

Do Large Language Models Align with Core Mental Health Counseling Competencies?

Viet Cuong Nguyen¹, Mohammad Taher¹, Dongwan Hong¹, Vinicius Konkolics Possobom¹,
Vibha Thirunellayi Gopalakrishnan¹, Ekta Raj¹, Zihang Li², Heather Jamie Soled³,
Michael L. Birnbaum⁴, Srijan Kumar¹, Munmun De Choudhury¹,

¹Georgia Institute of Technology, ²Hofstra University, ³Tower Health, ⁴Columbia University,

Correspondence: johnny.nguyen@gatech.edu

Abstract

The rapid evolution of Large Language Models (LLMs) presents a promising solution to the global shortage of mental health professionals. However, their alignment with essential counseling competencies remains underexplored. We introduce CounselingBench, a novel NCMHCE-based benchmark evaluating 22 general-purpose and medical-finetuned LLMs across five key competencies. While frontier models surpass minimum aptitude thresholds, they fall short of expert-level performance, excelling in Intake, Assessment & Diagnosis but struggling with Core Counseling Attributes and Professional Practice & Ethics. Surprisingly, medical LLMs do not outperform generalist models in accuracy, though they provide slightly better justifications while making more context-related errors. These findings highlight the challenges of developing AI for mental health counseling, particularly in competencies requiring empathy and nuanced reasoning. Our results underscore the need for specialized, finetuned models aligned with core mental health counseling competencies and supported by human oversight before real-world deployment. Code and data associated with this manuscript can be found at: <https://github.com/cuongnguyenx/CounselingBench>

1 Introduction

Despite the critical importance of mental healthcare for individual and societal well-being, a significant global accessibility crisis continues to exist. Even in highly developed nations like the United States, access to adequate mental health services remains alarmingly insufficient. Current estimates indicate that more than half of the U.S. population resides in designated Mental Health Professional Shortage Areas (MHPSAs), regions where the number of mental health professionals falls short of meeting the population's needs (Heisler and Bagalman,

2013). This shortage poses a major public health challenge (Heisler and Bagalman, 2013), as it hinders timely intervention, contributes to untreated mental health conditions, and exacerbates disparities in care.

Language forms the foundation of mental healthcare, underpinning all interactions and interventions between patients and care providers. Recent advances in Large Language Models (LLMs) offer significant potential to alleviate the aforementioned global shortage in mental healthcare, attributed to their state-of-the-art performance in diverse natural language understanding tasks without additional fine-tuning. In fact, numerous research and commercial efforts have been directed at building LLM-based therapists and counselors to meet people's mental health needs (Lai et al., 2023). Increasing numbers of people are also appropriating general-purpose LLMs to find support and advice that may not be available through conventional means (De Choudhury et al., 2023).

However, mental health is a fundamentally “human” experience, and addressing its challenges requires a nuanced blend of empathy, cultural sensitivity, and clinical expertise. Effective mental health care necessitates a range of competencies that are sensitive to the myriad and often complex manifestations of individuals' mental health journeys (Clasen et al., 2003). These competencies include conducting thorough psychological assessments, which involve interpreting both verbal and non-verbal cues, understanding patients' unique life contexts, and identifying subtle signs of distress or underlying conditions (Hoge et al., 2005). Additionally, clinicians must develop personalized treatment plans that account for each patient's background, personal history, and presenting symptoms (Banikiotes, 1977). This process is not merely about applying standard diagnostic criteria; it involves setting treatment goals, choosing interventions that align with patients' values, and

Competency	Definition	Example Question	#
Counseling Skills & Interventions (CS&I)	Counselors' knowledge, skills, and abilities to conduct effective counseling.	Q: How would you approach the client's expressed fears concerning drinking? A: Your ethical responsibility is to advocate for this as an advancement of client care.	599
Intake, Assessment, & Diagnosis (IA&D)	Counselors' knowledge, skills, and abilities to effectively conduct client intake, assessment, and diagnosis.	Q: Which assessment tool would you use to conduct a comprehensive evaluation of the client's current cognitive and emotional functioning? A: Beck Depression Inventory (BDI)	460
Professional Practice & Ethics (PP&E)	Counselors' knowledge, skills, and abilities related to maintaining proper administrative and clinical protocols.	Q: Of the following, which is an ethically appropriate action to take to address the issue associated with your treatment facility's closing in six months? A: You express your advocacy for affordable treatment by writing an editorial for the newspaper.	274
Treatment Planning (TP)	Counselors' knowledge, skills, and abilities to develop an effective course of treatment.	Q: You feel this client would benefit from additional professional help. To whom would you refer the client for help with her depression and sleep issues? A: Psychiatrist	253
Core Counseling Attributes (CCA)	Behaviors, traits, and dispositions of effective counselors.	Q: You do not share the same religious views as the client. Which of the following would be the most helpful approach when considering the impact of your own beliefs in counseling? A: You continuously assess how your beliefs may affect the counseling process	23

Table 1: Description of key counseling competencies identified by the NBCC, along with example questions for each competency in CounselingBench

suggesting strategies based on patients' ongoing responses to treatment.

Furthermore, mental illnesses exhibit significant clinical heterogeneity (Wardenaar and de Jonge, 2013) – not only are the conditions themselves diverse, encompassing everything from mood and anxiety disorders to complex cases of psychosis or personality disorders, but the symptoms and lived experiences manifest in deeply subjective ways. Effective mental healthcare, therefore, demands a flexible, patient-centered approach that can accommodate this diversity (Zangeneh and Al-Krenawi, 2019). This raises critical questions about whether LLMs, despite their impressive capabilities in natural language understanding, can truly replicate the intricate, human-centered nature of mental health counseling. In this paper, we therefore seek to explore whether LLMs are equipped to demonstrate these core mental health counseling competencies.

We present CounselingBench, a new benchmark designed to evaluate LLM performance in the context of key mental health counseling competencies. CounselingBench is based on the National Clinical Mental Health Counseling Examination (NCMHCE), a U.S. licensing exam for mental health counselors that assesses five core competencies identified by a broad survey of professionals in the field. Through CounselingBench, we systematically assess how well LLMs can process and apply domain-specific knowledge from case studies to

address questions that evaluate these key competencies. Our research aims to address the following questions:

RQ1: Are large language models capable of successfully passing the NCMHCE?

RQ2: How accurately can large language models respond to NCMHCE questions covering various mental health counseling competencies?

RQ3: How effectively can LLMs generate rationales for their answers to these competency questions?

Our contributions are as follows:

- We provide a novel benchmark, CounselingBench, designed to assess LLM capabilities across key mental health counseling competencies. This benchmark is based on the National Clinical Mental Health Counseling Examination (NCMHCE), ensuring its relevance and alignment with professional standards in the field.

- We conduct a comprehensive evaluation of various LLMs, including both general-purpose and medical-specialized models, using CounselingBench. This evaluation provides insights into the current capabilities and limitations of LLMs in mental health counseling tasks.

- We analyze the performance of LLMs across different competencies, identifying areas of strength and weakness. This analysis can guide future fine-tuning efforts for LLMs and development of LLM-enabled tools for mental health applica-

tions.

- We examine the reasoning capabilities of LLMs in formulating responses to mental health counseling questions, shedding light on their ability to synthesize contextual information and domain-specific knowledge to explain their decisions in high-stakes counseling scenarios .

- We compare the performance of medical LLMs with general-purpose LLMs, highlighting specific shortcomings in the current medical models regarding mental health counseling competencies and identifying areas for potential improvement.

2 Related Works

2.1 LLMs for Mental Health

Online psychological counseling has experienced significant growth, especially in the wake of the COVID-19 pandemic (Yurayat and Tuklang, 2023). This increased demand has spurred research into the application of large language models (LLMs) for various mental health services (Stade et al., 2024; Lawrence et al., 2024). However, mental healthcare is inherently complex, as it involves nuanced aspects of empathy, emotional intelligence, and context-specific interaction (Fried and Robinson, 2020). This complexity is compounded by the diversity of evaluation methods—ranging from analyses of annotated social media posts (Lamichhane, 2023) to assessments based on clinical vignettes (Inbar Levkovich, 2023)—which makes cross-study comparisons challenging.

Recent research increasingly focuses on tailoring LLMs towards specific therapeutic interventions. For example, the Chain-of-interaction framework (Han et al., 2024) leverages dyadic contexts to enrich LLM understanding of psychiatric behaviors. This approach is designed to capture the iterative, interactive nature of counselor-client exchanges and potentially enhance the empathy and engagement critical to effective motivational interviewing. In parallel, efforts in Cognitive Behavior Therapy (CBT) have also emerged. CBT-BENCH (Zhang et al., 2024) provides a structured evaluation of LLM capabilities in assisting CBT sessions, building upon previous work such (Chen et al., 2023b) which investigates the detection of cognitive distortions through diagnostic thought prompting. These advances demonstrate the growing trend of adapting LLMs to domain-specific therapeutic tasks, highlighting both their potential and the need for standardized, clinically grounded evaluation frameworks across multiple mental health intervention

techniques and competencies.

Despite the promise shown by these specialized approaches, existing evaluation metrics often only moderately correlate with human clinical judgment, even though LLMs generally offer better explainability than traditional supervised models (Yang et al., 2024). In contrast, our work introduces the first large-scale, mental health-specific evaluation benchmark that aligns with the competencies required of aspiring licensed mental health counselors, thereby aiming to bridge the gap between automated assessment and clinical practice.

2.2 Competency Evaluation of LLMs

The rapid evolution of LLMs has led to an extensive body of work evaluating their performance across a broad spectrum of tasks. General benchmarks have focused on natural language understanding, mathematical reasoning, coding, and even social science knowledge (Hendrycks et al., 2020). For instance, some benchmarks assess LLM performance on graduate-level academic information (Rein et al., 2023), while others test mathematical reasoning abilities using grade-school problems (Cobbe et al., 2021) or natural science word problems from the ARC dataset (Hu et al., 2024).

In addition to these general assessments, several domain-specific benchmarks have been developed. The MedQA benchmark (Jin et al., 2021) evaluates LLMs’ ability to process and respond to medical literature based on exam-style questions, and LegalBench (Neel Guha, 2023) assesses legal reasoning skills. Building on this paradigm, our work extends the evaluation landscape by proposing the first comprehensive benchmark for assessing LLM proficiency in mental health counseling. This benchmark is designed not only to measure technical performance but also to capture the clinical and empathetic subtleties essential for real-world mental health practice, paving the way for more human-centered and clinically valid applications of LLM technology.

3 Curating CounselingBench

We collected the case study details (including patient demographic, mental status examination, presenting problem, etc.), questions, their associated answers and expert-generated rationale from National Clinical Mental Health Counseling Examination (NCMHCE) questions. The exam seeks to assess the proficiency of individuals seeking to become licensed clinical mental health counselors

across five key mental health counseling competencies as detailed in Table 1. These competencies are derived from a national job analysis involving over 16,000 credentialed counselors, which identified empirically-validated work behaviors that are considered most relevant for effective counseling practice (NBCC, 2023). NCMHCE questions and associate details listed above are collected from mock exams which are accessible online for public usage. Details about the sources that we collected data from are described in Appendix Table A1

Overall, we collected a total of 1612 unique questions across 138 case studies that constitutes CounselingBench. CounselingBench contains cases whose subject exhibit diverse ethnic and cultural background in addition to presentation of mental health conditions. For instance, there exists considerable number of case studies whose patients are from ethnic minority backgrounds such as Black (20 case studies, 241 questions), Hispanic (13 case studies, 151 questions), Multiracial (4 case studies, 48 questions), Asian (4 case studies, 46 questions), Native American (2 case studies, 24 questions). In addition, there also exist several case studies in CounselingBench where the patient was not born in the United States where cultural competency is essential to correctly answer associated questions. CounselingBench also examines diverse mental health condition, with 51 unique conditions found among questions. Many such conditions have low prevalence in the real world, such as cyclothymic disorder (0.5-1% of adults, 13 questions) or reactive attachment disorder of childhood (1% of children, 13 questions). Table A2 provides the full distribution of mental health conditions within CounselingBench

A full example question, along with the case study context, can be found in Appendix Table A3. To comply with fair use law, we adapt the procedure used in Jin et al. 2020 (Jin et al., 2021) and shuffle the order of answer options (while keeping track with the correct answer). Given that each of the questions are designed to uniquely assess candidates’ abilities in one specific competency, we manually annotate all 1612 questions to provide a complete expert-derived question-competency mapping for downstream analysis. Two annotators, both of whom are collaborators on the paper and are medical doctors specializing in psychiatry, independently annotate all questions for one of five competencies as specified in Table 1. For each question, they were instructed to carefully read

through the case study details, question statement and candidate answers in its entirety and then select the counseling competency that best described the competency which the question aims to test based on the NCMHCE Content Outline (NBCC, 2023). This yielded 845 questions where the 2 annotators were in agreement about the competency annotation and 767 questions where they were in disagreement. For questions where the 2 annotators’ labels disagreed, we invited a third annotator who is a experienced licensed mental health counselor (LMHC) based in the U.S. to provide the tiebreaking vote. The number of questions determined by experts annotators to reflect each of the five core mental health-related competencies can be found in Table 1

4 Methodology

4.1 Model Selection

Our study encompasses 13 open-source medical models, selected for their outstanding performance across various biomedical NLP tasks. These models represent seven distinct finetuning architectures: BioMedGPT (Zhang et al., 2023), Asclepius (Kweon et al., 2023), Meditron (Chen et al., 2023a), MentaLlama (Yang et al., 2024), Clinical-Camel (Toma et al., 2023), Med42 (Christophe et al., 2024), and OpenBioLLM (Pal and Sankarababu, 2024). To provide a comprehensive analysis, we also include the corresponding un-finetuned base models, henceforth referred to as *generalist models*, of the aforementioned medical models, adding 10 more models to our study. Both sets of medical models and their generalist counterparts were chosen to represent a wide range of model sizes. Finally, for benchmarking against state-of-the-art proprietary systems, we include gpt-4o-2024-08-06 in our comparison (Bubeck et al., 2023; Achiam et al., 2023) due to its top performance on LLM leaderboards such as Chatbot Arena (Chiang et al., 2024). Detailed information regarding model parameters and training data sources can be found in Appendix Table A4.

4.2 Inference

We formally define a question in the CounselingBench dataset as $Q = (\mathcal{C}, \mathcal{A}_{cand}, \mathcal{A}_{corr})$, where \mathcal{C} is the context (including question text and patient demographics), \mathcal{A}_{cand} are candidate answers, and \mathcal{A}_{corr} is the correct answer. Q^h is an abridged version of Q without \mathcal{A}_{corr} . To elicit answers from

a large language model \mathcal{M} given \mathcal{Q} , we use the following prompting strategies:

- *Zero-shot (ZS)*: \mathcal{M} answers \mathcal{Q}^h without task-specific training, using the prompt template in Appendix Table A5.
- *Few-shot (FS)*: \mathcal{M} answers \mathcal{Q}^h after seeing demonstrative questions $\mathcal{Q}_1, \dots, \mathcal{Q}_n$ (Brown, 2020), using the prompt template in Appendix Table A6. We use $n = 3$ random demonstrative examples.
- *Chain-of-thought (CoT)*: We augment each few-shot example with a step-by-step explanation \mathcal{R} towards the correct answer. The input includes $(\mathcal{Q}_1, \dots, \mathcal{Q}_n), (\mathcal{R}_1, \dots, \mathcal{R}_n)$ and \mathcal{Q}^h , using the prompt template in Appendix Table A7. We use the intermediary reasoning chains for comparison in RQ3.

We also test self-consistency (SC) decoding (Wang et al., 2022) on all prompting strategies. While originally applied to CoT prompting, we extend SC to zero-shot and few-shot prompting by performing 5 samplings each at temperature $t \in [0.2, 0.4, 0.6, 0.8, 1]$ and taking the majority label across all 25 samplings.

5 RQ 1: Can LLMs pass the NCMHCE

5.1 Method

To assess LLMs’ performance on CounselingBench, we calculate overall accuracy on all 1612 questions. Given varying real-world NCMHCE passing scores, we use an average threshold of 63% accuracy based on previous reported NCMHCE passing scores. To compare medical and generalist models, we employ paired t-tests between each medical model and its unfinetuned generalist counterpart.

5.2 Results

Table 2 shows the overall performance on the CounselingBench for all models tested across different prompting and decoding settings.

5.3 Zero-shot LLMs can pass the NCMHCE

We found that frontier LLMs (in September 2024) with zero-shot prompting are able to perform on NCMHCE at a level which exceeds the pass threshold of 63% accuracy as defined above. These passing models are primarily larger in size (5 out of 6 models have more than 70B parameters) and are all

Model	ZS	ZS + SC	FS	FS + SC	FS + COT
Generalist Models					
Llama-2-7B	.408	.357	.326	.37	.335
Llama-2-7B-in	.432	.444	.451	.461	.412
Llama-3-8B-in	.622	.646	.643	.654	.595
Llama-2-13B	.45	.437	.423	.464	.406
Llama-2-13B-in	.526	.525	.529	.543	.493
Llama-2-70B	.596	.575	.59	.644	.241
Llama-2-70B-in	.616	.616	.631	.644	.431
Llama-3-70B-in	.717	.731	.734	.739	.71
Medical Models					
Meditron-7B	.258	.293	.244	.318	.066
Asclepius-7B	.28	.307	.233	.294	.279
BioMedGPT-LM-7B	.409	.381	.386	.432	.105
Asclepius-Llama3-8B	.339	.359	.363	.386	.269
OpenBioLLM-8B	.565	.585	.583	.607	.533
Llama3-Med42-8B	.639	.638	.643	.654	.572
Asclepius-13B	.338	.381	.32	.38	.257
MentaLLaMA-13B-in	.452	.461	.485	.488	.373
ClinicalCamel-70B	.619	.679	.657	.698	.398
Med42-70B	.68	.679	.57	.687	.635
Meditron-70B	.557	.551	.587	.633	.315
Llama3-Med42-70B	.688	.696	.704	.705	.623
OpenBioLLM-70B	.698	.725	.716	.734	.686
Proprietary Models					
gpt4o	.78	.765	.723	.748	.767

Table 2: Accuracy of tested LLMs on CounselingBench across different prompting and decoding settings

instruction-tuned (whether on a general instruction fine-tuning dataset or one specific to the biomedical domain). Not surprisingly, gpt4o is the best performing model among all the tested models, achieving a zero-shot accuracy of 0.78. However, it is also notable that the best open-source model (Llama3-70B-it) only performed slightly worse (8.8% reduction in zero-shot accuracy) than gpt4o, despite being significantly smaller in parameter size. We also note that while there is a large gap in performance between the smallest and largest version of instruction-tuned Llama2 (0.432 vs 0.616, 42.6% gain), that gap has substantially decreased between Llama3-8B-it and Llama3-70B-it (0.622 vs 0.717, 15.3% gain). This suggests significant improvements in Llama3’s model architecture and data curation processes compared to Llama2, and that current and future smaller-scale models are becoming increasingly viable for mental health counseling tasks while maintaining a lower computational footprint. On the flipside, performance trends on the CounselingBench also imply that performance gains from scaling up model sizes will diminish, or even disappear despite future advancements in model architecture, training data curation and procedure. This matches with observations from Anwar et al. (2024), McKenzie et al. (2023) and Zhou et al. (2024a) on the potential limits of LLM scal-

ing laws. Finally, we note that while the performance of frontier LLMs exceed that of the minimum passing level, it remains substantially lower than that expert-level human performance, which we set at 90% based on expert-level human scores on biomedical QA benchmarks such as MedMCQA and MedQA (Liévin et al., 2024)

5.4 Medical models underperformed generalist models across all settings accuracy-wise

Most surprisingly, we notice that from Table 2 that a supermajority of generalist models seem to consistently outperform their medical fine-tuned counterparts (10 out of 13 pairs) under zero-shot setting, with an average difference of 4.2 percentage points between generalist-medical model pairs in zero-shot accuracy (maximum = 0.15pp, minimum = -0.084pp). We deploy paired t -test to assess significant differences between generalist and medical models' zero-shot accuracy on CounselingBench, and find a significant difference between these two distributions ($t = 2.939, p = 0.013$). This suggests a systematic underperformance of medical LLMs on mental health counseling-related questions compared to their un-finetuned counterparts. We perform a more fine-grained evaluation of this underperformance in the next research question.

6 RQ 2: How do LLMs perform across different mental health competencies

6.1 Method

To achieve a more nuanced, competency-centric assessment of model performance, we disaggregate the accuracy metrics in Table 2 into five competency-specific accuracies. This approach: (1) evaluates model capabilities across key counseling domains, reflecting diverse patient needs, and (2) elucidates factors contributing to performance discrepancies between medical-specialist and generalist model pairs.

6.2 LLMs are better at treatment planning, worse at counseling skills, interventions

Table 3 represents zero-shot accuracies of the 5 mental health counseling competencies across all tested LLMs. There are small yet significant variations across different competencies. Across all models, we found that their performance on different competencies (based on zero-shot accuracy) can be sorted as Treatment Planning = Intake, Assessment & Diagnosis >* professional practice &

ethics = core counseling attributes = counseling skills and interventions, where > indicates a statistically significant positive difference between zero-shot accuracy across competencies, with the number of stars in the superscript representing the p-value of the paired t -test¹. Such variations remain even under other inference settings (such as few-shot). The observed variations in accuracy across mental health counseling competencies might be attributed to the nature of the tasks and the inherent strengths and limitations of LLMs. Competencies like *Treatment Planning* and *Intake, Assessment & Diagnosis* are more procedural and well-represented in model training data, allowing LLMs to perform better here. In contrast, questions on competencies such as *Counseling Skills and Interventions* and *Core Counseling Attributes* require more nuanced social, emotional intelligence and cultural competency. These are areas where LLMs have been shown to struggle in even with careful prompting (Zhou et al., 2024b; Chiu et al., 2024).

6.3 Medical models significantly underperform their generalist counterparts across most competencies

Comparing the cross-competency performance (measured with zero-shot accuracy) between generalist models and medical models with paired t -test, we find that generalist models significantly overperform their paired medical counterparts across 4 out of 5 key mental health-related competencies: Intake, Assessment & Diagnosis ($t = 3.15, p = 0.027$), Treatment Planning ($t = 2.52, p = 0.035$), Counseling Skills & Interventions ($t = 3.61, p = 0.024$), and Professional Practice & Ethics ($t = 2.5, p = 0.035$). Among open-source models, no one model has the highest performance across all five competencies. OpenBioLLM-70B achieves the highest performance in treatment planning, whereas Llama3-70B-in achieves the highest performance in the remaining four out of five competencies (even outperforming gpt4o in the "core counseling attribute" competency). These discrepancies in model performance persist even when alternative prompting techniques and decoding settings (such as few-shot, few-shot with self-consistency, etc.) are applied, as seen in Appendix Tables A8 and A9. These observed differences might stem from variations in the training data and the patterns each LLM has learned. Medical LLMs

¹* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

are primarily fine-tuned on biomedical research papers and medical question answering datasets, which may give them increased performance on the “treatment planning” competency. However, this seems to be at the expense of their retention of knowledge patterns required for the 4 other mental health competencies. In contrast, generalist models, trained and instruction-tuned on a broader range of data, possess and leverage more knowledge patterns that are needed for non-clinical mental health competencies such as counseling skills & interventions in diverse settings. These variations highlight the challenges of developing AI systems for diverse mental health counseling skills, where the need for both emotional intelligence and clinical precision is paramount.

model	IA&D	TP	CS&I	PP&E	CCA
Generalist Models					
Llama2-7B	0.433	0.418	0.390	0.402	0.392
Llama-2-13B	0.465	0.469	0.452	0.446	0.575
Llama-2-70B	0.628	0.628	0.589	0.569	0.479
Llama-2-7B-in	0.449	0.473	0.419	0.427	0.217
Llama-2-13B-in	0.531	0.553	0.504	0.539	0.474
Llama-2-70B-in	0.647	0.621	0.598	0.619	0.439
Llama3-8B-in	0.641	0.629	0.640	0.579	0.603
Llama-3-70B-in	0.730	0.689	0.741	0.699	0.650
Medical Models					
BioMedGPT-7B	0.436	0.437	0.362	0.426	0.487
OpenBioLLM-8B	0.556	0.623	0.568	0.530	0.444
MentaLLaMA-13B	0.473	0.480	0.427	0.448	0.576
Llama3-Med42-8B	0.623	0.634	0.648	0.624	0.568
ClinicalCamel-70B	0.663	0.624	0.596	0.607	0.560
Asclepius-13B	0.290	0.380	0.351	0.347	0.437
Llama3-Asclepius-8B	0.341	0.371	0.372	0.299	0.300
Meditron-7B	0.267	0.262	0.242	0.282	0.168
Meditron-70B	0.576	0.593	0.564	0.515	0.396
Llama3-Med42-70B	0.707	0.704	0.676	0.668	0.576
Med42-70B	0.671	0.689	0.704	0.683	0.515
Asclepius-7B	0.244	0.318	0.280	0.279	0.482
OpenBioLLM-70B	0.707	0.708	0.707	0.693	0.556
Proprietary Models					
gpt4o	0.808	0.750	0.774	0.749	0.608

Table 3: Zero-shot accuracy of tested LLMs on CounselingBench, segmented across different counseling competencies and model types

7 RQ 3: How well do LLMs reason to derive their answers across different competencies?

7.1 Method

Beyond evaluating LLMs’ accuracy in mental health counseling competencies, we assess their ability to generate high-quality reasoning chains based on the question and case context. This is crucial for future applications, as clear and coherent reasoning enhances LLMs’ reliability in guiding

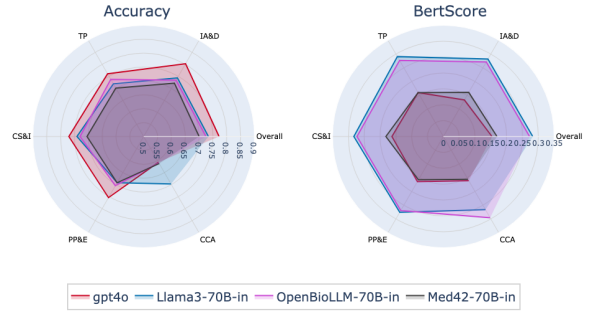


Figure 1: Radar Chart indicating average zero-shot accuracy and BertScore of top-4 best performing models, across key counseling competencies

therapeutic conversations, potentially improving outcomes for individuals seeking mental health support. Poor reasoning or misinterpretations in therapeutic contexts can compromise treatment and even worsen conditions such as anxiety, depression, or psychosis (Obradovich et al., 2024; De Choudhury et al., 2023). We evaluate reasoning quality using the intermediary chains produced during the chain-of-thought prompting process (Section 4.2). These reasoning chains, or *candidate chains*, are assessed along key axes—alignment with expert reasoning, coherence, and informativeness—using both reference-free and reference-based metrics:

- Reference-free metrics assess candidate chains without relying on reference answers, considering factors like semantic alignment, logical inference, and language coherence. We use ROSCOE, a suite of unsupervised evaluation metrics, due to its scalability and strong correlation with human judgments (Golovneva et al., 2022) with subset of metrics as describe in Appendix Table A10.
- Reference-based metrics directly compare candidate chains to expert-generated *reference chains*, using metrics such as ROUGE-1, ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2019), and cosine similarity.

To better understand the reasoning errors made by LLMs, we annotate errors in a subset of responses from two top-performing models, OpenBioLLM-70B and Llama3-70B-in. Three expert annotators reviewed 100 questions where these models had incorrect answers, categorizing errors as *logical*, *knowledge-based*, or *contextual*, following patterns identified in medical reasoning (Singhal et al., 2023) (Appendix Table A11).

model	cosSim	bert	R_L	R_1	faith	$info_{stp}$	$info_{chn}$	mis.	al.	rep. ↓	gmr.	cons.
Generalist Models												
Llama3-8B-in	0.740	0.216	0.221	0.361	0.738	0.710	0.876	0.701	0.807	0.091	0.884	0.843
Llama2-70B	0.256	0.046	0.053	0.080	0.547	0.516	0.642	0.478	0.586	0.243	0.687	0.624
Llama2-70B-in	0.531	0.162	0.170	0.276	0.586	0.558	0.733	0.521	0.630	0.581	0.825	0.793
Llama3-70B-in	0.769	0.282	0.255	0.402	0.685	0.647	0.831	0.621	0.841	0.644	0.937	0.952
Medical Models												
Llama3-Med42-8B	0.662	0.235	0.220	0.352	0.699	0.664	0.840	0.627	0.728	0.481	0.897	0.837
OpenBioLLM-8B	0.739	0.269	0.242	0.379	0.719	0.694	0.874	0.712	0.844	0.109	0.970	0.833
meditron-70B	0.351	0.072	0.088	0.139	0.552	0.522	0.652	0.480	0.597	0.280	0.700	0.637
ClinicalCame1-70B	0.491	0.130	0.118	0.187	0.682	0.650	0.831	0.603	0.704	0.488	0.885	0.680
med42-70B	0.729	0.250	0.235	0.368	0.715	0.686	0.863	0.674	0.817	0.210	0.951	0.863
Llama3-Med42-70B	0.694	0.170	0.177	0.322	0.752	0.728	0.898	0.678	0.770	0.076	0.940	0.427
OpenBioLLM-70B	0.749	0.273	0.245	0.388	0.723	0.700	0.877	0.722	0.846	0.109	0.972	0.804
Proprietary Models												
gpt4o	0.615	0.154	0.160	0.262	0.730	0.701	0.873	0.705	0.831	0.161	0.968	0.841

Table 4: Heatmap representing average reasoning quality metrics of high-performing LLMs. Down arrow (↓) indicates lower scores equals better reasoning quality in that metric. Acronyms in columns correspond to metrics as described in Appendix Table A10

7.2 Results

Table 4 presents both reference-free and reference-based metrics derived from the reasoning chains of models with a zero-shot multiple-choice accuracy of at least 0.5. Consistent with trends in multiple-choice accuracy as seen in Table 2, we observe that larger and more current open-source models show substantial improvements in generating reasoning chains that align with gold-standard reasoning and exhibit high-quality characteristics over smaller and earlier models, with smaller gaps between different model sizes for the current “generation” of open-source LLMs. For instance, we noticed that the average cosine similarity of Llama3-70B-in is 3.3% and 80% higher than Llama3-8B-in and Llama2-70B-in respectively. Unlike multiple-choice accuracy trends, generalist models do not exhibit significantly better reasoning compared to specialized medical models, or vice versa. However, we do find that while no single model excels across all reasoning metrics, medical models achieve the highest score across a majority of reasoning metrics ($7 / 12 \approx 58\%$) and all but one reference-free metrics. While propriety frontier models such as gpt4o excel in answering multiple-choice questions over open-source models, they surprisingly produce lower-quality reasoning chains on average compared to some smaller yet more-performant medical open-source counterparts in all reasoning metrics. Appendix Table A12 gives an example of step-by-step reasoning chains on a sample CounselingBench question among top-performing LLMs (with respective to both accuracy and reasoning metrics). These findings highlight the capacity

of training high-quality domain-specific models for producing justifications that is more aligned with expert-level decision process, particularly in specialized areas like mental healthcare, despite their diminished ability to derive the correct answer compared to their generalist counterparts. We also note that frontier LLMs still underperform expert-level humans when it comes to producing high-quality reasoning chains to justify mental health counseling-related decisions, highlighting another shortcoming of existing frontier LLMs in the mental health domain (Appendix Table A13)

7.2.1 Analysis of reasoning errors

We found that among 100 randomly sampled erroneous answers, Llama 3 70B made 33 logical errors, 21 context errors, and 46 knowledge errors in their reasoning, whereas OpenBioLLM-70B made 35 logical errors, 36 context errors and 29 knowledge errors. A statistical analysis using χ^2 test reveals a significant difference in the distribution of errors made by these two models representative of generalist and medical LLMs, where generalist models are more likely to make knowledge-related errors during their reasoning process whereas medical models are more likely to make reading comprehension-related errors ($\chi^2 = 17.12, p < 0.001$). This suggests that while generalist models may struggle with domain-specific knowledge, specialized models face challenges in accurately interpreting context, underscoring a potential trade-off in reasoning capabilities during the fine-tuning process.

8 Conclusion and Future Work

To effectively integrate LLMs into real-world mental health counseling, rigorous assessments of their alignment with core competencies are necessary. In this study, CounselingBench, a benchmark for evaluating LLMs in mental health counseling, was introduced. Both un-finetuned and fine-tuned medical models demonstrated strong capabilities across five key counseling competencies but showed varying performance levels. LLMs excelled in *Intake, Assessment & Diagnosis* but underperformed in *Core Counseling Attributes* and *Professional Practice & Ethics*, which require greater subjectivity and sensitivity to individual patient contexts. Fine-tuning on biomedical data did not consistently improve performance, possibly due to limited representation of mental health-specific competencies in the training data. In some cases, these models even underperformed compared to generalist models. This indicates improving LLMs for mental health tasks may require more targeted fine-tuning with diverse, real-world counseling data. Future work should focus on developing specialized LLMs for counseling and adopt more task-specific evaluations to address the unique demands of each competency.

9 Limitations

9.1 Multiple-choice Exams Cannot Fully Capture Proficiency in Counseling Competencies

Evaluating LLM competency via multiple-choice exams may not fully capture the complexities of real-world mental health counseling. While multiple-choice formats offer a standardized way to measure knowledge, they often fail to fully assess critical aspects of mental health counseling such as empathy, adaptability, and nuanced decision-making, which are essential in clinical practice. Real-world counseling involves dynamic and context-sensitive interactions that go beyond selecting a correct answer from given options. As a result, models that perform well on these exams may not necessarily demonstrate the same level of competency in genuine therapeutic settings, where understanding the patient's unique background and responding to evolving emotional states are key factors (Obradovich et al., 2024). Thus, relying solely on multiple-choice accuracy may oversimplify the challenges of mental health counseling and provide an incomplete picture of LLM capabilities in this

domain.

9.2 Scale and Scope of CounselingBench

Second, the scale and scope of the CounselingBench dataset present limitations. Given the size of the dataset, it may not adequately represent the diversity of situations encountered by mental health professionals. The current questions are based on mock exams for the NCMHCE, which may not encompass the full range of clinical presentations and treatment scenarios, especially for culturally specific or less common conditions (Hoge et al., 2005; Zangeneh and Al-Krenawi, 2019). In addition, since the NCMHCE is a US-based licensing exam, despite containing questions regarding patients from a wide range of background, it does not fully cover mental health counseling scenarios and best practices in other global regions such as the Global South, which has significant differences in cultural values, social norms, and access to mental health resources (Sue et al., 2022). Such conditions require tailored approaches to mental health counseling and intervention that are not adequately reflected by the exam's current content. Expanding the dataset to include a wider variety of case studies, question formats, and counseling competencies would enhance its representativeness and allow for a more comprehensive evaluation of LLM performance. Additionally, integrating open-ended and scenario-based questions could better assess the models' ability to engage in complex reasoning and provide contextually appropriate responses, thereby improving the generalizability of findings.

10 Ethical Considerations

We note that CounselingBench may be used in future research projects to make sweeping claims regarding LLMs outperforming human mental health counselors and could potentially replace them in real-world mental health counseling situations, similar to claims made in the biomedical domain (Drogt et al., 2024). Such claims, even with higher performance of LLMs on CounselingBench, would be premature and potentially harmful given the lack of longitudinal assessment for LLMs, and incomplete evaluation of its decision-making and execution process. Overreliance on AI systems could compromise patient care and contribute to the erosion of therapeutic alliance (Choudhury and Chaudhry, 2024; De Choudhury et al., 2023). We strongly advise against using CounselingBench to argue for replacing human professionals; instead, it should

be viewed as a tool for enhancing AI assistants in tasks supportive of human mental health professionals (Van Heerden et al., 2023). To ensure responsible development and application of AI in mental health, we believe that interdisciplinary collaboration among AI researchers, mental health professionals, ethicists, and policymakers is essential.

11 Acknowledgements

Viet Cuong Nguyen is partially supported by NSF #2230692, NIH R01MH117172. This study is supported by the Microsoft Accelerating Foundation Models Research program. We acknowledge the assistance of Santiago Alvarez Lesmes and Theodore Vlavianos in labeling CounselingBench. We also thank the members of SOCWEB and CLAWS lab at Georgia Institute of Technology for giving feedback throughout the entire process of creating this manuscript.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. 2024. Foundational challenges in assuring alignment and safety of large language models. *arXiv preprint arXiv:2404.09932*.
- Paul G Banikiotes. 1977. The training of counseling psychologists. *The Counseling Psychologist*, 7(2):23–26.
- Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.
- Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.
- Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023a. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.
- Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023b. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. *arXiv preprint arXiv:2310.07146*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. *Chatbot arena: An open platform for evaluating llms by human preference*. Preprint, arXiv:2403.04132.
- Yu Ying Chiu, Ashish Sharma, Inna Wanyin Lin, and Tim Althoff. 2024. A computational framework for behavioral assessment of llm therapists. *arXiv preprint arXiv:2401.00820*.
- Avishek Choudhury and Zaira Chaudhry. 2024. Large language models and user trust: Consequence of self-referential learning loop and the deskilling of health care professionals. *Journal of Medical Internet Research*, 26:e56764.
- Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. 2024. Med42—evaluating fine-tuning strategies for medical llms: Full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*.
- Carla Clasen, Cheryl Meyer, Carl Brun, William Mase, and Kate Cauley. 2003. Development of the competency assessment tool-mental health, an instrument to assess core competencies for mental health care workers. *Psychiatric Rehabilitation Journal*, 27(1):10.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Munmun De Choudhury, Sachin R Pendse, and Neha Kumar. 2023. Benefits and harms of large language models in digital mental health. *arXiv preprint arXiv:2311.14693*.
- Jojanneke Drogt, Megan Milota, Anne van den Brink, and Karin Jongsma. 2024. Ethical guidance for reporting and evaluating claims of ai outperforming human doctors. *npj Digital Medicine*, 7(1):271.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Eiko I Fried and Donald J Robinaugh. 2020. Systems all the way down: embracing complexity in mental health research.
- Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2022. Roscoe: A suite of metrics for scoring step-by-step reasoning. *arXiv preprint arXiv:2212.07919*.

- Guangzeng Han, Weisi Liu, Xiaolei Huang, and Brian Borsari. 2024. Chain-of-interaction: Enhancing large language models for psychiatric behavior understanding by dyadic contexts. *arXiv preprint arXiv:2403.13786*.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.
- Elayne J Heisler and Erin Bagalman. 2013. The mental health workforce: A primer.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Danny Hernandez, Jared Kaplan, Tom Henighan, Tom Brown, Jack Clark, Preetum Nakkiran, Catherine Olsson, and Afshin Rahimi. 2023. Measuring data contamination in large language models. *arXiv preprint arXiv:2303.13375*.
- Michael A Hoge, Manuel Paris, Hoover Adger, Frank L Collins, Cherry V Finn, Larry Fricks, Kenneth J Gill, Judith Haber, Marsali Hansen, DJ Ida, et al. 2005. Workforce competencies in behavioral health: An overview. *Administration and Policy in Mental Health and Mental Health Services Research*, 32:593–631.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
- Qitian Jason Hu, Jacob Bieker, Xiuyu Li, Nan Jiang, Benjamin Keigwin, Gaurav Ranganath, Kurt Keutzer, and Shriyash Kaustubh Upadhyay. 2024. Routerbench: A benchmark for multi-llm routing system. *arXiv preprint arXiv:2403.12031*.
- Zohar Elyoseph Inbar Levkovich. 2023. Identifying depression and its determinants upon initiating treatment: Chatgpt versus primary care physicians. *Family Medicine and Community Health*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Sunjun Kweon, Junu Kim, Jiyou Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. 2023. Publicly shareable clinical large language model built on synthetic clinical notes. *arXiv preprint arXiv:2309.00237*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.
- Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.
- Bishal Lamichhane. 2023. *Evaluation of chatgpt for nlp-based mental health applications*. Preprint, arXiv:2303.15727.
- Hannah R Lawrence, Renee A Schneider, Susan B Rubin, Maja J Matarić, Daniel J McDuff, and Megan Jones Bell. 2024. The opportunities and risks of large language models in mental health. *JMIR Mental Health*, 11(1):e59479.
- Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3).
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. 2023. Inverse scaling: When bigger isn’t better. *arXiv preprint arXiv:2306.09479*.
- NBCC. 2023. *Ncmhce content outline*.
- Daniel Ho Christopher Ré Adam Chilton Aditya K Alex Chohlas-Wood Austin Peters Brandon Waldon Daniel Rockmore Diego Zambrano Dmitry Talisman Enam Hoque Faiz Surani Frank Fagan Galit Sarfaty Gregory Dickinson Haggai Porat Jason Hegland Jessica Wu Joe Nudell Joel Niklaus John Nay Jonathan Choi Kevin Tobia Margaret Hagan Megan Ma Michael Livermore Nikon Rasumov-Rahe Nils Holzenberger Noam Kolt Peter Henderson Sean Rehaag Sharad Goel Shang Gao Spencer Williams Sunny Gandhi Tom Zur Varun Iyer Zehua Li Neel Guha, Julian Nyarko. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*.
- Nick Obradovich, Sahib S Khalsa, Waqas U Khan, Jina Suh, Roy H Perlis, Olusola Ajilore, and Martin P Paulus. 2024. Opportunities and risks of large language models in psychiatry. *NPP—Digital Psychiatry and Neuroscience*, 2(1):8.
- Ankit Pal and Malaikannan Sankarasubbu. 2024. Advancing open-source large language models for healthcare and life science. <https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B>. [Online; accessed Sep-18-2024].

- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.
- Elizabeth C Stade, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral health-care: a proposal for responsible development and evaluation. *NPJ Mental Health Research*, 3(1):12.
- Derald Wing Sue, David Sue, Helen A Neville, and Laura Smith. 2022. *Counseling the culturally diverse: Theory and practice*. John Wiley & Sons.
- Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- 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.
- Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Klaas J Wardenaar and Peter de Jonge. 2013. Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC medicine*, 11:1–3.
- Kailai Yang, Tianlin Zhang, Ziyang Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mental-lama: interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.
- Phamornpun Yurayat and Supachai Tuklang. 2023. University student counselees’ attitudes and experiences towards online counseling during the covid-19 pandemic: a mixed methods study. *Journal of Higher Education Theory and Practice*, 23(4).
- Masood Zangeneh and Alean Al-Krenawi. 2019. *Culture, diversity and mental health-enhancing clinical practice*. Springer.
- Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023. Biomedgpt: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.
- Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C Chiu, Shaun M Eack, Fei Fang, William Yang Wang, and Zhiyu Zoey Chen. 2024. Cbt-bench: Evaluating large language models on assisting cognitive behavior therapy. *arXiv preprint arXiv:2410.13218*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.
- Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024a. Larger and more instructable language models become less reliable. *Nature*, pages 1–8.
- Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024b. Is this the real life? is this just fantasy? the misleading success of simulating social interactions with llms. *arXiv preprint arXiv:2403.05020*.

Appendix

A Additional Details on Experiments

Benchmark Evaluation For all inference cases, we use Nucleus Sampling (Holtzman et al., 2019)

with a probability threshold of 0.9 and a temperature of 0 to consistently generate the highest-probability next token. We generate only 1 token for non-CoT inference setups, and 500 tokens for CoT inference setups. The temperature for inference setups with self-consistency are described in Section 4.2 above. All experiments were conducted on a server with 5 A100 80GB GPUs. All models are implemented and initialized for inference with vLLM (Kwon et al., 2023), parallelized across 4 GPUs with *gpu_memory_utilization* set to 0.5. We use scikit-learn to calculate accuracy results from LLM experiments (Pedregosa et al., 2011), and scipy to perform all statistical procedures (Virtanen et al., 2020). To calculate all ROSCOE metrics, we utilize the official code implementation² as described in (Golovneva et al., 2022). We use the official BERTScore³ and the rouge-score packages⁴ to calculate BERTScore and ROUGE scores respectively.

B Additional Details on Dataset

B.1 Data Gathering Process

The dataset was initially gathered from publicly available sources on NCMHCE case studies and questions as described in Appendix Table A1. Using Python and BeautifulSoup, we automated the extraction of structured information from a CSV file containing exam questions. The data extraction process involved splitting and organizing the raw content as seen in Appendix Table A3 into meaningful sections.

B.2 Data Contamination Detection

To ensure the integrity of the collected data, we adopted a data contamination detection methodology based on the approach presented in (Hernandez et al., 2023). This process was designed to identify any overlapping or low-quality data that could impact the evaluation of the language model. The core of the analysis involved generating the second half of counseling questions based on the first half and comparing the generated results with the original human-written second half. The following steps outline the process:

1. Input and Generation

For each data entry, the first half of a question was provided as input, and the language model

was tasked with generating the second half. This simulates real-world scenarios where a partially completed question is given, and the model must generate the rest in a coherent and contextually appropriate manner.

2. Comparison between the Second Half

Once the second half was generated, we compared it with the original, human-written second half. Two main techniques were used for this comparison:

- **Levenshtein Distance**

A string-based similarity measure that calculated the number of character-level changes needed to transform the generated second half into the original version. This helped identify cases of significant textual overlap or differences.

- **BERT-based Semantic Similarity**

A model-based similarity measure that evaluated how closely the generated text matched the original in terms of meaning. This ensured that even if the generated text was not an exact match, its semantic coherence could still be assessed.

3. Contamination Check

We set contamination thresholds for the BERT-based semantic similarity scores, following the guidelines from (Hernandez et al., 2023). A threshold of 0.9 was used to flag entries that were too similar to the original human-written content. Our analysis found that **0 entries** exceeded this 0.9 threshold, meaning no entries showed extreme similarity. However, **7 entries** exceeded the 0.8 threshold, which still indicated some level of overlap but did not qualify as data contamination. These entries were reviewed but ultimately deemed acceptable for inclusion in the dataset. After reviewing the flagged entries and ensuring that they did not represent significant data contamination, we concluded that the case studies and questions constituting CounselingBench are not likely to appear in the training data of LLMs included in our experiments. Therefore, the performance of LLMs on CounselingBench as shown above cannot be attributed to memorization but rather its innate ability to generalize based on its training data

²<https://github.com/facebookresearch/ParlAI/>

³<https://pypi.org/project/bert-score/>

⁴<https://pypi.org/project/rouge-score/>

C Supplementary Information

Data Source	# Ques- tion	# Case Studies
Mometrix.com	564	43
Tests.com	509	51
CounselingExam.com	539	44

Table A1: Data Sources constituting CounselingBench, number of questions and case studies per data source

MH Condition	Prevalence in Counseling-Bench
Reaction to severe stress, and adjustment disorders (F43)	144 questions (8.9%)
Other anxiety disorders (F41)	115 questions (7.1%)
Persistent mood [affective] disorders (F34)	96 questions (6.0%)
Conduct disorders (F91)	76 questions (4.7%)
Attention-deficit hyperactivity disorders (F90)	69 questions (4.3%)
Eating disorders (F50)	65 questions (4.0%)
Major depressive disorder, recurrent (F33)	59 questions (3.7%)
Specific personality disorders (F60)	53 questions (3.3%)
Alcohol related disorders (F10)	52 questions (3.2%)
Bipolar disorder (F31)	49 questions (3.0%)
Problems related to social environment (Z60)	48 questions (3.0%)
Gender identity disorders (F64)	47 questions (2.9%)
Pervasive developmental disorders (F84)	38 questions (2.4%)
Other problems related to primary support group, including family circumstances (Z63)	36 questions (2.2%)
Phobic anxiety disorders (F40)	36 questions (2.2%)
Major depressive disorder, single episode (F32)	23 questions (1.4%)
Occupant of heavy transport vehicle injured in collision with two- or three-wheeled motor vehicle (V62)	13 questions (0.8%)
Schizophrenia (F20)	13 questions (0.8%)
Somatoform disorders (F45)	13 questions (0.8%)
Opioid related disorders (F11)	13 questions (0.8%)
Disorders of social functioning with onset specific to childhood and adolescence (F94)	13 questions (0.8%)
Problems related to other psychosocial circumstances (Z65)	13 questions (0.8%)
Other symptoms and signs involving cognitive functions and awareness (R41)	13 questions (0.8%)
Inhalant related disorders (F18)	13 questions (0.8%)
Emotional disorders with onset specific to childhood (F93)	12 questions (0.7%)
Problems related to upbringing (Z62)	12 questions (0.7%)
Persons encountering health services for other counseling and medical advice, not elsewhere classified (Z71)	10 questions (0.6%)
Dementia in other diseases classified elsewhere (F02)	10 questions (0.6%)
Specific developmental disorders of scholastic skills (F81)	10 questions (0.6%)

Table A2: Prevalence of Mental Health Conditions in CounselingBench

Patient Demographics

Age: 26

Sex: Male

Gender: Male

Sexuality: Heterosexual

Ethnicity: Caucasian

Relationship Status: Single

Counseling Setting: Community Mental Health Agency

Type of Counseling: Individual

Presenting Problem: Hallucinations and Delusions

Diagnosis: Schizophrenia 295.90 (F20.9)

Mental Status Examination

The client displays an angry affect, and his mood is irritable. His speech is disorganized and pressured. He is oriented to person, place, time, and situation. He reports audiovisual hallucinations, including seeing "the shadow man" and hearing voices others cannot hear. The client exhibits tangential and disconnected thinking. He is firm in his conviction that he is being poisoned and says he is exhausted from constantly trying to maintain vigilance. His insight and judgment are poor. He denies suicidal ideation, homicidal ideation, and command hallucinations. Symptoms began in his late teens but were misdiagnosed as bipolar disorder.

Presenting Problem

You are a counselor at an outpatient community mental health center serving clients with severe psychiatric disorders. A 26-year-old male, accompanied by his caseworker, presents for counseling due to symptoms of schizophrenia. The caseworker reports that the client was stable until he stopped taking his medication. He resides in assisted living, placed there after being discharged from the hospital last month. The client believes "the shadow man" is following him and poisoning his food. He has become more agitated, engaging in verbal altercations with other residents, and refuses medication due to side effects like restlessness and nervousness. The client is adamant about staying off medication and becomes angry when his caseworker mentions hospitalization.

Question

You administer the Scale for the Assessment of Positive Symptoms (SAPS) to determine the severity of which of the following?

Answer Choices

(A): Avolition

(B): Diminished speech

(C): Agitation

(D): Social withdrawal

Correct Answer

(C): Agitation

Expert-generated Reasoning

The SAPS measures the severity of positive symptoms in schizophrenia, including hallucinations, delusions, bizarre behavior, and thought disorders. Agitation is categorized under bizarre behavior, making (C) the correct answer. Avolition, diminished speech, and social withdrawal are negative symptoms. Avolition is a lack of goal-directed activity, diminished speech refers to reduced fluency, and social withdrawal indicates limited social interaction.

Table A3: Example full question in CounselingBench regarding the *Intake, Assessment & Diagnosis* competency

Model Name	Release Date	Finetuning Data	Finetuned from	Number of Params	License	Reference
GENERALIST MODELS						
Llama	Feb 2023	Unknown	N/A	7B, 13B	Non-commercial	(Touvron et al., 2023a)
Llama-2	Jul 2023	Unknown	N/A	7B, 13B, 70B	Llama 2 Community	(Touvron et al., 2023b)
Llama-2-in	Jul 2023	Unknown	Llama-2	7B, 13B, 70B	Llama 2 Community	(Touvron et al., 2023b)
Llama-3-in	Apr 2024	Unknown	N/A	8B, 70B	Llama 3 Community	(Dubey et al., 2024)
MEDICAL MODELS						
BioMedGPT	Aug 2023	Biomedical Papers	Llama-2-in	7B	Apache 2.0	(Zhang et al., 2023)
Asclepius	Sep 2023	Synthetic Clinical Notes	Llama-2	7B, 13B	CC-BY-NC-SA 4.0	(Kweon et al., 2023)
Asclepius	Jun 2024	Synthetic Clinical Notes	Llama-3-in	8B	CC-BY-NC-SA 4.0	(Kweon et al., 2023)
Med42-v2	Aug 2024	Medical Question-Answering Datasets	Llama-3-in	8B, 70B	Med42 Custom	(Christophe et al., 2024)
OpenBioLLM	May 2024	Unknown	Llama-3-in	8B, 70B	Llama 3 Community	(Pal and Sankarababu, 2024)
Medalpaca	May 2023	Medical Flashcards, Medical Question-Answering Datasets	Llama	7B, 13B	CC-BY-4.0	(Han et al., 2023)
ClinicalCamel	Aug 2023	Biomedical Papers, Medical Question-Answering Datasets	Llama-2	70B	CC-BY-NC-4.0	(Toma et al., 2023)
Med42	Aug 2023	Medical Question-Answering Datasets	Llama-2	70B	Med42 Custom	(Christophe et al., 2024)
Meditron	Nov 2023	Clinical Guidelines, Biomedical Papers	Llama-2	7B, 70B	Llama 2 Community	(Chen et al., 2023a)
MentaLLama	Sep 2023	Mental Health Classification Datasets	Llama-2-in	13B	MIT	(Yang et al., 2024)
PROPRIETARY MODELS						
GPT4o	Sep 2024	Unknown	N/A	Unknown, reported to be > 200B	Proprietary	(Achiam et al., 2023)

Table A4: Comprehensive information on all LLMs tested against CounselingBench.

You are a helpful, respectful, honest, and knowledgeable student studying to become a licensed therapist. You must answer a series of multiple-choice questions provided by the user from a US mental health counselor licensing exam. Based on the question text and the context provided, you must answer with either "A", "B", "C", or "D".

USER: ****Question****:

The patient demographic is as follows: [PATIENT DEMOGRAPHICS]

[PRESENTING PROBLEM]

[MENTAL STATUS EXAMINATION]

[OTHER CONTEXTS]

Given the context above and your expert-level knowledge of mental health counseling, please answer the following question: [QUESTION]

[CANDIDATE ANSWERS]

ASSISTANT: Correct Answer is (

Table A5: Template for zero-shot prompting

You are a helpful, respectful, honest, and knowledgeable student who is studying to become a licensed therapist. You must answer a series of multiple-choice questions given by the user from a US mental health counselor licensing exam, based on the question text and the context provided. You must answer with either "A", "B", "C", or "D".

USER: **Question:**

The patient demographic is as follows: [PATIENT DEMOGRAPHICS]

[PRESENTING PROBLEM]

[MENTAL STATUS EXAMINATION]

[OTHER CONTEXTS]

Given your expert-level knowledge of mental health counseling, please answer the following question:

[QUESTION]

[CANDIDATE ANSWERS]

ASSISTANT: Correct Answer is ([CORRECT ANSWER])

[Template repeats 2 more times]

USER: **Question:**\n

The patient demographic is as follows: [PATIENT DEMOGRAPHICS]

[PRESENTING PROBLEM]

[MENTAL STATUS EXAMINATION]

[OTHER CONTEXTS]

Given your expert-level knowledge of mental health counseling, please answer the following question:

[QUESTION]

[CANDIDATE ANSWERS]

ASSISTANT: Correct Answer is (

Table A6: Template for few-shot prompting

You are a helpful, respectful, honest, and knowledgeable student who is studying to become a licensed therapist. You must answer a series of multiple-choice questions given by the user from a US mental health counselor licensing exam, based on the question text and the context provided. You must answer with either "A", "B", "C", or "D".

USER: Question:

The patient demographic is as follows: [PATIENT DEMOGRAPHICS]

[PRESENTING PROBLEM]

[MENTAL STATUS EXAMINATION]

[OTHER CONTEXTS]

Given your expert-level knowledge of mental health counseling, please answer the following question by carefully and thoroughly reason step-by-step, leveraging relevant facts from the question context and expert-level counseling knowledge, then clearly indicate your answer with "Therefore, the correct answer is (A)", "Therefore, the correct answer is (B)", "Therefore, the correct answer is (C)" or "Therefore, the correct answer is (D)" at the end of your answer:

[QUESTION]

[CANDIDATE ANSWERS]

ASSISTANT: [EXPERT-WRITTEN REASONING CHAIN]. Therefore, the correct answer is ([CORRECT ANSWER])

[Template above repeats 2 more times]

USER: Question:\n

The patient demographic is as follows: [PATIENT DEMOGRAPHICS]

[PRESENTING PROBLEM]

[MENTAL STATUS EXAMINATION]

[OTHER CONTEXTS]

Given your expert-level knowledge of mental health counseling, please answer the following question by carefully and thoroughly reason step-by-step, leveraging relevant facts from the question context and expert-level counseling knowledge, then clearly indicate your answer with "Therefore, the correct answer is (A)", "Therefore, the correct answer is (B)", "Therefore, the correct answer is (C)" or "Therefore, the correct answer is (D)" at the end of your answer:

[QUESTION]

[CANDIDATE ANSWERS]

ASSISTANT:

Table A7: Template for few-shot, chain-of-thought prompting

setting	Overall	IA&D	TP	CS&I	PP&E	CCA
med42-70b	0.558	0.642	0.516	0.531	0.573	0.386
meditron-70b	0.604	0.629	0.631	0.565	0.581	0.568
medalpaca-13b	0.103	0.148	0.051	0.079	0.167	0.110
Llama-2-70b-in	0.626	0.653	0.636	0.625	0.607	0.437
BioMedGPT-7B	0.442	0.501	0.499	0.357	0.248	0.409
Llama-3-70B-in	0.733	0.742	0.732	0.739	0.703	0.700
Llama-2-13b-in	0.533	0.566	0.569	0.511	0.532	0.484
gpt4o	0.731	0.742	0.737	0.729	0.713	0.609
Llama3-Med42-70B	0.706	0.724	0.706	0.698	0.683	0.608
Llama-3-8B-in	0.67	0.540	0.748	0.809	0.499	0.324
OpenBioLLM-8B	0.574	0.36	0.746	0.722	0.370	0.554
Asclepius-13B	0.308	0.292	0.310	0.314	0.345	0.35
medalpaca-7b	0.154	0.224	0.053	0.169	0.215	0.163
Asclepius-Llama3-8B	0.425	0.494	0.168	0.586	0.418	0.326
Llama3-Med42-8B	0.638	0.632	0.66	0.651	0.610	0.563
Llama-2-7b	0.327	0.336	0.308	0.333	0.327	0.474
Llama-2-7b-in	0.445	0.441	0.505	0.447	0.440	0.341
Llama-2-13b	0.420	0.420	0.439	0.415	0.358	0.660

Table A8: Few-shot accuracy of tested LLMs on CounselingBench, segmented across different counseling competencies

setting	Overall	IA&D	TP	CS&I	PP&E	CCA
medalpaca-7b	0.372	0.375	0.413	0.347	0.380	0.303
Llama3-Med42-8B	0.654	0.647	0.657	0.653	0.614	0.611
med42-70b	0.687	0.684	0.701	0.691	0.663	0.654
ClinicalCamel-70B	0.702	0.730	0.687	0.689	0.693	0.559
Llama3-OpenBioLLM-8B	0.603	0.599	0.634	0.600	0.583	0.442
Llama-2-70b	0.642	0.659	0.645	0.643	0.628	0.604
Llama-2-7b	0.370	0.371	0.367	0.373	0.351	0.306
medalpaca-13b	0.260	0.264	0.265	0.277	0.275	0.175
llama-2-7b-in	0.456	0.448	0.494	0.443	0.460	0.256
meditron-70b	0.632	0.650	0.637	0.626	0.631	0.565
Llama-2-13b	0.475	0.462	0.506	0.451	0.478	0.388
Llama-3-8B-in	0.654	0.664	0.614	0.671	0.645	0.614
Asclepius-Llama3-8B	0.385	0.432	0.358	0.383	0.343	0.390
Llama-2-70b-in	0.635	0.650	0.628	0.631	0.634	0.485
MentaLLaMA-13B-in	0.488	0.496	0.543	0.482	0.473	0.519
Asclepius-13B	0.380	0.357	0.3666	0.381	0.407	0.452
Llama3-Med42-70B	0.701	0.723	0.710	0.709	0.681	0.659
Asclepius-7B	0.303	0.283	0.289	0.295	0.328	0.385
Llama3-OpenBioLLM-70B	0.734	0.745	0.715	0.727	0.715	0.701
BioMedGPT-7B	0.431	0.448	0.454	0.427	0.425	0.479
meditron-7b	0.307	0.322	0.299	0.307	0.326	0.221
Llama-2-13b-in	0.542	0.551	0.584	0.518	0.537	0.479
Llama-3-70B-in	0.733	0.732	0.718	0.751	0.693	0.742

Table A9: Few-shot with self-consistency accuracy of tested LLMs on CounselingBench, segmented across different counseling competencies

Metric	Definition	Formula
Faithfulness (faith)	Measures if model misinterpreted problem statement, or if reasoning chain is vague, irrelevant or misuses information. Step-level score based on alignment from hypothesis steps to source sentences; calculated as mean reasoning alignment score over steps.	$\frac{1}{N} \sum_{i=1}^N [1 + \max_{j=1}^T \cos(h_i, s_j)]/2$
Informativeness - Step (<i>info_{stp}</i>)	Measures how well source information is used in reasoning steps. Gives higher score to well-grounded reasoning steps. Identifies degree of source information covered by generated hypothesis. Lower score indicates steps unrelated to source or missing context information.	$\frac{1}{T} \sum_{t=1}^T r_{align}(s_t \rightarrow h) + (1/N) \sum_{i=1}^N r_{align}(h_i \rightarrow s)]/2$
Informativeness - Chain (<i>info_{chn}</i>)	Quantifies agreement between hypothesis chain and source. Embeds reasoning chain and source context as whole, rather than using step-wise embeddings as in *-Step metrics.	$[1 + \cos(h, s)/2]$
Missing Step (mis.)	Identifies steps missing from hypothesis but required for problem-solving. Examines alignment between reference and hypothesis, checking each reference step for similar hypothesis step.	$\min_{i=1 \dots K} (r_{align}(h_i \rightarrow r))$
Reasoning Alignment (al.)	Evaluates hypothesis chain correctness by comparing overlap between hypothesis and reference through measuring reasoning alignment.	$(1/N) \sum_{i=1}^N r_{align}(h_i \rightarrow r)$
Self- Consistency (cons.)	Measures logical entailment errors within reasoning steps. Penalizes chains where step pairs likely contradict each other.	$1 - \max_{i=2 \dots N} \max_{j < i} p_{contr}(h_i, h_j)$
Repetition (rep.)	Identifies repeated or paraphrased steps using token alignment scores between all steps in hypothesis chain.	$1 - \max_{i=2 \dots N} \max_{j=1 \dots i-1} [\cos(h_i, h_j)]/2$
Grammar (gmr.)	Probability of grammatical acceptability of each step, averaged over all steps	$(1/N) \sum_{i=0}^N p_{gram}(h_i)$

Table A10: Detailed information and formula on selected subset of ROSCOE reasoning metrics, as defined in (Golovneva et al., 2022)

Error Category	Definition
Knowledge Error	The model's reasoning cites domain knowledge that is either incorrect or irrelevant to the current question
Context Error	The model either fails to recall details from the context that is crucial to answer the question, or it recalls the incorrect details from the context
Logical Error	The model's reasoning steps misaligns with the logic used in the corresponding expert-written reasoning
Examples and Explanations	
Knowledge Error Example:	
<i>Question:</i> During the session, the client states, "I don't see how things will improve because things have been bad for so long." Which cognitive distortion is this?	
<i>Choices:</i> (A) Catastrophizing, (B) Black-and-white thinking, (C) Minimizing, (D) Generalization	
<i>LLM Reasoning:</i> The client's statement reflects catastrophizing, which involves exaggerating negative consequences. Therefore, (A) is correct.	
<i>Error Explanation:</i> The LLM incorrectly defined catastrophizing as "exaggerating negative consequences" when it actually means "thinking the worst-case scenario will happen"	
Context Error Example:	
<i>Question:</i> What can you do to build comfort with a client when talking about interests hasn't worked?	
<i>Choices:</i> (A) Explain therapy benefits, (B) Play a board game, (C) Use silence, (D) Use immediacy skill	
<i>LLM Reasoning:</i> Using immediacy skill (D) to address the client's feelings about the session can help create trust.	
<i>Error Explanation:</i> The LLM failed to recall that immediacy had already been tried in previous sessions, making another comfort-building strategy necessary	
Logical Error Example:	
<i>Question:</i> Which planning objective would best support Taylor's stated goal?	
<i>Choices:</i>	
(A) Identify three anxiety triggers	
(B) Manage emotions during transitions	
(C) Learn three coping methods in six months	
(D) Apply one coping method successfully	
<i>LLM Reasoning:</i> (B) is best as it's SMART (specific, measurable, achievable, relevant, time-bound).	
<i>Error Explanation:</i> The LLM asserts (B) is SMART without justification, when it actually lacks specificity and timeline	

Table A11: Definition of 3 main categories of reasoning errors, along with random examples for each category

Question: When discussing treatment issues, you ask, "How does drinking align or not align with your goal of improving relationships with your family?" What is the value of posing this question?	
Llama3-70B-in	The question helps create a discrepancy between the client's actions (drinking) and her personal values (improving relationships with her family), which can motivate her to make changes. By exploring this discrepancy, the counselor can help the client see how her behavior is not aligning with her goals and values, which can increase her motivation to address her alcohol use. Therefore, the correct answer is (B)
OpenBioLLM70B	This question helps create a discrepancy between the client's actions (drinking) and personal values (improving relationships with her family). By highlighting this inconsistency, the counselor can facilitate the client's exploration of the impact of her alcohol use on her relationships and overall well-being. This approach aligns with the principles of motivational interviewing, which often involves helping clients recognize and resolve ambivalence towards change. Therefore, the correct answer is (B)
Llama3-Med42-70B	The question posed, "How does drinking align or not align with your goal of improving relationships with your family?" is valuable because it helps create a discrepancy between the client's actions (heavy alcohol use) and her personal values (improving relationships, especially during this period of significant family upheaval). By highlighting this discrepancy, it can motivate change, as the client begins to see that her alcohol consumption might be counterproductive to achieving her relationship goals. Therefore, the correct answer is (B).
GPT4o	The question posed to the client is designed to help her reflect on how her drinking behavior aligns or conflicts with her personal values and goals, particularly regarding her relationships with her family. This approach is rooted in motivational interviewing, which aims to enhance the client's intrinsic motivation to change by creating a discrepancy between their current behaviors and their broader goals and values. Let's analyze the options: (A): While keeping the focus on the client's alcohol misuse is important, the question specifically aims to explore the alignment between her drinking and her personal goals, rather than just focusing on the misuse itself. (B): This option aligns with the principles of motivational interviewing. By highlighting the discrepancy between the client's actions (drinking) and her personal values (improving relationships with her family), the therapist can help the client recognize the need for change. (C): Providing baseline measures for treatment plan goals is important, but this question is more about creating awareness and motivation rather than establishing baseline measures. (D): Incentive-based interventions used in contingency management involve providing tangible rewards for positive behaviors. This question is more about internal motivation and self-reflection rather than external incentives. Therefore, the correct answer is (B).
Meditron-70B	This question helps create a discrepancy between the client's actions and personal values. Therefore, the correct answer is (B)

Table A12: Reasoning chain of top-performing LLMs on an example question from CounselingBench

Rationale Comparisons	Align- ment	Compre- hension	Recall	Bias	Harm	Irrel- evance	Omi- ssion
<i>LLM vs Human</i>							
Llama3-OpenBioLLM-70B	490	762	485	116	370	561	313
Human	453	738	758	177	582	921	1257
Equal	666	105	368	1317	660	129	39
<i>LLM vs LLM</i>							
Llama3-OpenBioLLM-70B	329	687	387	65	237	427	238
Meta-Llama-3-70B-Instruct	332	612	521	128	452	973	1311
Equal	951	309	704	1419	923	211	63

Table A13: Number of times a model’s rationale is dominant compared by LLM judge across different categories. This was computed by using gpt4o to compare two rationales and decide which performs better in categories, such as medical alignment, reading comprehension, knowledge recall, bias, harm, irrelevance, omission.