Published in final edited form as:
KDD. 2025 August ; 2025: 5367–5378. doi:10.1145/3711896.3737393.

MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance

Jia Xu[*,†],
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Tianyi Wei[*,‡],
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Bojian Hou[*],
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Patryk Orzechowski[§],
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Shu Yang,
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Ruochen Jin,
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Rachael Paulbeck,
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Joost Wagenaar,
University of Pennsylvania, Philadelphia, Pennsylvania, USA

George Demiris,
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Li Shen[¶]
University of Pennsylvania, Philadelphia, Pennsylvania, USA

[†]The Roux Institute, Northeastern University, Portland, Maine, USA (second affiliation)

[‡]The Rockefeller University, New York, NY, USA (second affiliation)

[§]AGH University of Krakow, Kraków, Poland (second affiliation)

Abstract

We introduce MentalChat16K, an English benchmark dataset combining a synthetic mental health counseling dataset and a dataset of anonymized transcripts from interventions between Behavioral

[¶]Corresponding author: li.shen@pennmedicine.upenn.edu.
[*]Equal contributions

Health Coaches and Caregivers of patients in palliative or hospice care. Covering a diverse range of conditions like depression, anxiety, and grief, this curated dataset is designed to facilitate the development and evaluation of large language models for conversational mental health assistance. By providing a high-quality resource tailored to this critical domain, MentalChat16K aims to advance research on empathetic, personalized AI solutions to improve access to mental health support services. The dataset prioritizes patient privacy, ethical considerations, and responsible data usage. MentalChat16K presents a valuable opportunity for the research community to innovate AI technologies that can positively impact mental well-being. The dataset is available at https://huggingface.co/datasets/ShenLab/MentalChat16K and the code and documentation are hosted on GitHub at https://github.com/PennShenLab/MentalChat16K.

Keywords

Conversational AI; Large Language Models; LLM Fine-tuning; Mental Health; QLoRA

1 Introduction

The proliferation of Large Language Models (LLMs) has transformed artificial intelligence’s capability to understand and generate human-like text, unlocking new opportunities for applications in various domains, including mental health support [25, 64]. This advancement is particularly timely, given the rising global prevalence of mental health disorders such as depression and anxiety [5, 47], highlighting an urgent need for innovative and accessible support solutions [22, 24, 35, 60]. Notably, most recent studies have demonstrated that LLMs, even at the 7B scale, are capable of generating empathic responses that can be more empathic than human-written responses, thus enhancing human peer support in contexts where empathy is crucial [36, 68].

Recent years have witnessed the emergence of several AI models aimed at addressing mental health challenges, including Psy-LLM [32], Mental-LLM [67], ChatPsychiatrist [39], and MentalBERT [27]. Despite the advancements in the field, there is a notable paucity of LLMs that concentrate on mental health counseling. Among the aforementioned works, ChatPsychiatrist is the only open-source English LLM focusing on question-answering in psychological consultation settings. The majority of existing work has primarily emphasized mental health detection, diagnosis, and prediction [27, 67]. This disparity can be attributed to several obstacles, including language limitations, a scarcity of domain-specific training data, and privacy concerns surrounding the use of such data.

To overcome these challenges and advance research in this critical domain, we introduce *MentalChat16K*, an English benchmark dataset consisting of 16K question-answer pairs of synthetic mental health counseling conversations and anonymized interview conversations from interventions between Behavioral Health Coaches and Caregivers of patients in palliative or hospice care. This curated dataset is available at https://huggingface.co/datasets/ShenLab/MentalChat16K. It covers a diverse range of conditions, including depression, anxiety, and grief, enabling the development and evaluation of LLMs tailored for conversational mental health assistance. Notably, MentalChat16K is twice the size of the training data for ChatPsychiatrist [39], providing a broader and deeper coverage of real-life mental health issues to enhance the capabilities of AI models in offering comprehensive and empathetic support.

MentalChat16K presents a valuable opportunity for the research community to innovate AI technologies that can positively impact mental well-being. By leveraging this dataset, researchers can fine-tune and evaluate LLMs, enabling the development of empathetic and personalized AI solutions that can engage in warm and nuanced interactions, simulating the communication typically expected in human counseling sessions. The dataset prioritizes patient privacy, ethical considerations, and responsible data usage, ensuring that the development of AI technologies in this sensitive area is conducted with the utmost care and responsibility.

To demonstrate the effectiveness of the MentalChat16K dataset in tailoring LLMs for mental health counseling, We fine-tuned seven state-of-the-art LLMs. We employed the Quantized Low-Rank Adapter (QLoRA) technique [15] for efficient fine-tuning of state-of-the-art LLMs, thereby reducing computational demands without sacrificing model performance. We mainly focus on 7B local open-source models because we want to demonstrate a fine-tuned pipeline with limited resources (such as one single A40 or A100 GPU). There are mainly two families of local LLMs in the market, one is the LLaMA family including LLaMA, LLaMA2, Alpaca, Vicuna etc. and the other is the Mistral family including Mistral, Mixtral, Zephyr, etc. Mistral 7B[1] has been validated as one of the most powerful local open-sourced models.

To evaluate the ability of LLMs fine-tuned on MentalChat16K, 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 [45] and Gemini Pro 1.0 [57] as impartial judges. We also incorporated real human evaluations for a more comprehensive and convincing comparison. Our four evaluators include one senior Postdoc, and three Master students, providing interdisciplinary expertise in both computer science and medical sciences, making them well-suited to assess the technical and healthcare aspects of the models' responses. The human evaluation results are consistent with the evaluation results of GPT-4 and Gemini Pro. The complete architecture is summarized in Figure 1.

In summary, our contributions are three-fold:

- We introduce MentalChat16K, a benchmark dataset that contains anonymized transcripts from interventions between Behavioral Health Coaches and Caregivers of patients in palliative or hospice care. This dataset can be used for fine-tuning pre-trained large language models to provide empathetic, personalized AI solutions to improve access to mental health support services.

[1] https://mistral.ai/news/announcing-mistral-7b/

- We curate a synthetic counseling conversation dataset covering a broad range of topics in mental health such as depression and anxiety. This synthetic data works as a complimentary of the real dataset and composes the complete MentalChat16K benchmark together with the real dataset.
- We provide a pipeline for data collection, data filtering, LLMs fine-tuning, and evaluation. Our extensive experiments demonstrate that the fine-tuned LLMs on the MentalChat16K dataset outperform existing models in providing mental health support, validating the effectiveness of MentalChat16K. This pipeline also serves as a valuable demo for institutions lacking computing resources, enabling them to fine-tune their own large language models.

2 Related Work

Mental Health

Mental health disorders like depression and anxiety have a profound impact, leading to substantial challenges and socio-economic consequences. The global economy faces an estimated annual productivity loss of $1 trillion due to these disorders [58]. Depression prevalence among older adults ranges from 7.2% to 49% [16], even higher than dementia [1]. AI integration in healthcare, especially through LLMs like GPT, LLaMA, and BERT, offers promising prospects for innovative mental health solutions [20, 67, 69].

LLMs in Mental Health Care

Depression is the leading cause of disability globally [59]. LLMs, including GPT3.5, GPT4, LLaMA1, and LLaMA2, have transformed mental health care with their ability to grasp natural language context and produce human-like outputs [14]. Researchers have integrated open-source LLMs into mental health chatbots like ChatPsychiatrist [39], MentalBERT [27], Mental-LLM [67], and Psy-LLM [32]. LLMs have been employed in various mental health tasks, such as suicide risk detection [6], psychotherapy homework assignment [48], and emotion recognition [70]. They have also aided non-professional counselors [19] and supported depression diagnosis and treatment [65].

Benchmark Datasets

Benchmark datasets are crucial for advancing NLP research in mental health. Althoff et al. (2016) presented a large-scale, quantitative study on the SNAP dataset [2], a text-message-based counseling conversation dataset containing over 13 million messages. The PsyQA dataset contains Chinese counseling conversations, and Na et al. used CBT prompts with GPT-3.5-turbo-16k to generate CBT-informed responses [44, 55]. CounselChat [7] includes 3.6k questions and answers from online counseling platforms. The HOPE dataset [42] includes 12.9k annotated utterances from counseling session videos for dialog-act classification, and the MEMO dataset annotates these for mental health counseling summarization [54]. ChatPsychiatrist's Psych8K dataset [39] comprises data from 260 real counseling recordings. Our MentalChat16K dataset includes face-to-face or video conference conversations, encompassing verbal and non-verbal interactions. These datasets support advancing NLP applications for mental health [39].

3 Approach

This section outlines the methodologies utilized for curating the MentalChat16K datasets, as well as the approaches used for fine-tuning and evaluating LLMs using MentalChat16K. Figure 1 illustrates our pipeline from data collection to model evaluation.

3.1 Data Collection and Processing

MentalChat16K consists of two datasets. One is the real anonymized interview transcripts between behavioral health coaches and caregivers, and the other is a synthetic mental health counseling conversation dataset generated by GPT-3.5 Turbo [46]. We have provided detailed statistics of both datasets to facilitate understanding and utilization. Table 1 summarizes the key statistics of the MentalChat16K dataset.

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 behavior health coaches and caregivers of individuals in palliative or hospice care. The clinical trial, entitled PISCES (Problem-solving Intervention to Support Caregivers in End of Life Care Settings), aims to test a problem-solving therapy intervention to support the emotional needs of hospice caregivers. Upon consent to participate in the study, 514 caregivers were enrolled to randomly receive intervention either face-to-face with a behavioral health coach or via video conference. Figure 2 shows that each caregiver has three formal and one exit visit with a behavioral health coach, generating interview audio files transcribed into text by human experts, ranging from brief greetings to dialogues with filler words. As a token of appreciation for their participation, caregivers receive a $50 reloadable gift card upon completion of the intervention, and $25 upon completion of the 40-day follow-up assessment.

Demographic information was collected from 421 caregivers who completed the information survey, providing insights into their backgrounds. Specifically, the majority of caregivers are female, with 415 out of 421 total participants of the survey. Among female caregivers, White Caucasians constitute the largest group, making up approximately 88% (366 out of 415), with a small proportion identifying as Hispanic (less than 1%). Male caregivers are a small minority, totaling 6, with White Caucasians again being the predominant group. Other racial categories include Asian American, Black/African American, Multi-racial, and others, each comprising a small proportion of the caregiver population. For more details, please refer to Table 4 in the Appendix. The demographic distribution also reflects a real situation in the real world [10]. Additionally, the skew observed in our dataset aligns with broader trends [9] in hospice care and research participation.

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 [28] model, which is a state-of-the-art lightweight LLM to paraphrase and summarize interview transcripts documents. We fed each page of the 378 transcripts into the model and provided instructions to summarize the page into a single round of conversation between the caregiver and the behavioral health coach. Subsequently, we filtered out any conversations with less than 40 words in the question and answer, resulting in a total of 6,338 question-answer pairs. To ensure privacy and confidentiality, we conducted a manual inspection of the paraphrased transcript to remove any sensitive and identifiable information such as name, address, financial information, and etc. The consent to release the paraphrased transcript data is obtained from the anonymous research group upon removal of all sensitive and identifiable information.

3.1.2 Synthetic Data. To enrich our dataset with diverse therapeutic dialogues, we used the OpenAI GPT-3.5 Turbo [46] model to generate 9,775 question-answer pairs with a customized adaptation of the Airoboros self-generation framework[2]. Under the Airoboros framework, we customized a new prompt (see Table 6) to provide clear instructions to generate patient queries using GPT-3.5 Turbo. 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 (Appendix A.1.2) that typically arises in a counseling session according to the CounselChat [7] platform was specified in the prompt. We additionally conducted a random review of 100 synthetic QA pairs to ensure that each question and answer displayed an authentic counseling style and provided guidance suitable for the domain. This combined approach ensures the synthetic conversations authentically mimic the complexity and diversity of therapist-client interactions, thereby exposes our models to a wide spectrum of psychological conditions and therapeutic strategies.

3.2 Fine-tuning and Inference

To perform efficient fine-tuning by using only one A40 or A100 GPU that is more affordable, we adopt Quantized Low Rank Adaptation (QLoRA) [15].

The inference stage involved using both the fine-tuned and base models, alongside baseline models (Samantha v1.11 and v1.2 [13], ChatPsychiatrist [39]), to generate responses to 200 sampled questions. These questions were collected from Reddit [26] and the Mental Health Forum [18], representing a wide range of real-world inquiries in a therapeutic setting. In addition to the questions, the models were given explicit instructions as follows.

> “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”.

3.3 Evaluation

To ensure a comprehensive and unbiased evaluation of our model in mental health counseling, we employed both LLM evaluators and human evaluators. Combining these approaches mitigates individual biases and leverages the strengths of both automated and human judgment. The use of powerful third-party LLMs for evaluation is a well-established methodology in the field [4, 11, 66]. This dual evaluation framework enhances the reliability of our results and ensures a robust assessment of counseling responses.

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

We employed GPT-4 Turbo [45] and Gemini Pro 1.0 [57] as robust and scalable judges for automated LLM evaluation. We deliberately chose two distinct and state-of-the-art LLMs for evaluation as divergent architectures help mitigate model-specific biases.. We utilized the LLM Judge framework [71] to generate judgments and ratings that assess the quality of the models' responses to the benchmark questions we collected. We instructed GPT-4 Turbo and Gemini Pro 1.0 to be objective and assess the response based on 7 devised mental health metrics (see Table 2). Our proposed seven metrics for evaluating therapeutic dialogue systems are grounded in both established therapeutic practices and recent advancements in AI-based mental health support evaluation. These metrics synthesize multiple validated frameworks, primarily building upon ChatCounselor's [39] evaluation methodology while incorporating insights from contemporary research. The Active Listening metric, derived from Miller and Moyers' [43] foundational work and validated in AI contexts through PsyQA [55], assesses the system's comprehension and reflection capabilities. Empathy & Validation draws from EPITOME's [53] empirically validated empathy metrics and Rogers' [52] therapeutic principles. Safety & Trustworthiness incorporates criteria from Dialogue Safety [49] and PsyEval [38], addressing critical aspects of therapeutic interaction safety. The remaining metrics—Open-mindedness & Non-judgment, Clarity & Encouragement, Boundaries & Ethical, and Holistic Approach—integrate established therapeutic principles [30, 52] with recent frameworks from PsyQA [55]. This comprehensive framework ensures thorough evaluation of AI systems' therapeutic capabilities while maintaining alignment with professional standards [3] and empirically validated assessment approaches.

The judge models were tasked to rate each response for each metric on a scale ranging from 1 to 10. In addition, we asked the judge models to justify their ratings and make comments on the model responses. To demonstrate the significance of our results, we conducted statistical analyses by randomly selecting 50 questions from the original 200 test questions and running five rounds of inference on both fine-tuned and base models. We compared the average scores across all five rounds on each of the seven metrics between the fine-tuned and base models through a two-sample t-test with a 0.95 confidence interval. The null hypothesis is that the scores for the fine-tuned and the base models have identical average (expected) values across the specified metrics.

To incorporate human evaluation of the model performance, we invited a senior Postdoc and three Master's students who possessed interdisciplinary expertise in both computer science and medical sciences to compare the responses generated by the base models, fine-tuned models, and baseline models. For each input question, the participants ranked the responses from the following models: (1) the base model, (2) the base model fine-tuned on synthetic data, (3) the base model fine-tuned on interview data, (4) the base model fine-tuned on both synthetic and interview data, and three baseline models: (5) Samantha-1.1, (6) Samantha-1.2, and (7) ChatPsychiatrist. The responses were ranked from 1 to 7, with 1 being the most effective response and 7 being the least effective. For each model, the final ranking results were calculated as the average ranking across 50 input questions randomly sampled using a fixed seed of 42 to ensure reproducibility from the 200-question evaluation dataset described in Section 3.2. The difference in scoring scales between LLM-based and human evaluators arises from the nature of their distinct evaluation methodologies. LLM scoring provides objective and granular insight into response quality, while human ranking captures intuitive and comparative judgments, which also help minimize bias and simplify the evaluation tasks. Together, they offer a holistic view of model performance.

To make the human evaluation more reliable, we calculate the Cohen's Kappa score [34] to assess the consistency among the evaluators. Specifically, we randomly selected 30 questions from the 200 evaluation questions and had responses generated by 7 models (3 global baselines including Samantha-1.11, Samantha-1.2 and ChatPsychiatrist, 1 randomly selected base model before fine-tuning such as Zephyr-Alpha, Vicuna-7B-V1.5, LLaMA2–7B, Mistral-7B-V0.1, Mistral-7B-Instruct-V0.2, Mixtral-8×7B-V0.1, Mixtral-8×7B-Instruct-V0.1, and its corresponding 3 fine-tuned models fine-tuned on synthetic, interview and both data respectively). Four human evaluators ranked these responses from 1 (best) to 7 (worst). By treating each rank as the prediction target, we make the task become a 7-class classification problem and each human evaluator will generate a list of predictions for the 30 questions. We calculate Cohen's Kappa score among all the human evaluators' lists of predictions. The resulting agreement is 0.441, which is larger than the acceptable threshold of 0.4 as indicated by [34].

4 Experiments

Our study aims to investigate the effectiveness of fine-tuning LLMs using MentalChat16K for mental health counseling. By fine-tuning LLMs with this specialized dataset, we aim to enhance the models' capacity to generate empathetic, relevant, and contextually appropriate responses in mental health counseling scenarios. This section details the methodology, implementation, and evaluation metrics employed to assess the performance improvements of the fine-tuned models.

4.1 Baseline Models

We selected three baseline models, chosen for their relevance and pioneering contributions to AI-assisted mental health support, setting a benchmark for our fine-tuned models' comparative analysis.

ChatPsychiatrist [39] is an instruction-tuned LLM fine-tuned on LLaMA-7B [61] using the Psych8k dataset, composed of authentic dialogues between clients and psychologists. This model outperformed other open-source solutions such as Alpaca-7B [56], LLaMA-7B [62], and ChatGLMv2-6B [17] on the counseling Bench the authors devised.

Samantha-v1.11/v1.2 [13] are open-source models hosted on Hugging Face, fine-tuned on the LLaMA-2–7B [62] and Mistral-7B [28] respectively. Unique for their training in philosophy, psychology, and personal relationships, Samantha models are designed as sentient companions.

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 [62] 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* [28] is a pre-trained LLM engineered for superior performance and efficiency. It outperforms LLaMA2–13B across all tested benchmarks. *Mixtral-8×7B-v0.1* [29] is an advanced generative Sparse Mixture of Experts model. It outperforms LLaMA2–70B on most benchmarks tested. *Mistral-7B-Instruct-v0.2* [28] and *Mixtral-8×7B-Instruct-v0.1* [29] are instruction fine-tuned versions of *Mistral-7B-v0.1* and *Mixtral-8×7B-v0.1*, trained on a variety of publicly available conversation datasets.

Vicuna-7B-v1.5 [41] 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 [63] 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 [50].

4.3 Metrics

In the current landscape of LLM evaluation, several metrics dominate the literature. Common performance measures include perplexity, accuracy [12, 23], semantic similarity [8, 51], and human evaluation metrics such as fluency, coherence, and relevance [11]. While these metrics offer valuable insights into general-purpose 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 [37]. Current metrics lack the specificity and sensitivity required to gauge these aspects accurately. To address this gap, we devised seven metrics (shown in Table 2) for evaluating mental health LLMs. These novel metrics 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 types of data from MentalChat16K: a real interview dataset and a synthetic dataset. We hypothesized that models fine-tuned on MentalChat16K would exhibit enhanced performance on various mental health counseling evaluation metrics, indicative of a more nuanced understanding of patient interactions. Each base model underwent fine-tuning under three distinct configurations:

- Fine-tuning with Synthetic Data: Models were fine-tuned exclusively on the synthetic dataset 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 the entire MentalChat16K dataset, testing the hypothesis that a diverse training input could yield superior performance.

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 the 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).

4.5 Results

Main Results: In Table 3, the evaluation scores reveal distinct patterns in model performance when assessed by GPT 4 Turbo, Gemini Pro, and human experts. The results show clear patterns in model performance for both evaluation methods. GPT 4, Gemini Pro and human experts evaluation 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, validating the effectiveness of our MentalChat16K.

Discussion on GPT-4 Evaluation: GPT-4'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 bias towards models fine-tuned on synthetic data may be attributed to GPT-4's alignment with data generated by GPT-3.5 Turbo, possibly introducing an intrinsic preference for similar language patterns and styles. This bias indicates that it may not be fair to merely use GPT-4 as the evaluator. Therefore, we incorporate Gemini Pro from Google as another LLM evaluator.

Discussion on Gemini Evaluation: In contrast, Gemini's evaluations, while also acknowledging the improvements brought 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, and 6 out of seven cases in terms of the seven metrics separately. Gemini Pro's evaluations suggest that the interview data contributes significantly to the models' performance in aspects that require real-world contextual understanding and nuanced human interactions. 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 Pro. Gemini Pro's evaluations suggest that while synthetic data can contribute to conversational diversity, the integration of real-world dialogues is crucial for achieving the depth of engagement and empathy required in mental health support. However, the models fine-tuned on the combined data did not consistently outperform those fine-tuned on individual datasets, indicating that simply combining synthetic and interview data may not be the most effective approach. These discrepancies between GPT-4 and Gemini Pro evaluations highlight the potential biases and limitations of relying solely on LLM-based evaluations, especially when different LLMs may have inherent preferences based on their training data and architectures.

Discussion on Average Scores: The average scores across all 7 metrics evaluated by both GPT and Gemini (14 scores in total for each row) demonstrate the superiority of synthetic data fine-tuning. Among the 7 base models evaluated, models fine-tuned exclusively on synthetic data (*) achieve the highest average scores in 6 cases, with scores ranging from 7.85 to 8.04 compared to 4.02–7.99 for other variants. This consistent advantage observed in diverse architectures, including LLaMA2, Mistral and Mixtral families, indicates that synthetic counseling conversations provide highly effective training data for mental health support capabilities.

Discussion on Human Evaluation: The human ranking results, shown in the last column of Table 3, clearly indicate that fine-tuned models significantly outperform their base models and the baseline model in the context of mental health counseling. Notably, human evaluators often preferred models fine-tuned on synthetic data, but in several cases, models fine-tuned on interview data or both datasets also performed well. This aligns with GPT-4's and Gemini Pro's evaluations. This trend underscores the effectiveness of our MentalChat16K dataset and fine-tuning pipeline in enhancing the LLMs' capabilities for mental health applications.

Significance Analysis: Additionally, we conducted statistical analyses to demonstrate the significance of our results. We randomly selected 50 questions from the original 200 test questions and ran five rounds of inference on both fine-tuned and base models. Using GPT-4 Turbo and Gemini Pro for evaluation, we compared the average scores across all five rounds on each of the seven mental health metrics between the fine-tuned models and their base model by running a two-sample t-test with a 0.95 confidence interval. For Gemini Pro evaluation, 18 of 21 fine-tuned models showed significant differences from their base model in at least 6 of 7 metrics. For GPT-4 Turbo, 17 of 21 fine-tuned models showed significant differences from their base model in at least 6 of 7 metrics. We use "•" to mark the scores with significant P-values.

5 Ethical Considerations

Participants in the study signed an informed consent document outlining the risks and benefits of the study. All anonymous study sessions were conducted in a private environment to ensure confidentiality and privacy. These sessions were recorded using team-provided devices and stored on secure institutionally backed cloud servers. Study data was captured and stored on institutionally backed data management software. Audio files and study data were labeled with unique identifiers and without the inclusion of personal identifying information. Participants' personal identifying information was stored on a separate database linked to the study data through the unique identifier. Only members of the research team had access to this 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 Limitations

Despite the promising advancements facilitated by MentalChat16K, several limitations need to be acknowledged. First, the reliance on synthetic data generated by GPT-3.5 Turbo may introduce biases or lack the depth of real human interactions. Table 3 showed that combining synthetic and interview data during fine-tuning did not consistently improve model performance; in some cases, it led to performance degradation. Moreover, the interview data focuses on specific caregivers and patients, which differs significantly from the broad profile of the synthetic data. This suggests that users should handle the synthetic and interview datasets separately and exercise caution when combining them.

Second, the anonymization process, while crucial for privacy, may inadvertently strip conversations of contextual nuances essential for effective mental health support. Breaking down interviews into individual question-answer pairs may also result in the loss of broader conversational context, potentially affecting the quality of generated responses. Additionally, the paraphrasing process may introduce minor deviations or potential hallucinations. To mitigate this effect, paraphrasing was designed to preserve the essence of each interaction while making the QA pairs as self-contained as possible. The original order of the QA pairs is maintained, allowing downstream processes to reconstruct context if needed.

Additionally, the dataset primarily focuses on English, limiting its applicability to non-English-speaking regions where cultural and linguistic differences play a significant role in mental health counseling. Furthermore, the interview data were collected from a specific group of caregivers offering care to patients in palliative or hospice care, which may limit the generalizability of the findings to other populations.

Finally, the evaluation framework, though robust, relies heavily on automated assessments and selected human evaluations, which might not fully capture the complexities of real-world counseling efficacy. The use of LLMs as evaluators introduces potential biases, and inconsistencies were observed between different evaluators. Advanced evaluation frameworks like G-Eval [40] and Prometheus [31] were not utilized and could provide more nuanced assessments in future work.

It is also important to recognize that, while human evaluation is crucial for assessing the quality of LLM output, it inevitably introduces subjectivity and bias from personal preferences. Although we calculated inter-rater agreement to assess consistency among human evaluators, the moderate agreement level indicates room for improvement in evaluation methodologies. Addressing these limitations in future work will be crucial for further enhancing the utility and impact of AI-driven mental health support systems.

More state-of-the-art models, such as DeepSeek [21], were not incorporated into our experiments because they were not available at the time of the study. Future research can evaluate these models, including DeepSeek and its distilled variants across multiple sizes, to further assess their potential in enhancing our system.

Finally, because our pipeline breaks interviews into isolated QA pairs, MentalChat16K does not preserve multi-turn conversational flow (e.g., follow-up questions, clarifications). In real counseling, exchanges often span multiple turns. We plan to release a multi-turn subset in a future version to support dialog-flow research.

7 Conclusion

In this paper, we propose MentalChat16K, a benchmark dataset that significantly advances the field of AI-driven mental health support. By incorporating both synthetic counseling conversations and anonymized real-life intervention interview data, MentalChat16K addresses the critical need for domain-specific training data, facilitating the development of large language models capable of empathetic and personalized interactions. Our comprehensive evaluation framework, utilizing state-of-the-art models and advanced metrics, demonstrates the superior performance of LLMs fine-tuned on this dataset in delivering nuanced and compassionate mental health assistance. This work not only highlights the transformative potential of AI in augmenting mental health services but also establishes a new standard for ethical and effective AI development in this sensitive and vital domain.

Acknowledgments

Research reported in this publication was supported in part by the National Institute On Aging of the National Institutes of Health under Award Numbers R01AG069936 and P30AG073105. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

We also gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017811.

Finally, we want to thank Ann Muramatsu, Stacey Brown, Cathryn Koplitz for their help with data collection and data filtering, as well as Aaron Wang, Boning Tong, Jiayu Zhang, Jiong Chen, Qipeng Zhan, Ruiming Wu, Tianhua Zhai, Weiqing He, Yanbo Feng, Zexuan Wang, Zhuoping Zhou, Zixuan Wen for their contributions to data quality checking.

A: Appendix

A.1 Dataset Metadata

A.1.1 Demographic Statistics of Caregivers in the Anonymous Study.

Demographic information was collected from 421 caregivers who completed the information survey. Specifically, the majority of caregivers are female. Among female caregivers, White

Caucasians constitute the largest group, making up approximately 88%, with less than 1% identifying as Hispanic. Male caregivers are a small minority, totaling 6, with White Caucasians being the predominant group. Other racial categories include Asian American, Black/African American, Multi-racial, and others, each comprising smaller proportions of the caregiver population.

The data is skewed toward white female caregivers in hospice care but the skew is not entirely unexpected as the national hospice population itself is known to have a similar demographic trend. According to a systematic review by [10], 9 out of 14 articles had a 70% or greater female population, with 4 being above 80%. Additionally, 5 out of 8 US-based studies had white/Caucasian populations above the national census average of 75%. Similarly, another review by [9] found that in 25 studies, 19 had 76% or more female caregivers, and 11 US-based studies reported Caucasian populations above the national average of 72%. These findings suggest that the skew in our dataset aligns with broader trends in hospice care and research participation.

Table 4:

Demographic Statistics of Caregivers in the Anonymous Study

Gender	Race	Non-Hispanic	Hispanic	Decline to answer	Total
Female	American Indian or Alaska Native	1		1	2
	Asian American	14		1	15
	Black/African American	19		1	20
	Decline to answer		1		1
	Multi-racial	6	1	2	9
	Other		2		2
	White Caucasian	339	3	24	366
Female Total		379	7	29	415
Male	Other			1	1
	White Caucasian	5			5
Male Total		5		1	6

A.1.2 Topic Distribution For Synthetic Data.

The Synthetic Data covered 33 mental health topics, including Relationships, Anxiety, Depression, Intimacy, Family Conflict, etc. The proportion of each topic (Figure 3) that typically arises in a counseling session according to the CounselChat [7] platform was specified in the prompt. This method ensures the synthetic conversations authentically mimic the complexity and diversity of therapist-client interactions, thereby exposing our models to a wide spectrum of psychological conditions and therapeutic strategies.

A.2 Human and LLM-Based Evaluation Correlation

To assess the alignment between LLM-based evaluations and human judgments, we conducted a correlation analysis using Pearson’s correlation coefficients across seven key evaluation metrics (Table 5). This analysis, performed with GPT-4 Turbo and Gemini Pro, revealed several compelling findings.

Holistic Approach exhibited the strongest correlation across both models (GPT-4: 0.489, Gemini: 0.489), followed by **Empathy & Validation** (GPT-4: 0.433, Gemini: 0.355) and **Active Listening** (GPT-4: 0.411, Gemini: 0.324). These metrics demonstrated particularly robust alignment with human evaluations, underscoring the capability of LLMs to reliably assess critical aspects of mental health interactions. **Clarity & Encouragement** also maintained strong correlations (GPT-4: 0.413, Gemini: 0.364), while **Open-mindedness & Non-judgment** (GPT-4: 0.328, Gemini: 0.338) and **Safety & Trustworthiness** (GPT-4: 0.272, Gemini: 0.291) displayed moderate alignment. Although **Boundaries & Ethical Considerations** showed comparatively lower correlations (GPT-4: 0.203, Gemini: 0.291), they were still meaningful and consistent with expectations for this dimension.

To contextualize these results, we compared them with benchmarks from recent work, such as the EMNLP 2023 paper *"Towards Interpretable Mental Health Analysis with Large Language Models"*. That study reported correlations between human judgments and automated metrics primarily in the 0.10–0.40 range, with ChatGPT$_{\text{true}}$ achieving 0.15–0.35 and BART-Score exceeding 0.40 only for specific metrics. In contrast, our correlation coefficients (ranging from 0.20 to 0.49) not only align with these benchmarks but often exceed them. For instance, **Holistic Approach** (0.489) performs comparably to their best metric (BART-Score: 0.428), while several of our metrics outperform their BERT-based methods (0.172–0.373). Crucially, the consistent performance across both GPT-4 Turbo and Gemini Pro highlights the robustness and reliability of our evaluation framework.

The overall results suggest that LLM-based evaluation is highly reliable for assessing key aspects of response quality. This evidence supports the adoption of LLMs for targeted evaluation tasks in mental health contexts, bridging a critical gap in scalable, automated evaluation methodologies.

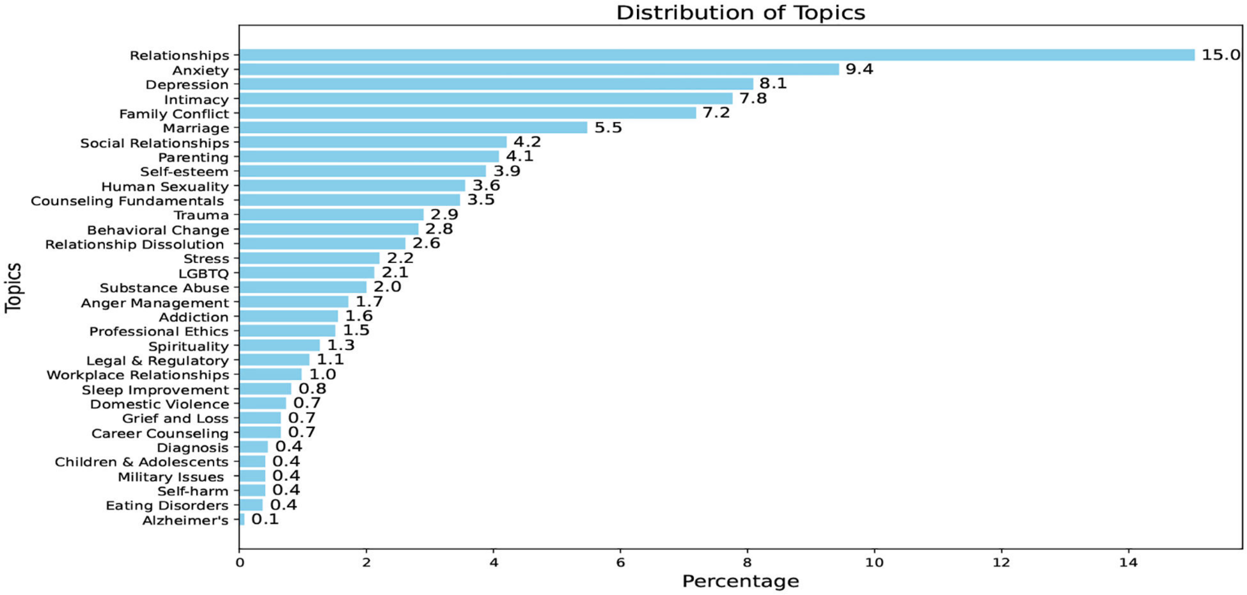

Figure 3:

Topic Distribution for Synthetic Data.

Table 5:

Pearson’s correlation coefficients between human evaluation and LLM evaluation results.

Metrics	GPT-4 Turbo (*r*)	Gemini Pro 1.0 (*r*)
Active Listening	0.411	0.324
Empathy & Validation	0.433	0.355
Safety & Trustworthiness	0.272	0.291
Open-mindedness & Non-judgment	0.328	0.338
Clarity & Encouragement	0.413	0.364
Boundaries & Ethical	0.203	0.291
Holistic Approach	0.489	0.489

A.3 Synthetic Query Generation

This section presents the prompt used for synthetic query generation (Table 6).

Table 6:

Prompt 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’ descriptions of experiences, feelings, behaviors, questions, and all the details that may be shared with a mental health professional. Make the messages as specific and detailed as possible. Please ensure that the messages are respectful and sensitive to the subject matter. Each message must cover all of the following requirements:

Prompt for Generating Mental Health Counseling Conversations

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 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.

The output must strictly follow the format:
Caregiver: [[caregiver's query from first-person view]]
Coach: [[coach's response from first-person view]]

Topics: {topics}

References

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

[2]. Althoff Tim, Clark Kevin, and Leskovec Jure. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. Transactions of the Association for Computational Linguistics (TACL) 4 (2016), 463–476. [PubMed: 28344978]

[3]. American Psychological Association. 2017. Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code Retrieved November 26, 2024.

[4]. Anil Rohan, Dai Andrew M, Firat Orhan, Johnson Melvin, Lepikhin Dmitry, Passos Alexandre, Shakeri Siamak, Taropa Emanuel, Bailey Paige, Chen Zhifeng, et al. 2023 Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023).

[5]. Arias Daniel, Saxena Shekhar, and Verguet Stéphane. 2022. Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine 54 (2022).

[6]. Bantilan Niels, Malgaroli Matteo, Ray Bonnie, and Hull Thomas D. 2021. Just in time crisis response: suicide alert system for telemedicine psychotherapy settings. Psychotherapy research 31, 3 (2021), 289–299. [PubMed: 32366192]

[7]. Bertagnolli Nicolas. 2023. Counsel Chat: Bootstrapping High-Quality Therapy Data. https://towardsdatascience.com/counsel-chat-bootstrapping-high-quality-therapy-data-971b419f33da.

[8]. Bulian Jannis, Buck Christian, Gajewski Wojciech, Boerschinger Benjamin, and Schuster Tal. 2022. Tomayto, Tomahto. Beyond Token-level Answer Equivalence for Question Answering Evaluation. arXiv:2202.07654 [cs.CL]

[9]. Chi Nai-Ching, Barani Emelia, Fu Ying-Kai, Nakad Lynn, Gilbertson-White Stephanie, Herr Keela, and Saeidzadeh Seyedehtanaz. 2020. Interventions to support family caregivers in pain management: a systematic review. Journal of pain and symptom management 60, 3 (2020), 630–656. [PubMed: 32339651]

[10]. Chi Nai-Ching, Demiris George, Lewis Frances M, Walker Amy J, and Langer Shelby L. 2016. Behavioral and educational interventions to support family caregivers in end-of-life care: a systematic review. American Journal of Hospice and Palliative Medicine® 33, 9 (2016), 894–908. [PubMed: 26157046]

[11]. Chiang Cheng-Han and Lee Hung yi. 2023. Can Large Language Models Be an Alternative to Human Evaluations? arXiv:2305.01937 [cs.CL]

[12]. Clark Peter, Cowhey Isaac, Etzioni Oren, Khot Tushar, Sabharwal Ashish, Schoenick Carissa, and Tafjord Oyvind. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457 [cs.AI]

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

[14]. Demszky D, Yang D, Yeager DS, et al. 2023. Using large language models in psychology. Nat Rev Psychol (2023). doi:10.1038/s44159-023-00241-5

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

[16]. Djernes Jens Kronborg. 2006. Prevalence and predictors of depression in populations of elderly: a review. Acta Psychiatrica Scandinavica 113, 5 (2006), 372–387. [PubMed: 16603029]

[17]. Du Zhengxiao, Qian Yujie, Liu Xiao, Ding Ming, Qiu Jiezhong, Yang Zhilin, and Tang Jie. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. arXiv:2103.10360 [cs.CL]

[18]. Mental Health Forum. 2023. Mental Health Forum. https://www.mentalhealthforum.net/. Accessed: 2023-12-23.

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

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

[21]. Guo Daya, Yang Dejian, Zhang Haowei, Song Junxiao, Zhang Ruoyu, Xu Runxin, Zhu Qihao, Ma Shirong, Wang Peiyi, Bi Xiao, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025).

[22]. He Weiqing, Hou Bojian, Demiris George, and Shen Li. 2024. Interpretability study for long interview transcripts from behavior intervention sessions for family caregivers of dementia patients. AMIA Summits on Translational Science Proceedings 2024 (2024), 201.

[23]. Hendrycks Dan, Burns Collin, Basart Steven, Zou Andy, Mazeika Mantas, Song Dawn, and Steinhardt Jacob. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY]

[24]. Hou Bojian, Zhang Hao, Ladizhinsky Gur, Yang Stephen, Kuleshov Volodymyr, Wang Fei, and Yang Qian. 2021. Clinical evidence engine: proof-of-concept for a clinical-domain-agnostic decision support infrastructure. arXiv preprint arXiv:2111.00621 (2021).

[25]. Hua Yining, Liu Fenglin, Yang Kailai, Li Zehan, Sheu Yi-han, Zhou Peilin, Moran Lauren V, Ananiadou Sophia, and Beam Andrew. 2024. Large Language Models in Mental Health Care: a Scoping Review. arXiv preprint arXiv:2401.02984 (2024).

[26]. InFamousCoder. 2022. Depression: Reddit Dataset (Cleaned), Version 1. https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned/data. Retrieved December 23, 2023.

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

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

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

[30]. Kabat-Zinn Jon. 2013. Full catastrophe living: Using the wisdom of your body and mind to face stress, pain, and illness. Bantam.

[31]. Kim Seungone, Shin Jamin, Cho Yejin, Jang Joel, Longpre Shayne, Lee Hwaran, Yun Sangdoo, Shin Seongjin, Kim Sungdong, Thorne James, et al. 2023. Prometheus: Inducing fine-grained

evaluation capability in language models. In The Twelfth International Conference on Learning Representations.

[32]. Lai Tin, Shi Yukun, Du Zicong, Wu Jiajie, Fu Ken, Dou Yichao, and Wang Ziqi. 2023. Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. arXiv preprint arXiv:2307.11991 (2023).

[33]. Lambert Michael J and Barley Dean E. 2001. Research summary on the therapeutic relationship and psychotherapy outcome. Psychotherapy: Theory, Research, Practice, Training 38, 4 (2001), 357.

[34]. Landis J Richard and Koch Gary G. 1977. The measurement of observer agreement for categorical data. biometrics (1977), 159–174. [PubMed: 843571]

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

[36]. Lee Yoon Kyung, Suh Jina, Zhan Hongli, Li Junyi Jessy, and Ong Desmond C. 2024. Large language models produce responses perceived to be empathic. arXiv preprint arXiv:2403.18148 (2024).

[37]. Li Anqi, Lu Yu, Song Nirui, Zhang Shuai, Ma Lizhi, and Lan Zhenzhong. 2024. Automatic Evaluation for Mental Health Counseling using LLMs. arXiv:2402.11958 [cs.CL]

[38]. Li Xiang, Xu Yue, Che Wanxiang, and Liu Tianyu. 2023. PsyEval: A Suite of Mental Health Related Tasks for Evaluating Large Language Models. arXiv preprint arXiv:2311.09189 (2023).

[39]. Liu Cheng, Wu Qingyu, Liu Jiayi, and Yu Hong. 2023. ChatCounselor: A Large Language Model for Mental Health Counseling. arXiv preprint arXiv:2309.11473 (2023).

[40]. Liu Yang, Iter Dan, Xu Yichong, Wang Shuohang, Xu Ruochen, and Zhu Chenguang. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634 (2023).

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

[42]. Malhotra G, Waheed A, Srivastava A, Akhtar MS, and Chakraborty T. 2022. Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations. In Proceedings of the fifteenth ACM international conference on web search and data mining. 735–745.

[43]. Miller William R and Moyers Theresa B. 2006. Eight stages in learning motivational interviewing. Journal of Teaching in the Addictions 5, 1 (2006), 3–17.

[44]. Na H. 2024. CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2930–2940.

[45]. OpenAI. 2024. New embedding models and API updates. https://openai.com/blog/new-embedding-models-and-api-updates.

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

[47]. 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.

[48]. Gal Peretz, Taylor C Barr, Ruzek Josef I, Jefroykin Samuel, and Sadeh-Sharvit Shiri. 2023. Machine Learning Model to Predict Assignment of Therapy Homework in Behavioral Treatments: Algorithm Development and Validation. JMIR Formative Research 7 (2023), e45156. [PubMed: 37184927]

[49]. Qiu Yujia, Zhang Haoqin, Deng Yang, Zhang Xiaoying, Li Yanran, and Wang Minlie. 2023. Dialogue Safety Assessment with LLMs. arXiv preprint arXiv:2311.00206 (2023).

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

[51]. Risch Julian, Möller Timo, Gutsch Julian, and Pietsch Malte. 2021. Semantic Answer Similarity for Evaluating Question Answering Models. arXiv:2108.06130 [cs.CL]

[52]. Rogers Carl R. 1957. The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology 21, 2 (1957), 95–103. [PubMed: 13416422]

[53]. Sharma Ashish, Lin I-Hsiang, Ko Albert, Dwivedi Pragati, and Baker Ryan. 2020. EPITOME: An Empathy-based Platform for Therapeutic Online Mental Health Education. International Conference on Artificial Intelligence in Education (2020), 284–288.

[54]. Srivastava Aseem, Suresh Tharun, Sarah P Lord Md Shad Akhtar, and Chakraborty Tanmoy. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 3920–3930.

[55]. Sun H, Lin Z, Zheng C, Liu S, and Huang M. 2021. PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 1489–1503.

[56]. Taori Rohan, Gulrajani Ishaan, Zhang Ting, Dubois Yann, Li Xuechen, Guestrin Carlos, Liang Percy, and Hashimoto Tatsunori B.. 2023. Stanford Alpaca: An Instruction-Following Llama Model. https://github.com/tatsu-lab/stanford_alpaca.

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

[58]. National Alliance on Mental Illness. 2023. Mental Health By the Numbers. https://nami.org/mhstats Accessed: 2024-10-01.

[59]. World Health Organization. 2021. Depression. https://www.who.int/news-room/fact-sheets/detail/depression

[60]. Torous John, Bucci Sandra, Bell Imogen H, Kessing Lars V, Faurholt-Jepsen Maria, Whelan Pauline, Carvalho Andre F, Keshavan Matcheri, Linardon Jake, and Firth Joseph. 2021. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. World Psychiatry 20, 3 (2021), 318–335. [PubMed: 34505369]

[61]. Touvron Hugo, Lavril Thibaut, Izacard Gautier, Martinet Xavier, Lachaux Marie-Anne, Lacroix Timothée, Rozière Baptiste, Goyal Naman, Hambro Eric, Azhar Faisal, Rodriguez Aurelien, Joulin Armand, Grave Edouard, and Lample Guillaume. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[62]. Touvron Hugo, Martin Louis, Stone Kevin, Albert Peter, Almahairi Amjad, Babaei Yas-mine, Bashlykov Nikolay, Batra Soumya, Bhargava Prajjwal, Bhosale Shruti, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023).

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

[64]. van Heerden Alastair C, Pozuelo Julia R, and Kohrt Brandon A. 2023. Global mental health services and the impact of artificial intelligence–powered large language models. JAMA psychiatry 80, 7 (2023), 662–664. [PubMed: 37195694]

[65]. Wang Xiao, Liu Kai, and Wang Chunlei. 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). IEEE, 532–536.

[66]. Wang Yizhong, Kordi Yeganeh, Mishra Swaroop, Liu Alisa, Smith Noah A, Khashabi Daniel, and Hajishirzi Hannaneh. 2022. Self-instruct: Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560 (2022).

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

[68]. Zhan Hongli, Zheng Allen, Lee Yoon Kyung, Suh Jina, Li Junyi Jessy, and Ong Desmond C. 2024. Large Language Models are Capable of Offering Cognitive Reappraisal, if Guided. arXiv preprint arXiv:2404.01288 (2024).

[69]. Zhang Tianlin, Schoene Annika M, Ji Shaoxiong, and Ananiadou Sophia. 2022. Natural language processing applied to mental illness detection: a narrative review. NPJ digital medicine 5, 1 (2022), 46. [PubMed: 35396451]

[70]. Zhang Xinyao, Tanana Michael, Weitzman Lauren, Narayanan Shrikanth, Atkins David, and Imel Zac. 2023. You never know what you are going to get: Large-scale assessment of therapists' supportive counseling skill use. Psychotherapy 60, 2 (2023), 149. [PubMed: 36301302]

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

CCS Concepts

• **Computing methodologies → Natural language generation**; *Cognitive science*; **Natural language generation**; • **Applied computing → Psychology**; Psychology.

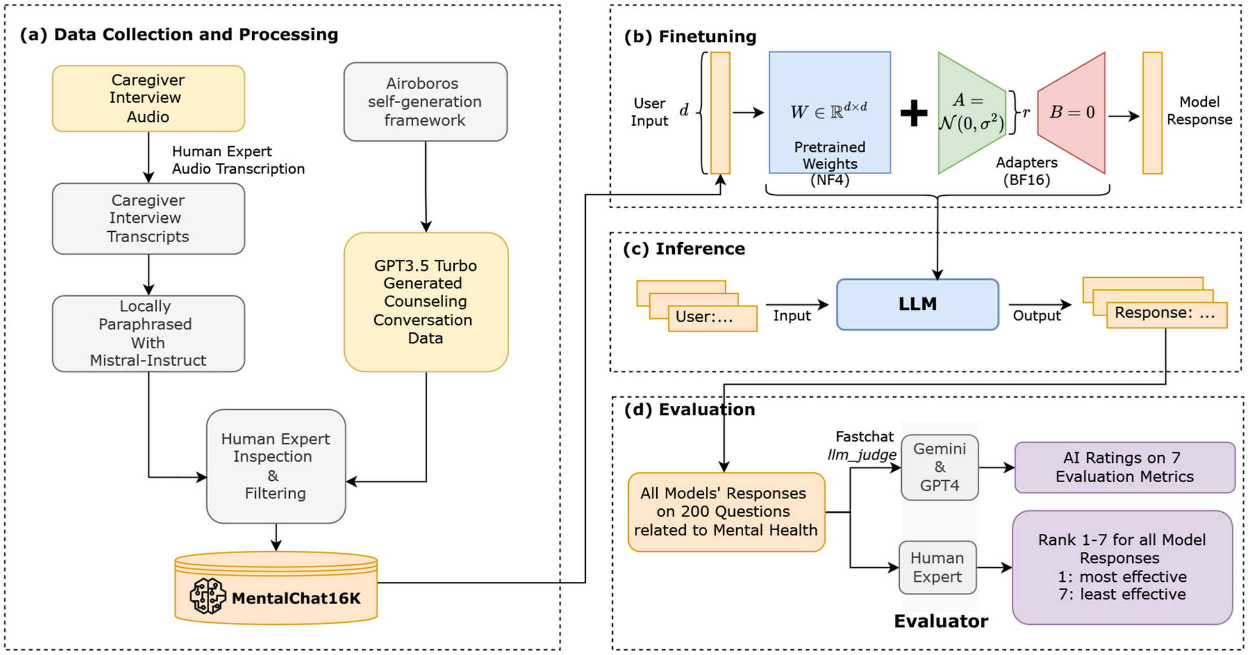
(a) Data Collection and Processing
Caregiver Interview Audio
Human Expert Audio Transcription
Caregiver Interview Transcripts
Locally Paraphrased With Mistral-Instruct
Airoboros self-generation framework
GPT3.5 Turbo Generated Counseling Conversation Data
Human Expert Inspection & Filtering
MentalChat16K
(b) Finetuning
User Input
d
$W \in \mathbb{R}^{d \times d}$
Pretrained Weights (NF4)
$A = \mathcal{N}(0, \sigma^2)$
r
$B = 0$
Adapters (BF16)
Model Response
(c) Inference
User:...
Input
LLM
Output
Response: ...
(d) Evaluation
All Models' Responses on 200 Questions related to Mental Health
Fastchat
llm_judge
Gemini & GPT4
Human Expert
Evaluator
AI Ratings on 7 Evaluation Metrics
Rank 1-7 for all Model Responses
1: most effective
7: least effective

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, Mistral-Instruct. (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, GPT-4 Turbo and human experts as the judges to score those responses.

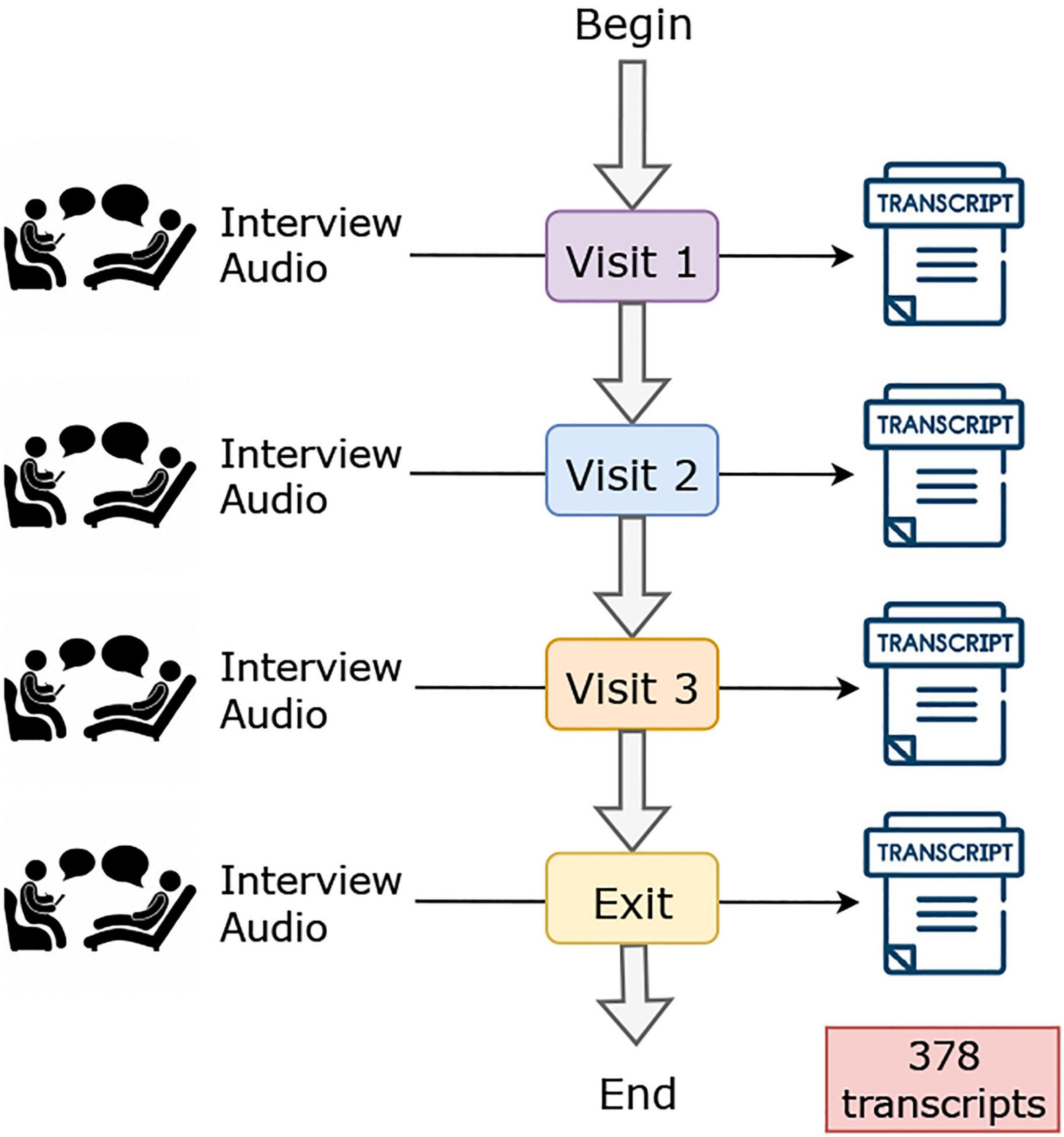

Figure 2:

Illustration of behavioral intervention interview data. A caregiver has three formal visits and an exiting visit. Each visit will generate an audio file that will be transcribed into transcripts.

Table 1:

Summary of MentalChat16K Dataset Statistics

	Dataset Size (Rows)	Columns	Avg. Input Word Count	Avg. Output Word Count	Number of Sessions	Avg. QA Pairs Per Session	Number of Topics
Interview Data	9,775	instruction, input, output	69.94	235.85	378	16.8	-
Synthetic Data	6,338	instruction, input, output	111.24	363.94	-	-	33

Table 2:

LLMs evaluation metrics on mental health.

Strategy	Description	References
Active Listening	Responses demonstrate careful consideration of user concerns, reflecting understanding and capturing the essence of the issue. Avoid assumptions or jumping to conclusions.	[43], [55], [39]
Empathy & Validation	Convey deep understanding and compassion, validating feelings and emotions without being dismissive or minimizing experiences.	[53], [52], [39]
Safety & Trustworthiness	Prioritize safety, refrain from harmful or insensitive language. Ensure the information provided is consistent and trustworthy.	[49], [33], [38]
Open-mindedness & Non-judgment	Approach without bias or judgment. Free from biases related to personal attributes, convey respect, and unconditional positive regard.	[52], [55], [30]
Clarity & Encouragement	Provide clear, concise, and understandable answers. Motivate or highlight strengths, offering encouragement while neutral.	[39], [38]
Boundaries & Ethical	Clarify the response’s role, emphasizing its informational nature. In complex scenarios, guide users to seek professional assistance.	[49], [39]
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.	[55], [39]

Table 3:

Comparison of evaluation scores rated by GPT-4 Turbo and Gemini Pro across all models on 7 mental health metrics, as well as human rankings of model responses on a scale of 1 to 7. In each two-column cell, the best score evaluated by GPT-4 Turbo is highlighted in red, and the best score evaluated by Gemini Pro is highlighted in blue. The score with a significant P-value (< 0.05) is marked with •. The last column contains the average ranking of each model evaluated by humans. The bold numbers represent the highest average rank among the four models in one block and the three baseline models. The ranking procedure is described in Section 3.3. This result showed that fine-tuning on MentalChat16K significantly improved the performance of base LLMs. The model fine-tuned on synthetic data alone outperforms the other three cases most of the time according to GPT-4 and human evaluations. The model fine-tuned on real interview data alone outperforms the other three cases most of the time when using Gemini Pro for evaluation. Models fine-tuned on the entire MentalChat16K also significantly outperform their base model most of the time.

Model (7B)	Active Listening ↑		Empathy & Validation ↑		Safety & Trustworthiness ↑		Open-mindedness & Non-judgment ↑		Clarity & Encouragement ↑		Boundaries & Ethical ↑		Holistic Approach ↑		Average Score ↑	Human Rank ↓
	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	4.02	4.45
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•	7.66	3.79
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•	**7.87**	3.65
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•	7.83	**3.55**
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	7.99	3.20
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•	7.76	2.55
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	**8.04**	**2.35**
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•	7.89	3.10
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	6.32	5.15
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•	7.78	3.30
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	**7.92**	**1.90**
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•	7.87	2.60
Mixtral-8×7B-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	5.55	6.25
Mixtral-8×7B-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•	7.79	3.10

Model (7B)	**Active Listening ↑**		**Empathy & Validation ↑**		**Safety & Trustworthiness ↑**		**Open-mindedness & Non-judgment ↑**		**Clarity & Encouragement ↑**		**Boundaries & Ethical ↑**		**Holistic Approach ↑**		**Average Score ↑**	**Human Rank ↓**
	GPT	**Gemini**	**GPT**	**Gemini**	**GPT**	**Gemini**	**GPT**	**Gemini**	**GPT**	**Gemini**	**GPT**	**Gemini**	**GPT**	**Gemini**		
Mixtral-8×7B-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•	**8.04**	**1.55**
Mixtral-8×7B-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•	7.88	2.70
Mixtral-8×7B-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	6.74	5.60
Mixtral-8×7B-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•	7.77	2.55
Mixtral-8×7B-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•	**8.03**	**1.90**
Mixtral-8×7B-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•	7.83	3.05
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	7.20	3.75
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•	7.80	3.85
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•	**7.85**	4.00
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•	7.78	**3.37**
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	7.63	4.20
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•	7.80	2.90
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•	7.89	**2.05**
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•	**7.90**	2.85
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	6.91	5.74
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	7.39	4.61
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	7.46	4.22

Notes. No label: Base Model,

$^{\dagger:}$Model fine-tuned on Interview Data (6K),

$^{*:}$Model fine-tuned on Synthetic Data (10K),

$^{*\dagger:}$Model fine-tuned on both Synthetic and Interview Data (16K),

$^{\S:}$Baseline Model.