Ψ-Arena: Interactive Assessment and Optimization of LLM-based Psychological Counselors with Tripartite Feedback

Shijing Zhu[1], Zhuang Chen[1*], Guanqun Bi[2], Binghang Li[3], Yaxi Deng[1], Dazhen Wan[3], Libiao Peng[3], Xiyao Xiao[3], Rongsheng Zhang[4], Tangjie Lv[4], Zhipeng Hu[4], FangFang Li[1*], Minlie Huang[2]

[1]School of Computer Science and Engineering, Central South University
[2]CoAI Group, DCST, IAI, BNRIST, Tsinghua University
[3]Lingxin AI
[4]Netease
zhchen18@foxmail.com

Abstract

Large language models (LLMs) have shown promise in providing scalable mental health support, while evaluating their counseling capability remains crucial to ensure both efficacy and safety. Existing evaluations are limited by the static assessment that focuses on knowledge tests, the single perspective that centers on user experience, and the open-loop framework that lacks actionable feedback. To address these issues, we propose Ψ-ARENA, an interactive framework for comprehensive assessment and optimization of LLM-based counselors, featuring three key characteristics: (1) Realistic arena interactions that simulate real-world counseling through multi-stage dialogues with psychologically profiled NPC clients; (2) Tripartite evaluation that integrates assessments from the client, supervisor, and counselor perspectives; (3) Closed-loop optimization that iteratively improves LLM counselors using diagnostic feedback. Experiments across eight state-of-the-art LLMs show significant performance variations in different real-world scenarios and evaluation perspectives. Moreover, reflection-based optimization results in up to a 141% improvement in counseling performance. We hope Ψ-ARENA provides a foundational resource for advancing reliable and human-aligned LLM applications in mental healthcare.

Extended version — https://arxiv.org/abs/2505.03293

Introduction

Mental health disorders affect over 1 billion people globally, with the World Health Organization noting their significant societal and economic impacts (WHO 2023). However, there is a severe shortage of mental health professionals, with approximately 100,000 people per counselor. This shortage has driven the exploration of AI-based counseling systems as a potential solution. In the 1960s, early rule-based AI systems like ELIZA (Weizenbaum 1966) showed the feasibility of automated counseling. Today, large language models (LLMs) like GPT-4 (Achiam et al. 2023) and Claude (Anthropic 2023) exceed human abilities in certain tasks, prompting increasing efforts to use LLMs for scalable counseling and make mental services more accessible (Chen et al. 2023; Iftikhar et al. 2024; Xu et al. 2025). This trend highlights the urgent need for rigorous evaluation to ensure that these systems meet clinical standards for effectiveness, control, and safety.

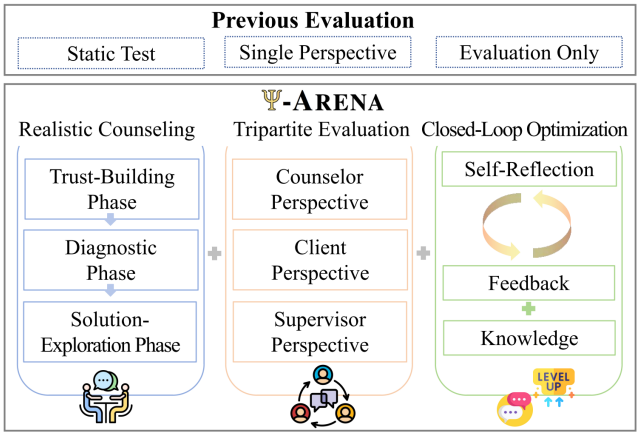


Figure 1: The comparison between Ψ-ARENA and existing studies on evaluating LLM-base counselors.

Although pioneering studies have attempted to evaluate LLM counselors, three key challenges remain that prevent comprehensive and in-depth assessments: **1) Gap between understanding and application**. Existing studies tend to focus on static assessments, such as multiple-choice questions or diagnostic accuracy metrics, which measure knowledge rather than practical application (Jin et al. 2023; Zhang et al. 2024b). **2) Limited user-centric metrics**. While Zhao et al. (2024) and Wang et al. (2024b) try to simulate counseling interactions between clients and counselors, they primarily focus on client satisfaction and subjective feelings, ignoring evaluations from supervisors and counselors themselves. **3) Lack of feedback loops**. Most existing frameworks lack actionable feedback for model improvement, which should be a key objective of any evaluation system.

In this paper, we propose Ψ-ARENA, an interactive plat-

*Corresponding author.

form for assessing and optimizing LLM-based psychological counselors. In Ψ-ARENA, LLM counselors engage with virtual NPC clients and receive assessments from three perspectives: the client, the supervisor, and the counselor. These evaluations provide targeted feedback that helps optimize the counseling process. Specifically, Ψ-ARENA features three key elements: **1) Realistic counseling scenarios**. To ensure the arena simulates real-world counseling, we focus on client profiles and behaviors. For profiles, we analyze real counseling records to identify key attributes for virtual clients' psychological profiles and create 10,000 virtual client profiles (NPC cards) across 100 topics for use. For behaviors, we base the simulation on professional counseling knowledge, ensuring meaningful interactions across different phases: "*trust-building*→ *diagnosis* → *solution exploration*". **2) Tripartite evaluation metrics**. We integrate evaluations from clients (subjective experience), supervisors (professional competency), and counselors (reflective awareness), enabling a 360° competency analysis across 33 dimensions. **3) Closed-loop optimization**. We introduce a feedback and optimization cycle, where evaluation results are combined with professional counseling guides to generate specific feedback, allowing LLM counselors to self-reflect and iteratively improve their responses.

In Ψ-ARENA, we evaluate eight state-of-the-art LLMs, including closed-source models like Claude-3.5-Sonnet and open-source models like DeepSeek-671B. Our results show significant performance disparities across these LLM counselors when evaluated from different perspectives, emphasizing the need for arena simulations and multi-source evaluations. We also compare the automatic evaluation results with those of human experts, revealing high consistency and validate the effectiveness. Additionally, through specific feedback and optimization, we achieve up to a 141% improvement in counseling performance, showcasing the potential of a closed-loop evaluation system.

Our key contributions are: (1) Introducing Ψ-ARENA, which features realistic counseling scenarios, tripartite evaluation metrics, and closed-loop optimization. (2) Evaluating the counseling performance of state-of-the-art LLMs and demonstrating consistency with human experts, achieving performance improvements based on feedback. (3) Conducting in-depth analysis of LLM performance across various dimensions and topics. We hope Ψ-ARENA to serve as an efficient and effective evaluation framework, advancing the responsible development of LLMs in mental healthcare.

Ψ-ARENA

Framework Overview

As shown in Figure 2, Ψ-ARENA is an interactive framework for assessing and optimizing LLM-based psychological counselors. Ψ-ARENA encompasses virtual clients with diverse psychological profiles who engage in multi-stage counseling dialogues with LLM counselors. Then Ψ-ARENA evaluates counselor performance from three perspectives: client, supervisor, and counselor. Based on these evaluations, Ψ-ARENA generates feedback to guide the counselor's self-reflection and iterative improvement.

Client in Ψ-ARENA

In Ψ-ARENA, virtual NPC clients are created with rich psychological profiles and realistic behaviors to ensure that the simulation of counseling scenarios is both diverse and authentic.

Client Profiles The construction of profiles is based on real-world counseling records, ensuring that the virtual clients reflect authentic psychological concerns (Wenhua 2020). Each profile includes several key attributes that define the client's background, emotional state, and the issues they seek counseling for. The below attributes are extracted and incorporated into the client profiles: *demographics*, *cultural background*, *personality trait*, *emotional state*, *current distress*, *detailed distress description*, and *core theme*. Details of attributes can be found in Appendix.

To ensure high-quality and diverse client profiles, we resort to the real-world PsyQA dataset (Sun et al. 2021) which contains conversations from real clients, covering common mental health disorders across nine themes, including self-growth, emotional issues, relationships, behavior, family, therapy, marriage, and career. Each theme contains several subtopics, ultimately generating 100 distinct topics for client profiles. Details of topics can be found in Appendix.

To build these profiles, we first use GPT-4o to extract initial psychological profiles from PsyQA. The extraction process is guided by carefully designed prompts to ensure that each profile accurately captured key psychological characteristics while strictly adhering to real-world dialogue content and mitigating potential biases. Detailed prompts can be found in Appendix. Subsequently, we conducted manual verification of the extracted results to ensure coverage across diverse genders, age groups, and cultural backgrounds. We ultimately generated 10,000 high-quality client psychological profiles. Building upon this, we further manually selected the most representative high-quality profiles for each specific topic to serve as the foundational dataset for constructing the NPC clients.

Client Behaviors To simulate realistic counseling interactions, we design client behaviors that match the different stages of a typical counseling process. These behaviors help ensure the virtual clients engage in meaningful conversations with LLM counselors. Based on counseling models from existing research, we focus on three main phases of client behavior during the simulation.

- **Trust-Building Phase** (Sachse 2024) In the beginning, the virtual client works on building trust by being open to the counselor's questions, sharing personal feelings, and offering context about their struggles. The client might show vulnerability, helping establish a connection and encourage a safe space for further discussions.
- **Diagnostic Phase** (Zhang et al. 2024a) During this phase, the client begins to share more personal information, such as their background, emotional state, and the deeper causes of their distress. They may reflect on past experiences and feelings, providing the counselor with insights into what might be influencing their current struggles.

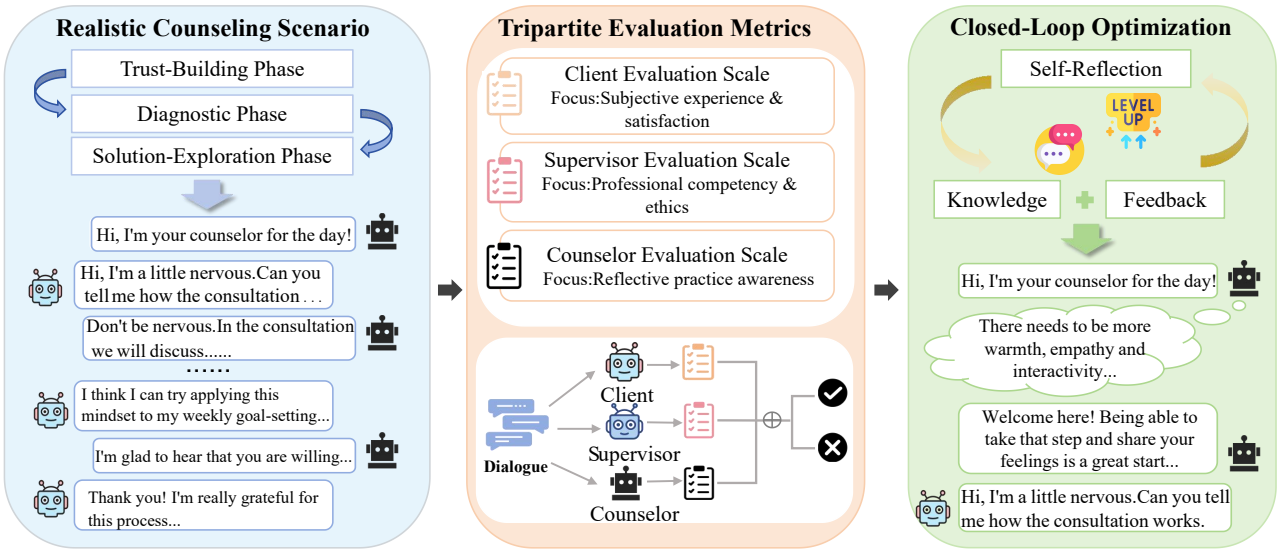


Figure 2: In Ψ-ARENA, LLM-based counselors interact with NPC clients, receive multi-source evaluations, and improve counseling performance through self-reflection.

- **Solution-Exploration Phase** (Hill 2020) In the final phase, the client actively explores possible solutions and coping strategies. They may express hope, consider different options, or reflect on past efforts to resolve their issues. The client may also ask for advice and discuss potential next steps for moving forward.

Client Simulation To ensure authenticity, the client's responses are simultaneously guided by their defined psychological profiles and behavior patterns. To align the virtual client's dialogue with real counseling scenarios, we follow five core principles when instructing GPT-4o for client simulation (Tu et al. 2024; Shao et al. 2023; Chen et al. 2024): realism (ensuring the conversation matches the client's language style and emotional expression), fluency (maintaining logical and natural dialogue flow), completeness (covering key tasks across all counseling stages), personalization (reflecting the client's unique background and traits), and behavioral consistency (ensuring stable behavior patterns across different conversation turns). Detailed prompts for client simulation can be found in Appendix.

Counselor in Ψ-ARENA

In Ψ-ARENA, the evaluated LLMs act as counselors and engage in conversations with different clients. To ensure a fair evaluation of LLM-based counselors, we provide each model with a standardized `psycho prompt` that clearly defines its role as a psychological counselor, including the role positioning, interaction rounds, and the response constraints. No additional guiding information is given to avoid artificially boosting performance. Further details of counselor prompts are provided in Appendix. For comparison, we also include the default `system prompt` (e.g., "you are a helpful assistant") to observe the vanilla performance without any specific instruction.

Tripartite Evaluation Metrics

To ensure a comprehensive evaluation of LLM-based psychological counselors, we introduce a tripartite evaluation system that assesses the counseling dialogue from three distinct perspectives: the client, the supervisor, and the counselor. This approach, inspired by established frameworks in psychology (Lockyer 2003; Kuzmits et al. 2004; Tham 2007), provides a holistic assessment of the counselor's professional abilities, as shown in Table 1. Below, we briefly describe each evaluation scale and its core focus areas. The detailed scoring items are available in the Appendix.

Client-Oriented Scale The client-oriented scale, developed by Joel Black in "*Who Stole Your Trust and Confidence*"? (Black 2003), gathers feedback from clients on the counselor's effectiveness and the quality of the counselor-client relationship. It includes 16 dimensions such as trust, empathy, and communication clarity, rated on a scale from 0 to 4. The focus is on the client's emotional experience and satisfaction with the counselor's approach, offering insights into the counseling process.

Supervisor-Oriented Scale The Supervisor-Oriented Scale, based on the "*Consultant Competency Assessment Tool*" developed by the American Psychological Association (APA 2023), evaluates the counselor's professional competence and adherence to ethical standards. It includes 8 dimensions, such as therapeutic techniques and cultural competence, rated from 0 to 4. This scale emphasizes the counselor's technical skills, ethical conduct, and adherence to professional practices.

Counselor-Oriented Scale The counselor-oriented scale, developed by Yang and Xiong (2018), allows counselors to self-assess their practices. It covers 20 dimensions, with 9 focused on practical counseling abilities like empathy and client response. Counselors rate their alignment with these

Perspective	Evaluation Focus	Scale Characteristics	Realistic Threshold
Client	Subjective experience & satisfaction	16-dimension scale (0-4 per item)	>42 (Total 64)
Supervisor	Professional competency & ethics	8-dimension APA scale (0-4 per item)	>24 (Total 32)
Counselor	Reflective practice awareness	9-dimension ability scale (0-5 per item)	>35 (Total 45)

Table 1: Evaluation criteria and thresholds of tripartite scales.

abilities on a scale from 0 to 5. This scale encourages self-awareness and supports ongoing professional development.

During the evaluation phase, we use GPT-4o to simulate the roles of clients, supervisors, and consultants for automated assessment. This phase does not rely on the inherent expertise of LLMs but is instead based on clearly defined standardized scale dimensions and scoring criteria. The design strictly confines the role of the LLM to a rule-based scoring executor, thereby ensuring the robustness of the evaluation even when the model lacks in-depth domain knowledge. Detailed prompts for these role-playing scenarios are provided in Appendix. To validate the effectiveness of the automatic scoring system, we also recruit psychological experts to score a sample of 30 sessions conducted by various LLM counselors. We then compare the consistency between the automatic evaluations and the human expert scores. The results of this comparison are presented in Section to demonstrate the reliability and effectiveness of the tripartite evaluation system.

Closed-Loop Optimization

Ψ-ARENA employs an iterative self-reflection mechanism to enhance LLM-based counselors. Specifically, we first use GPT-4o to automatically generate detailed feedback for the low-scoring evaluation dimensions through the counseling dialogue, and draw from established psychological frameworks, such as "*Practice of Counseling and Psychotherapy*" (Corey 2013), to construct a knowledge base, including 11 methods such as Cognitive Behavioral Therapy (CBT), Humanistic Therapy, and Psychodynamic Therapy. Then, we restart the dialogue, and each round of the optimization process followed two steps. (1) The counselor reflects on the strengths and weaknesses of the current response based on the feedback. (2) The counselor rewrites the response based on the knowledge base and reflection results to improve it. Finally, the new version of the response is re-evaluated. By comparing the scores before and after the revision, we assess whether Ψ-ARENA can help counselors improve in a way similar to real-world supervision. Detailed prompts for feedback generation and self-reflection are provided in Appendix.

Experiment

Settings & Metrics

In the experiment, each LLM-based counselor engages in individual dialogues with 100 virtual clients. Each counseling session consists of 25 conversational rounds. The dialogue content is then evaluated using a tripartite scoring system, and feedback is generated to improve the model.

For metrics, we calculate the average scores on each of the three scales and determine an overall mean score. We also introduce the "pass rate," a metric commonly used in real-world counselor assessments, to provide a clear view of counseling performance. Specifically, each scale has a threshold to determine whether the counselor meets the required standards: the client scale (Total >42), the counselor scale (Total >35), and the supervisor scale (Total >24). We analyze the pass rate on each scale and also compute the overall pass rate, which considers a counselor as passing only if they meet the thresholds on all three scales.

To validate the effectiveness of the automated evaluation, we also recruit two graduate students with a background in psychology to manually label the model's performance. These two experts are very familiar with the research content, and we provide them with the complete tripartite scales and necessary instructions to ensure they are fully equipped to carry out the labeling task. We pay each expert an hourly rate of $13.78.

Models

We evaluate eight state-of-the-art large language models. The closed-source models selected are GPT-3.5 Turbo (OpenAI 2023a), GPT-4o (OpenAI 2023b), Claude-3.5-Sonnet (Anthropic 2025), GLM-4-Plus (Zhipu 2023), and MiniMax-Text-01 (Li et al. 2025). For open-source models, we use LLaMA-3.3 (70B) (Meta 2023), Qwen-2.5 (72B) (Yang et al. 2024), and Deepseek-v3 (671B) (Liu et al. 2024).

For all open-source models, we deploy and experiment on two 8×H20 GPU servers. For all closed-source models, we gain access through the official APIs. For experiments, we keep all default hyperparameters (such as temperature, top-p, etc.) unchanged for all models. We strictly follow the license requirements of each model during usage.

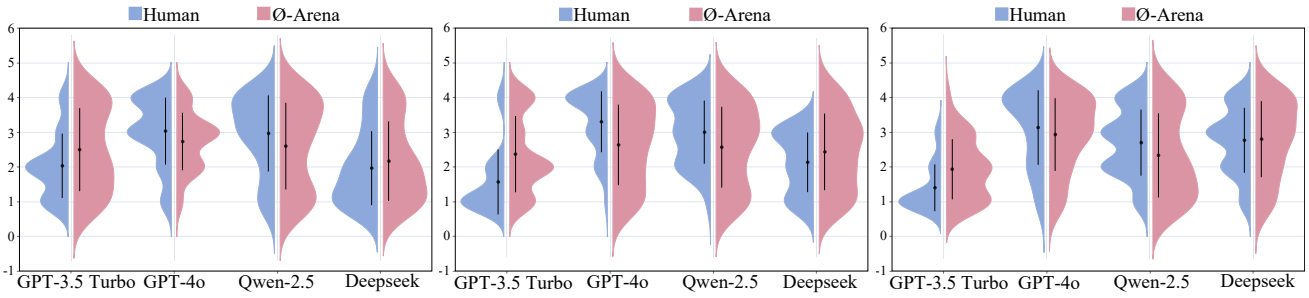
Evaluation Results

Table 2 shows the counseling performance of the evaluated LLMs. We now break down the results and highlight several key observations. For clarity, we primarily focus on the pass rates.

Overall Counseling Performance Overall, Claude-3.5-Sonnet, GPT-4o, and Deepseek are the top-performing models, with Claude-3.5-Sonnet (84%) leading in both client and supervisor evaluations. GPT-4o and Deepseek also perform well, consistently earning high marks across all evaluation dimensions, especially in counselor self-assessments.

Model	Vanilla Prompt				Psycho Prompt				δ
	Client	Supervisor	Counselor	Overall	Client	Supervisor	Counselor	Overall	Pass Rate
GPT-3.5 Turbo	2.68 (57%)	2.60 (13%)	3.43 (10%)	2.90 (33%)	2.61 (50%)	2.66 (20%)	3.60 (22%)	2.96 (39%)	+6%
GLM-4-Plus	2.58 (47%)	2.41 (9%)	3.32 (12%)	2.77 (16%)	2.53 (43%)	2.64 (14%)	3.71 (38%)	2.96 (39%)	+23%
MiniMax-Text-01	2.61 (56%)	**2.71 (23%)**	**3.59** (19%)	2.97 (35%)	2.64 (60%)	2.77 (27%)	3.62 (25%)	3.01 (46%)	+11%
LLaMA-3.3	2.47 (21%)	2.53 (9%)	3.46 (17%)	2.82 (15%)	2.63 (47%)	2.65 (18%)	3.78 (45%)	3.02 (48%)	+33%
GPT-4o	**2.73 (73%)**	2.69 (17%)	3.54 (19%)	**2.99 (44%)**	2.69 (64%)	2.80 (30%)	3.78 (42%)	3.09 (57%)	+13%
Qwen-2.5	2.62 (64%)	2.68 (15%)	3.56 **(22%)**	2.95 (33%)	2.71 (65%)	2.74 (26%)	3.72 (37%)	3.06 (59%)	+26%
Deepseek	2.58 (53%)	2.60 (9%)	3.43 (9%)	2.87 (28%)	2.64 (58%)	2.81 (34%)	4.03 (60%)	3.16 (65%)	+37%
Claude 3.5 Sonnet	2.66 (59%)	2.65 (17%)	3.53 (20%)	2.95 (37%)	**2.75 (72%)**	**2.87 (47%)**	**4.08 (77%)**	**3.23 (84%)**	**+47%**

Table 2: Evaluation results of Ψ-ARENA of all LLMs including both scores (best in bold) and pass rates (best in underline). The score reflects the model's average performance across all tests, while the pass rate measures the number of passed tests. Therefore, the highest score and the highest pass rate may be achieved by different models.



(a) Client scale. (b) Supervisor scale. (c) Counselor scale.

Figure 3: Comparison of consistency results between Ψ-ARENA and human experts.

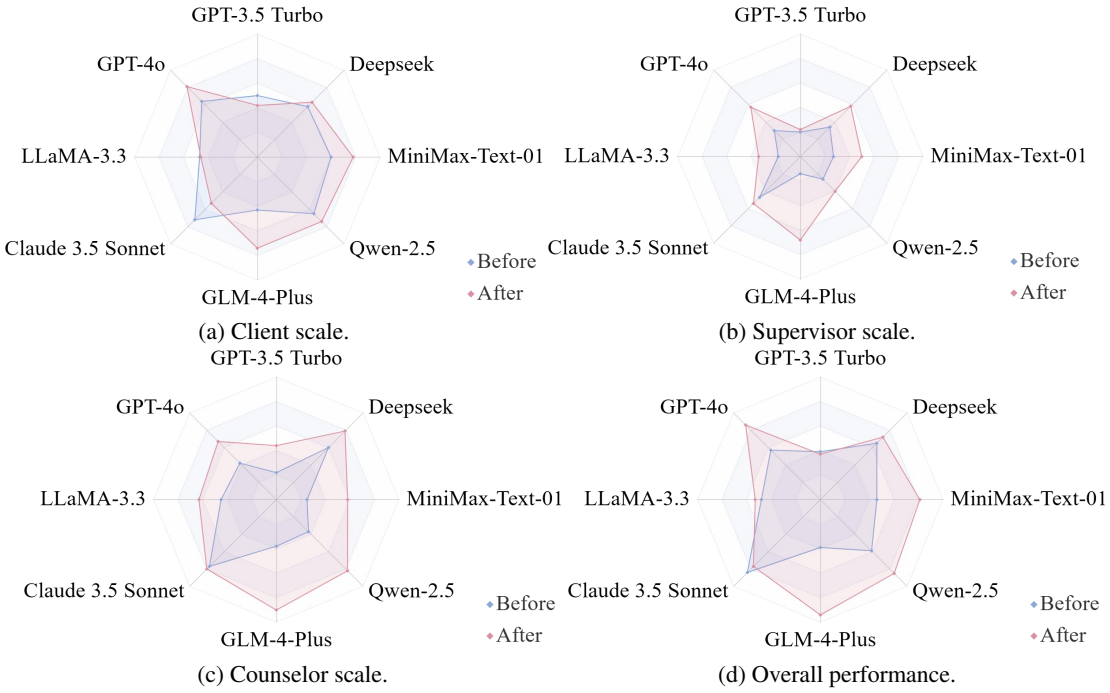
Necessity of Tripartite Evaluation The tripartite evaluation show significant variability. Client-side evaluations consistently yield higher scores, reflecting their focus on empathy and the quality of personal interaction. In contrast, supervisor evaluations, which emphasize professional standards, techniques, and ethics, are more stringent. Counselor-side evaluations lie between those of the client and supervisor, suggesting that self-assessment can uncover insights beyond the immediate client experience.

Impact of the Psycho Prompt The introduction of the `Psycho Prompt` improves counselor performance across the board, with the effect being more pronounced in stronger models. For instance, GPT-3.5-Turbo shows a modest increase of 6 percentage points in its overall pass rate (from 33% to 39%), while Claude-3.5-Sonnet sees a substantial increase of 47 points (from 37% to 84%). This suggests that stronger models benefit more from simple instructions and guidance.

Consistency with Human Experts We validate the effectiveness of automated evaluation through manual assessment.A pilot study reveals that manual scores fluctuate due to subjective interpretation, emotional swings, and fatigue. Therefore,we employ a relative ranking approach. Two psychological experts conduct consensus-based rankings of 30 dialogues from four models (GPT-3.5-Turbo, GPT-4o, Qwen-2.5, Deepseek) according to three evaluation scales.The consensus-based expert annotation had a pre-consensus Spearman correlation of 0.997. We convert rankings into scores (1st = 4 points, 2nd = 3 points, etc.) and analyze score distributions. As shown in Figure 3, the results indicate a high overall consistency between automated and manual evaluations, especially in the client scale. Although minor discrepancies occur in supervisor and counselor scales, the overall trends remain consistent. For quantitative verification, we implement a pairwise comparison strategy to calculate Cohen's Kappa coefficients. As shown in Table 3.

Cohen's Kappa	Client	Supervisor	Counselor
3.5 vs 4o	0.842	0.615	0.793
3.5 vs qwen	0.545	0.553	0.667
3.5 vs deepseek	0.933	1.000	0.535
4o vs qwen	0.867	0.865	0.862
4o vs deepseek	0.545	0.553	0.667
qwen vs deepseek	0.667	0.587	0.587
Avg	0.733	0.696	0.685

Table 3: Cohen's Kappa coefficients between Ψ-ARENA and human experts.



(a) Client scale. (b) Supervisor scale.

(c) Counselor scale. (d) Overall performance.

Figure 4: Comparison of model pass rates before and after optimization.

Optimization Results

As shown in Figure 4, we visually demonstrate the differences in model pass rates before and after self-reflection. We here have three key observations.

Counseling Performance Improvement Most models show significant improvements after incorporating feedback, with GLM-4-Plus showing the largest increase of 55% points in its overall pass rate (from 39% to 94%, relatively 141%). This highlights that models with lower initial performance benefit the most from iterative feedback and optimization. In contrast, models with better starting performance, such as GPT-4o, show more moderate but still substantial improvements.

Discrepancy of Tripartite Feedback The improvements primarily stem from supervisor and counselor metrics, underscoring the tripartite evaluation system's value. While the client score, which focuses on emotional responses and satisfaction, already shows good performance, the main improvements are seen in the application of professional knowledge and its validation. This indicates that supervisor/-counselor feedback critically improves models' counseling capabilities beyond emotional response management.

Diminishing Returns and Marginal Effects The feedback process exhibits diminishing returns, not only for high-performance models but also for those with moderate initial capabilities. Strong models, like Claude-3.5-Sonnet, which start with high scores, experience limited improvements. This phenomenon can be attributed to the inherent high quality of their initial responses, which leaves limited room for further enhancement. Meanwhile, models with lower capabilities, such as GPT-3.5-Turbo, struggle to fully grasp and apply feedback, also resulting in slower progress.

Analysis

Thematic Analysis of Model Performance

Figure 5 shows comprehensive average performance of eight LLMs across nine themes. All models demonstrate strong performance in *Emotion* and *Self-growth*, indicating significant potential for emotional support and personal development. However, some models exhibit relatively weaker performance in *Treatment* and *Career*, suggesting these areas require specialized knowledge and complex reasoning.

It is noteworthy that Claude-3.5-Sonnet excels in the *Treatment* and *Career*, demonstrating its proficiency in complex, professional tasks. In contrast, Deepseek exhibits balanced performance across all themes, particularly in *Emotion* and *Love Problem*, indicating its strong generalization performance.

Dimensional Analysis of Performance

To thoroughly investigate the performance differences of the models across various dimensions, we systematically calculate the average scores for each dimension across three scales: client, supervisor, and counselor, as shown in Appendix.

In the client scale, the models perform well in *Equality*, *Inclusiveness*, and *Consultant*. However, the models underperform in *Humor* and *Candor*, suggesting limitations in emotional expression and interactive flexibility. In the supervisor scale, the models are good at *Active Listening* and

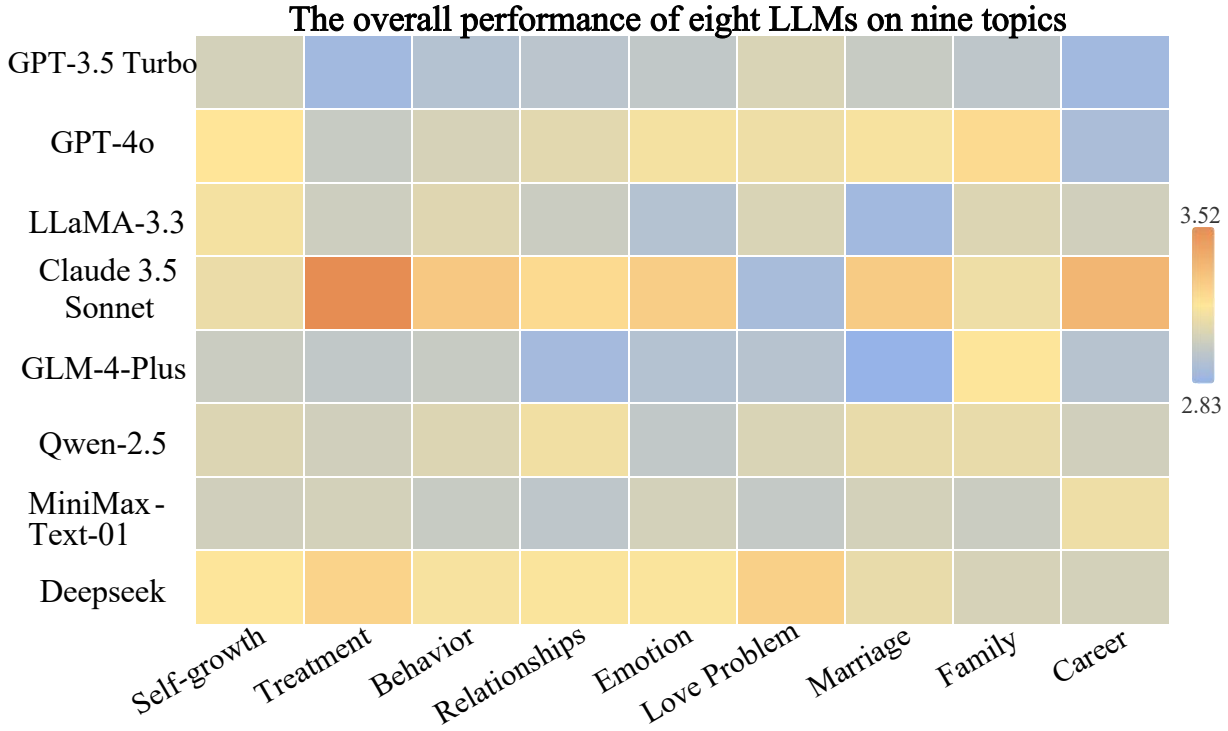


Figure 5: Fine-grained counseling performance of all LLMs on different topics.

Self-Regulation, but show deficiencies in *Professional Application* and *Flexible Intervention*, indicating a need for improvement in professional judgment and dynamic adjustment in complex scenarios. In the counselor scale, the models perform well in *Respect* and *Appropriate Response*, demonstrating their ability to effectively show respect and provide appropriate responses to clients. However, further optimization is needed in higher-level emotional support and strategic intervention to enhance their overall performance in complex counseling scenarios.

Related Work

With the rapid development of language models, they have been increasingly applied in the field of psychology, covering various areas such as mental health detection (Ji et al. 2021; Vajre et al. 2021; Zhang, Schoene, and Ananiadou 2021; Xu et al. 2024; Lan et al. 2024) and emotional support (Kang et al. 2024; Wang et al. 2024a; Xie and Peng 2025). However, their accuracy, ethical compliance, and scientific rigor remain critical challenges.

To address these issues, researchers have proposed various evaluation frameworks. Jin et al. (2023) focuses on evaluating LLMs' performance in mental health knowledge, diagnostic accuracy, and emotional support capabilities. Zhang et al. (2024b) systematically assesses LLMs' application abilities in cognitive behavioral therapy (CBT) across three dimensions. Both studies employ static evaluation methods, such as knowledge understanding tests and multiple-choice questions, emphasizing theoretical mastery over dynamic interaction capabilities. Zhao et al. (2024) assesses emotional support dialogues using role cards, role-playing models, and seven dimensions such as fluency and empathy. Wang et al. (2024b) evaluates LLMs as therapists from the client's perspective, focusing on conversational effectiveness, therapeutic alliance, and self-reported experiences. Despite utilizing dynamic interaction processes, both frameworks are confined to single-perspective evaluations, lacking a holistic multi-dimensional assessment approach. Moreover, these studies lack a feedback mechanism, remaining solely at the evaluation stage. In contrast, our approach incorporates a three-stage dynamic interaction process, assessing model performance from the perspectives of the client, supervisor, and counselor, and introduces a feedback loop to iteratively optimize the performance.

Conclusion

We present Ψ-ARENA, a framework for evaluating and optimizing LLM-based psychological counselors. By combining dynamic, real-world simulations with a tripartite evaluation from clients, supervisors, and counselors, Ψ-ARENA addresses key limitations of existing evaluation methods. The framework's closed-loop optimization improves LLM performance by up to 141%, demonstrating the effectiveness of feedback-driven enhancement. Our experiments with eight leading LLMs reveal significant performance disparities, emphasizing the need for multi-perspective evaluation. This work establishes Ψ-ARENA as a foundation for advancing reliable and scalable LLM applications in mental healthcare.

Ethical Statement

We here elaborate on the potential ethical issues.

Data Privacy and Confidentiality One of the key ethical concerns in using Ψ-ARENA lies in ensuring the privacy and confidentiality of virtual client data. Although the client profiles are based on de-identified real-world data and constructed to simulate typical psychological issues, the complexity of managing sensitive psychological data poses significant challenges. We aim to implement stringent safeguards for data usage, but constraints related to data storage, encryption, and usage policies may limit the full implementation of desired privacy protections at this stage.

Cultural Sensitivity and Bias Ψ-ARENA aims to simulate a broad range of psychological profiles to account for diverse cultural backgrounds and personal issues. However, due to the limitations in current LLM capabilities and available datasets, it is challenging to ensure perfect cultural sensitivity. While we strive for inclusivity in our simulations, we acknowledge that the diversity of profiles may not fully represent all cultural nuances, and unintended biases might arise in the counseling interactions.

Model Accountability and Responsibility As Ψ-ARENA operates by evaluating LLM-based counselors, it is important to consider the accountability for actions taken by these models. In case of harmful interactions or incorrect advice, responsibility may be unclear, especially as these models do not possess human understanding or judgment. Despite our best efforts to design the system to avoid such outcomes, the inherent limitations of AI systems in handling complex emotional and ethical situations raise concerns about responsibility and accountability.

Dependency and Human Interaction While Ψ-ARENA can optimize LLM counselors' performances, we acknowledge the limitations of relying on AI for psychological support. Despite the potential advantages of AI-based systems, human interaction remains irreplaceable for the most effective psychological care. While we aim to improve the model's ability to simulate human-like responses, we are constrained by the inability of AI to fully replicate the empathy, intuition, and ethical judgment of human therapists. This highlights the need for AI to function as a supplementary tool rather than a replacement for trained mental health professionals.

Acknowledgements

This research was supported by the Open Project of Xiangjiang Laboratory [No.25XJ03020], National Natural Science Foundation of China [62172449,72374070], Hunan Provincial Natural Science Foundation of China [2022JJ3021, 2025JJ20071].

This work was supported in part by the High Performance Computing Center of Central South University.

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*.

Anthropic. 2023. Claude: A Family of Language Models. Accessed: 2025-02-15.

Anthropic. 2025. Claude 3.5 Sonnet Announcement. Accessed: 2025-02-16.

APA. 2023. American Psychological Association Website. Accessed: 2025-02-16.

Black, J. 2003. *Who stole your trust and confidence - a psychiatrist clinic diary(Chinese Edition)*. Shantou University Press.

Chen, J.; Wang, X.; Xu, R.; Yuan, S.; Zhang, Y.; Shi, W.; Xie, J.; Li, S.; Yang, R.; Zhu, T.; et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231*.

Chen, S.; Wu, M.; Zhu, K. Q.; Lan, K.; Zhang, Z.; and Cui, L. 2023. LLM-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Corey, G. 2013. *Theory and practice of counseling and psychotherapy*. Cengage learning.

Hill, C. E. 2020. *Helping skills: Facilitating exploration, insight, and action*. American Psychological Association.

Iftikhar, Z.; Ransom, S.; Xiao, A.; and Huang, J. 2024. Therapy as an NLP Task: Psychologists' Comparison of LLMs and Human Peers in CBT. *arXiv preprint arXiv:2409.02244*.

Ji, S.; Zhang, T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Jin, H.; Chen, S.; Wu, M.; and Zhu, K. Q. 2023. PsyEval: A Comprehensive Large Language Model Evaluation Benchmark for Mental Health. *arXiv preprint arXiv:2311.09189*.

Kang, D.; Kim, S.; Kwon, T.; Moon, S.; Cho, H.; Yu, Y.; Lee, D.; and Yeo, J. 2024. Can Large Language Models be Good Emotional Supporter? Mitigating Preference Bias on Emotional Support Conversation. *arXiv preprint arXiv:2402.13211*.

Kuzmits, F. E.; Adams, A. J.; Sussman, L.; and Raho, L. E. 2004. 360-feedback in health care management: a field study. *The Health Care Manager*, 23(4): 321–328.

Lan, K.; Jin, B.; Zhu, Z.; Chen, S.; Zhang, S.; Zhu, K. Q.; and Wu, M. 2024. Depression Diagnosis Dialogue Simulation: Self-improving Psychiatrist with Tertiary Memory. *arXiv preprint arXiv:2409.15084*.

Li, A.; Gong, B.; Yang, B.; Shan, B.; Liu, C.; Zhu, C.; Zhang, C.; Guo, C.; Chen, D.; Li, D.; et al. 2025. Minimax-01: Scaling foundation models with lightning attention. *arXiv preprint arXiv:2501.08313*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lockyer, J. 2003. Multisource feedback in the assessment of physician competencies. *Journal of Continuing education in the Health Professions*, 23(1): 4–12.

Meta. 2023. LLaMA 3.3 Model Card and Prompt Formats. Accessed: 2025-02-16.

OpenAI. 2023a. GPT-3.5 Turbo Documentation. Accessed: 2025-02-16.

OpenAI. 2023b. GPT-4 Technical Report. Accessed: 2025-02-16.

Sachse, R. 2024. *Relationship Building*, 83–87. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-662-69780-1.

Shao, Y.; Li, L.; Dai, J.; and Qiu, X. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

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. *ArXiv*, abs/2106.01702.

Tham, K.-Y. 2007. 360 feedback for emergency physicians in Singapore. *Emergency Medicine Journal*, 24(8): 574–575.

Tu, Q.; Fan, S.; Tian, Z.; and Yan, R. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Vajre, V.; Naylor, M.; Kamath, U.; and Shehu, A. 2021. PsychBERT: a mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1077–1082. IEEE.

Wang, C.; Liao, M.; Huang, Z.; Wu, J.; Zong, C.; and Zhang, J. 2024a. BLSP-Emo: Towards Empathetic Large Speech-Language Models. *arXiv preprint arXiv:2406.03872*.

Wang, J.; Xiao, Y.; Li, Y.; Song, C.; Xu, C.; Tan, C.; and Li, W. 2024b. Towards a Client-Centered Assessment of LLM Therapists by Client Simulation. *arXiv preprint arXiv:2406.12266*.

Weizenbaum, J. 1966. ELIZA–A Computer Program For the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1): 36–45.

Wenhua, Y. 2020. How to understand the interpersonal relationship patterns of clients? *Popular Psychology*, (07): 4–6.

WHO. 2023. Mental health: strengthening our response. Technical report, World Health Organization.

Xie, Q.; and Peng, W. 2025. MAGO: Multi-Knowledge Aware and Global Strategy Sequence Optimizing Network for Emotional Support Conversation. *Neurocomputing*, 618: 128888.

Xu, A.; Yang, D.; Li, R.; Zhu, J.; Tan, M.; Yang, M.; Qiu, W.; Ma, M.; Wu, H.; Li, B.; et al. 2025. AutoCBT: An Autonomous Multi-agent Framework for Cognitive Behavioral Therapy in Psychological Counseling. *arXiv preprint arXiv:2501.09426*.

Xu, X.; Yao, B.; Dong, Y.; Gabriel, S.; Yu, H.; Hendler, J.; Ghassemi, M.; Dey, A. K.; and Wang, D. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1): 1–32.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang, W.; and Xiong, L. 2018. Development and Psychometric Evaluation of the Self-Assessment Scale for Psychological Counseling Competence. In *Proceedings of the 21st National Conference on Psychology*, 727–728. Department of Psychology, Hunan Normal University.

Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; and Hu, X. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.

Zhang, M.; Yang, X.; Zhang, X.; Labrum, T.; Chiu, J. C.; Eack, S. M.; Fang, F.; Wang, W. Y.; and Chen, Z. Z. 2024b. CBT-Bench: Evaluating Large Language Models on Assisting Cognitive Behavior Therapy. *arXiv preprint arXiv:2410.13218*.

Zhang, T.; Schoene, A. M.; and Ananiadou, S. 2021. Automatic identification of suicide notes with a transformer-based deep learning model. *Internet interventions*, 25: 100422.

Zhao, H.; Li, L.; Chen, S.; Kong, S.; Wang, J.; Huang, K.; Gu, T.; Wang, Y.; Jian, W.; Liang, D.; et al. 2024. ESC-Eval: Evaluating Emotion Support Conversations in Large Language Models. *arXiv preprint arXiv:2406.14952*.

Zhipu. 2023. GLM-4 Usage Guide. Accessed: 2025-02-16.