

MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance

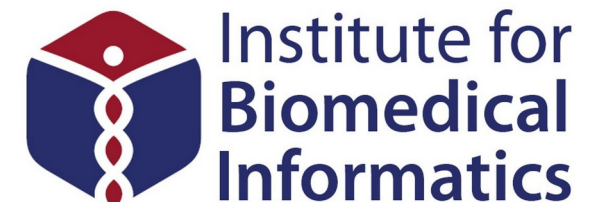
Jia Xu^{*,†}, Tianyi Wei^{*}, Bojian Hou^{*}, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, Li Shen [†]

^{*}University of Pennsylvania, Philadelphia, Pennsylvania, USA

[†]The Roux Institute, Northeastern University, Portland, Maine, USA



DEPARTMENT of
BIostatISTICS
EPIDEMIOLOGY &
INFORMATICS



Motivation

Global Impact

Mental and substance use disorders affect nearly a billion individuals, with anxiety and depression alone impacting over 300 million.

Access Barriers

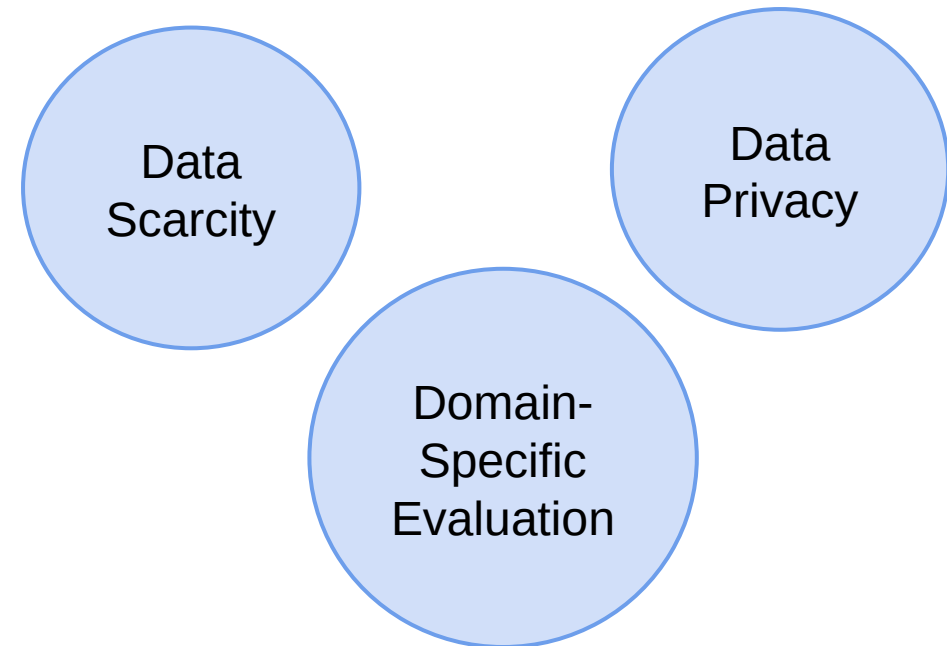
Millions face limited access due to high costs, widespread stigma, and a severe global therapist shortage.

LLM Limitations

General LLMs lack the empathy, nuance, and critical safety protocols essential for effective and ethical therapeutic dialogue.

Key Challenge

The advancement of LLMs for mental health is constrained by the absence of public, high-quality benchmarks for **therapeutic counseling conversations**. General-purpose datasets fail to address the specific safety, privacy, and nuanced interaction requirements of this domain.



Introducing MentalChat16K

- We introduce **MentalChat16K**, a novel dataset that combines **anonymized real transcripts** from behavioral health sessions with **diverse synthetic conversations**, covering a broad range of topics in the mental health domain.
- We design **7 evaluation metrics** and curate a **evaluation dataset** to systematically evaluate the response quality of mental health-focused LLMs.
- We provide a practical framework encompassing **data collection, efficient LLM fine-tuning (using QLoRA), and LLM judges evaluation.**



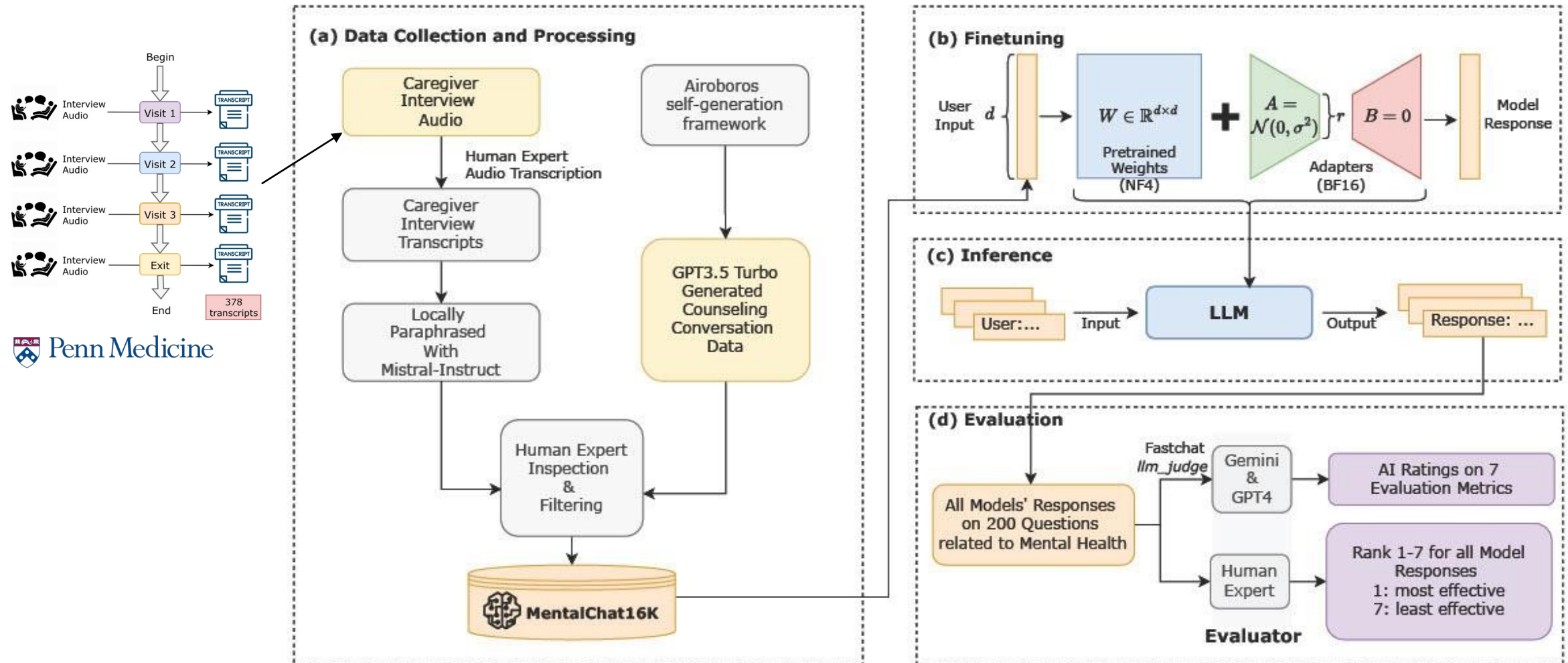
Anonymized Interview Transcripts

6.3K question–answer pairs derived from 378 human expert-transcribed interviews between behavioral health coaches and caregivers of individuals in palliative or hospice care, collected from the ongoing PISCES clinical trial.

Synthetic Counseling Conversations

9.7K question-answer pairs generated by GPT-3.5 Turbo, covering 33 diverse mental health topics like relationships, anxiety, and depression. This ensures broad exposure to psychological conditions and therapeutic strategies.

Overall Architecture



Anonymized Interview Data

- **Data Source:** Collected 378 interview transcripts from a clinical trial involving behavioral health coaches and caregivers in palliative or hospice care; transcriptions were done by human experts from audio recordings.
- **Processing:** Used **Mistral-7B-Instruct-v0.2**, a local lightweight LLM, to paraphrase and summarize each transcript page into a single QA pair; filtered for quality and removed entries under 40 words.
- **Privacy & Filtering:** Manually inspected all outputs to remove sensitive information (e.g., names, addresses), resulting in 6,338 anonymized QA pairs for public release with proper consent.

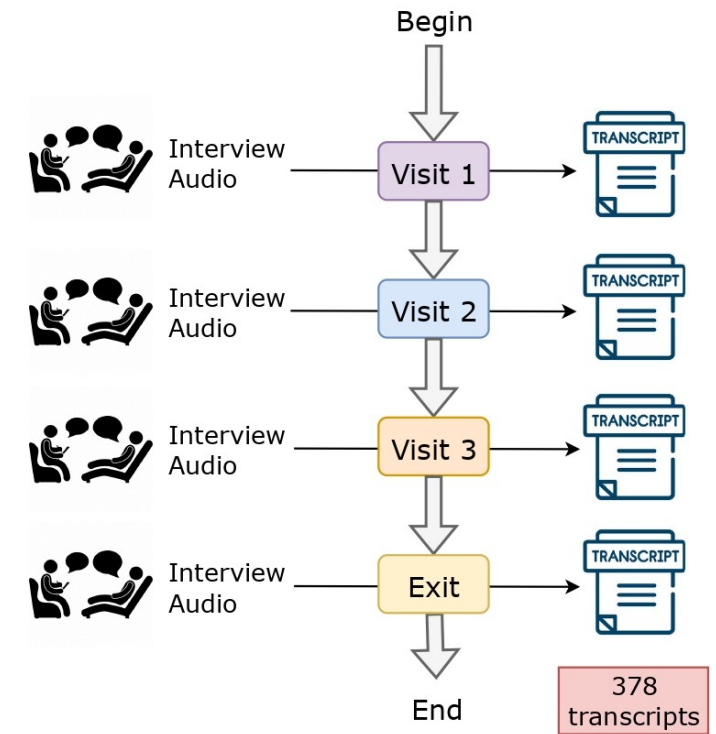
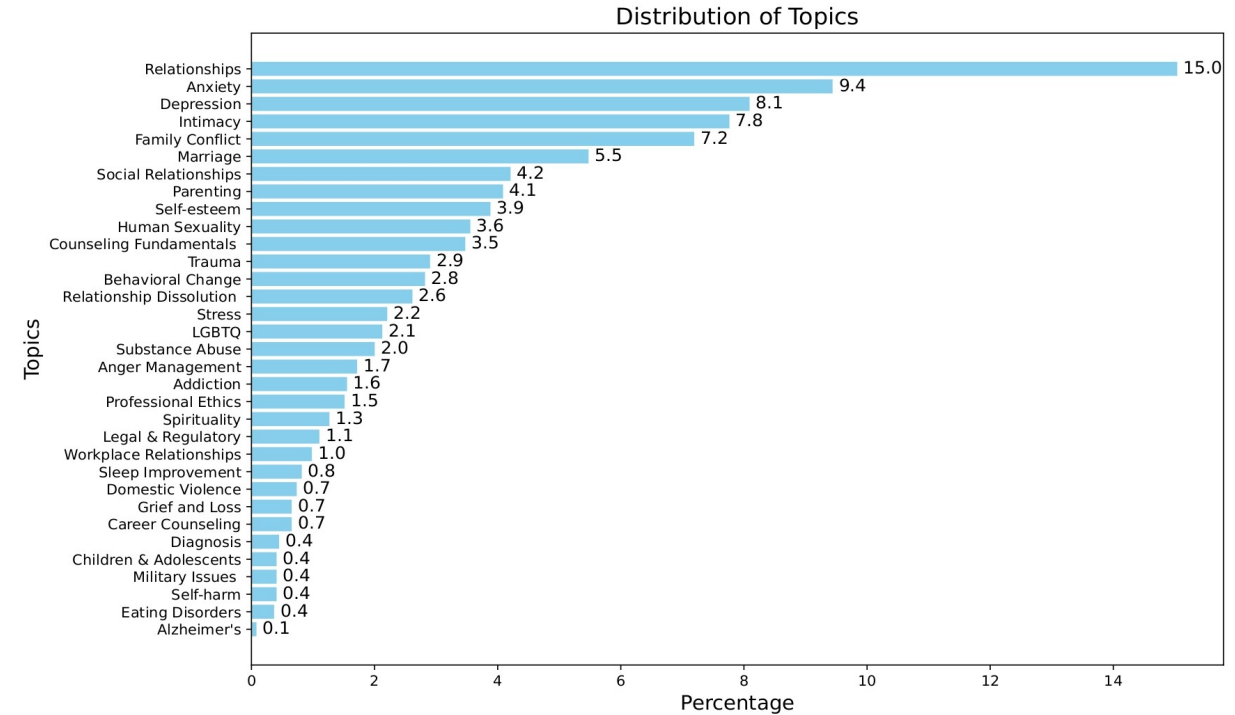


Table 1: Summary of MentalChat16K Dataset Statistics

	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

Synthetic Data

- **Generation Framework:** Generated 9,775 QA pairs using GPT-3.5 Turbo with a customized Airoboros framework, including manual review to ensure authentic counseling style.
- **Topic Coverage:** Covered 33 mental health topics (e.g., anxiety, depression) with distributions aligned to real-world counseling data from CounselChat.



Fine-tuning and Inference

- **Baseline Models:** ChatPsychiatrist, Samantha-v1.11 & v1.2
- **Base Models for Finetuning:** LLaMA2-7b, Mistral-7b-v0.1, Mixtral-8x7b-v0.1, Mistral-7b-Instruct-v0.2, Mixtral-8x7b-Instruct-v0.1, Vicuna-7b-v1.5, Zephyr-7b-alpha.
- We fine-tuned base models using QLoRA on three configurations of the MentalChat16K dataset: synthetic-only, interview-only, and a hybrid of both.

Evaluation Metrics

- Traditional LLM metrics (e.g., perplexity) lack the nuance to assess empathy, emotional sensitivity, and ethics in mental health counseling.
- We derived 7 metrics are grounded in psychotherapy literature and prior LLM evaluation benchmarks (e.g., PsyQA, EPITOME, ChatCounselor) to capture core dimensions of effective and ethical mental health dialogue.



Active Listening

Reflecting understanding and capturing the essence of user concerns.



Empathy & Validation

Conveying deep understanding and compassion, validating feelings.



Safety & Trustworthiness

Prioritizing safety, refraining from harmful language, ensuring Consistent information.



Open-mindedness & Non-judgment

Approaching without bias, conveying respect and positive regard.



Clarity & Encouragement

Providing clear, concise answers, motivating, and highlighting strengths.



Boundaries & Ethical

Clarifying the response's role, guiding users to professional help when needed.



Holistic Approach

Addressing concerns from various angles, considering broader context.

Evaluation Framework

Evaluation Data

- The 200-question evaluation dataset consists of real-world mental health inquiries sourced from Reddit and the Mental Health Forum
- All models are tasked to generate response to the 200 inquires.

Evaluation Method

- Both **LLM** and **human evaluators** were used to ensure comprehensive, unbiased assessment of mental health counseling responses.
- **LLM Judges (Fastchat)**: GPT-4 Turbo and Gemini-Pro-1.0 were used to rate responses from 1 to 10 across the seven devised metrics.
 - Two-sample t-test to compare base and finetuned model scores.
- **Human Evaluators**: Four students with interdisciplinary expertise ranked model responses (1 = best, 7 = worst) for 50 randomly selected questions.
 - 1 base model, 3 finetuned models, 3 baseline models
 - Inter-rater agreement: Cohen's Kappa score = 0.441

Evaluation Results

Overall performance: Our experiments demonstrate that fine-tuning LLMs on MentalChat16K (synthetic, interview, or both) consistently outperform base models across all evaluation methods (GPT4, Gemini Pro, and Human Experts).

LLM Evaluators (GPT-4 & Gemini)

Fine-tuned models consistently outperformed base models across all metrics. **GPT-4** showed a bias towards synthetic data fine-tuning, while **Gemini Pro** valued the realism of interview data, especially for safety and ethical considerations.

Models fine-tuned on synthetic data achieved the highest average scores in 6 out of 7 cases, ranging from 7.85 to 8.04, indicating its effectiveness across diverse architectures.

Human Evaluation

Human evaluators consistently ranked fine-tuned models higher than base and baseline models. While synthetic data fine-tuning was often preferred, models fine-tuned on interview data or combined datasets also performed well.

Statistical analyses confirmed significant performance differences between fine-tuned and base models, validating MentalChat16K's effectiveness.

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-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	5.55	6.25
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●	7.79	3.10
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●	8.04	1.55
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●	7.88	2.70
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	6.74	5.60
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●	7.77	2.55
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●	8.03	1.90
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●	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, †: Model fine-tuned on Interview Data (6K), *: Model fine-tuned on Synthetic Data (10K),

*†: Model fine-tuned on both Synthetic and Interview Data (16K), §: Baseline Model.

Conclusion

- **MentalChat16K:** Introduced a new benchmark combining synthetic and anonymized real-life counseling data to support domain-specific LLM training.
- **Improved LLM Performance:** Fine-tuning on MentalChat16K significantly enhances LLMs' ability to provide empathetic, nuanced, and effective mental health support.

Thank you for listening!

MentalChat16K is open-source and ready to use on Github and Hugging Face. It's designed to help the community build safer, more trustworthy AI for high-stakes use cases like mental health. We welcome you to explore, build, and improve.

Visit our GitHub and Hugging Face pages to access the dataset and code.

 github.com/PennShenLab/MentalChat16K
 huggingface.co/datasets/ShenLab/MentalChat16K
 li.shen@pennmedicine.upenn.edu

Acknowledgments: This work is supported in part by NIH R01AG069936 and NIH P30AG073105.

