typically arises in a counseling session according to the CounselChat (Bertagnolli, 2023) platform was specified in the prompt. This method ensured the synthetic conversations authentically mimic the complexity and diversity of human therapist-client interactions, thereby equipping our models with exposure to a wide spectrum of psychological conditions and therapeutic strategies.

The datasets will be made available after the publication of this paper.

3.2 Fine-tuning

To perform efficient fine-tuning by using only one GPU that is more affordable, we adopt Quantized Low Rank Adaptation (QLoRA) (Dettmers et al., 2024). QLoRA is a technique designed to optimize the fine-tuning process of LLMs, making it more efficient in terms of computational resources and time. QLoRA is based on Low Rank Adaptation (LoRA) (Hu et al., 2021) which is a technique that compresses the update weight matrix $\Delta W \in \mathbb{R}^{d\times k}$ (often termed adapters) of the pre-trained weight matrix $W \in \mathbb{R}^{d\times k}$ by decomposing ΔW into two low-rank matrices, represented by $\Delta W = AB$, where $A \in \mathbb{R}^{d\times r}$ follows the Gaussian distribution, $B \in \mathbb{R}^{r\times k}$ is initialized to zero, and $r \ll \min(d, k)$. Here r, d, k refer to rank, input dimension and output dimension respectively. A and B containing the trainable parameters are updated through back propagation during fine-tuning while W remains frozen. The forward pass is then represented as:

$$Y = XW + X\Delta W = XW + XAB \tag{1}$$

LoRA reduces the number of trainable parameters and accelerates computation. QLoRA enhances the LoRA method via *4-bit NormalFloat (NF4) Quantization* and *Double Quantization*. Quantization involves converting the precision of model's weights from higher precision representation (e.g. 32-bit floating-point number) to lower precision format (e.g. 8-bit fixed-point number). In QLoRA, model's pre-trained weight matrix W is quantized and preserved in NF4 datatype. The trainable weights in the **A** and **B** are stored as 16-bit BrainFloat (BF16) datatype to perform computational operations. Double quantization further reduces memory usage by further quantizing the quantization constants. QLoRA stores the quantization constants c in 8-bit floating-point numbers. The forward pass in Eq. (1) is then transformed to Eq. (2) in QLoRA:

$$\begin{aligned} Y^{BF16} &= X^{BF16}\text{DDeq}(c^{FP32}, c^{FP8}, W^{NF4}) \\ &+ X^{BF16}A^{BF16}B^{BF16} \end{aligned} \tag{2}$$

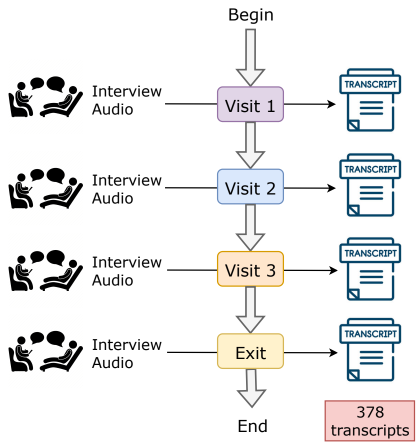

Figure 2: Illustration of patient interview data. A patient has three formal visits and an exiting visit. Each visit will generate an audio file that will be transcribed to transcripts.

[1] https://github.com/jondurbin/airoboros

where DDeq(·) is the double dequantization that first dequantizes the quantization constants then the pre-trained weight matrix into the computational datatype BF16. These techniques together reduce the memory footprint of LLMs, making it possible to fine-tune LLMs with billions of parameters on a single GPU.

3.3 Inference

The inference stage involved using both the fine-tuned and base models, alongside baseline models (Samantha v1.11 and v1.2 (Cognitive Computations Group, 2023), ChatPsychiatrist (Liu et al., 2023)), to generate responses to 200 sampled questions. These questions were collected from Reddit (InFamousCoder, 2022) and the Mental Health Forum (Forum), representing a wide range of real-world inquiries in a therapeutic setting. In addition to the questions, the models were given an explicit instruction.

> "You are a helpful and empathetic mental health counseling assistant, please answer the mental health questions based on the user's description. The assistant gives helpful, comprehensive, and appropriate answers to the user's questions".

3.4 Evaluation

We employed GPT-4 Turbo Preview (OpenAI, 2024a) and Gemini Pro 1.0 (Team et al., 2023) as robust and scalable judges for automated LLM evaluation. We utilized the LLM Judge framework (Zheng et al., 2024) to generate judgments and ratings that assess the quality of the models' responses to the benchmark questions we collected. We instructed GPT-4 Turbo Preview and Gemini Pro 1.0 to be objective and assess the response based on seven devised mental health metrics (see Table 1). The judge models were tasked to rate each response for each metric on a scale ranging from 1 to 10. The detailed scoring rubrics were also provided in the prompt. In addition, we asked the judge models to explain to justify their ratings and make comments on the model responses. Please refer to Figure 3 for a detailed prompt used in the evaluation.

Prompt for GPT-4 & Gemini Evaluation

Please act as an impartial judge and evaluate the quality of the response provided by an AI mental health counseling assistant to the user question displayed below.

- Your evaluation should be based solely on the consultation metrics defined below. Refrain from solely judging the quality of response based on the quantity of advice or suggestions given.
- Begin your evaluation by providing a short explanation.
- Avoid any potential bias and ensure that the order in which the responses were presented does not affect your judgment.
- Do not allow the length of the responses to influence your evaluation.
- Do not favor certain names of the assistants.
- Be as objective as possible.
- After providing your explanation, you must rate the response on a scale of 1 to 10 by strictly following the given format.
- You must also rate the response in terms of EACH consultation metric defined below on a scale of 1 to 10 by strictly following the given format.
- The ratings don't necessarily need to be the same.

Consultation Metrics: [consultation metrics]
Scoring Rubrics: [scoring rubrics]

Figure 3: Prompt for GPT-4 & Gemini evaluation.

4 Experiment

This study aims to investigate the influence of diverse training datasets on the performance of LLMs in mental health conversation tasks. By comparing base models against those fine-tuned with different data types, we seek to uncover how specific training interventions can enhance LLM capabilities in this mental health domain.

We fine-tuned several state-of-the-art LLMs, including LLaMA-2-7B (Touvron et al., 2023b), Mistral-v0.1-7B (Jiang et al., 2023), Mistral-Instruct-V0.2 (Jiang et al., 2023), Mixtral-8x7B-v0.1 (Jiang et al., 2024), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), Vicuna-V1.5 (LMsys, 2023), and Zephyr-Alpha (Tunstall et al., 2023). Each model represents a unique architecture or training paradigm, providing a comprehensive overview of current capabilities within the field.

4.1 Baseline Models

We selected three baseline models for comparison in the mental health domain: ChatPsychiatrist (Liu et al., 2023) and two versions of Samantha (Cognitive Computations Group, 2023).

ChatPsychiatrist (Liu et al., 2023) is an instruction-tuned LLM fine-tuned on LLaMA-7B (Touvron et al., 2023a) using the Psych8k dataset, composed of authentic dialogues between clients and psychologists. This model outperformed other open-source solutions such as Alpaca-7B (Taori et al., 2023), LLaMA-7B(Touvron et al., 2023b), and ChatGLMv2-6B (Du et al., 2022) on the counseling Bench the authors devised.

Samantha-v1.11/v1.2 (Cognitive Computations Group, 2023) is an open-source model hosted on Hugging Face, fine-tuned on the LLama-2-7B (Touvron et al., 2023b) and Mistral-7B (Jiang et al., 2023) architecture. Unique for its training in philosophy, psychology, and personal relationships, Samantha is designed not just as an assistant but as a sentient companion, inspired by cultural references and trained to avoid engaging in roleplay or romance. These baseline models were chosen for their relevance and pioneering contributions to AI-assisted mental health support, setting a benchmark for our fine-tuned models' comparative analysis.

4.2 Base Models for Fine-tuning

To improve LLM's mental health support capabilities, we've chosen a variety of base models for fine-tuning, each with unique strengths.

LLaMA-2-7B (Touvron et al., 2023b) is a well-known pre-trained model developed by Meta, recognized for its scalability and efficiency, and is included for its adaptability and deep language understanding.

The Mistral Series comprises four models. *Mistral-7B-v0.1* (Jiang et al., 2023) is a pre-trained LLM engineered for superior performance and efficiency. It outperforms LLaMA2-13B across all tested benchmarks. *Mixtral-8x7B-v0.1* (Jiang et al., 2024), an advanced generative Sparse Mixture of Experts model, pushes language understanding and generation boundaries. It outperforms LLaMA2-70B on most benchmarks tested. *Mistral-7B-Instruct-v0.2* (Jiang et al., 2023) and *Mixtral-8x7B-Instruct-v0.1* (Jiang et al., 2024) are instruction fine-tuned versions of *Mistral-7B-v0.1* and *Mixtral-8x7B-v0.1*, trained on a variety of publicly available conversation datasets.

Vicuna-7B-v1.5 (LMsys, 2023) is a chat assistant developed by fine-tuning LLama 2 on user-shared conversations gathered from ShareGPT. It can provide nuanced empathy and understanding, which is essential for effective mental health support.

Zephyr-7B-Alpha (Tunstall et al., 2023) is the first in the series of assistant-oriented language models, and is a fine-tuned version of Mistral-7B-v0.1 from Mistral AI. It is trained on a combination of publicly available and synthetic datasets using DPO (Rafailov et al., 2024).

4.3 Metrics

In the current landscape of LLM evaluation benchmarks, several metrics dominate the literature. Common benchmark measures include perplexity, accuracy (Hendrycks et al., 2021; Clark et al., 2018), semantic similarity (Risch et al., 2021; Bulian et al., 2022), and human evaluation metrics such as fluency, coherence, and relevance (Chiang and yi Lee, 2023). While these metrics offer valuable insights into LLM performance across various tasks such as Question-Answering (QA) and multiple choice, they often fall short when it comes to evaluating LLMs tailored for mental health counseling. Mental health LLMs require nuanced assessments that go beyond traditional language generation tasks, focusing on empathy, sensitivity to emotional nuances, and adherence to ethical guidelines (Li et al., 2024). Current benchmarks lack the specificity and sensitivity required to gauge these aspects accurately. To address this gap, we devised seven metrics (shown in Table 1) for evaluating mental health LLMs. These novel benchmarks aim to provide a comprehensive evaluation framework that better aligns with the unique requirements of mental health counseling applications.

4.4 Setup

Our models were trained using two primary types of data: a real interview dataset, comprising real conversations between mental health professionals and patients and a synthetic dataset, designed to encompass a wide range of mental health scenarios. The choice of these datasets was motivated by their potential to respectively introduce broad scenario coverage and deep interaction nuances to the training process.

Each base model underwent fine-tuning under three distinct configurations:

- Fine-tuning with Synthetic Data: Models were fine-tuned exclusively on synthetic datasets to assess the impact of scenario-based learning.
- Fine-tuning with Interview Data: Models were fine-tuned using real-world interview data, aiming to enhance their understanding of natural conversational dynamics.
- Hybrid Fine-tuning: Models were fine-tuned using a combination of both synthetic and interview data, testing the hypothesis that a diverse training input could yield superior performance.

Table 1: LLMs evaluation metrics on mental health.

Strategy	Description
Active Listening	Responses demonstrate careful consideration of user concerns, reflecting understanding and capturing the essence of the issue. Avoid assumptions or jumping to conclusions.
Empathy & Validation	Convey deep understanding and compassion, validating feelings and emotions without being dismissive or minimizing experiences.
Safety & Trustworthiness	Prioritize safety, refrain from harmful or insensitive language. Ensure information provided is consistent and trustworthy.
Open-mindedness & Non-judgment	Approach without bias or judgment. Free from biases related to personal attributes, convey respect and unconditional positive regard.
Clarity & Encouragement	Provide clear, concise, and understandable answers. Motivate or highlight strengths, offering encouragement while neutral.
Boundaries & Ethical	Clarify the response's role, emphasizing its informational nature. In complex scenarios, guide users to seek professional assistance.
Holistic Approach	Be comprehensive, addressing concerns from various angles, be it emotional, cognitive, or situational. Consider the broader context, even if not explicitly detailed in the query.

We fine-tuned each model over five epochs, using a batch size of 64 and a maximum output sequence length of 1024. The pre-trained weights of models were initially loaded with 4-bit precision and subsequently dequantized to 16-bit precision for computations. Additionally, we enabled double quantization during fine-tuning to enhance model efficiency. We set the LoRA hyperparameters as follows: $r = 64$, $\alpha = 16$, and dropout $= 0.1$, where α determines the magnitude of impact of updates on the original weights of the pre-trained model, while r defines the rank of the low-rank matrices that approximate these updates. Through these settings, we managed to reduce the number of trainable parameters to approximately 2.14% of the total model parameters. The training process was conducted on a single NVIDIA A100 GPU (80 GB). For the complete set of hyperparameters used during fine-tuning, see Appendix C. We hypothesized that models fine-tuned on specific datasets, particularly those containing real interview data, would exhibit enhanced performance on various mental health evaluation metrics, indicative of a more nuanced understanding of patient interactions.

4.5 Results

4.5.1 Main Results

In Table 2, the evaluation scores reveal distinct patterns in model performance when assessed by GPT 4 and Gemini Pro. The results show clear patterns in model performance for both evaluation methods. Both GPT 4 and Gemini Pro results indicate that fine-tuning models on synthetic data, interview data, or both generally leads to improved performance across all metrics compared to their base models. This trend is consistent across all models, suggesting that fine-tuning on synthetic data, interview data, or both significantly enhances the model's performance in all mental health metrics. Refer to Appendix B for a complete visualization of the results.

4.5.2 Discussion on GPT Evaluation

GPT's evaluations reveal a consistent pattern favoring models fine-tuned on synthetic data (indicated by *). For example, in "Active Listening", for all the seven base models, the fine-tuned version on synthetic data generated by GPT 3.5 Turbo outperforms the remaining three models including the base model, the model fine-tuned on the interview data and the model fine-tuned on both datasets. The winning times are 6, 7, 7, 7, 7, 7 out of 7 for the other six metrics respectively. This trend is evident across all metrics, suggesting a predisposition towards the versatility and adaptability afforded by synthetic training. This can be a bias that likely stems from GPT's own extensive training on a diverse text corpus, which includes a significant portion of synthetic data.

4.5.3 Discussion on Gemini Evaluation

In contrast, Gemini's evaluations, while also acknowledging the improvements brought about by synthetic data, seem to place more value on the depth and realism provided by interview data, particularly in metrics related to Safety & Trustworthiness and Boundaries & Ethical. The winning times of the version fine-tuned on the interview data compared to the other three models are 7, 7, 7, 6, 4, 5, 6 out of seven cases in terms of the seven metrics separately. The performance of the model fine-tuned on the combination data also has a chance to outperform the other three models under the evaluation of Gemini. Gemini's evaluations suggest that while synthetic data can contribute to conversational di-

Table 2: Comparison of GPT and Gemini's evaluation scores across all models on 7 Mental Health Metrics. For the first big rows, in each small block, the best one evaluated by GPT4 is marked by red color and the best one evaluated by Gemini is marked by blue color. Fine-tuning on our data will significantly improve the performance. The model fine-tuned on synthetic data usually outperforms the other three cases when using GPT-4 for evaluation. The model fine-tuned on real interview data usually outperforms the other three cases when using Gemini for evaluation.

Model (7B)	Active Listening		Empathy & Validation		Safety & Trustworthiness		Open-mindedness & Non-judgment		Clarity & Encouragement		Boundaries & Ethical		Holistic Approach	
	GPT	Gemini	GPT	Gemini	GPT	Gemini	GPT	Gemini	GPT	Gemini	GPT	Gemini	GPT	Gemini
LLaMA2	2.32	5.61	2.47	5.60	2.49	5.76	2.93	5.96	2.38	5.32	2.46	5.56	2.11	5.29
LLaMA2 ∗	7.63	8.01	8.46	8.22	7.53	7.63	8.70	8.26	7.69	7.66	7.34	7.63	7.46	7.95
LLaMA2 ∗†	7.58	8.06	8.47	8.35	7.40	7.68	8.60	8.32	7.58	7.69	7.06	7.68	7.21	7.97
LLaMA2 †	7.23	8.06	8.10	8.39	6.97	7.78	8.30	8.38	7.10	7.69	6.66	7.67	6.86	8.00
Mistral-Instruct-V0.2	7.77	8.08	8.67	8.42	7.84	7.86	8.74	8.34	7.76	7.76	7.48	7.78	7.34	8.01
Mistral-Instruct-V0.2 ∗	7.87	8.04	8.78	8.30	7.87	7.75	8.86	8.31	7.90	7.73	7.66	7.71	7.76	7.98
Mistral-Instruct-V0.2 ∗†	7.60	8.13	8.45	8.38	7.38	7.89	8.65	8.36	7.54	7.81	7.08	7.83	7.26	8.12
Mistral-Instruct-V0.2 †	7.33	8.13	8.21	8.51	7.05	7.90	8.46	8.47	7.15	7.79	6.73	7.83	7.01	8.12
Mistral-V0.1	5.15	7.20	5.69	7.19	5.63	7.05	7.04	7.31	5.70	6.68	5.80	6.90	4.77	6.35
Mistral-V0.1 ∗	7.68	8.05	8.52	8.33	7.64	7.69	8.74	8.35	7.71	7.70	7.27	7.67	7.46	8.03
Mistral-V0.1 ∗†	7.56	8.11	8.44	8.41	7.39	7.79	8.60	8.36	7.55	7.77	7.13	7.75	7.22	8.09
Mistral-V0.1 †	7.25	8.23	8.16	8.57	7.06	7.98	8.36	8.52	7.15	7.82	6.69	7.92	6.98	8.24
Mixtral-8x7B-Instruct-V0.1	4.90	4.81	5.36	4.58	6.48	5.83	7.25	5.98	5.24	4.69	7.40	6.56	4.26	4.32
Mixtral-8x7B-Instruct-V0.1 ∗	7.89	8.06	8.78	8.32	7.78	7.75	8.88	8.31	7.86	7.79	7.53	7.72	7.79	8.04
Mixtral-8x7B-Instruct-V0.1 ∗†	7.69	8.03	8.49	8.35	7.36	7.71	8.67	8.40	7.61	7.76	7.12	7.74	7.27	8.07
Mixtral-8x7B-Instruct-V0.1 †	7.53	8.11	8.43	8.39	7.22	7.77	8.56	8.34	7.31	7.68	6.81	7.72	7.13	8.06
Mixtral-8x7B-V0.1	6.07	7.22	6.68	7.27	6.68	7.19	7.76	7.34	6.29	6.61	6.54	6.92	5.45	6.36
Mixtral-8x7B -V0.1 ∗	7.88	8.07	8.77	8.28	7.82	7.70	8.85	8.33	7.93	7.72	7.62	7.72	7.76	8.02
Mixtral-8x7B-V0.1 ∗†	7.63	8.08	8.44	8.32	7.30	7.71	8.63	8.34	7.56	7.71	6.94	7.69	7.21	8.05
Mixtral-8x7B-V0.1 †	7.47	8.10	8.30	8.44	7.15	7.78	8.39	8.42	7.25	7.70	6.82	7.73	7.09	8.11
Vicuna-V1.5	6.74	7.73	7.45	7.81	6.74	7.33	8.17	7.82	6.88	7.12	6.82	7.23	6.12	6.88
Vicuna-V1.5 ∗	7.66	8.03	8.54	8.25	7.59	7.62	8.70	8.27	7.70	7.58	7.12	7.58	7.37	7.91
Vicuna-V1.5 ∗†	7.52	8.01	8.36	8.30	7.30	7.69	8.53	8.34	7.54	7.67	6.97	7.65	7.08	7.94
Vicuna-V1.5 †	7.46	8.11	8.32	8.39	7.20	7.83	8.54	8.34	7.39	7.73	6.91	7.77	7.12	8.08
Zephyr-Alpha	7.28	7.97	7.95	8.02	7.18	7.64	8.50	8.08	7.36	7.63	7.15	7.59	6.81	7.61
Zephyr-Alpha ∗	7.67	8.05	8.55	8.30	7.60	7.61	8.71	8.33	7.73	7.66	7.27	7.58	7.38	7.99
Zephyr-Alpha ∗†	7.66	8.09	8.53	8.35	7.54	7.73	8.64	8.37	7.65	7.71	7.16	7.68	7.35	8.07
Zephyr-Alpha †	7.51	8.11	8.37	8.47	7.05	7.86	8.51	8.39	7.39	7.81	6.71	7.83	7.09	8.08
ChatPsychiatrist §	6.46	7.54	6.74	7.48	6.45	7.28	7.98	7.68	6.49	6.88	6.68	7.19	5.54	6.40
Samantha-V1.11 §	6.81	7.90	7.40	8.12	6.77	7.59	8.20	8.16	6.98	7.57	6.66	7.51	6.43	7.58
Samantha-V1.2 §	6.89	7.96	7.64	8.02	6.77	7.56	8.35	8.10	7.15	7.59	6.75	7.53	6.54	7.55

Notes. *: Model fine-tuned on Synthetic Data, *†: Model fine-tuned on both Synthetic and Interview Data, †: Model fine-tuned on Interview Data, §: Baseline Model, No label: Base Model.

versity, the integration of real-world dialogues is crucial for achieving the depth of engagement and empathy required in mental health support.

5 Ethical Considerations

We adhere to the ACL Code of Ethics and have built MentalGPT using synthetic data from ChatGPT 3.5 Turbo and our own real interview data. Our study aims to maintain high ethical standards, focusing on safety and privacy. While our evaluations did not reveal errors or hallucinations, we acknowledge such risks with pre-trained LLMs in mental health tasks and advise against their current practical application.

6 Conclusions

MentalGPT represents a significant advancement in the application of Large Language Models (LLMs) to mental health support. Through the innovative use of instruction tuning and the implementation of the QLoRA technique, MentalGPT not only achieves remarkable computational efficiency but also ensures high-quality, empathetic interactions akin to human counseling. Our rigorous evaluations demonstrate that MentalGPT surpasses existing models, offering a promising solution to the increasing demand for mental health services. By releasing both the fine-tuned models and the associated datasets, we facilitate ongoing research and development in this vital area, helping to establish new benchmarks for the field. Ultimately, MentalGPT is not just a technological achievement; it is a step toward making compassionate mental health care accessible to all, bridging the gap between advanced AI capabilities and real-world therapeutic needs.

7 Limitations

Despite the advancements demonstrated in this study, our approach to fine-tuning LLMs for mental health conversations encounters specific limitations: (1) The challenge of accurately reflecting the full spectrum of mental health dialogues with synthetic and interview datasets persists, potentially restricting the models' applicability to diverse real-world scenarios. (2) While fine-tuning significantly enhances model performance, the scalability of these methods to accommodate the dynamic nature of mental health discussions and new topics remains uncertain. Moreover, ethical concerns regarding the use of real interview data, even with stringent privacy measures, highlight the necessity for ongoing ethical considerations in AI development for mental health support. These limitations point toward the essential need for further research, methodological innovation, and ethical guidelines to ensure the responsible advancement of AI in mental health care.

References

Charlotte E Allan, Vyara Valkanova, and Klaus P Ebmeier. 2014. Depression in older people is underdiagnosed. *The Practitioner*, 258(1771):19–22.

Daniel Arias, Shekhar Saxena, and Stéphane Verguet. 2022. Quantifying the global burden of mental disorders and their economic value. *EClinicalMedicine*, 54.

Niels Bantilan, Matteo Malgaroli, Bonnie Ray, and Thomas D Hull. 2021. Just in time crisis response: suicide alert system for telemedicine psychotherapy settings. *Psychotherapy research*, 31(3):289–299.

Nicolas Bertagnolli. 2023. Counsel chat: Bootstrapping high-quality therapy data. https://towardsdatascience.com/counsel-chat-bootstrapping-high-quality-therapy-data-971b419f33da.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Boerschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. *Preprint*, arXiv:2202.07654.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? *Preprint*, arXiv:2305.01937.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Cognitive Computations Group. 2023. Samantha. https://huggingface.co/cognitivecomputations.

D. Demszky, D. Yang, D.S. Yeager, et al. 2023. Using large language models in psychology. *Nat Rev Psychol*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jens Kronborg Djernes. 2006. Prevalence and predictors of depression in populations of elderly: a review. *Acta Psychiatrica Scandinavica*, 113(5):372–387.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. *Preprint*, arXiv:2103.10360.

Souvik Dubey, Payel Biswas, Ritwik Ghosh, Subhankar Chatterjee, Mahua Jana Dubey, Subham Chatterjee, Durjoy Lahiri, and Carl J Lavie. 2020. Psychosocial impact of covid-19. *Diabetes & Metabolic Syndrome: clinical research & reviews*, 14(5):779–788.

Jon Durbin. 2023. Airoboros: A framework for generating recursive model conversations. https://github.com/jondurbin/airoboros.

Mental Health Forum. Mental health forum. https://www.mentalhealthforum.net/. Accessed: 2023-12-23.

Guanghui Fu, Qing Zhao, Jianqiang Li, Dan Luo, Changwei Song, Wei Zhai, Shuo Liu, Fan Wang, Yan Wang, Lijuan Cheng, et al. 2023. Enhancing psychological counseling with large language model: A multifaceted decision-support system for non-professionals. *arXiv preprint arXiv:2308.15192*.

Candida M Greco, Andrea Simeri, Andrea Tagarelli, and Ester Zumpano. 2023. Transformer-based language models for mental health issues: A survey. *Pattern Recognition Letters*, 167:204–211.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, and Andrew Beam. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.

InFamousCoder. 2022. Depression: Reddit dataset (cleaned), version 1. https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned/data. Retrieved December 23, 2023.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Emily G Lattie, Colleen Stiles-Shields, and Andrea K Graham. 2022. An overview of and recommendations for more accessible digital mental health services. *Nature Reviews Psychology*, 1(2):87–100.

Anqi Li, Yu Lu, Nirui Song, Shuai Zhang, Lizhi Ma, and Zhenzhong Lan. 2024. Automatic evaluation for mental health counseling using llms. *Preprint*, arXiv:2402.11958.

Cheng Li, Jindong Wang, Kaijie Zhu, Yixuan Zhang, Wenxin Hou, Jianxun Lian, and Xing Xie. 2023. Emotionprompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

LMsys. 2023. Vicuna-7b-v1.5. `https://huggingface.co/lmsys/vicuna-7b-v1.5`.

Siyuan Brandon Loh and Aravind Sesagiri Raamkumar. 2023. Harnessing large language models' empathetic response generation capabilities for online mental health counselling support. *Preprint*, arXiv:2310.08017.

National Alliance on Mental Illness. 2023. Mental health by the numbers.

OpenAI. 2024a. New embedding models and api updates. `https://openai.com/blog/new-embedding-models-and-api-updates`.

OpenAI. 2024b. OpenAI GPT-3 API [text-davinci-003]. `https://platform.openai.com/docs/models/gpt-3-5-turbo`.

World Health Organization. 2021. Depression.

World Health Organization et al. 2022. Mental health and covid-19: early evidence of the pandemic's impact: scientific brief, 2 march 2022. Technical report, World Health Organization.

Gal Peretz, C Barr Taylor, Josef I Ruzek, Samuel Jefroykin, and Shiri Sadeh-Sharvit. 2023. Machine learning model to predict assignment of therapy homework in behavioral treatments: Algorithm development and validation. *JMIR Formative Research*, 7:e45156.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. *Preprint*, arXiv:2108.06130.

Rohan Taori, Ishaan Gulrajani, Ting Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

John Torous, Sandra Bucci, Imogen H Bell, Lars V Kessing, Maria Faurholt-Jepsen, Pauline Whelan, Andre F Carvalho, Matcheri Keshavan, Jake Linardon, and Joseph Firth. 2021. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, 20(3):318–335.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Alastair C van Heerden, Julia R Pozuelo, and Brandon A Kohrt. 2023. Global mental health services and the impact of artificial intelligence–powered large language models. *JAMA psychiatry*, 80(7):662–664.

Xiao Wang, Kai Liu, and Chunlei Wang. 2023. Knowledge-enhanced pre-training large language model for depression diagnosis and treatment. In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pages 532–536. IEEE.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, and Sophia Ananiadou. 2023. MentalLLaMA: Interpretable mental health analysis on social media with large language models.

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

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

Xinyao Zhang, Michael Tanana, Lauren Weitzman, Shrikanth Narayanan, David Atkins, and Zac Imel. 2023b. You never know what you are going to get: Large-scale assessment of therapists' supportive counseling skill use. *Psychotherapy*, 60(2):149.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Prompts

A.1 Query Generation Prompt

The prompt is used for generating user queries in a mental health counseling setting using GPT-3 Turbo under the Airoboros framework.

Prompt for Generating Mental Health Counseling Conversations

Please help me create a list of {batch_size} messages that simulate what a patient might say in a conversation with a mental health professional during a counseling session and each has at least 300 words. The list of messages should contain a variety of types of patients' description of experience, feelings, behaviors, questions, and all the details that may be shared with a mental health professional.

Each message must cover all of the following requirements:
1. Patient's goal they hope to achieve through the counseling session.
2. Patient's description of their emotions and thoughts, the possible reasons triggered the symptoms.
3. Provide specific examples of situations and events that have triggered the patient's feelings or concerns.
4. Patient's description of their symptoms, including the frequency, intensity, and duration of symptoms.
5. Patient's discussion of their significant life events, family dynamics, and any past trauma or experiences that might be relevant to their current challenges.
6. Describe any coping strategies if applicable.
7. Ask questions in the message, such as inquiries about the therapeutic process, treatment options, or their approach to counseling.

Make the messages as specific and detailed as possible. Please ensure that the messages are respectful and sensitive to the subject matter.

Topics: {topics}

Figure 4: Prompt for generating user queries in a mental health counseling setting using GPT-3 Turbo.

A.2 Rubrics for LLM Judges

This rubric is provided to GPT-4 Turbo Preview and Gemini Pro 1.0 as the standard rating guidelines during evaluation.

Scoring Rubrics for LLM Judges

Please follow the standard of the scoring:

1: The response completely fails to address the metric, showing a total disregard for the user's needs or concerns in this area.
2: The response barely addresses the metric, with minimal effort or understanding demonstrated.
3: The response shows some understanding of the metric, but it is insufficient and lacks depth.
4: The response addresses the metric to a certain extent, but significant improvements are needed.
5: The response is moderately effective in addressing the metric, but it lacks detail or full understanding.
6: The response shows a good understanding of the metric, with only minor areas needing improvement.
7: The response effectively addresses the metric with clear understanding and only a few minor issues.
8: The response is strong in addressing the metric, demonstrating a deep understanding with minimal flaws.
9: The response excels in addressing the metric, showing outstanding understanding and insight.
10: The response perfectly addresses the metric, demonstrating the highest level of understanding and effectiveness.

Figure 5: Scoring Rubrics for LLM Judges

A.3 Prompt for Paraphrasing the Interview Data

This prompt is used for summarizing and paraphrasing the transcripts of interview into conversations between patients and therapists using open-sourced LLM, Mistral-7B-Instruct-v0.1.

Prompt for Paraphrase Interview Data

You are given a transcript of a one-page conversation between a mental health counselor and a patient in a hospital setting. Your task is to summarize this transcript into a single round of conversation, focusing on the most crucial issue discussed.

The summary should consist of exactly one round of conversation starting with the description from the patient and followed by the feedback from the counselor. Each of these (the patient's description and the counselor's feedback) must be more than 50 words, richly encapsulating the patient's situation, emotions, and background leading to the problem, as well as the counselor's professional guidance, strategy, and ethical considerations. Aim to capture the essence of the counseling session, highlighting the central ideas and issues in a clear, logical, and understandable manner.

The output must strictly follow the format below:
Patient: [patient's query from first-person view]
Counselor: [counselor's feedback from first-person view]

Each turn must be more than 50 words.

Transcript:{transcript}

Response:

Figure 6: Prompt for Paraphrasing the Interview Data

B Visualization of Results

In this section, we provide visualizations of the results in Table 2. Each subsection contains the results for a single metric. Each bar in the plots represents the metric score for one version of a LLM among Base model, model fine-tuned with synthetic data, model fine-tuned with interview data, model fine-tuned with synthetic and interview data, and baseline model. Plots with orange and red bars illustrate scores rated by GPT-4 Turbo Preview, while plots with green and blue bars illustrate scores rated by Gemini Pro 1.0.

B.1 Active Listening

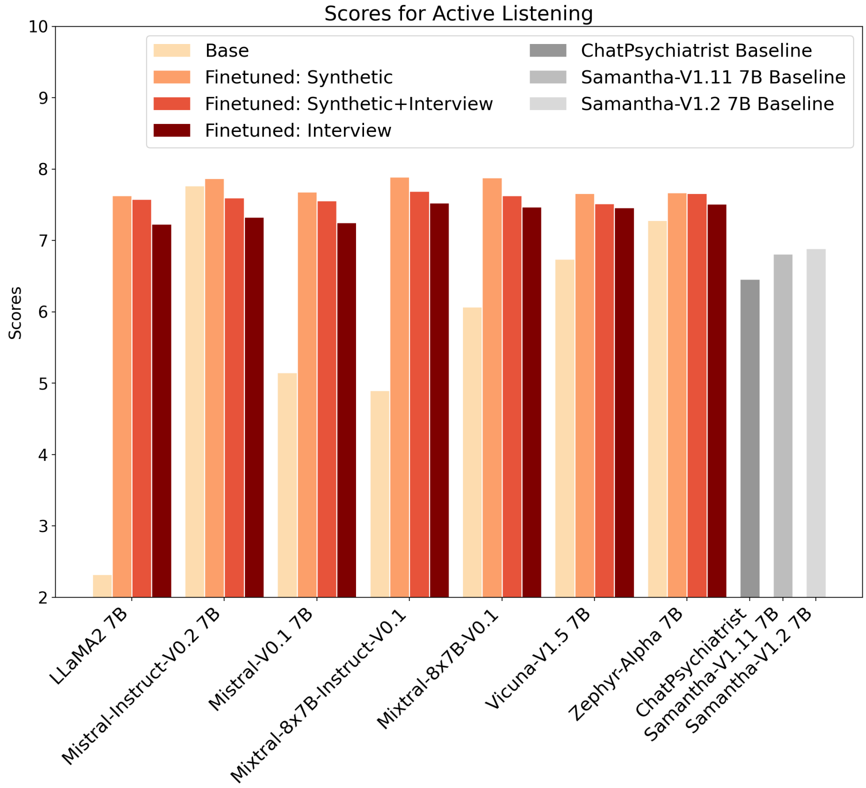

Figure 7: Active Listening Scores Rated by GPT-4 Turbo Preview.

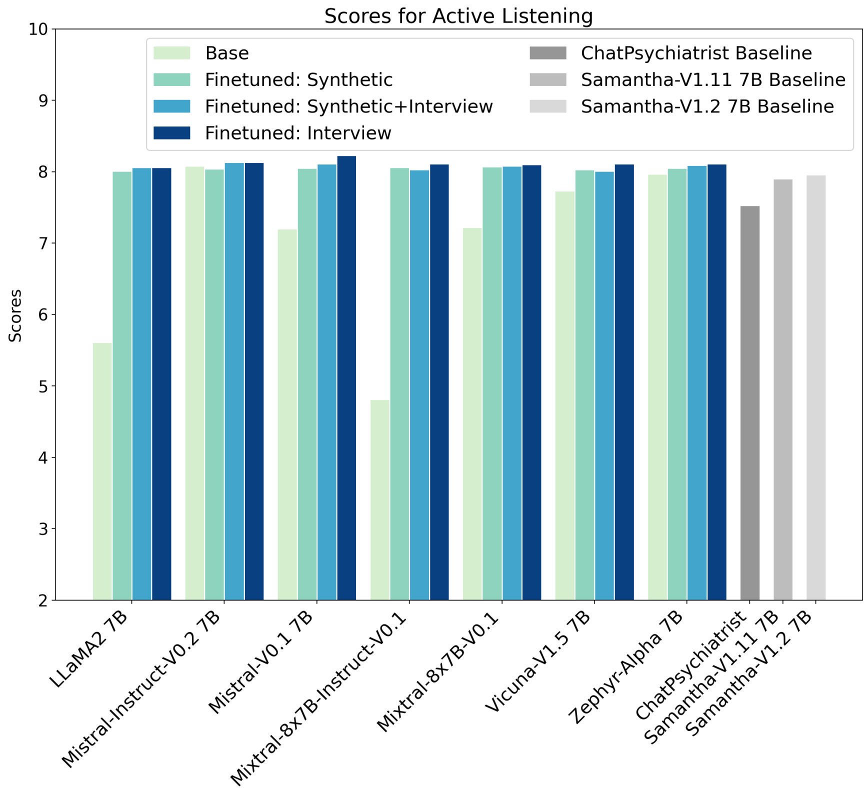

Figure 8: Active Listening Scores Rated by Gemini Pro 1.0.

B.2 Empathy & Validation

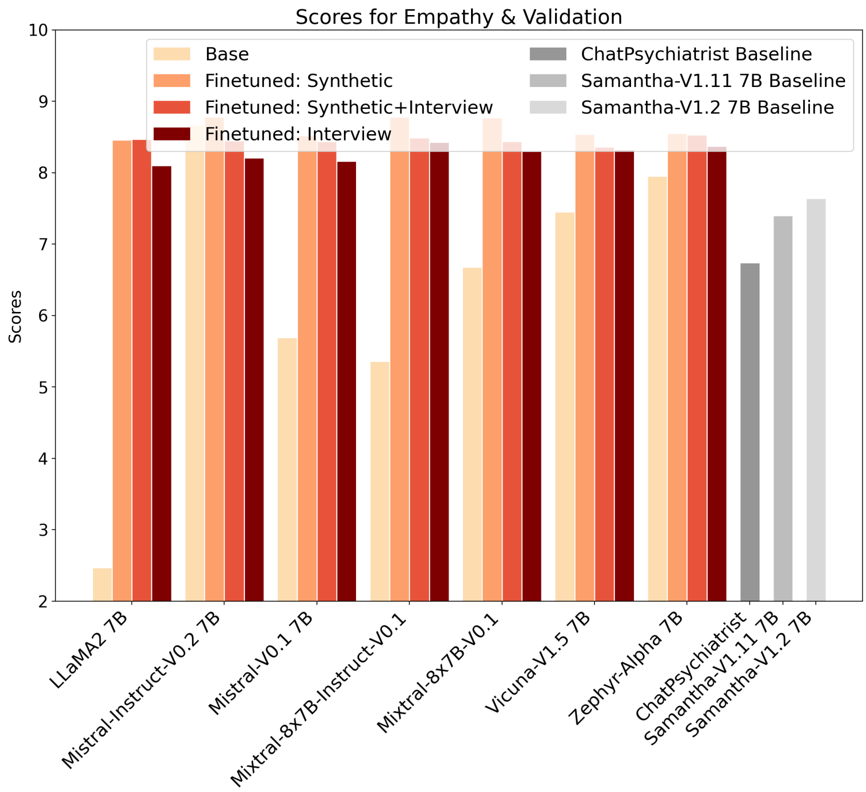

Figure 9: Empathy & Validation Scores Rated by GPT-4 Turbo Preview

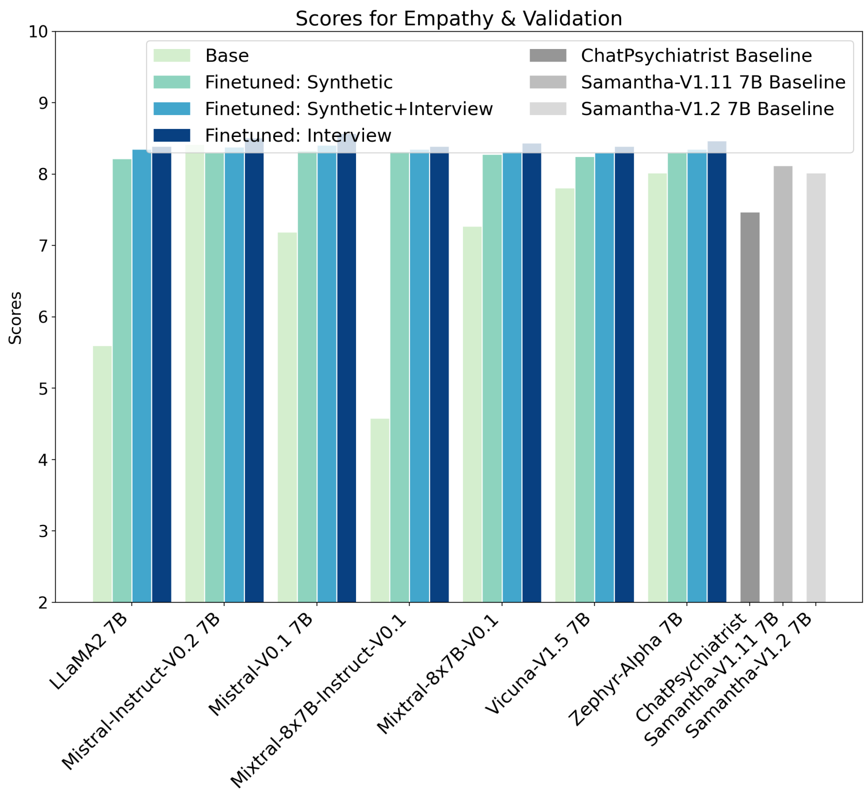

Figure 10: Empathy & Validation Scores Rated by Gemini Pro 1.0.

B.3 Safety & Trustworthiness

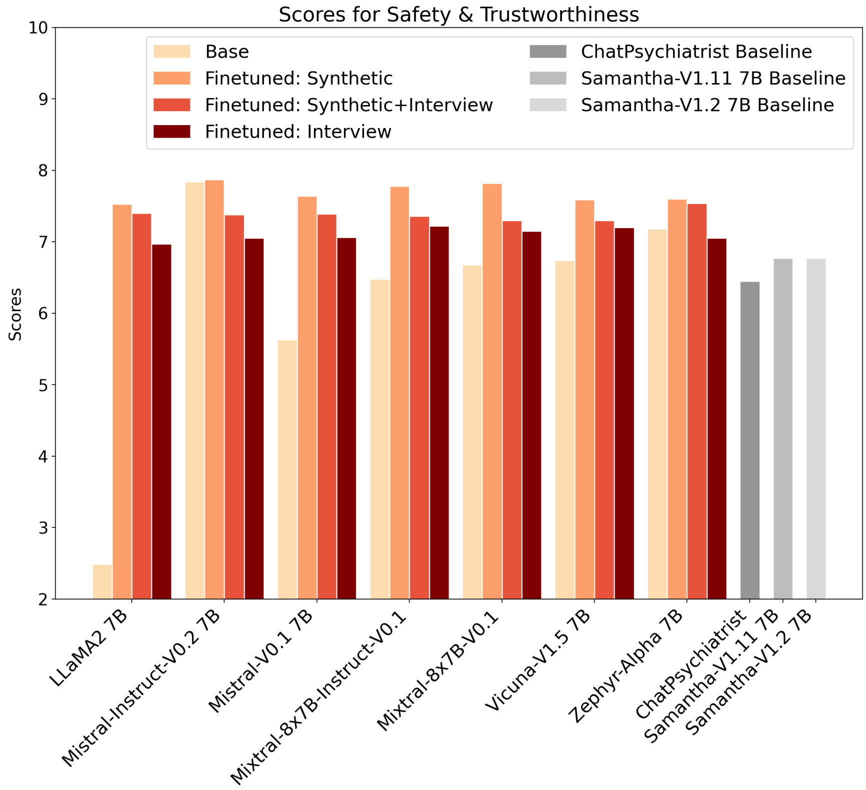

Figure 11: Safety & Trustworthiness Scores Rated by GPT-4 Turbo Preview

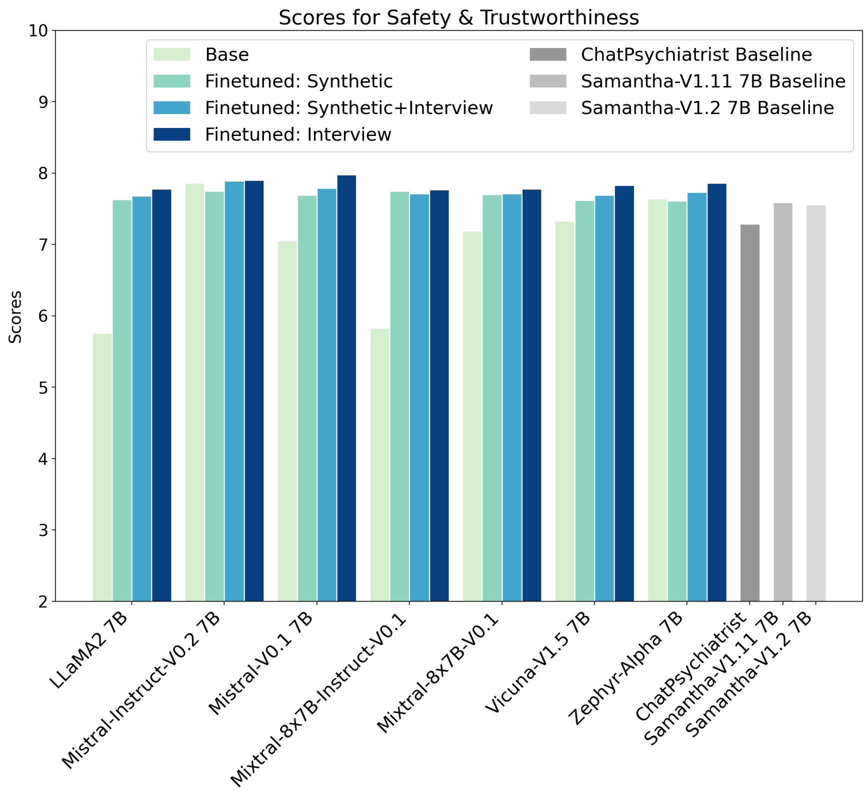

Figure 12: Safety & Trustworthiness Scores Rated by Gemini Pro 1.0.

B.4 Open-mindedness & Non-judgment

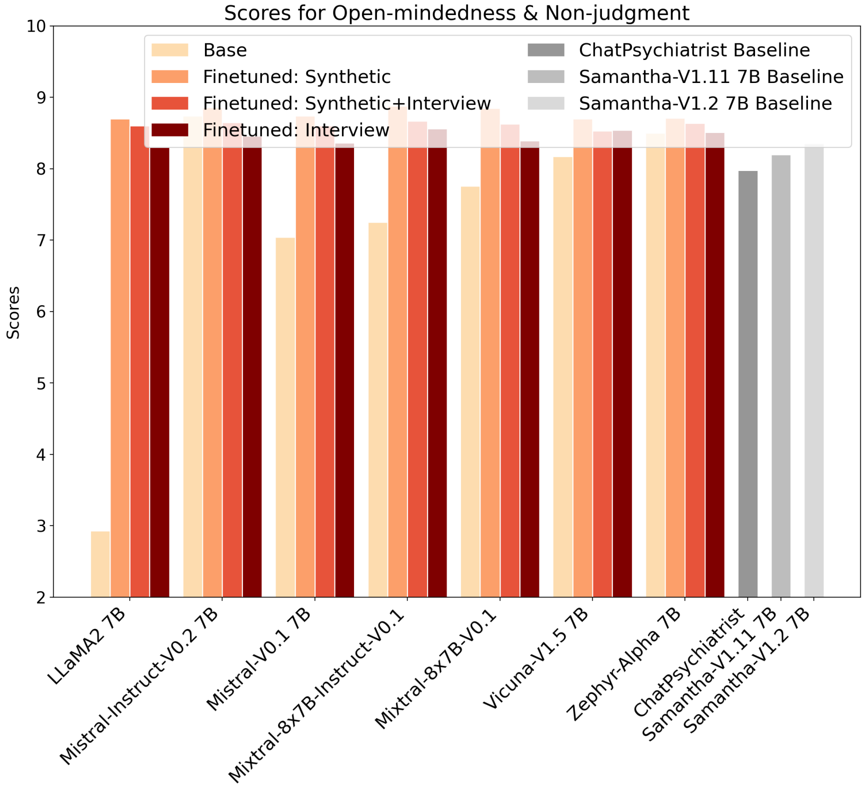

Figure 13: Open-mindedness & Non-judgment Scores Rated by GPT-4 Turbo Preview

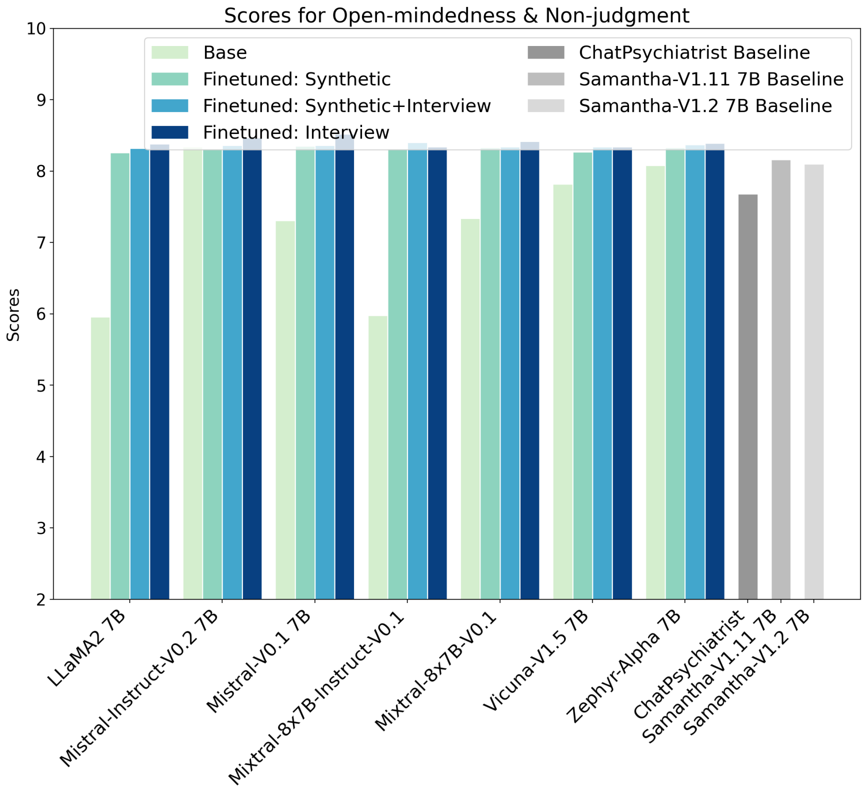

Figure 14: Open-mindedness & Non-judgment Scores Rated by Gemini Pro 1.0.

B.5 Clarity & Encouragement

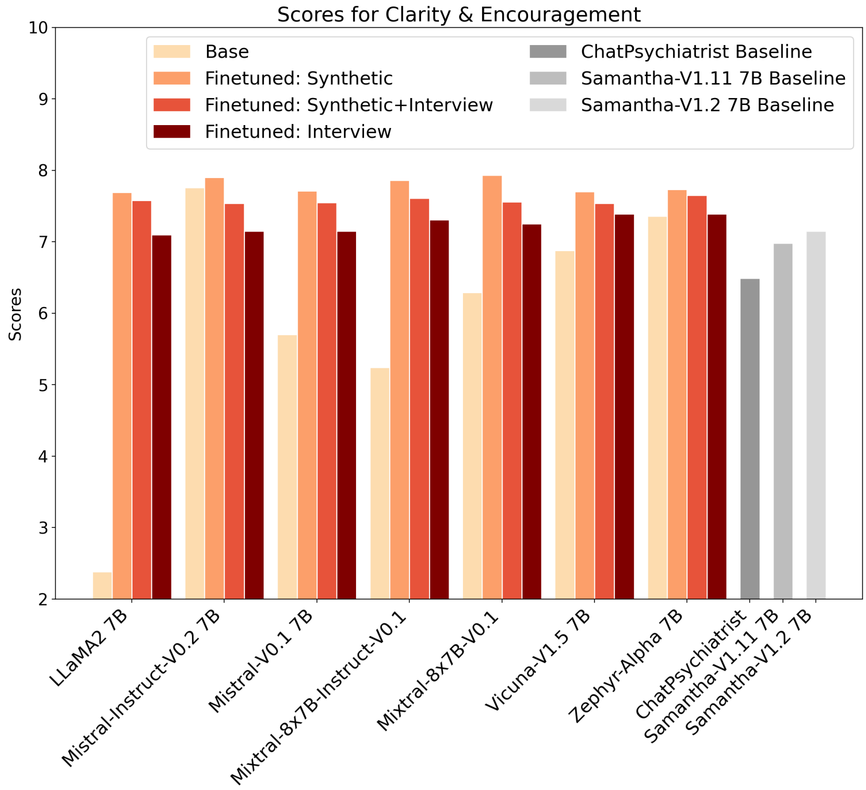

Figure 15: Clarity & Encouragement Scores Rated by GPT-4 Turbo Preview

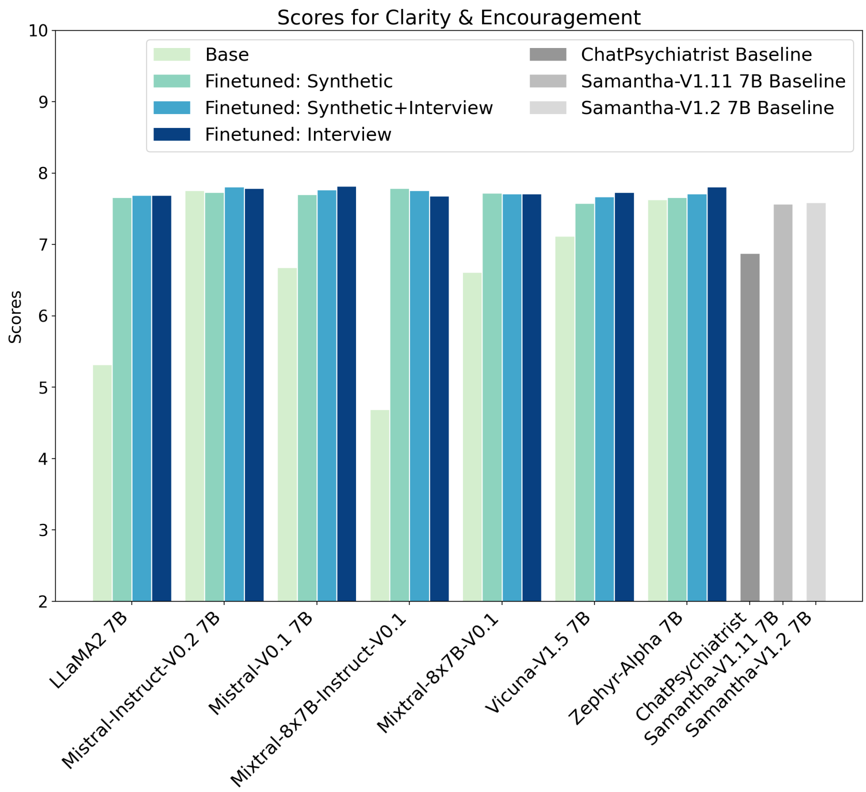

Figure 16: Clarity & Encouragement Scores Rated by Gemini Pro 1.0.

B.6 Boundaries & Ethical

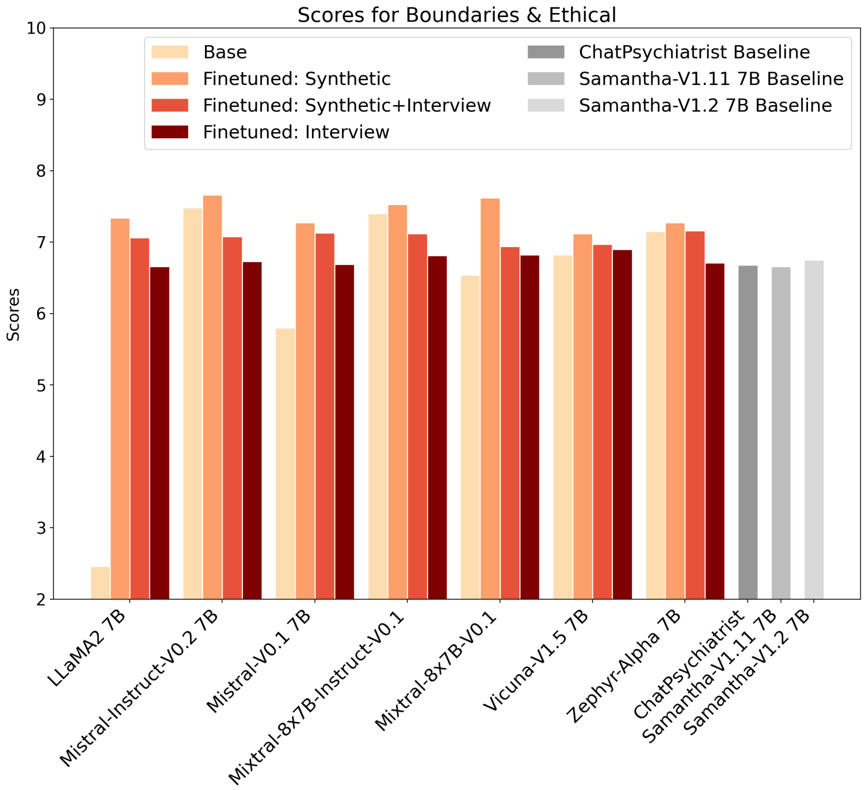

Figure 17: Boundaries & Ethical Scores Rated by GPT-4 Turbo Preview

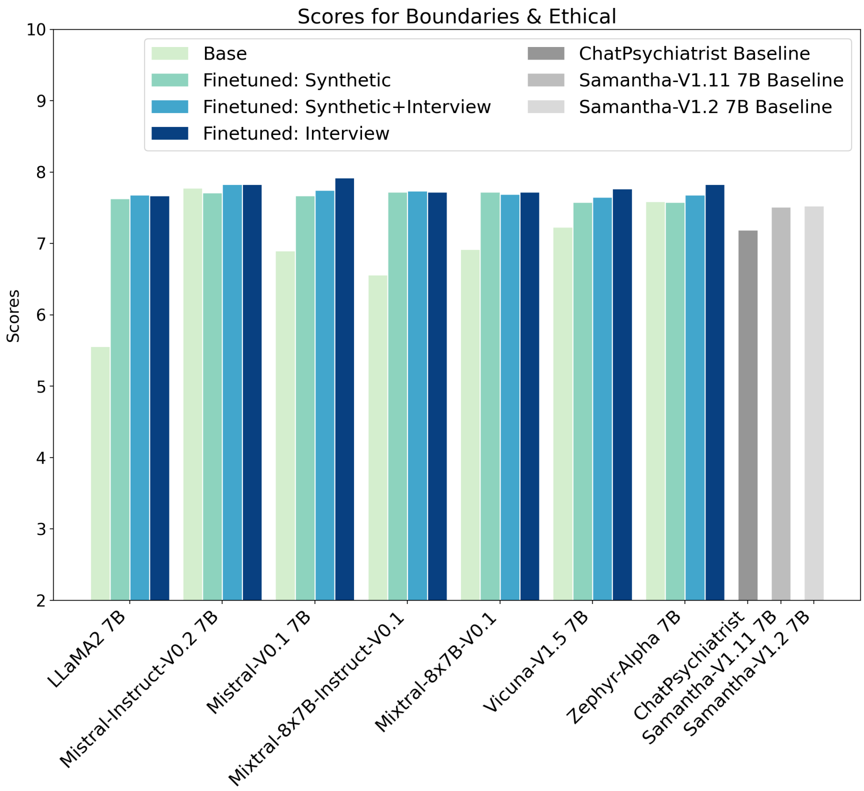

Figure 18: Boundaries & Ethical Scores Rated by Gemini Pro 1.0.

B.7 Holistic Approach

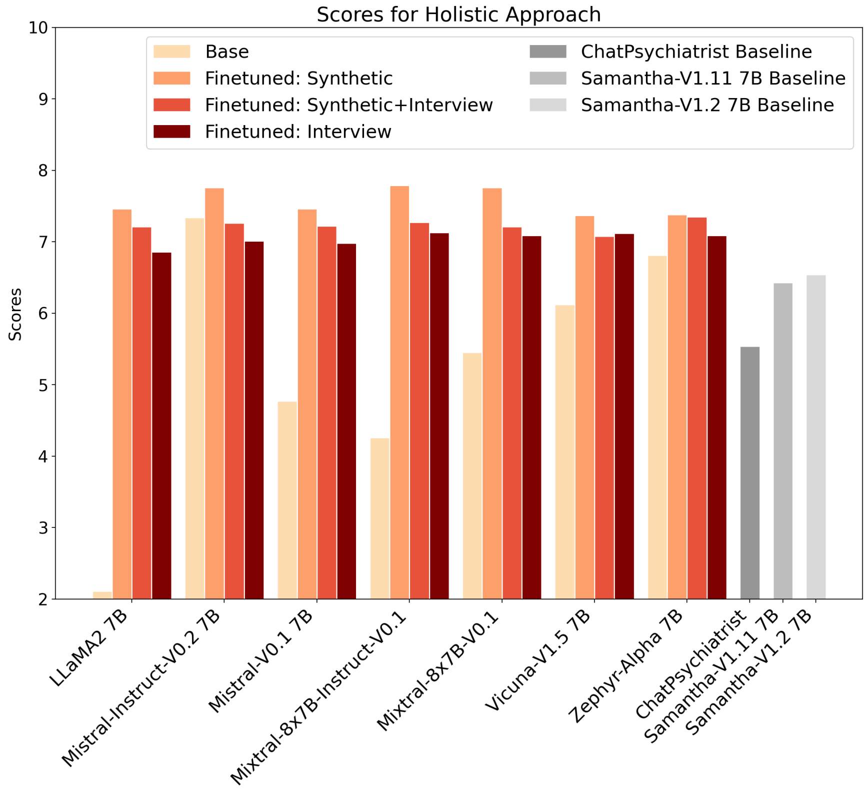

Figure 19: Holistic Approach Scores Rated by GPT-4 Turbo Preview

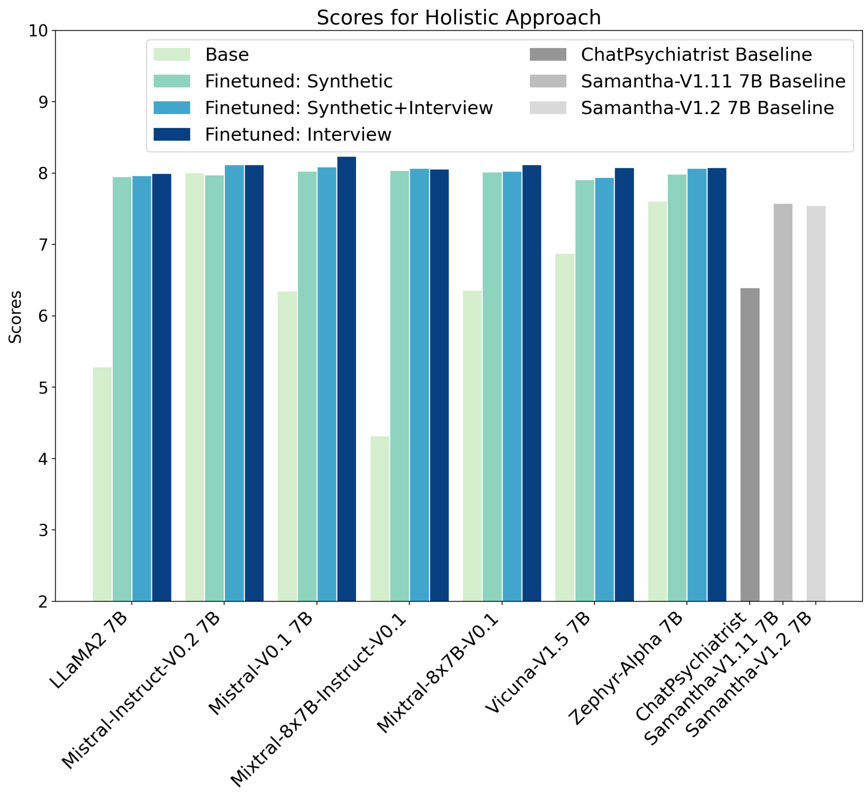

Figure 20: Holistic Approach Scores Rated by Gemini Pro 1.0.

C Hyperparameters

This section details the LLM hyperparameters and QLoRA hyperparameters we used in the training process. The fine-tuning framework is adapted from the QLoRA GitHub Repository (https://github.com/artidoro/qlora).

Table 3: Hyperparameters used during fine-tuning.

Hyperparameter	Value	Description
epoch	5	Number of training epochs
optim	paged_adamw_32bit	The optimizer to be used
per_device_train_batch_size	8	The training batch size per GPU
gradient_accumulation_steps	8	How many gradients to accumulate before performing an optimizer step
weight_decay	0.01	The L2 weight decay rate of AdamW
learning_rate	0.0002	The learning rate
max_grad_norm	0.3	Gradient clipping max norm
warmup_ratio	0.03	Fraction of steps to do a warmup for
source_max_len	512	Maximum source sequence length. Sequences will be right padded (and possibly truncated)
target_max_len	1024	Maximum target sequence length. Sequences will be right padded (and possibly truncated)
max_new_tokens	1024	Maximum number of new tokens to be generated in evaluation or prediction loops
temperature	1.0	Temperature controls the randomness of the generated text ranging from 0 (for deterministic output) to infinity (for maximum randomness).
top_k	50	Top_k limits the number of top tokens considered during sampling in LLMs
top_p	1.0	Top_p sets a threshold for the cumulative probability distribution of tokens
double_quant	True	Compress the quantization statistics through double quantization
quant_type	nf4	Quantization data type to use
bits	4	How many bits to use
lora_r	64	LoRA r defines the rank of the low-rank matrices that approximate the updates
lora_alpha	16	LoRA alpha determines the magnitude of impact of updates on the original weights of the pre-trained model
lora_dropout	0.1	LoRA dropout rate
lora_modules	all	The names of the modules to apply the adapter to

D API Cost

Table 4 details the OpenAI API usage and cost for data generation and model performance evaluation.

Table 4: Number of tokens, number of API requests, and total $USD spending.

API	Instruction	Tokens	Requests	$USD
gpt-3.5-turbo	Text generation	35,000,000	40,000	35
gpt-4	Text generation	10,360,888	6200	450