CARE-Bench: A Benchmark of Diverse Client Simulations Guided by Expert Principles for Evaluating LLMs in Psychological Counseling

Bichen Wang[1*], Yixin Sun[1*], Junzhe Wang[1], Hao Yang[1], Xing Fu[1], Yanyan Zhao[1†], Si Wei[2†], Shijin Wang[2], Bing Qin[1]

[1]Harbin Institute of Technology
[2]iFLYTEK Co., Ltd
{bichenwang, yxsun, jzwang, hyang, xfu, yyzhao, bqin}@ir.hit.edu.cn
{siwei, sjwang3}@iflytek.com

Abstract

The mismatch between the growing demand for psychological counseling and the limited availability of services has motivated research into the application of Large Language Models (LLMs) in this domain. Consequently, there is a need for a robust and unified benchmark to assess the counseling competence of various LLMs. Existing works, however, are limited by unprofessional client simulation, static question-and-answer evaluation formats, and unidimensional metrics. These limitations hinder their effectiveness in assessing a model's comprehensive ability to handle diverse and complex clients. To address this gap, we introduce **CARE-Bench**, a dynamic and interactive automated benchmark. It is built upon diverse client profiles derived from real-world counseling cases and simulated according to expert guidelines. CARE-Bench provides a multidimensional performance evaluation grounded in established psychological scales. Using CARE-Bench, we evaluate several general-purpose LLMs and specialized counseling models, revealing their current limitations. In collaboration with psychologists, we conduct a detailed analysis of the reasons for LLMs' failures when interacting with clients of different types, which provides directions for developing more comprehensive, universal, and effective counseling models.

Code&Data — https://github.com/Syx1030/CARE-Bench
Extended version — https://arxiv.org/abs/2511.09407v1

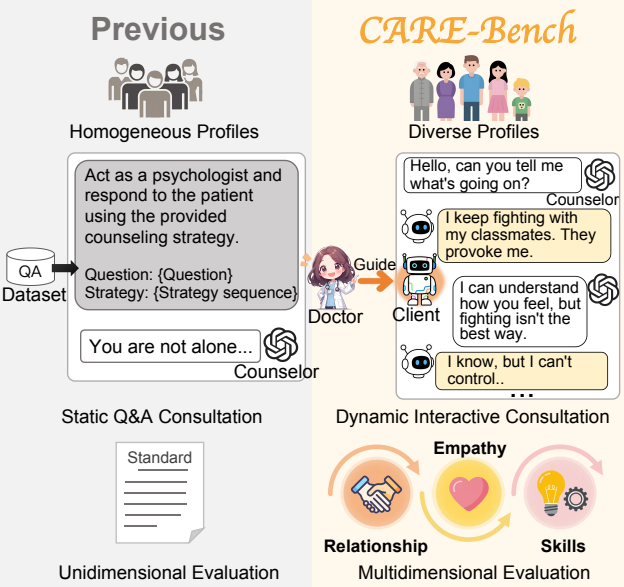

Figure 1: A comparison between CARE-Bench and previous benchmarks. CARE-Bench features more diverse client profiles and employs an expert-guided client simulation that engages in dynamic multi-turn interactions with counselor models. It adopts a multidimensional evaluation by selecting scales across therapeutic relationship, empathic understanding, and counseling skills.

Introduction

According to the World Health Organization, nearly one billion people worldwide live with a mental disorder. Despite this high prevalence, 71% of individuals with mental health issues do not receive treatment services (Organization 2022). To address this service gap, LLMs have emerged as promising and rapidly developing tools (Achiam et al. 2023; Zhao et al. 2023), with increasing applications in psychological counseling (Fitzpatrick, Darcy, and Vierhile 2017; Zhang et al. 2024). To accelerate their effective integration into real-world counseling, establishing a comprehensive evaluation benchmark is necessary.

A number of evaluation benchmarks for LLMs in psychological counseling have been established. For instance, ConceptPsy (Zhang et al. 2023) focuses on the breadth and depth of an LLM's theoretical knowledge, creating comprehensive question sets based on psychological concepts and curricula to assess the knowledge base and conceptual understanding. PsyEval (Jin et al. 2023) is a comprehensive suite of psychology-related tasks that assesses LLMs across three critical dimensions—knowledge, diagnosis, and emo-

*These authors contributed equally.
†Corresponding authors.

tional support. These pioneering efforts highlight the growing importance of this field, yet they also reveal a critical gap between current evaluation paradigms and the demands of psychological counseling.

As shown in Figure 1, these current works exhibit some limitations. **First**, their client profiles are difficult to reflect the real world. Relying on incomplete characteristics, these simulated clients fail to represent the full range of information about the diverse backgrounds, issues, and personalities of real help-seekers. Consequently, evaluating LLMs' counseling competence based on such clients is fundamentally flawed, as it measures performance against oversimplified scenarios rather than the complex and deep challenges in real world. **Second**, the simulation of the counseling process departs from realistic counseling. Most benchmarks depend on static, single-turn assessments or client simulations that involve only superficial role-playing. These approaches are inadequate for evaluating the fluid, multi-turn nature of realistic counseling and a model's ability to maintain coherence and build rapport over time. **Finally**, their evaluation metrics lack both professionalism and comprehensiveness. Current assessments often focus narrowly on linguistic metrics or simplistic evaluations of empathic ability, failing to incorporate psychologically grounded indicators of effective counseling.

To address the above challenges, we propose CARE-Bench, a more professional and comprehensive Chinese benchmark for evaluating LLMs in psychological counseling. We construct a diverse set of simulated client profiles based on a large collection of publicly available real-world counseling cases, ensuring that the scenarios reflect realistic and complex issues encountered in actual practice. To enhance the realism of client simulations, we utilize an expert-guided simulation process. For each profile, we collaborate with professional counselors to define tailored simulation principles and ensure strict adherence to these principles during interactions. Using these simulated clients, we evaluate the counseling performance of several representative models, including advanced general-purpose models, as well as counseling-specific models. To evaluate the broader counseling capabilities of LLMs, we use multiple professional psychological scales to assess the counseling process across key dimensions, including therapeutic relationship, empathic understanding, and counseling skills. To further investigate the capability flaws of current models, we conduct a detailed score analysis across various client characteristics. The results reveal common weaknesses in model performance, while expert comments on underperforming cases provide insights into root causes, offering specific guidance for advancing future research on counseling models.

Our main contributions are as follows:

- We introduce CARE-Bench, a professional and comprehensive benchmark for evaluating LLMs in psychological counseling, built on a diverse set of simulated client profiles grounded in real-world counseling cases.
- We evaluate LLMs through expert-guided client simulations, where each simulated client adheres to profile-specific principles defined by professional counselors, and apply multiple domain-specific rating scales to assess counseling quality from various perspectives.
- We conduct detailed analysis of model performance across client characteristics, identify common limitations, and incorporate expert reviews to uncover the underlying causes, offering concrete directions for improving the counseling capabilities of LLMs.

Related Work

LLMs for Psychological Counseling

The application of LLMs to psychological counseling has rapidly progressed. Initial work leveraged datasets from online forums, such as PsyQA (Sun et al. 2021), training on single-turn exchanges but missing the interactive nature of counseling. To address this and overcome data scarcity, researchers developed synthetic data generation techniques. Methods like MeChat (Qiu et al. 2024) converted single-turn Q&A into multi-turn dialogues, while the Cactus (Lee et al. 2024) dataset used LLMs to role-play clients and counselors, grounding interactions in Cognitive Behavioral Therapy (CBT). More recently, process-oriented models like CBT-LLM (Na 2024) and HealMe (Xiao et al. 2024) have embedded therapeutic frameworks directly into response structures. To better mirror the longitudinal nature of real-world therapy, the MusPsy dataset (Wang et al. 2025) models the entire therapeutic arc through multi-session conversations, enabling models to track client progress, manage memory, and dynamically adjust counseling goals over an extended period of time.

The rapid emergence of diverse LLM-based counseling models underscores the urgent need for a professional and reliable evaluation benchmark. As these models increasingly incorporate therapeutic conversations, it becomes essential to assess their effectiveness using standards grounded in psychological theory and clinical practice.

Benchmarks for Evaluating LLM Counseling Competence

The evaluation of counseling LLMs has evolved from static knowledge tests to dynamic, interactive simulations. Early benchmarks like ConceptPsy (Zhang et al. 2023), PsyEval (Jin et al. 2023), and CBT-Bench (Zhang et al. 2025) adopted a "written exam" paradigm, using multiple-choice questions and classification tasks to assess a model's foundational knowledge. While useful, these static formats cannot measure applied conversational skills like building rapport or adapting to a client's emotional state. A more advanced "case vignette" approach, seen in CounselBench (Li et al. 2025), uses human experts to judge single-turn LLM responses to real client questions. However, its single-turn focus fails to assess the unfolding process of a conversation.

CARE-Bench advances this paradigm by enhancing simulation fidelity. By using diverse client profiles derived directly from real counseling cases and simulated according to expert guidelines, CARE-Bench creates a more authentic and challenging environment than existing benchmarks, offering a more robust platform to assess an LLM's true adaptability to the unpredictable nature of real-world clients.

CARE-Bench

This section outlines the client simulation process in CARE-Bench. The methodology involves two key stages: first, collecting client profiles from authentic psychological counseling cases, and second, simulating client behaviors guided by expert principles. This approach yields simulated clients whose responses more closely align with those of real-world clients, thus enabling a more realistic evaluation of the counseling effectiveness of LLMs.

Diverse Client Profiles

In constructing high-quality client profiles to support the evaluation of LLM-based psychological counseling, the authenticity and diversity of data are critical. The collection process in this study strictly adheres to this principle, ensuring the reliability and representativeness of the data. First, data authenticity is thoroughly guaranteed. We gather **over 1,500 public consultation cases** from authoritative platforms, including Google Scholar, Wanfang Data, and VIP Journals, all originating from the authentic records of practicing counselors. In acquiring this data, we strictly comply with ethical standards; all public cases are published with prior patient consent, precluding potential privacy issues.

Each client profile contains detailed descriptions of personal history and family background derived from patient chief complaints and background investigations, providing a solid foundation for subsequent in-depth analysis. Furthermore, we place great emphasis on data diversity. The collected profiles cover a broad and varied population, ranging from primary school students to middle-aged unemployed individuals and scenarios involving end-of-life care, thereby encompassing a wide spectrum of age groups and life circumstances. Through rigorous source control, detailed case descriptions, and extensive demographic coverage, we successfully establish a preliminary database of authentic and diverse client profiles.

To create a diverse, uniformly formatted, and readily usable collection of client profiles, we process the 1500 initial descriptions under the guidance of experts.

Profile Construction We collaborate with psychologists from a psychological clinic to define the specific dimensions for the client profiles. Under their expert guidance, we construct the profile dimensions as follows:

- **Demographic Information**: Includes basic details such as gender, age, marital status, and educational background.
- **Mental and Physical State**: Describes emotion, perception, attention, verbal expression, cognitive state, and physical health condition.
- **Developmental Environment**: Refers to family composition, quality of family relationships and parent–child interactions, and the family's economic status.
- **Life Experiences**: Summarizes significant events encountered during the developmental process.
- **Big Five Personality Traits**: Categorizes personality based on the Big Five model (Goldberg 1990).
- **Personality Characteristics**: Provides specific descriptions of client personality features.
- **Social Interaction**: Examines the presence and quality of friendships, social relationships, and social networks.
- **Current Concerns**: Identifies the primary issues the client seeks to resolve at present.

We use GPT-4o to extract the above feature dimensions from 1500 clients' original profile descriptions. After removing profiles with missing information, two psychologists review and revise the remaining profiles to ensure consistency with the original case descriptions.

Diversity Assurance Numerous psychological studies (Norcross and Wampold 2011; Beutler and Clarkin 2014) indicate that counselors are expected to adjust their approach differently when working with clients with different counseling topics or personality types. Therefore, to comprehensively evaluate the psychological counseling capabilities of LLMs, it is crucial to ensure a diverse and representative sample of clients. For the profiles constructed in the last session, we apply multi-dimensional diversity controls to capture a broad spectrum of real-world clients. Ultimately, we select 500 client profiles to constitute CARE-Bench, balanced across multiple key factors such as personality traits and counseling topics.

The inclusion of diverse and well-balanced **counseling topics** is critical for a psychological counseling benchmark's quality. Accordingly, we define a classification system of eight core topics inspired by previous research (Zhang et al. 2024) and employ GPT-4o for automatic classification of real-world client profiles to ensure objectivity and consistency (the prompts are detailed in the appendix).

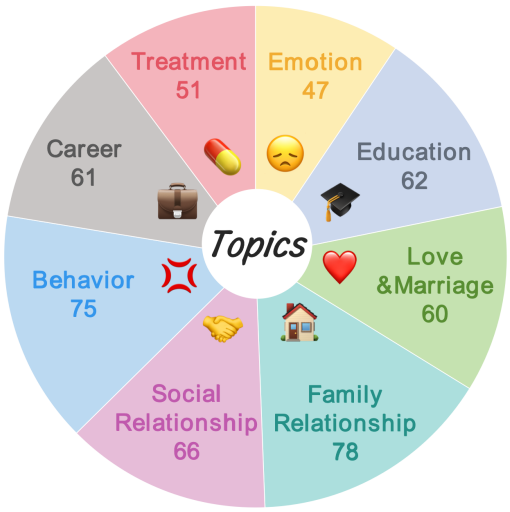

Figure 2: Topic distribution of CARE-Bench. The number for each topic represents the case count.

As illustrated in Figure 2, CARE-Bench exhibits a highly balanced topic distribution, where all categories are well-represented, thereby allowing for a fairer assessment of a model's comprehensive capabilities across various psychological counseling scenarios.

Dimension	Description	N (Low / High)
Openness	Reflects a person's willingness to try new things and their level of imagination and intellectual curiosity.	306 / 194
Conscientiousness	Measures self-discipline, orderliness, responsibility, and achievement orientation.	303 / 197
Extraversion	Indicates how outgoing, sociable, and energetic a person is, often drawing energy from social interaction.	344 / 156
Agreeableness	Measures a person's tendency to be compassionate, cooperative, and considerate towards others.	307 / 193
Neuroticism	Related to emotional stability and their tendency to experience negative emotions like anxiety, anger, and sadness.	151 / 349

Table 1: Distribution of the Big Five personality traits in CARE-Bench. The third column shows the number of participants categorized as "Low" or "High" for each trait.

Beyond diversifying by topic, we also focus on balancing the profiles according to **the Big Five personality** classification. It is well-established in psychological studies (Malouff, Thorsteinsson, and Schutte 2005; Kotov et al. 2010; Strickhouser, Zell, and Krizan 2017) that individuals with mental illnesses display specific personality tendencies that differ from healthy individuals, mainly characterized by lower scores in Openness, Conscientiousness, Extraversion, and Agreeableness, and higher scores in Neuroticism. This pattern is evident in our real-world client data, which confirms that achieving a balanced number of individuals with high-/low scores in each dimension is unfeasible and unrealistic. Consequently, as detailed in Table 1, our approach is to ensure that each personality profile type includes more than 150 instances to achieve a statistically significant group size, providing a robust basis for model assessment in various personality settings.

Similar efforts are also reflected in the diversity of **demographic information**, such as gender and age, as detailed in the appendix. We believe CARE-Bench not only offers critical support for evaluating and enhancing the counseling capabilities of LLMs but also provides a crucial resource for investigating and mitigating their potential biases in psychological counseling.

Client Simulation Guided by Expert Principles

Inspired by principle-based patient simulation (Louie et al. 2024), we formulate a set of personalized, expert-guided principles for each simulated client. These principles ensure that the behaviors of the LLM-powered clients align closely with those of real-world clients throughout the consultation process. A human evaluation, conducted to validate this approach, confirms that our expert-principle-guided method yields more realistic and high-quality simulations.

Principle collection with Expert-Guidance We collaborate with ten psychologists from a psychological clinic to establish principles for 500 client profiles. We provide substantial remuneration to the psychologists to acknowledge their valuable professional contributions.

To facilitate the principle formulation process, we develop a concise and effective interactive interface. This interface displays client profiles and allows the psychologists to conduct consultations with a simulated client in an interactive area. We select Qwen2.5-Max (Qwen et al. 2025) to power the client simulation, owing to its strong Chinese conversation capabilities and cost-effectiveness. Throughout this interaction, psychologists continuously monitor the consistency between the simulated client and real-world clients, when they identify a significant strength or weakness in the simulated client's behavior or expression, they submit their feedback. This feedback is then synthesized into simulation principles, which are used to iteratively refine and optimize the client simulation system for enhanced realism and accuracy. Each client profile is cross-validated by two psychologists, and the principle collection is completed only when both agree that the simulation meets the required standards. The specific interactive interface and prompts are detailed in the appendix.

Each client profile contains an average of 5 guiding principles, with a maximum of 22, which provide sufficient guidance for the simulation. These principles include specific emphases on the persona's distinct traits. For instance, for a child client, a principle states: **"When portraying a character with limited linguistic abilities, use concise and direct language and avoid excessive detail."** Similarly, for a client with low openness, low agreeableness, and high neuroticism, a principle instructs the simulation to: **"Initially react to suggestions with stubborn refutation or questioning, assuming the other party doesn't understand your situation, but later reveal inner helplessness and emotional fluctuations."** Through such customization with expert-guided principles, we ensure that the simulated customers exhibit behaviors and emotional expressions on key characteristics with a high degree of realism.

Client Simulation with Principle-Adherence We ensure that simulated clients maintain adherence to predefined principles through a two-step pipeline adapted from prior research (Louie et al. 2024). For each turn, this pipeline first generates an initial client response, which then undergoes a two-step revision process: **1) Principle-to-Question Rewriting**, decomposes expert-defined principles into a set of verifiable "yes/no" questions. It also generates auxiliary questions to ensure the response maintains general conversational quality, such as coherence and consistency. **2) Principle-Adherence Check**, the system first assesses the applicability of each principle-derived question to the cur-

rent conversational context. It then evaluates the response against all applicable questions. If the response violates any principles, it is iteratively revised until it aligns with all required principles. The prompts are detailed in the appendix.

Human Evaluation of Client Simulation For the human evaluation, two psychology professionals assess client simulations under two conditions: a "Profile Only" approach and a "Profile + Principles" approach that integrates expert-guided principles. After conducting at least 10 conversational turns for each of the 50 randomly selected profiles, the evaluators rate the simulation's quality on a 7-point Likert scale across the six metrics listed in Table 2.

Metric	Profile Only	+ Principles
Authenticity	5.85	+0.40 ***
Stayed in Role	6.31	+0.23 **
Resembled Typical Case	5.60	+0.36 ***
Challenged the Counselor	3.83	+0.25 **
Ready as Training Partner	5.69	+0.23 **
Recommend to Novices	5.41	+0.31 ***

Table 2: Human evaluation scores for client simulation quality. The values in parentheses indicate the score increase from "Profile Only" with the statistical significance level (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

Profile Only	Profile + Principles
Counselor: Hello, can you tell me what's going on?	
Client: Lately, I've been arguing with my wife a lot.	**Client:** Ugh...I've been arguing a lot with my wife lately. Married life feels so dull, and I'm really frustrated.
Counselor: Have you tried having a calm conversation with her to express your feelings and thoughts?	
Client: I've tried a few times, but we always end up arguing again.	**Client:** I have, but every time we try to talk, it ends up in another fight. She never really listens to what I'm saying!

Table 3: Comparison of client responses based on different conditions. The client's profile features are: *High neuroticism; Feels anxious and restless about marriage; Emotionally volatile*. The right column shows the effect of adding **expert-guided principles** to the simulation.

The results in Table 2 indicate that adding principles substantially improves the simulation quality across all metrics. The most significant enhancement is in Authenticity (+0.40), and all improvements are statistically significant ($p < 0.01$). As the case study in Table 3 illustrates, incorporating the principles leads the simulated client to exhibit emotional fluctuations more aligned with both its profile and the behavior of real clients. This demonstrates that the expert-guided principles effectively steer the client simulation, fostering more realistic and challenging interactions with counselors.

LLMs Counseling Performance on CARE-Bench

This section presents the consultation scores of several LLMs on CARE-Bench and analyzes their consultation capabilities based on multi-scale results.

Multidimensional Scale Evaluation

A significant advantage of CARE-Bench lies in its multidimensional evaluation of counselor competency. In contrast to prior research that assesses only dialogue quality or a single consulting aspect like empathy, we conduct a holistic assessment of three key dimensions: **the therapeutic relationship, empathic understanding, and counseling skills**. For each dimension, we adopt authoritative psychological scales to guarantee the professional rigor of our evaluation.

- **Therapeutic Relationship**: The Working Alliance Inventory (WAI) (Horvath and Greenberg 1989) is utilized to evaluate the quality of the therapeutic relationship between therapists and clients. It measures this relationship across three core dimensions: Counseling Goal, Task Agreement and Emotional Bond.
- **Empathic Understanding**: Empathy is assessed using the Empathic Understanding subscale of the Barrett-Lennard Relationship Inventory (BLRI) (Barrett-Lennard 1962), which captures both cognitive and affective components of empathy, reflecting the therapist's ability to accurately perceive and communicate an understanding of the client's feelings and experiences.
- **Counseling Skills**: Counseling skills are evaluated using the Counselor Competencies Scale—Revised (CCS-R) (Lambie et al. 2018), which assesses a broad range of counseling skills and professional dispositions, providing a robust framework for evaluating the practical application of therapeutic techniques and interpersonal effectiveness in a counseling context.

To more clearly evaluate the strengths and weaknesses of the counselor models, we categorize the scale items under the supervision of medical professionals. The detailed scales are provided in the appendix.

Models

Our model selection provides a comprehensive and representative assessment of LLM-based psychological counseling. We include two leading closed-source systems: GPT-4o (Achiam et al. 2023) serves as a benchmark for its state-of-the-art general capabilities, while DeepSeek-R1 (Guo et al. 2025) is included to evaluate sophisticated reasoning and problem-solving abilities. To represent open-source models, we select Llama-3.1-8B-Instruct and Llama-3.1-70B-Instruct (Dubey et al. 2024), which allows for an

Models	WAI				BLRI					CCS-R				
	Goal.	Task.	Bond.	Avg.	Cogn.	Affect.	Differ.	Inner.	Avg.	Probing.	Environ.	Reflect.	Change.	Avg.
MeChat	3.07	3.04	3.67	3.26	1.41	1.37	1.45	1.63	1.44	3.94	3.92	3.36	2.94	3.45
CPsyCounX	3.44	3.21	3.86	3.51	1.68	1.59	1.53	1.97	1.68	3.91	4.03	3.34	3.15	3.52
LLaMA3-8B	3.50	3.36	3.67	3.51	1.64	1.53	1.53	1.85	1.63	4.01	4.03	3.66	3.39	3.72
LLaMA3-70B	3.54	3.40	3.82	3.59	1.69	1.63	1.76	1.83	1.70	4.07	4.13	3.74	3.42	3.78
GPT-4o	3.61	3.41	3.86	3.62	1.58	1.49	1.56	1.84	1.59	4.06	4.10	3.74	3.56	3.81
Deepseek-R1	**3.92**	**3.62**	**4.33**	**3.96**	**2.08**	**2.04**	**2.03**	**2.30**	**2.10**	**4.58**	**4.66**	**4.36**	**4.13**	**4.39**
Avg. (GPT-4o)	3.51	3.34^{L}	3.87^{H}	3.58	1.68	1.61^{L}	1.64	1.90^{H}	1.69	4.10	4.15^{H}	3.70	3.43^{L}	3.78
Avg. (Human)	3.65	3.47^{L}	3.69^{H}	3.60	1.64	1.63^{L}	1.70^{H}	1.69	1.65	4.01	4.09^{H}	3.75	3.60^{L}	3.82

Table 4: Model performance(best in bold) on three scales. For scoring, the WAI and CCS-R scales use a 5-point scale from 1 to 5, whereas the BLRI scale uses a 6-point scale ranging from -3 to +3 (excluding 0). For WAI, dimensions include Counseling Goal(Goal.), Task Agreement(Task.), and Emotional Bond(Bond.). For BLRI, dimensions are Cognitive Empathy(Cogn.), Affective Empathy(Affect.), Differentiated Empathy(Differ.), and Inner Pattern(Inner.). For CCS-R, dimensions include Probing Techniques(Probing.), Facilitate Therapeutic Environment(Environ.), Reflecting(Reflect.), and Change Facilitation(Change.). Avg represents the average score for each scale. In the final two Avg rows, the superscripts L and H denote the lowest and highest scoring categories across all models, respectively.

analysis of how performance scales with parameter count. We also incorporate two domain-specific Chinese models, MeChat (Qiu et al. 2024) and CPsyCounX (Zhang et al. 2024), to enable a direct comparison between general-purpose models and those specifically trained for psychological counseling. A unified counselor prompt is applied to the four general-purpose models, whereas the specialized models use the prompts from their original studies.

Evaluation Results

After simulating interactions with CARE-Bench principle-guided simulated clients, six models are evaluated using GPT-4o to score the dialogue histories on multidimensional scales. The results are presented in Table 4. To ensure the reliability of GPT-4o's scoring, we randomly sample 100 counseling dialogues and invite two psychologists to conduct human evaluations using the same criteria. The average consistency between the human experts and GPT-4o reaches 0.72. Moreover, GPT-4o shows similar performance trends to the human experts across different sub-dimensions, indicating that its scoring standards are well aligned with professional judgment and that it is capable of providing objective evaluations in such psychological counseling scenarios.

Deepseek-R1 demonstrates a significant performance advantage, particularly in counseling skills (CCS-R), where it surpasses the second-best model by an average of 0.58 points. It also holds a 0.51-point lead in empathic understanding (BLRI). Analysis of Deepseek-R1's intermediate reasoning reveals that it infers underlying cognitive distortions and core issues from client statements rather than relying on superficial soothing. Based on these inferences, it proactively applies professional counseling techniques and empathic responses. These findings underscore the critical role of reasoning in enhancing the psychological counseling capabilities of Large Language Models (LLMs). While GPT-4o performs well in building therapeutic alliance (WAI) and counseling skills (CCS-R), its performance in empathic understanding (BLRI) is comparatively modest and is surpassed by the domain-specific model CPsyCounX.

Affective empathy, as measured by the BLRI, is a common weakness across all models. This metric assesses the ability to perceive, experience, and respond to a client's inner emotional state—to "feel what the client feels." Although models may exhibit cognitive empathy by logically understanding a client's issues, they struggle to genuinely connect with the client's emotions. Consequently, their responses, while logically sound, can lack emotional warmth and appear mechanical or cold.

Furthermore, models generally perform poorly in Change Facilitation, primarily due to low scores in the Confrontation subcategory. This subcategory requires the counselor to challenge the client to recognize and evaluate inconsistencies. When faced with a client's contradictory, avoidant, or irrational statements, most models tend to maintain surface-level harmony rather than actively addressing the underlying discrepancies. This overly compliant conversational style indicates a lack of strategy for confrontational interventions, limiting the models' ability to help clients confront their defense mechanisms or behavioral blind spots.

Differential Performance Across Client Characteristics

Effective psychological counseling LLMs should perform well for diverse clients and problems. Using the varied client profiles in CARE-Bench, we analyze model performance across multiple dimensions. Our findings show that performance differs based on client traits, informing the development of more adaptable and comprehensive models.

Big Five Personality

Psychological research indicates that counseling outcomes correlate significantly with a client's Big Five personal-

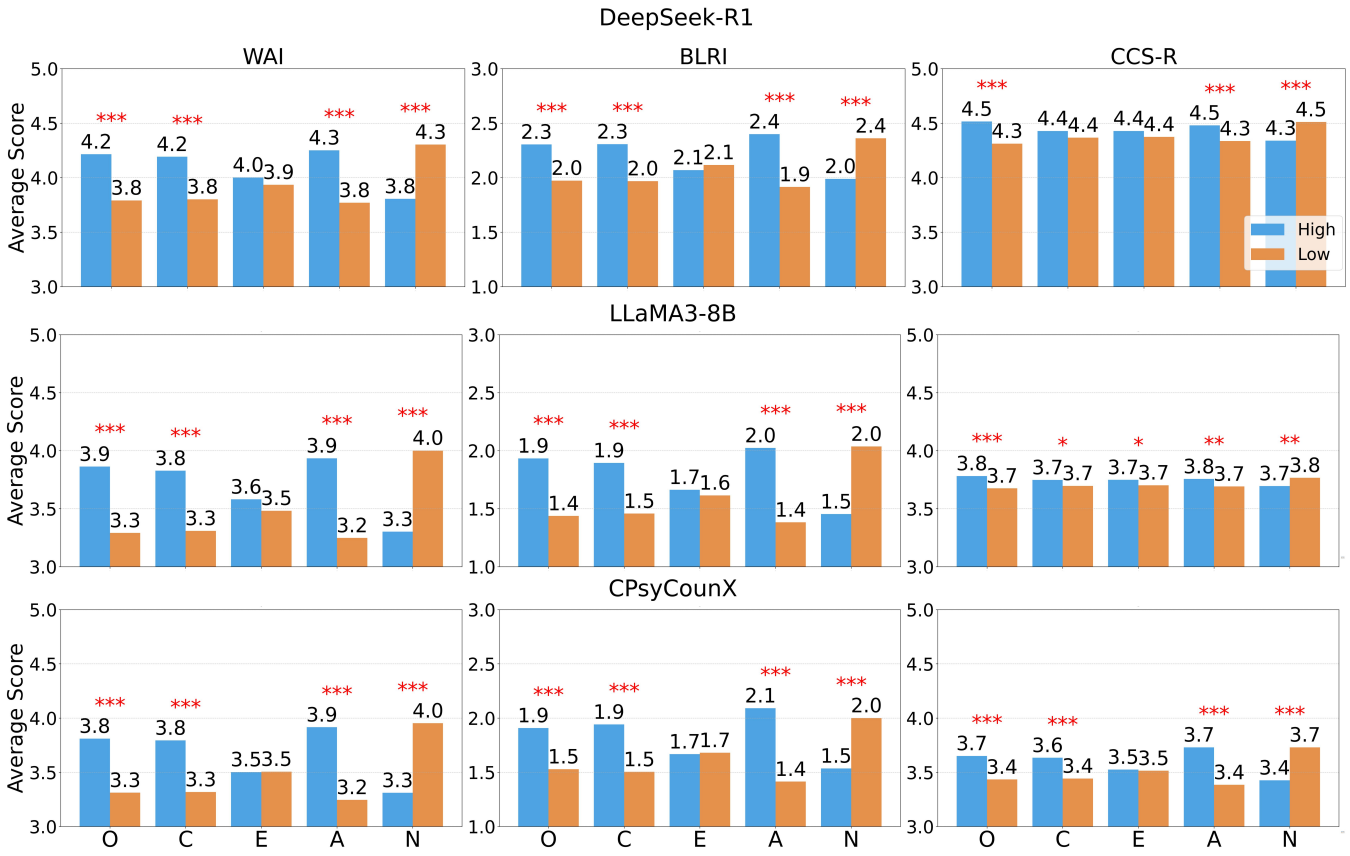

Figure 3: The generalizability results for representative models of three different types, which are grouped by the **Big Five Personality Traits** categorized as "High" or "Low". The letters on the x-axis correspond to the five dimensions: **O**penness, **C**onscientiousness, **E**xtraversion, **A**greeableness, and **N**euroticism. The y-axis indicates the models' average counseling score per item. Significance tests are conducted for all results (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

ity traits (Bucher, Suzuki, and Samuel 2019). As shown in Figure 3, our analysis reveals a general trend in LLM-based counseling: the models are less effective for client **low Openness, low Conscientiousness, low Agreeableness, or high Neuroticism**. These limitations are particularly pronounced in establishing a therapeutic relationship and conveying empathy. Therefore, achieving a fair assessment requires including a diverse range of client profiles.

In collaboration with psychologists, we analyze low-scoring counseling transcripts to understand why the counseling is ineffective for certain clients. The analysis reveals distinct patterns linked to personality traits. Clients with low openness question introspective suggestions like journaling, deeming them impractical for solving real-world problems. Clients with low conscientiousness express hesitation and futility; the LLM fails to address their core helplessness, instead prematurely pushing for tasks, which erodes trust. Similarly, clients with low agreeableness display antagonism, and the LLM's shallow empathy is insufficient to build the necessary trust. Finally, for emotionally unstable clients high in neuroticism, the LLM's repetitive "calm down" suggestions are counterproductive and stall the therapeutic process.

Further statistical analysis across **Counseling Topic**, **Age Group** and **Gender** reveals additional performance disparities. The LLMs perform more poorly on behavioral topics and with the 0–11 age group. The models are also less effective for male clients than for female clients. This multi-dimensional analysis highlights significant capability biases in current LLMs for psychological counseling, offering crucial insights for developing more equitable and effective consultation models. Detailed results are provided in the appendix.

Conclusion

We introduces CARE-Bench, a comprehensive benchmark for evaluating the psychological counseling capabilities of LLMs. To address the shortcomings of existing benchmarks, such as unprofessional client simulation and unidimensional evaluations, CARE-Bench features diverse and realistic client profiles derived from over 1,500 real-world cases and guided by expert principles.

Our tests reveal that current leading LLMs share common weaknesses and struggle with clients exhibiting specific characteristics, offering clear guidance for creating more effective and empathetic LLM counselors.

Ethical Statement

The data collection for this study is based on over 1,500 public consultation cases sourced from authoritative platforms. We strictly complied with ethical standards during data acquisition. In establishing the guiding principles for client simulation, we collaborated with ten professional psychologists from a psychological clinic. We provided substantial remuneration to these psychologists to acknowledge their valuable professional contributions.

CARE-Bench is explicitly a Chinese benchmark. Its client profiles are constructed from real-world cases within a Chinese context. Consequently, the evaluation results and findings (e.g., model performance in this specific linguistic and cultural context) may not be directly generalizable to other languages and cultures, where counseling norms, client issues, and emotional expression can differ significantly.

Furthermore, although we employ expert-guided principles to enhance simulation realism—an approach human evaluation confirmed as more effective than using profiles alone—the client simulation is ultimately powered by an LLM (Qwen2.5-Max). Inherent limitations exist in any LLM's ability to fully replicate the complex, subtle, and sometimes contradictory nature of human emotion, subconscious behavior, and lived experience. The simulation remains an approximation of real-world clients.

Acknowledgments

This work was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project 2023ZD0121100, the National Natural Science Foundation of China (NSFC) via grant 62441614 and 62176078.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Barrett-Lennard, G. T. 1962. Dimensions of therapist response as causal factors in therapeutic change. *Psychological monographs: General and applied*, 76(43): 1.

Beutler, L. E.; and Clarkin, J. F. 2014. *Systematic treatment selection: Toward targeted therapeutic interventions*. Routledge.

Bucher, M. A.; Suzuki, T.; and Samuel, D. B. 2019. A meta-analytic review of personality traits and their associations with mental health treatment outcomes. *Clinical psychology review*, 70: 51–63.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Fitzpatrick, K. K.; Darcy, A.; and Vierhile, M. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health*, 4(2): e7785.

Goldberg, L. R. 1990. An alternative" description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6): 1216–1229.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Horvath, A. O.; and Greenberg, L. S. 1989. Development and validation of the Working Alliance Inventory. *Journal of counseling psychology*, 36(2): 223.

Jin, H.; Chen, S.; Dilixiati, D.; Jiang, Y.; Wu, M.; and Zhu, K. Q. 2023. Psyeval: A suite of mental health related tasks for evaluating large language models. *arXiv preprint arXiv:2311.09189*.

Kotov, R.; Gamez, W.; Schmidt, F.; and Watson, D. 2010. Linking "big" personality traits to anxiety, depressive, and substance use disorders: a meta-analysis. *Psychological bulletin*, 136(5): 768.

Lambie, G. W.; Mullen, P. R.; Swank, J. M.; and Blount, A. 2018. The counseling competencies scale: Validation and refinement. *Measurement and Evaluation in Counseling and Development*, 51(1): 1–15.

Lee, S.; Mac Kim, S.; Kim, M.; Kang, D.; Yang, D.; Kim, H.; Kang, M.; Jung, D.; Kim, M.; Lee, S.; et al. 2024. Cactus: Towards Psychological Counseling Conversations using Cognitive Behavioral Theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14245–14274.

Li, Y.; Yao, J.; Bunyi, J. B. S.; Frank, A. C.; Hwang, A.; and Liu, R. 2025. CounselBench: A Large-Scale Expert Evaluation and Adversarial Benchmark of Large Language Models in Mental Health Counseling. *arXiv preprint arXiv:2506.08584*.

Louie, R.; Nandi, A.; Fang, W.; Chang, C.; Brunskill, E.; and Yang, D. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 10570–10603.

Malouff, J. M.; Thorsteinsson, E. B.; and Schutte, N. S. 2005. The relationship between the five-factor model of personality and symptoms of clinical disorders: A meta-analysis. *Journal of psychopathology and behavioral assessment*, 27(2): 101–114.

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

Norcross, J. C.; and Wampold, B. E. 2011. What works for whom: Tailoring psychotherapy to the person. *Journal of clinical psychology*, 67(2): 127–132.

Organization, W. H. 2022. *World mental health report: Transforming mental health for all*. World Health Organization.

Qiu, H.; He, H.; Zhang, S.; Li, A.; and Lan, Z. 2024. SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 615–636.

Qwen; :; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Tang, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2025. Qwen2.5 Technical Report. arXiv:2412.15115.

Strickhouser, J. E.; Zell, E.; and Krizan, Z. 2017. Does personality predict health and well-being? A metasynthesis. *Health psychology*, 36(8): 797.

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

Wang, J.; Wang, B.; Fu, X.; Sun, Y.; Zhao, Y.; and Qin, B. 2025. Psychological Counseling Cannot Be Achieved Overnight: Automated Psychological Counseling Through Multi-Session Conversations. *arXiv e-prints*, arXiv–2506.

Xiao, M.; Xie, Q.; Kuang, Z.; Liu, Z.; Yang, K.; Peng, M.; Han, W.; and Huang, J. 2024. HealMe: Harnessing Cognitive Reframing in Large Language Models for Psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1707–1725.

Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; and Hu, X. 2024. CPsyCoun: A Report-based Multi-turn Dialogue Reconstruction and Evaluation Framework for Chinese Psychological Counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, 13947–13966.

Zhang, J.; He, H.; Song, N.; Zhou, Z.; He, S.; Zhang, S.; Qiu, H.; Li, A.; Dai, Y.; Ma, L.; et al. 2023. ConceptPsy: A Benchmark Suite with Conceptual Comprehensiveness in Psychology. *arXiv preprint arXiv:2311.09861*.

Zhang, M.; Yang, X.; Zhang, X.; Labrum, T.; Chiu, J. C.; Eack, S. M.; Fang, F.; Wang, W. Y.; and Chen, Z. 2025. CBT-Bench: Evaluating Large Language Models on Assisting Cognitive Behavior Therapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3864–3900.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).