

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

* Equal contribution, Email: jiaxu@alumni.upenn.edu, tywei@alumni.upenn.edu, bojian.hou@Pennmedicine.upenn.edu

Introduction

- **Challenge:** Progress in mental health LLMs is limited by the scarcity of real-world, domain-specific conversational data due to privacy concerns.
- 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.**

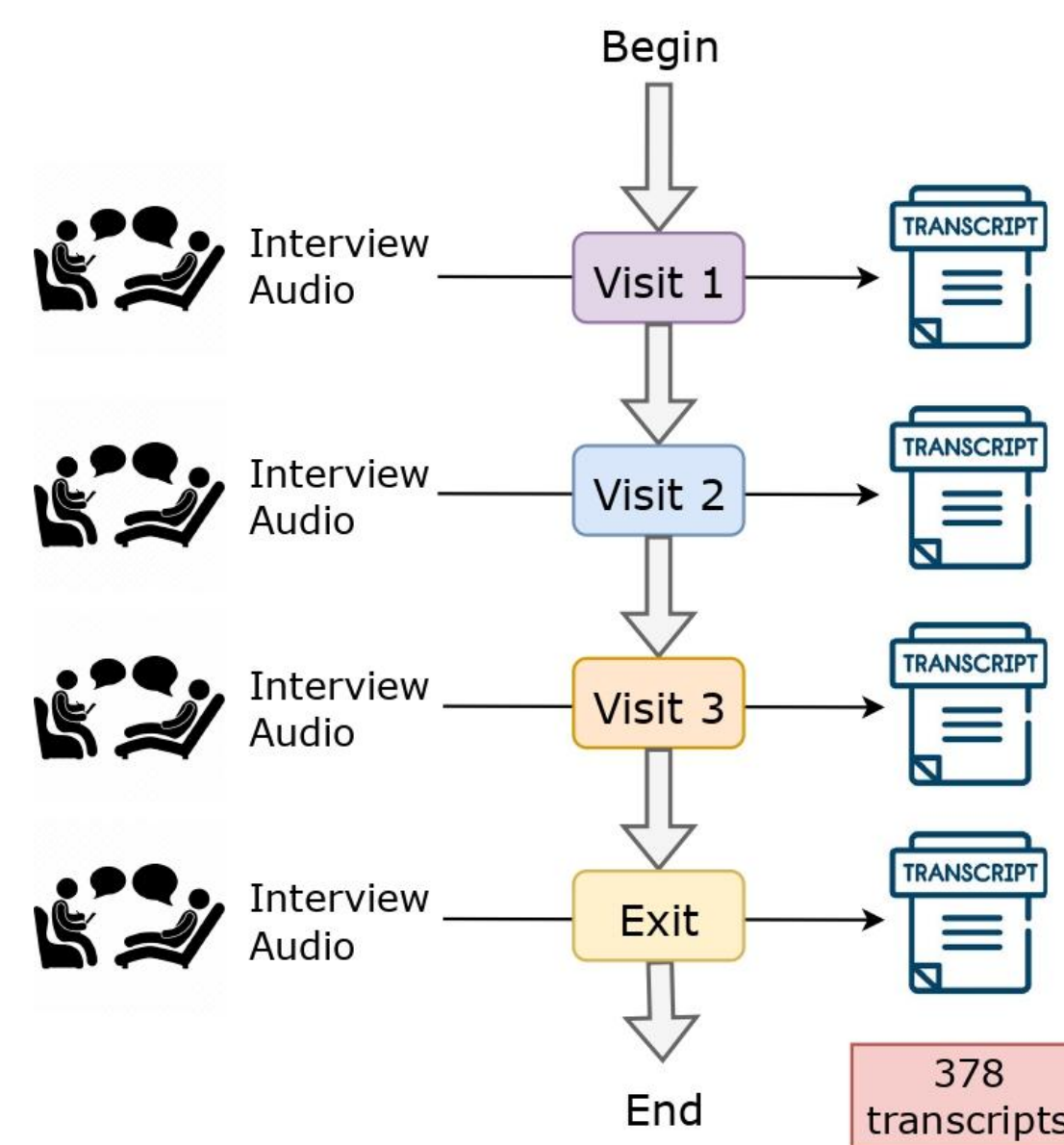
Data Collection and Processing

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

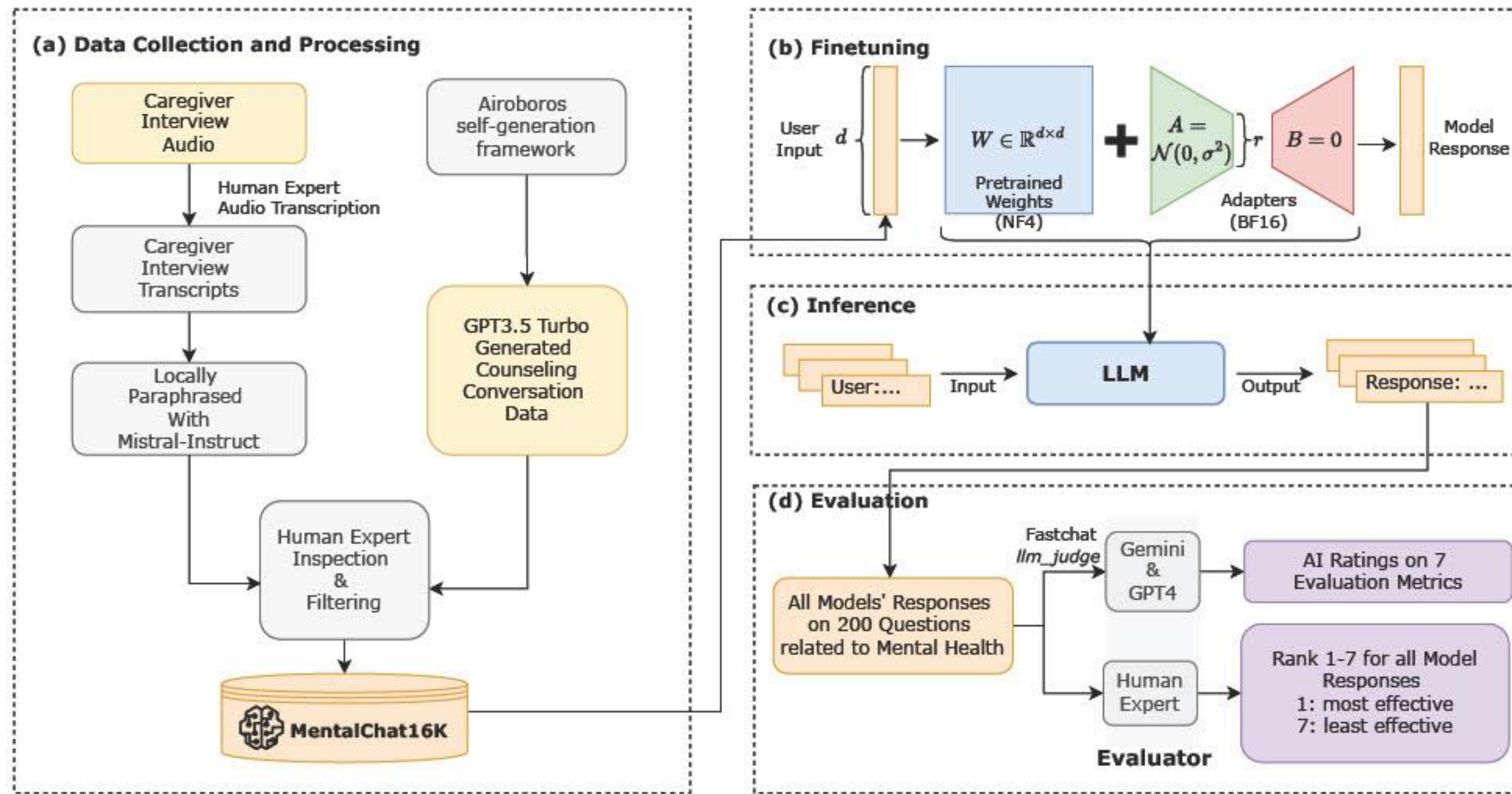
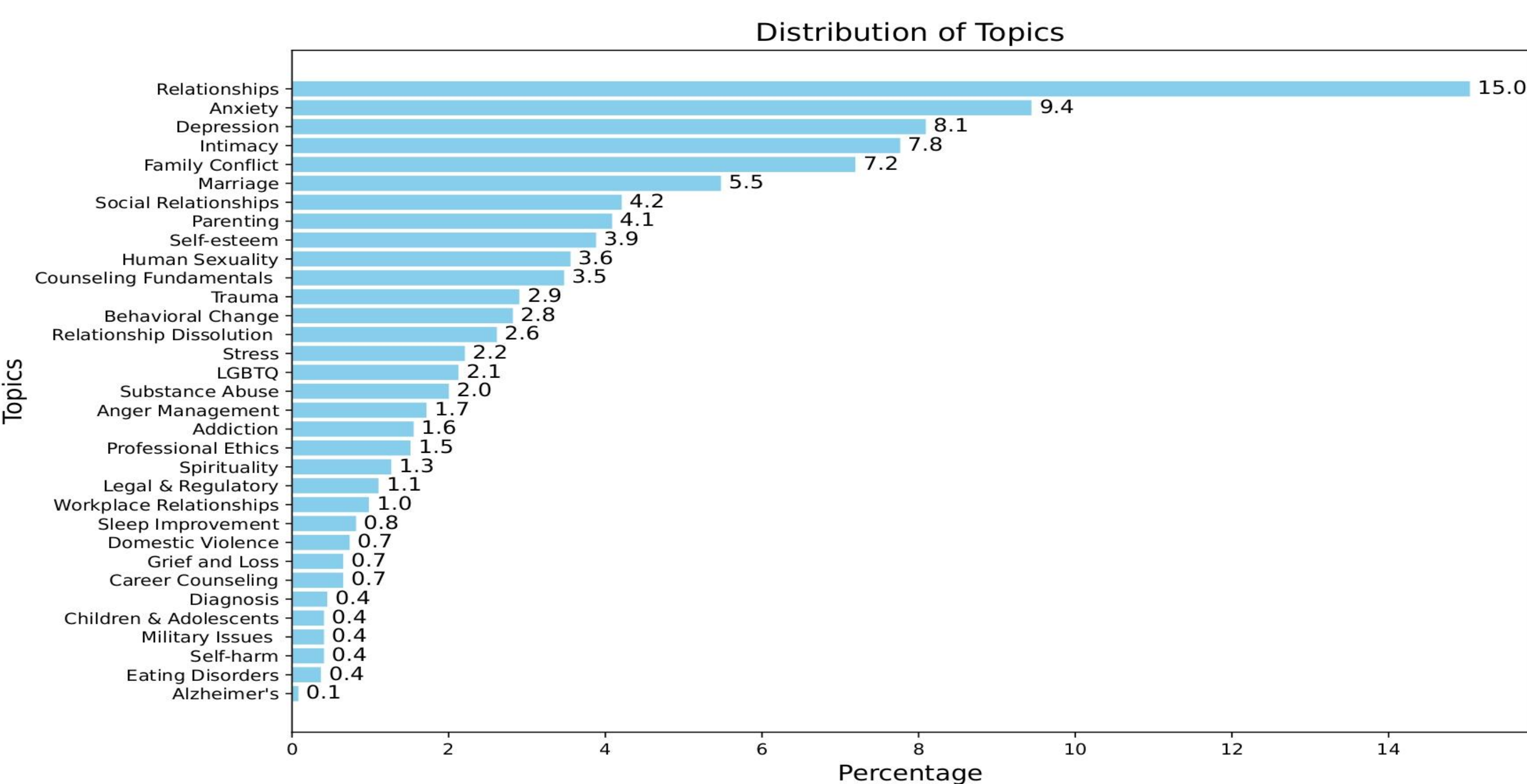
i. Real-World Interview Data

- **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.



ii. 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.



Finetuning 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.

Results

Model (7B)	Active Listening $\uparrow$		Empathy & Validation $\uparrow$		Safety & Trustworthiness $\uparrow$		Open-mindedness & Non-judgment $\uparrow$		Clarity & Encouragement $\uparrow$		Boundaries & Ethical $\uparrow$		Holistic Approach $\uparrow$		Average Score $\uparrow$	Human Rank $\downarrow$
	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 $\dagger$	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*	7.66*	7.34*	7.63*	7.46*	7.95*	7.46*	7.95*	7.87	3.65
LLaMA2 $\dagger$ *	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 $\dagger$	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 $\dagger$ *	7.60	8.15*	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	8.15	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 $\dagger$	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 $\dagger$ *	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 $\dagger$	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*	8.85*	8.31*	7.86*	7.76*	7.53	7.72*	7.79*	8.04*	8.04	1.53
Mixtral-8x7b-Instruct-V0.1 $\dagger$ *	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 $\dagger$	7.47*	8.10*	8.30*	8.44*	7.15*	7.78*	8.39*	8.34*	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 $\dagger$ *	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 $\dagger$	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 $\dagger$ *	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 $\dagger$	7.51*	8.11*	8.37*	8.47*	7.05	7.86*	8.51	8.34*	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 $\dagger$ *	7.66*	8.09*	8.53*	8.35*	7.51*	7.73*	8.64*	8.37*	7.65*	7.71*	7.16*	7.68*	7.35*	8.07*	7.90	2.85
ChatPsychiatrist $\S$	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 $\S$	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 $\S$	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), $\ast$: Model fine-tuned on Synthetic Data (10K), $\dagger\ast$: Model fine-tuned on both Synthetic and Interview Data (16K), $\S$: Baseline Model.

- **Overall Performance:** Models fine-tuned on MentalChat16K (synthetic, interview, or both) consistently outperformed base models across all evaluation methods (GPT-4 Turbo, Gemini Pro, and human experts), validating the dataset's effectiveness.
- **GPT-4 Turbo Evaluation:** Showed a strong preference for models fine-tuned on synthetic data, with synthetic variants winning in most metrics across all base models, likely due to alignment with GPT-3.5 Turbo-generated data.
- **Gemini Pro Evaluation:** Favored interview data for metrics requiring realism and contextual sensitivity, though synthetic and combined data also showed strong results in some areas.
- **Human Evaluation:** Human judges ranked fine-tuned models significantly higher than base and baseline models, with strong preference for synthetic fine-tuning, though interview and combined variants also performed well in some cases.

Evaluation

i. Mental Health LLM Evaluation Metrics

- Traditional LLM metrics (e.g., perplexity) lack the nuance to assess empathy, emotional sensitivity, and ethics in mental health counseling.
- We devised seven new domain-specific metrics to better evaluate mental health LLMs, offering more targeted and ethically aligned assessments.

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 the 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.

ii. Evaluation Dataset

- 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 inquiries.

iii. Evaluation Framework

- 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

- **Human and LLM Evaluation Correlation:** Pearson's correlation coefficients across 7 metric scores (Normal Range: $r = 0.10-0.40$)

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

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.

API Cost

	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

Github: <https://github.com/PennShenLab/MentalChat16K>

Huggingface Dacard: <https://huggingface.co/datasets/ShenLab/MentalChat16K>

Correspondence to Li.Shen@pennmedicine.upenn.edu