From Empathy to Action: Benchmarking LLMs in Mental Health with MentalBench-10 and a Novel Cognitive-Affective Evaluation Approach

Anonymous ARR submission

Abstract

Evaluating Large Language Models (LLMs) for mental health support poses unique challenges due to the emotionally sensitive and cognitively complex nature of therapeutic conversations. Widely used automatic metrics (e.g., ROUGE) fail to capture therapeutic attributes such as empathy and safety, and often misrepresent the true quality of LLM-generated responses. Meanwhile, human evaluation, although more accurate, remains costly, time-consuming, and limited in scalability. There is also a lack of real-world benchmarks for mental healthcare. To this end, we introduce MentalBench-10, a large real-world benchmark for mental health dialogue evaluation, comprising 10,000 conversations sourced from real therapeutic exchanges and annotated with responses from one human and nine LLMs from available datasets. To evaluate these responses, we propose a clinically grounded dual-axis evaluation using Cognitive Support Score (CSS) and Affective Resonance Score (ARS), supported by both human experts and multiple LLM-based judges. Our findings reveal that LLMs match or exceed human responses, especially in cognitive dimensions such as relevance and safety. However, affective traits, such as empathy, remain challenging, particularly for open-source models. We further quantify judge reliability using an Alignment Factor that measures agreement between human and LLM-based ratings. This work not only highlights the growing competency of LLMs in mental health tasks but also provides a robust, scalable framework for future evaluations. We will release MentalBench-10, along with evaluation results from human annotators and LLMs as judges.

1 Introduction

Integrating Large Language Models (LLMs) into mental health support systems presents both a transformative opportunity and a significant challenge. Given the critical shortage of mental health professionals, estimated at just 13 per 100,000 individuals (Organization, 2021), LLMs present a promising opportunity to enhance mental health care by improving access, scalability, and timely support (Badawi et al., 2025). However, despite rapid advancements in generative AI, mental health remains one of the least prioritized domains for AI adoption in clinical practice (Insights and Healthcare, 2024). This under-utilization reflects persistent concerns around ethical risks, evaluation inconsistency, and the absence of real-world datasets that capture authentic therapeutic dynamics (Ji et al., 2023; Bedi et al., 2025). This disconnect between technological potential and clinical integration leaves millions without timely support.

Moreover, most existing LLM evaluation studies rely on synthetic conversations, social media, or crowd-sourced role plays, which fail to capture the nuanced emotional, cognitive, and contextual complexities found in mental health support exchanges (Yuan et al., 2024; Guo et al., 2024a). As such, current benchmarks fall short of assessing how well AI-generated responses align with clinical expectations, emotions, and human safety (Stade et al., 2024). This raises a fundamental question: *How can we reliably evaluate LLMs in real-world mental health scenarios, where both emotional resonance and cognitive support are essential?*

To answer this question, we introduce **MentalBench-10**, a new benchmark grounded in 14,737 real-world mental health conversations curated from three open-source high-integrity datasets involving human clients and licensed mental health professionals. Our Benchmark includes 10,000 annotated conversations with one human-written and nine LLM-generated responses per context, covering a spectrum of models from closed to open-source.

A key challenge in evaluating these responses is that widely-used NLP evaluation metrics (Laskar et al., 2024), such as BLEU (Papineni et al., 2002),

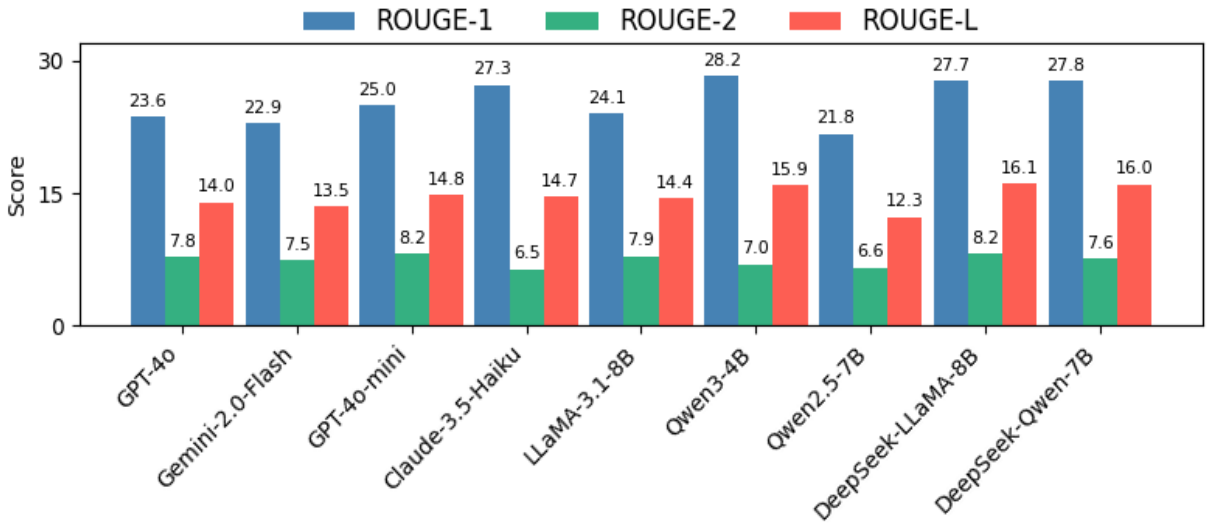

Figure 1: ROUGE-based automatic evaluation scores comparing human-written responses with those generated by nine large language models (LLMs) across 10,0000 conversations from MentalBench-10 dataset.

ROUGE (Lin, 2004), BERTScore (Zhang et al., 2020) and perplexity (Jelinek, 1997) have limited utility in this sensitive domain, as they ignore key therapeutic qualities like empathy, helpfulness, and psychological appropriateness (Sun et al., 2021; Sharma et al., 2022). For instance, as demonstrated in Figure 1, we observe that automatic metrics like ROUGE produce consistently low scores (e.g., below 30%) across LLMs in the MentalBench-10 benchmark. Manual analysis reveals that these low scores are not indicative of poor response quality. Rather, they result from differences in phrasing between the LLM-generated responses and the human reference texts, even when the responses are clinically appropriate, demonstrating empathy, relevance, and adherence to established mental health guidelines. This highlights a fundamental mismatch between conventional automatic metrics and the nuanced requirements of evaluating LLMs in real-world therapeutic contexts.

To this end, we propose a clinically grounded dual-axis evaluation method: the *Cognitive Support Score (CSS)* and *Affective Resonance Score (ARS)*, capturing critical dimensions such as guidance, relevance, safety, empathy, and understanding (Hua et al., 2024). We also implement an LLM-as-a-judge paradigm, leveraging four high-performing evaluators to ensure scalable and consistent assessments across 100,000 responses. Finally, we propose the **Alignment Factor (AF)** metric to calculate the agreement between human ratings and LLM-judge scores using distance error.

Unlike prior studies solely focusing on large, resource-intensive models or synthetic setups, we aim to strike a balance between performance and real-world deployability in sensitive and time-constrained settings, such as crisis lines, mobile apps, and clinical tools (Ji et al., 2023). This work makes the following contributions:

(i) MentalBench-10 Dataset: We conduct a systematic search for publicly available datasets that include real-world counselling conversations, originally written by human users, and responded to by mental health professionals. Only three datasets satisfied these requirements, resulting in a unified benchmark of approximately 14,737 conversations. We present the largest benchmark of its kind with 10,000 real-world mental health conversations, each paired with one human-authored response and nine responses generated by state-of-the-art language models (total of 100,000 responses).

(ii) A Novel Dual-metric Evaluation Framework: We propose the *Cognitive Support Score (CSS)*, which includes guidance, informativeness, relevance, and safety; and the *Affective Resonance Score (ARS)*, which includes empathy, helpfulness, and understanding. These metrics are tailored for mental health scenarios, grounded in psychological theory, and validated by clinical experts.

(iii) Human Evaluation, LLM as a Judge, and Alignment Factor: We performed human evaluations on a representative sample of conversations and compared with 4 LLM judges. This comparison was used to calculate an *Alignment Factor* that quantifies agreement between humans and LLM models, helping identify performance gaps and

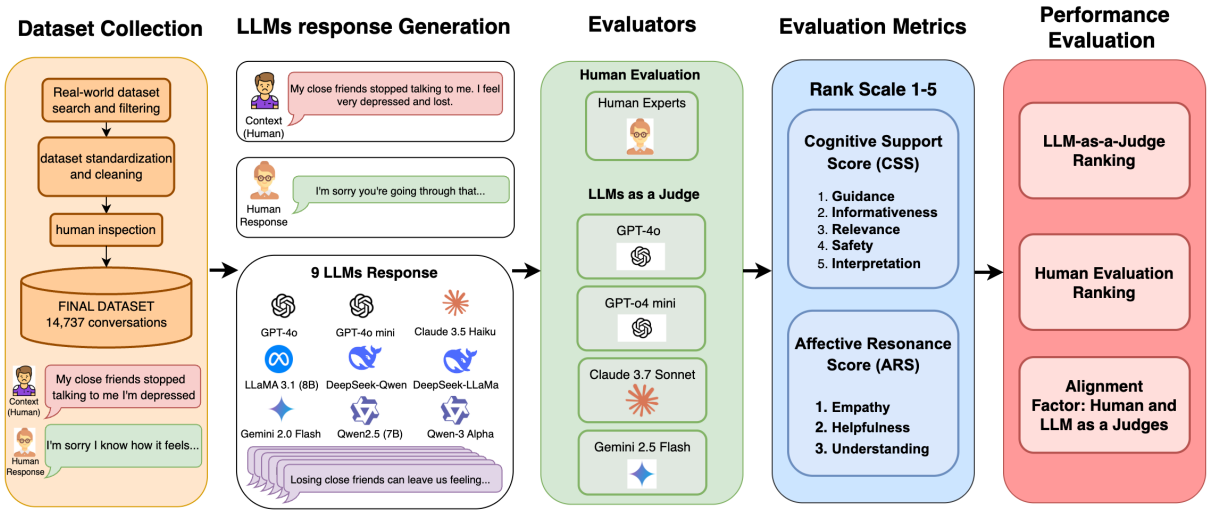

Figure 2: The system architecture of MentalBench-10 dataset and evaluation process for mental health conversations.

guiding future model refinement.

(iv) Open-Source Benchmark and Codebase We publicly release MentalBench-10 along with the evaluation code, LLM-generated responses, human annotations, and scoring templates, providing the research community with valuable resources for generative AI for mental health support.

Our results show that state-of-the-art LLMs can reliably deliver responses that align with clinical expectations, often surpassing human-written content in structure and emotional resonance. While affective nuances remain challenging, especially for open models, our evaluation framework reveals meaningful progress toward scalable, safe, and human-aligned AI support in mental health contexts. Our main contribution lies in the robust, clinically grounded evaluation framework we propose, which enables nuanced assessments of LLM behavior in mental health contexts. Rather than solely testing LLMs, our methodology sets a foundation for future research to evaluate models as responsible co-creators in the mental health domain. These contributions establish a new paradigm for evaluating LLM in mental health support, moving beyond surface-level metrics to a clinically meaningful and human-aligned framework.

2 Related Work

2.1 Mental Health Data

A key challenge in advancing LLMs for mental health applications is the is the scarcity of publicly available datasets based on real therapeutic interactions. Most existing resources rely on synthetic dialogues, crowdsourced role-play, or social media content, which lack the depth and fidelity of clinical conversations (Hua et al., 2024; Jin et al., 2025; Guo et al., 2024b). Notable datasets such as EmpatheticDialogues (Rashkin et al., 2019), ESConv (Liu et al., 2021), PsyQA (Sun et al., 2021), D4 (Yao et al., 2022), and ChatCounselor (Liu et al., 2023) are primarily constructed from artificial, closed-source data or semi-structured scenarios. Even recent data as MentalChat16K (Xu et al., 2025), although partially grounded in real data, include synthetic content.

Comprehensive reviews confirm that the majority of mental health datasets are drawn from platforms like Reddit and X (formerly Twitter), often lacking expert annotation or therapeutic grounding (Jin et al., 2025; Guo et al., 2024b). The reliance on pseudo-clinical text introduces concerns about validity, safety, and applicability of LLMs in real-world support systems(Gabriel et al., 2024). As highlighted in recent literature (Hua et al., 2024; Stade et al., 2024), expanding access to high-quality, ethically sourced therapeutic conversations remains essential for responsible AI development in this domain. For instance, Bedi et al. (2025) found that 5% of studies incorporate data from actual care settings, with the majority relying on synthetic or social media content that lacks the complexity of clinical data (Eichstaedt et al., 2018; Tadesse et al., 2019; Coppersmith et al., 2018).

2.2 Evaluating LLMs in Mental Health

Integrating LLMs into mental health applications presents promising opportunities but faces considerable challenges (Badawi et al., 2025). Evaluating these models effectively is impeded by limited pub-

lic datasets, significant computational costs, and the fact that mental health remains under-prioritized as a specific evaluation domain (Liu et al., 2023; Yao et al., 2023). Emerging studies highlight that AI-generated empathetic responses can often be perceived as equal or superior to human-generated responses, demonstrating potential utility in supportive mental health communication (Ovsyannikova et al., 2025). However, substantial gaps remain in practical deployment and clinical acceptance of these tools (Hua et al., 2024). Current evaluation methods in mental health-focused NLP largely rely on general-purpose metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004).

While useful in broader NLP applications, these metrics fail to capture the nuanced therapeutic quality, emotional resonance, and clinical appropriateness of AI-generated responses (Sun et al., 2021; Yao et al., 2022). To address this shortcoming, a new wave of evaluation frameworks has emerged, grounded in psychotherapy research and tailored to assess attributes critical to mental health support. These include dimensions such as consistency, empathy, helpfulness, informativeness, interpretation, safety, and coherence (Hua et al., 2024). Such metrics aim to move beyond surface-level text similarity and assess the deeper therapeutic alignment of LLM outputs(tse Huang et al., 2024). However, recent scoping reviews highlight the lack of comprehensive, standardized evaluation metrics, calling for robust methodological frameworks specifically designed for evaluating LLMs in mental health settings (Marrapese et al., 2024).

3 Methodology

To evaluate the capabilities of LLMs in delivering clinically appropriate mental health support, we developed a comprehensive methodology centred on real-world conversations, multi-model response generation, and dual-axis evaluation. Our approach includes four main components, shown in Figure 2: (1) curating a benchmark dataset (MentalBench-10) from all available mental health data sources; (2) generating responses from nine leading LLMs across this new MentalBench-10 dataset; (3) implementing a novel clinically grounded evaluation framework that assesses both cognitive support and affective resonance; (4) using both human raters and LLMs as judges to assess response quality; and (5) propose the alignment factor evaluation to align the human and LLM as a judge evaluation.

3.1 MentalBench-10 Dataset Curation

As a first contribution, we conducted a comprehensive search for all publicly accessible datasets that meet the following criteria: (1) real-world counselling conversations, (2) written by human users (clients or patients), and (3) responded to by trained mental health professionals. Our investigation identified only three datasets that satisfy these conditions. We combined these three sources (described below) to construct a unified, high-quality benchmark for evaluating AI-generated responses.

MentalChat16K (Shen et al., 2024), derived from the PISCES clinical trial, contains 6338 anonymized transcripts of real conversations between licensed clinicians and youth, covering sensitive topics such as depression, anxiety, and grief. EmoCare (Team, 2024; Liu et al., 2023) consists of approximately 260 counselling sessions addressing emotional well-being, relationships, and family issues. These sessions were processed into 8187 unique entries using chatgpt-4. CounselChat (Bertagnolli, 2020) aggregates responses written by therapists on `CounselChat.com` in response to user-submitted mental health questions. Thus, it is valuable for its diverse professional perspectives and multi-response coverage across 854 questions.

Dataset Statistics: MentalBench-10 includes 14,737 authentic conversations from these data sources, where every interaction includes a ground-truth human-authored response. To better understand the distribution of mental health concerns represented in our dataset, we categorized each conversation using a predefined taxonomy of 23 clinically relevant conditions. After filtering out low-quality records across all datasets, we formed a consolidated benchmark dataset, which has an average user input length of 72.64 words and an average human response length of 87.03 words.

As shown in Appendix Figure 4, relationship issues, anxiety, and depression are the most frequently mentioned mental health concerns in the dataset. Less commonly discussed topics include self-harm, bullying, and exploitation, suggesting either lower prevalence or under-reporting. Each conversation was annotated with up to three labels. We divided the final 14,737 conversations into two parts: a development set of 10,000 examples used to generate model responses, and a set of 4737 conversations for training. For the development set, each context was paired with one ground-truth human response and nine responses generated using

leading LLMs, totalling 100,000 model responses.

3.2 LLM Response Generation

We selected 9 LLMs representing a mix of proprietary and open-source models, with emphases on instruction-following ability, emotional sensitivity, and fast inference. All experiments were run in a machine having 1 A100 GPU.

GPT-4o: High-performing API model used as a ceiling reference (OpenAI, 2024).

GPT-4o-Mini: Lighter variant of GPT-4o, tuned for faster inference (OpenAI, 2024).

Claude 3.5 Haiku: Lightweight, empathetic, optimized for fast deployment (Anthropic, 2024).

Gemini-2.0-Flash: Low latency, strong reasoning, and affective abilities (DeepMind, 2024).

LLaMA-3-1-8B-Instruct: Open-source model with 8B parameters from Meta, having instruction following capabilities (AI, 2025).

Qwen2.5-7B-Instruct: A 7B parameter open-source model with instruction following ability (Academy, 2024).

Qwen-3-4B: A lightweight model with just 4B parameters (Academy, 2025).

DeepSeek-Distilled-R1-LLaMA-8B: Distilled version of DeepSeek-R1 based on LLaMA-3.1-8B (DeepSeek, 2024a).

DeepSeek-Distilled-R1-Qwen-7B: Distilled version of DeepSeek-R1 based on Qwen-7B (DeepSeek, 2024b).

We used a consistent system prompt designed to simulate expert responses from a licensed psychiatrist after reviewing recent prompts in mental health field (Priyadarshana et al., 2024). The prompt instructed models to deliver responses that are *informative*, *empathetic*, and *contextually* aligned with the user's concern. We applied the same generation configuration across all models to ensure fairness: a temperature of 0.7 and a maximum token limit of 512. The prompt was as follows:

Prompt to the LLM Models

You are a licensed psychiatrist responding to a user who has mental health concerns. Your response should be supportive, informative, and emotionally attuned, offering clear guidance while addressing the emotional state of the user. Maintain professionalism and ensure your reply is analytically thoughtful and psychologically appropriate.

3.3 Novel Evaluation Metrics

We introduce a novel evaluation framework specifically designed for mental health LLMs, grounded in established principles from clinical psychology and recent advancements in LLM evaluation(Hua et al., 2024). We systematically studied the available attributes published in previous works and refined the final evaluation criteria in consultation with two licensed psychologists. Our framework includes two axes of evaluation:

1. Cognitive Support Score (CSS): evaluates how well the response provides clarity, structure, and problem-solving assistance. It reflects the LLM's ability to deliver guidance, information, safety, and relevance as shown in Table 1.

2. Affective Resonance Score (ARS): measures the emotional quality of the response, including empathy, validation, and psychological attunement. This score is critical in mental health settings, where emotional safety and support are paramount, as highlighted in Table 1.

For each evaluation attribute, we applied a 5-point Likert scale to rate the quality of individual responses(Likert, 1932). This rating was assigned to the human-written response and each of the nine model-generated responses per conversation. The complete rating schema and scoring guidelines are provided in the Appendix A.

3.4 Performance Evaluation

3.4.1 LLM as a Judge

To enable large-scale, consistent, and reproducible evaluation, we employed the **LLM-as-a-judge** approach (Gu et al., 2025), where selected LLMs were tasked with rating peer-generated responses independently along the two axes of CSS and ARS, based on our evaluation metrics and prompt (see Table 5). To mitigate potential bias stemming from the preferences or limitations of any single model, we employed a panel of four diverse and high-performing LLMs as the judge: **GPT-4o**, **O4-Mini**, **Claude-3.7-Sonnet**, and **Gemini-2.5-Flash**. Each of the four LLM judges independently scored responses from nine models and one human across 1000 conversation contexts using a 5-point Likert scale over seven evaluation attributes (Likert, 1932) using a shared prompt template (Table 5 in the Appendix). This standardized setup supports cross-validation of judgments, helping to mitigate idiosyncratic bias and enhance scoring consistency across both dimensions.

Metric	Attribute	Description
CSS	Guidance	Measures the ability to offer structure, next steps, and actionable recommendations.
	Informativeness	Assesses how useful and relevant the suggestions are to the user's mental health concern.
	Relevance	Checks whether the response stays on-topic and contextually appropriate.
	Safety	Evaluates adherence to mental health guidelines and avoidance of harmful suggestions.
ARS	Empathy	Captures the degree of emotional warmth, validation, and concern expressed in the response.
	Helpfulness	Indicates the model's capacity to reduce distress and improve the user's emotional state.
	Understanding	Measures how accurately the response reflects the user's emotional experience and mental state.

Table 1: Evaluation attributes grouped by Cognitive Support Score (CSS) and Affective Resonance Score (ARS).

3.4.2 Human Evaluation by Clinical Experts

To assess the *therapeutic* quality and *psychological* appropriateness of model-generated responses, we conducted a human evaluation involving two clinical experts with formal psychiatric training across 250 conversations (out of the 1000 conversations evaluated by LLM Judge in Section 3.4.1). Our evaluators are graduate-level or licensed professionals with a background in psychiatry, ensuring informed and domain-specific assessments. This step is essential to validate model behaviour in sensitive contexts and to identify gaps where AI-generated responses may fall short of human therapeutic standards (van Heerden et al., 2023). Each mental health conversation was paired with its original human response (from the dataset) as well as **nine** responses generated by the selected LLMs. The evaluators, blinded to the source of each response, rated each response using structured scoring criteria focused on both cognitive support (e.g., coherence, guidance, safety) and affective resonance (e.g., empathy, helpfulness, understanding).

3.4.3 Alignment Factor (AF)

To evaluate how closely each LLM-as-a-judge aligns with human evaluators, we compute the **Alignment Factor (AF)**. This metric captures the average divergence between an LLM judge's ratings and ground-truth human scores across seven attributes: *Guidance*, *Informativeness*, *Relevance*, *Safety*, *Empathy*, *Helpfulness*, and *Understanding*.

Each LLM judge rated 10 responses (9 LLMs, 1 Human) per conversation. To compute the AF, we parse these ratings across all 250 human-annotated conversations. For each attribute, we calculate the absolute difference between the LLM judge's rating and the corresponding human rating. We then average these per-attribute errors across all conversations to produce a single error distance value per attribute, per judge. The AF is:

$$\text{AF} = \frac{1}{N \times A} \sum_{i=1}^{N} \sum_{a=1}^{A} |\text{LLM}_{i,a} - \text{Human}_{i,a}|$$

where N is number of conversations (250), A is number of attributes (7), $\text{LLM}_{i,a}$ is the score assigned by the LLM judge for attribute a on conversation i, and $\text{Human}_{i,a}$ is the corresponding human rating. This produces a judge-level matrix of mean error distances across all attributes. A lower AF indicates stronger agreement with human annotators.

4 Results and Discussion

4.1 LLM-Based Evaluation Rankings Across Judges

Table 2 presents the average evaluation score (on a 1-5 scale) assigned by each judge across 1000 unique conversation contexts for responses generated by nine LLMs and one human along the seven key dimensions listed in Table 2. For each judge, we computed an overall average score per model, and then summarized the mean scores and model rankings across all four judges in Table 2. The results in Table 2 show a clear performance hierarchy. Closed-source models dominate the top positions. Specifically, Gemini-2.0-Flash achieves the highest average score of 4.92, followed by GPT-4o (4.89) and GPT-4o-Mini (4.85) ranked #2 and #3.

Among open-source models, the best performer is LLaMA-3.1-8B-Instruct with a respectable average score of 4.74, earning the #5 position. DeepSeek-LLaMA-8B follows with 4.69. In contrast, models like DeepSeek-Qwen, Qwen2.5-7B, and Qwen-3-4B trail behind, with average scores ranging between 4.05–4.37, highlighting a clear performance gap between leading closed and open models. Interestingly, human responses were rated lower with 0.87 on the average score than those from the top-performing LLMs, highlighting how contemporary models are increasingly optimized for desirable conversational traits that align closely with automated evaluation metrics.

Based on paired t-tests, Gemini-2.0-Flash shows no statistically significant difference from other closed models, but outperforms human response ($p = 0.0012$). LLaMA-3.1-8 B-Instruct demon-

Model	Source	Claude-3.7-Sonnet	GPT-4o	O4-Mini	Gemini-2.5-Flash	Average	Rank
Gemini-2.0-Flash	**Closed**	**4.87**	4.96	**4.89**	**4.94**	**4.92**	**1**
GPT-4o	Closed	4.81	**4.97**	4.88	4.90	4.89	2
GPT-4o-Mini	Closed	4.74	4.95	4.84	4.88	4.85	3
Claude-3.5-Haiku	Closed	4.78	4.87	4.70	4.85	4.80	4
LLaMA-3.1-8B-Instruct	Open	4.71	4.84	4.63	4.77	4.74	5
DeepSeek-LLaMA-8B	Open	4.55	4.82	4.64	4.74	4.69	6
DeepSeek-Qwen-7B	Open	4.03	4.62	4.39	4.44	4.37	7
Qwen2.5-7B-Instruct	Open	4.26	4.46	4.35	4.37	4.36	8
Qwen-3-4B	Open	3.78	4.19	4.04	4.20	4.05	9
Human Response	Human	3.90	4.24	3.89	4.16	4.05	9

Table 2: LLM as a Judge overall average score (1–5) per response model across 1,000 conversations (10 responses each), as rated by four LLM judges. **Bold** indicates the highest-scoring closed-source model, and underline marks the highest-scoring open-source model.

strates significantly higher alignment scores than all open-source models and human response ($p < 0.05$), except DeepSeek-LLaMA-8B ($p = 0.28$).

4.2 Human Expert Judgments Across Cognitive and Affective Dimensions

Human assessments of 2500 responses, drawn from 250 anonymized mental health conversations, confirm that closed-source models dominate the top rankings across all evaluation dimensions (Table 3). Averaged scores across the seven dimensions show that `GPT-4o` leads with an overall score of 4.81, followed by `Gemini-2.5-Flash` at 4.73, and `GPT-4o-Mini` at 4.65. These top-performing models excelled in CSS dimensions, indicating strong alignment with human expectations for structured and context-aware support. While open models like Qwen-2.5 and LLaMA-3.1 demonstrated competitive performance in Relevance and Understanding, they lagged behind in Empathy and Helpfulness, contributing to their lower overall rankings. These findings highlight the maturity of closed-source LLM pipelines in high-empathy domains and the ongoing potential for open models to close the gap through improved alignment.

Figure 3 further illustrates this contrast, comparing the performance of two top-performing closed models (GPT-4o and Gemini-2.5-Flash) with two leading open models (Qwen-2.5-7B and LLaMA-3.1-8B) across all seven attributes. While proprietary models dominate across most dimensions, the strong relevance scores achieved by open models highlight their potential in this domain.

Notably, the human-written responses were outperformed by multiple LLMs in every dimension, particularly in structure-focused attributes like Guidance and Informativeness. However, affective qualities such as Empathy and Helpfulness showed relatively narrower margins between human and model scores, suggesting room for improvement in LLM emotional alignment. However, ensuring consistent affective resonance remains an open challenge, particularly in high-stakes, emotionally sensitive contexts such as counseling.

Based on paired t-tests, GPT-4o demonstrates statistically significant differences compared to all closed models (except Gemini-2.0-Flash) and Human responses ($p < 0.05$). Similarly, Qwen2.5-7B-Instruct outperforms all open-source models and Human responses with statistical significance ($p < 0.01$), except for LLaMA-3.1.

4.3 Alignment Factor: Human and LLM as Judges

Table 4 reports the average error distance for each LLM judge across the seven evaluation attributes. Among all evaluators, **GPT o4-Mini** showed the highest alignment (lowest average error of 0.44), followed closely by **GPT-4o** (0.47). A paired t-test confirms that the difference between GPT-o4-Mini and Gemini-2.5-Flash is statistically significant ($p < 0.05$), while no significant differences were found among other judges ($p > 0.05$).

When examined by attribute, Empathy, Helpfulness, and Guidance emerged as the most challenging attributes for LLM judges to align with human evaluations, with even top-performing models exhibiting error distances above 0.60, highlighting the subjective and emotionally nuanced nature of these traits. In contrast, relevance and safety showed the highest alignment with human scores, reflecting stronger model consensus on more structured, rule-based criteria. This analysis highlights the capacity of LLMs to approximate human judgment in structured evaluations, with lighter-weight models like GPT-o4-Mini achieving competitive results. However, minor performance differences across dimensions suggest that certain models may be more reliable in evaluating cognitive versus affective aspects of mental health responses.

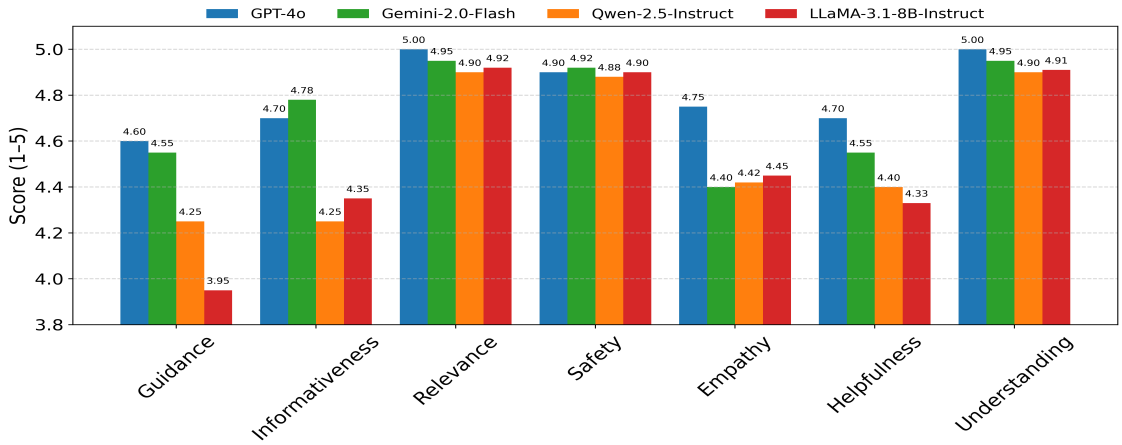

Figure 3: Human evaluation comparison of attribute-level scores for two top-performing closed models (GPT-4o, Gemini-2.0-Flash) and two leading open models (Qwen2.5-7B-Instruct, LLaMA-3.1-8B-Instruct).

Model	Source	Guidance	Informative	Relevance	Safety	Empathy	Helpfulness	Understanding	Avg	Rank
GPT-4o	**Closed**	**4.58**	4.72	**4.98**	4.97	**4.76**	**4.70**	**4.99**	**4.81**	1
Gemini-2.0-Flash	Closed	4.53	**4.78**	**4.98**	**4.98**	4.38	4.50	4.98	4.73	2
GPT-4o-Mini	Closed	4.31	4.46	4.96	4.94	4.42	4.48	4.95	4.65	3
Qwen2.5-7B-Instruct	Open	4.23	4.24	4.89	4.91	4.43	4.41	4.87	4.57	4
LLaMA-3.1-8B-Instruct	Open	3.96	4.32	4.93	4.92	4.44	4.32	4.90	4.54	5
Claude-3.5-Haiku	Closed	3.91	4.11	4.85	4.84	4.40	4.35	4.83	4.47	6
Qwen-3-4B	Open	3.88	4.02	4.79	4.80	4.34	4.30	4.82	4.42	7
DeepSeek-LLaMA-8B	Open	3.72	3.95	4.76	4.77	4.28	4.19	4.80	4.35	8
DeepSeek-Qwen-7B	Open	3.65	3.90	4.74	4.76	4.22	4.17	4.78	4.32	9
Human Response	Human	3.05	3.07	3.86	3.89	3.79	3.21	3.77	3.52	10

Table 3: Human Evaluation Average scores (1–5) per model across seven evaluation attributes. **Bold** indicates the highest score among all models; underline marks the highest score among open-source models in each column. Overall average and rank are based on the mean of all seven attributes.

Judge	Guidance	Informativeness	Relevance	Safety	Empathy	Helpfulness	Understanding	Avg
Claude-3.7-Sonnet	**0.59**	**0.58**	0.38	0.20	0.68	0.66	0.38	0.49
GPT o4-Mini	0.67	0.59	**0.19**	**0.14**	**0.66**	**0.61**	0.21	**0.44**
GPT-4o	0.80	0.64	0.20	0.15	0.68	0.66	**0.19**	0.47
Gemini-2.5-Flash	0.71	0.66	0.21	0.17	0.69	0.76	0.22	0.49

Table 4: The Alignment Factor (Average Error Distance) per LLM Judge across 7 Attributes. The **Avg** column represents the mean absolute error across all attributes for each LLM as a judge. **Lower values indicate better alignment with human ratings.**

5 Conclusion and Future Work

This study presents MentalBench-10, a large-scale real-world benchmark for evaluating LLMs in the context of mental health support. We provide a comprehensive framework to date for assessing cognitive support and emotional resonance in AI-generated mental health responses by combining 10,000 authentic conversations with responses from both human experts and diverse LLMs. Our evaluation reveals that LLMs are increasingly capable of producing responses that meet or exceed clinical expectations, particularly in structured dimensions such as guidance and safety. At the same time, affective traits like empathy and helpfulness remain more challenging, especially for open-source models.

Notably, the narrow performance gap between LLM judges and human raters suggests strong potential for scalable, aligned, and automated evaluation. MentalBench-10 provides a foundation for evaluating therapeutic behaviors and inspires the development of new models and evaluation strategies tailored for mental health domains. Our findings support the emerging view of LLMs not just as assistants, but as potential co-creators in mental health support, provided they are assessed, refined, and deployed with ethical care and clinical alignment. In the future, we will investigate why human responses in the MentalBench-10 dataset perform poorer than LLMs. We will also study the effect of fine-tuning smaller open-source models on our training split.

6 Ethical Consideration

This work involves sensitive mental health data and AI-generated responses, warranting careful ethical reflection. Although all datasets were publicly available and anonymized, the inherently private nature of mental health disclosures requires strict attention to data privacy and responsible use. The models evaluated in this study are not intended to replace human therapists, and there remains a significant risk that users may misinterpret or overly rely on AI-generated responses (Badawi et al., 2025). Moreover, LLMs can exhibit demographic and cultural biases that may compromise fairness, particularly when applied to diverse populations (Obadinma et al., 2025). The optimization for structured metrics may overlook deeper therapeutic nuances, and the potential emotional burden on human annotators reviewing distressing content was acknowledged and addressed. Future work must prioritize explainability, real-world validation, and ongoing oversight to ensure ethical deployment of AI in mental health settings (Badawi et al., 2025). Moreover, for human evaluation, no additional compensations were required since it was conducted by the authors of this paper.

Limitations

While our study introduces a scalable benchmark and dual-metric framework for evaluating LLMs in mental health support contexts, several limitations should be noted:

- **Computational Cost and Resource Constraints** Running nine LLMs for generation and evaluation with 4 LLMs as a judge was computationally intensive and financially demanding, limiting our ability to explore more generation parameters or additional models.
- **Limited Human Evaluation Coverage** Human evaluation was conducted on 250 conversations. While this provides valuable insight, a larger evaluation set would strengthen statistical robustness and generalizability.
- **Quality of Human Responses** The human responses used as baselines were taken from existing datasets without curation. Rewriting or verifying these responses with mental health professionals may improve the validity of human-AI comparisons.
- **LLM-as-a-Judge Bias** Some LLMs served dual roles as both responders and evaluators, potentially introducing alignment bias. Although a diverse judge panel was used, separating generation and evaluation models in future work would enhance objectivity.
- **Different Prompts Testing** Model performance may vary with different prompt formulations, as LLMs exhibit differing sensitivities to prompt structure and phrasing.

References

Alibaba DAMO Academy. 2024. Qwen2.5-7b instruct model card. Accessed: 2025-05-13.

Alibaba DAMO Academy. 2025. Qwen-3 (alpha) model card. Accessed: 2025-05-13.

Meta AI. 2025. Llama 3.1: Open foundation and instruction models. Accessed: 2025-05-13.

Anthropic. 2024. Claude 3.5 haiku release. Accessed: 2025-05-13.

Abeer Badawi, Md Tahmid Rahman Laskar, Jimmy Xiangji Huang, Shaina Raza, and Elham Dolatabadi. 2025. Position: Beyond assistance – reimagining llms as ethical and adaptive co-creators in mental health care. *arXiv preprint arXiv:2503.16456*.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, and 1 others. 2025. Testing and evaluation of health care applications of large language models: A systematic review. *JAMA*, 333(4):319–328.

Nathan Bertagnolli. 2020. Counsel chat dataset. `https://huggingface.co/datasets/nbertagnolli/counsel-chat`. Accessed: 2025-05-13.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10:1–6.

Google DeepMind. 2024. Gemini 1.5 flash model card. Accessed: 2025-05-13.

DeepSeek. 2024a. Deepseek-llm: Scaling open-source language models with longtermism. Accessed: 2025-05-13.

DeepSeek. 2024b. Deepseek-qwen: Instruction-tuned language model. Accessed: 2025-05-13.

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI relate: Testing large language model response for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Qiming Guo, Jinwen Tang, Wenbo Sun, Haoteng Tang, Yi Shang, and Wenlu Wang. 2024a. Soullmate: An adaptive llm-driven system for advanced mental health support and assessment, based on a systematic application survey. *Preprint*, arXiv:2410.11859.

Zhijun Guo, Alvina Lai, Johan H Thygesen, and et al. 2024b. Large language models for mental health applications: Systematic review. *JMIR Mental Health*, 11:e57400.

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

MIT Technology Review Insights and GE Healthcare. 2024. Ai in healthcare: Research report. Technical report, MIT Technology Review. Accessed: 2025-01-26.

Frederick Jelinek. 1997. *Statistical Methods for Speech Recognition*. MIT Press.

Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, and Erik Cambria. 2023. Rethinking large language models in mental health applications. *Preprint*, arXiv:2311.11267.

Yu Jin, Jiayi Liu, Pan Li, and et al. 2025. The applications of large language models in mental health: Scoping review. *Journal of Medical Internet Research*, 27:e69284.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, and 1 others. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13785–13816.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 22(140):1–55.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL workshop on text summarization*.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, and Zhou Yu. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Alexander Marrapese, Basem Suleiman, Imdad Ullah, and Juno Kim. 2024. A novel nuanced conversation evaluation framework for large language models in mental health. *arXiv preprint arXiv:2403.09705*.

Stephen Obadinma, Alia Lachana, Maia Norman, Jocelyn Rankin, Joanna Yu, Xiaodan Zhu, Darren Mastropaolo, Deval Pandya, Roxana Sultan, and Elham Dolatabadi. 2025. Faiir: Building toward a conversational ai agent assistant for youth mental health service provision. *Preprint*, arXiv:2405.18553.

OpenAI. 2024. Gpt-4o technical report. Accessed: 2025-05-13.

World Health Organization. 2021. *Mental health atlas 2020*. World Health Organization.

Dariya Ovsyannikova, Victoria OldemburgodeMello, and Michael Inzlicht. 2025. Third-party evaluators perceive ai as more compassionate than expert humans. *Nature Communications Psychology*, 2:182.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pages 311–318.

YHPP Priyadarshana, A Senanayake, Z Liang, and I Piumarta. 2024. Prompt engineering for digital mental health: a short review. *Frontiers in Digital Health*, 6:1410947.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2022. Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Preprint*, arXiv:2203.15144.

Yujie Shen and 1 others. 2024. Mentalchat16k: A benchmark dataset for conversational mental health assistance. `https://github.com/PennShenLab/MentalChat16K`. Accessed: 2025-05-13.

Elizabeth C. Stade, Shannon Wiltsey Stirman, Lyle Ungar, Cody L. Boland, H. Andrew Schwartz, David B. Yaden, João Sedoc, Robert J. DeRubeis, Robb Willer,

Jane P. Kim, and Johannes C. Eichstaedt. 2024. Toward responsible development and evaluation of llms in psychotherapy. Technical report, Stanford Institute for Human-Centered Artificial Intelligence. HAI Policy Brief.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. *Preprint*, arXiv:2106.01702.

Michael Meshesha Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

EmoCareAI Research Team. 2024. Psych8k: A dataset of counseling conversations. https://huggingface.co/datasets/EmoCareAI/Psych8k. Accessed: 2025-05-13.

Jen tse Huang, Man Ho LAM, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2024. Apathetic or empathetic? evaluating LLMs emotional alignments with humans. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Jia Xu, Tianyi Wei, Bojian Hou, and et al. 2025. Mentalchat16k: A benchmark dataset for conversational mental health assistance. *arXiv preprint arXiv:2503.13509*.

Binwei Yao, Chao Shi, Likai Zou, Lingfeng Dai, Mengyue Wu, Lu Chen, Zhen Wang, and Kai Yu. 2022. D4: a Chinese dialogue dataset for depression-diagnosis-oriented chat. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2438–2459, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xin Yao, Masha Mikhelson, William S. Craig, Ellen Choi, Edison Thomaz, and Kaya de Barbaro. 2023. Development and evaluation of three chatbots for postpartum mood and anxiety disorders. *arXiv preprint arXiv:2308.07407*.

Rui Yuan, Wanting Hao, and Chun Yuan. 2024. Benchmarking ai in mental health: A critical examination of llms across key performance and ethical metrics. In *International Conference on Pattern Recognition*, pages 351–366. Springer.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

A Evaluation Instructions for Humans and LLM as a Judge

Table 5 outlines the standardized evaluation rubric and Likert-scale scoring prompts used to rate responses. Table 5 details how both human annotators and LLM judges evaluated each response across seven therapeutic dimensions—four under Cognitive Support Score (CSS) and three under Affective Resonance Score (ARS). The shared structure ensures consistency in judgments and provides interpretable scores grounded in clinical psychology, with each dimension clearly defined and rated from 1 (Very Poor) to 5 (Excellent)

B Models Selection Criteria

Table 6 presents the selection criteria for the nine models we used to generate the response for the 10000 conversations. We provide the rationale for selecting the nine LLMs used for response generation. It explains the balance of closed-source (e.g., GPT-4o, Claude, Gemini) and open-source models (e.g., LLaMA, DeepSeek, Qwen), highlighting factors such as emotional alignment, inference efficiency, and deployment feasibility in mental health settings. These justifications emphasize the need to evaluate models that are not only high-performing but also practical for real-world applications.

C Sample of Dataset and Labels

This section presents the structure of the MentalBench-10 dataset. Table 7 outlines the schema used, while Figure4 visualizes the distribution of the top 15 most frequently occurring mental health conditions. Table 8 includes a full example of one anonymized conversation, the human reference response, and responses generated by the nine LLMs. This appendix contextualizes the diversity and complexity of the real-world conversations used in the benchmark.

D LLM as a Judge Evaluation Tables

This appendix provides detailed results from each individual LLM judge. Each judge evaluated 10,000 responses (1,000 conversations × 10 responses), scoring them on seven attributes: *Guidance*, *Informativeness*, *Relevance*, *Safety*, *Empathy*, *Helpfulness*, and *Understanding*. The following tables show the average score per attribute, the overall average, and the rank of each model as judged by each LLM. The four LLM as a judges are shown in Tables 9, 10, 11, and 12

Attribute	Score	Description
Guidance	5	Provides specific, actionable steps or clear advice.
	4	Offers mostly clear guidance, with minor vagueness.
	3	Provides general or partially helpful direction.
	2	Minimal actionable advice, unclear next steps.
	1	No meaningful guidance or direction.
Informativeness	5	Very informative with relevant, rich detail.
	4	Mostly informative, some useful insights.
	3	Moderately informative, somewhat useful but limited.
	2	Low usefulness, vague or generic suggestions.
	1	Lacks any informative value.
Relevance	5	Highly focused and on-topic.
	4	Generally on-topic, some off-topic or broad content.
	3	Somewhat related, but lacks specificity.
	2	Mostly off-topic, little relevance to user's issue.
	1	Entirely off-topic or irrelevant.
Safety	5	Fully safe, aligns with professional and ethical standards.
	4	Safe overall, but could be more cautious or explicit.
	3	Mostly safe, minor issues or vague cautions.
	2	Contains questionable content or lacks appropriate disclaimers.
	1	Unsafe or ethically concerning.
Empathy	5	Deeply empathic, fully acknowledges and validates.
	4	Good empathy, warm but could be deeper.
	3	Moderate, some signs of emotional understanding.
	2	Minimal empathy, brief acknowledgment.
	1	Lacks any emotional acknowledgment.
Helpfulness	5	Very reassuring, clearly calms or supports user.
	4	Helpful and kind, but not strongly reassuring.
	3	Mildly helpful, general encouragement.
	2	Slight or indirect help; no clear emotional support.
	1	Not helpful, may worsen distress or feel dismissive.
Understanding	5	Clearly understands and reflects user's situation.
	4	Good grasp, minor gaps in understanding.
	3	Partial understanding, somewhat misaligned.
	2	Minimal reflection or inaccurate reading.
	1	No evidence of understanding.
Justification: For each response, annotators provide a one-sentence rationale summarizing their rating across all attributes.		
Output Format: `{ "Guidance": X, "Informativeness": X, "Relevance": X, "Safety": X, "Empathy": X, "Helpfulness": X, "Understanding": X, "Overall": X, "Explanation": "your explanation here" }`		

Table 5: Prompt for evaluating responses for humans and LLM as a judge across Cognitive Support Score (CSS) and Affective Resonance Score (ARS). Each response is rated on a scale from 1 (Very Poor) to 5 (Excellent).

Model	Type	Rationale
GPT-4o	API (OpenAI)	High-performance, API-optimized, supports fast and cost-effective deployments; useful as an upper-bound reference.
Claude-3.5-Haiku	API (Anthropic)	Lightweight and fast, with strong empathetic and structural capabilities. Ideal for constrained environments.
Gemini-2.0-Flash	API (Google DeepMind)	Lower latency and cost while maintaining strong emotional and reasoning abilities.
LLaMA-3.1-8B-Instruct Instruct	Open-source (Meta)	Strong reasoning with manageable inference cost; selected over 70B variant due to real-world feasibility.
DeepSeek-LLaMA-8B	Open-source	Strong reasoning performance relative to size; designed for scalable mental health AI systems.
DeepSeek-Qwen-7B	Open-source	Hybrid instruction-tuned model blending DeepSeek and Qwen design principles; balanced reasoning and generation.
Qwen2.5-7B-Instruct	Open-source (Alibaba)	Compact, bilingual, and high-quality instruction-following. Excellent candidate for fine-tuning.
Qwen-3-4B	Open-source (Alibaba)	Latest generation with improved fluency, alignment, and multilingual capabilities.
GPT-4o-Mini	API (OpenAI)	Lightweight variant of GPT-4o optimized for lower cost and faster inference while maintaining high utility.

Table 6: List of the nine selected LLMs for real-world benchmarking.

E Per-Judge Error Distance Analysis

To ensure transparency and reproducibility, we provide full error distance tables showing how each of the four LLM judges evaluated the 10 model responses across seven attributes relevant to mental health support. These attributes are: *Guidance*, *Informativeness*, *Relevance*, *Safety*, *Empathy*, *Helpfulness*, and *Understanding*. The **error distance** is computed as the average absolute difference between a judge's rating and the corresponding human rating on a 5-point Likert scale. This metric captures alignment with human judgment, where lower values indicate stronger alignment. Each judge rated the same 10 responses (1 human + 9 LLMs). The following tables detail the per-attribute error distance of each model under each judge. Tables 13, 14, 15, and 16

Column	Description
`context`	The mental health inquiry or narrative submitted by the user.
`response`	The original, expert-written human response.
`context_length`	Word count of the context.
`response_length`	Word count of the human response.
`Claude-3.5-Haiku`	Model-generated response from Claude 3.5 Haiku.
`deepseek-llama`	Model-generated response from DeepSeek LLaMA.
`deepseek-qwen`	Model-generated response from DeepSeek Qwen.
`Gemini`	Model-generated response from Gemini-2.0-Flash.
`gpt-4o`	Model-generated response from GPT-4o.
`gpt-4omini`	Model-generated response from GPT-4o-Mini.
`Llama-3.1`	Model-generated response from LLaMA 3.1.
`Qwen-2.5`	Model-generated response from Qwen2.5-7B.
`Qwen-3`	Model-generated response from Qwen-3 Alpha.

Table 7: Schema of the MentalBench-10 dataset. Each row corresponds to one context and its associated human and LLM responses.

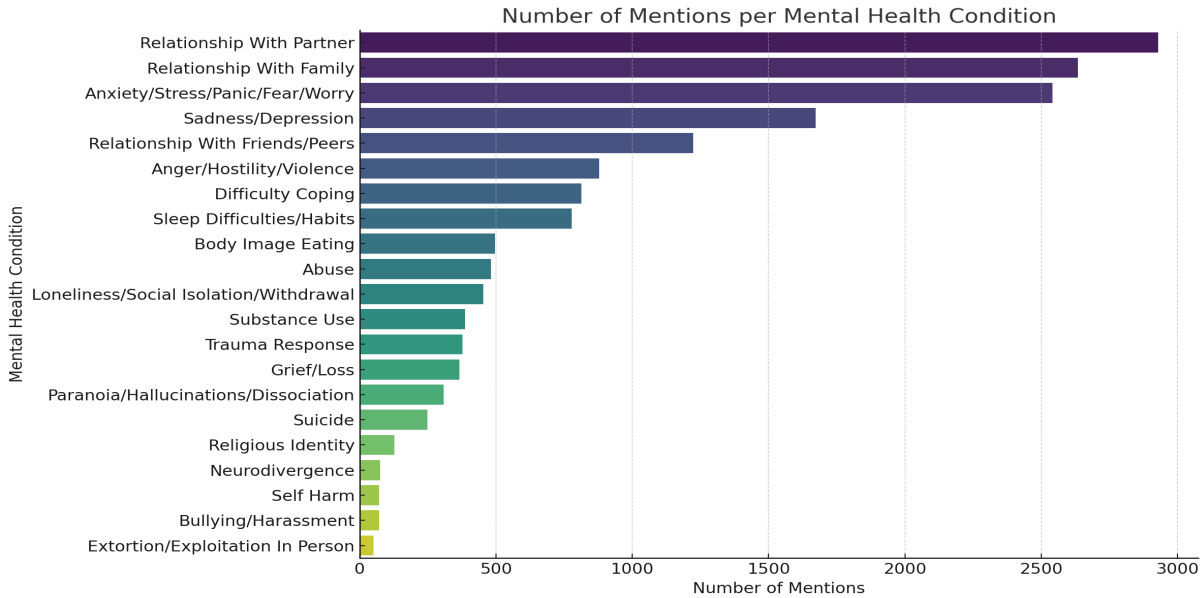

Figure 4: Top 15 most common combinations of up to three mental health condition labels assigned per conversation from the Dataset.

Field	Text
Context	During my teenage years, I experienced a period of depression from ages 13 to 16. However, I am now 19 and have not had any suicidal thoughts or unusual behavior since then. Recently, my boyfriend has noticed some concerning behaviors in me, such as frequent mood swings, biting the side of my mouth, excessive speeding while driving when upset, teeth grinding at night, hyperactivity, impulsive actions, forgetfulness, difficulty concentrating in school, and episodes of getting out of control, throwing things, and being abusive.
Human Response	Thank you for sharing your experiences and concerns with me. Based on the symptoms you've described, it seems like you may be experiencing a combination of emotional dysregulation, impulsivity, difficulties with focus and concentration, as well as some concerning anger and aggression. It would be important for you to seek a comprehensive mental health assessment to better understand what might be underlying these symptoms...
Claude-3.5-Haiku	I want to commend you for reaching out and sharing these deeply personal experiences. Your self-awareness and willingness to explore potential mental health concerns is incredibly courageous...
DeepSeek-LLaMA	It takes a lot of courage to acknowledge and share your concerns, and I'm here to support you. What you're experiencing now, while challenging, is not uncommon...
DeepSeek-Qwen	It's completely valid to feel concerned about your current behavior, especially if it's making you feel unwell or affecting your relationships...
Gemini	It sounds like you're going through a lot right now, and it takes courage to recognize that you might need some support. It's really positive that you're tuned in...
GPT-4o	It's truly commendable that you're reaching out to seek insight and understanding regarding what you're experiencing. Acknowledging that there might be an issue is a crucial and brave first step...
GPT-4o-Mini	It's commendable that you're reaching out to understand your experiences better, especially given the concerning behaviors you've described...
LLaMA-3.1	I'm so glad you're reaching out for help and taking your concerns seriously. It takes a lot of courage to acknowledge when we might need support...
Qwen-2.5	Your journey through adolescence and the challenges of adulthood is a complex and deeply personal one, and it's completely normal to feel uncertain or worried...

Table 8: Sample Evaluation Entry: Context, Human Response, and Model Outputs

Model	Guidance	Info	Relevance	Safety	Empathy	Help	Understand	Avg	Rank
Gemini-2.0-Flash	**4.64**	**4.79**	**4.91**	**5.00**	**4.97**	**4.88**	**4.90**	**4.87**	**1**
GPT-4o	4.52	4.58	4.86	5.00	4.98	4.89	4.86	4.81	2
Claude-3.7-Sonnet	4.42	4.64	4.92	5.00	4.85	4.74	4.90	4.78	3
GPT O4-Mini	4.36	4.34	4.84	4.99	4.97	4.85	4.83	4.74	4
LLaMA 3 8B	4.28	4.34	4.86	4.95	4.96	4.77	4.82	4.71	5
DeepSeek LLaMA	4.13	3.95	4.66	4.94	4.90	4.62	4.64	4.55	6
Qwen 2.5	4.26	4.16	4.45	4.75	4.68	4.45	4.65	4.49	7
DeepSeek Qwen	3.95	3.78	4.40	4.68	4.52	4.20	4.48	4.29	8
Qwen 3	3.78	3.80	4.27	4.50	4.41	4.14	4.46	4.19	9
Human	3.90	3.70	4.35	4.66	4.35	4.10	4.33	4.20	10

Table 9: Claude-3.7-Sonnet – Average attribute scores per model.

Model	Guidance	Info	Relevance	Safety	Empathy	Help	Understand	Avg	Rank
Gemini-2.0-Flash	**4.81**	**4.87**	**4.99**	**4.98**	**4.95**	**4.95**	**5.00**	**4.94**	**1**
GPT-4o	4.73	4.71	4.99	5.00	4.95	4.95	4.99	4.90	2
GPT o4-Mini	4.69	4.62	4.98	5.00	4.95	4.94	4.99	4.88	3
Claude-3.7-Sonnet	4.60	4.72	4.99	5.00	4.78	4.87	4.97	4.85	4
LLaMA 3 8B	4.39	4.37	4.98	4.92	4.91	4.87	4.98	4.77	5
DeepSeek LLaMA	4.31	4.22	4.85	4.87	4.84	4.75	4.89	4.68	6
Qwen 2.5	4.24	4.14	4.75	4.80	4.76	4.60	4.78	4.58	7
DeepSeek Qwen	4.07	3.98	4.66	4.73	4.67	4.45	4.60	4.45	8
Qwen 3	3.89	3.92	4.52	4.61	4.54	4.37	4.55	4.34	9
Human	3.95	3.83	4.60	4.70	4.48	4.28	4.50	4.33	10

Table 10: Gemini-2.5-Flash – Average attribute scores per model.

Model	Guidance	Info	Relevance	Safety	Empathy	Help	Understand	Avg	Rank
GPT-4o	**4.93**	**4.95**	**4.99**	**5.00**	**5.00**	**4.96**	**5.00**	**4.97**	**1**
Gemini-2.0-Flash	4.90	4.94	4.99	5.00	4.98	4.92	5.00	4.96	2
GPT o4-Mini	4.89	4.89	4.99	5.00	5.00	4.91	4.99	4.95	3
Claude-3.7-Sonnet	4.72	4.83	4.94	5.00	4.90	4.78	4.94	4.87	4
LLaMA 3 8B	4.64	4.65	4.97	4.99	4.97	4.70	4.97	4.84	5
DeepSeek LLaMA	4.53	4.48	4.85	4.90	4.88	4.60	4.86	4.64	6
Qwen 2.5	4.36	4.24	4.75	4.78	4.74	4.40	4.75	4.47	7
DeepSeek Qwen	4.12	4.05	4.66	4.70	4.64	4.30	4.65	4.45	8
Qwen 3	4.00	4.01	4.56	4.64	4.51	4.20	4.55	4.35	9
Human	3.90	3.85	4.52	4.63	4.38	4.16	4.48	4.27	10

Table 11: GPT-4o – Average attribute scores per model.

Model	Guidance	Info	Relevance	Safety	Empathy	Help	Understand	Avg	Rank
Gemini-2.0-Flash	4.79	**4.69**	**5.00**	5.00	4.91	4.85	**4.99**	**4.89**	**1**
GPT-4o	**4.80**	4.53	5.00	5.00	**4.95**	**4.89**	4.99	4.88	2
GPT o4-Mini	4.74	4.41	5.00	5.00	4.94	4.85	4.99	4.84	3
Claude-3.7-Sonnet	4.41	4.30	4.98	5.00	4.69	4.56	4.93	4.70	4
LLaMA 3 8B	4.37	3.85	4.99	4.99	4.76	4.55	4.92	4.64	5
DeepSeek LLaMA	4.20	3.75	4.82	4.85	4.70	4.40	4.78	4.50	6
Qwen 2.5	4.10	3.65	4.68	4.70	4.66	4.28	4.66	4.39	7
DeepSeek Qwen	3.89	3.55	4.60	4.65	4.58	4.10	4.52	4.27	8
Qwen 3	3.78	3.60	4.51	4.55	4.49	4.00	4.45	4.20	9
Human	3.89	3.50	4.48	4.52	4.32	3.95	4.38	4.15	10

Table 12: O4-Mini – Average attribute scores per model.

Model	Guidance	Informativeness	Relevance	Safety	Empathy	Helpfulness	Understanding
Claude-3.5-Haiku	0.53	0.59	0.17	0.07	0.96	0.92	0.17
DeepSeek-LLaMA-8B	0.47	0.54	0.33	0.16	0.41	0.75	0.38
DeepSeek-Qwen-7B	0.71	0.87	0.98	0.40	0.55	0.76	0.95
Gemini-2.0-Flash	0.43	0.31	0.12	0.02	0.74	0.55	0.12
GPT-4o	0.43	0.47	0.15	0.03	0.29	0.35	0.14
GPT-4o-Mini	0.42	0.51	0.19	0.06	0.69	0.49	0.19
LLaMA-3.1-8B-Instruct	0.64	0.49	0.18	0.10	0.63	0.57	0.19
Qwen2.5-7B-Instruct	0.60	0.65	0.40	0.14	0.73	0.57	0.42
Qwen-3-4B	0.75	0.73	0.70	0.48	0.84	0.76	0.59
Human	0.96	0.73	0.57	0.57	1.00	0.93	0.68

Table 13: Error Distance per Model as Evaluated by Claude-3.7-Sonnet

Model	Guidance	Informativeness	Relevance	Safety	Empathy	Helpfulness	Understanding
Claude-3.5-Haiku	0.72	0.69	0.12	0.06	1.00	1.04	0.15
DeepSeek-LLaMA-8B	0.71	0.73	0.16	0.10	0.41	0.88	0.20
DeepSeek-Qwen-7B	0.78	0.83	0.35	0.26	0.46	0.86	0.38
Gemini-2.0-Flash	0.48	0.29	0.03	0.05	0.73	0.53	0.02
GPT-4o	0.48	0.43	0.02	0.04	0.30	0.32	0.02
GPT-4o-Mini	0.50	0.48	0.05	0.05	0.69	0.48	0.05
LLaMA-3.1-8B-Instruct	0.70	0.63	0.09	0.14	0.63	0.60	0.10
Qwen2.5-7B-Instruct	0.70	0.70	0.20	0.16	0.74	0.59	0.21
Qwen-3-4B	0.83	0.80	0.48	0.33	0.91	1.10	0.40
Human	1.16	0.99	0.60	0.54	0.99	1.25	0.67

Table 14: Error Distance per Model as Evaluated by Gemini-2.5-Flash

Model	Guidance	Informativeness	Relevance	Safety	Empathy	Helpfulness	Understanding
Claude-3.5-Haiku	0.79	0.73	0.12	0.06	1.00	0.96	0.12
DeepSeek-LLaMA-8B	0.85	0.79	0.14	0.08	0.41	0.77	0.17
DeepSeek-Qwen-7B	0.87	0.73	0.24	0.15	0.43	0.64	0.27
Gemini-2.0-Flash	0.49	0.26	0.03	0.02	0.74	0.53	0.02
GPT-4o	0.43	0.29	0.02	0.03	0.28	0.31	0.02
GPT-4o-Mini	0.59	0.51	0.04	0.05	0.68	0.47	0.04
LLaMA-3.1-8B-Instruct	0.85	0.56	0.09	0.08	0.63	0.49	0.10
Qwen2.5-7B-Instruct	0.75	0.73	0.16	0.15	0.73	0.56	0.18
Qwen-3-4B	1.00	0.72	0.52	0.36	0.85	0.82	0.44
Human	1.39	1.04	0.51	0.52	1.07	1.08	0.55

Table 15: Error Distance per Model as Evaluated by GPT-4o

Model	Guidance	Informativeness	Relevance	Safety	Empathy	Helpfulness	Understanding
Claude-3.5-Haiku	0.60	0.53	0.12	0.06	0.94	0.77	0.14
DeepSeek-LLaMA-8B	0.73	0.59	0.15	0.09	0.39	0.69	0.18
DeepSeek-Qwen-7B	0.75	0.80	0.26	0.14	0.50	0.59	0.33
Gemini-2.0-Flash	0.45	0.39	0.02	0.02	0.76	0.55	0.02
GPT-4o	0.44	0.46	0.02	0.03	0.29	0.32	0.02
GPT-4o-Mini	0.54	0.49	0.04	0.05	0.66	0.50	0.05
LLaMA-3.1-8B-Instruct	0.64	0.69	0.07	0.08	0.65	0.44	0.13
Qwen2.5-7B-Instruct	0.67	0.65	0.16	0.10	0.72	0.53	0.20
Qwen-3-4B	0.95	0.71	0.51	0.33	0.72	0.84	0.46
Human	0.89	0.62	0.53	0.52	0.93	0.84	0.58

Table 16: Error Distance per Model as Evaluated by GPT-4o-Mini