MentalGuide: Towards Multi-Turn, State-Aware and Strategy-Driven Conversations for Mental Health Support

Jinwei He, Feng Lu*

State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University
{hejinwei, lufeng}@buaa.edu.cn

Abstract

The global shortage of psychiatrists has become a critical issue, and the advent of large language models (LLMs) presents new opportunities to address this challenge. However, existing approaches continue to underperform in multi-turn mental health counseling, particularly in the arrangement of counseling strategies. To overcome these limitations, we propose MentalGuide, a state-aware and strategy-driven conversation framework designed for multi-turn mental health support. Our method integrates expert-derived prior probabilities of counseling strategies tailored to the target client's state with the reasoning capabilities of LLMs. This enables effective strategy formulation and strategy-driven response generation, without the need for additional training. Experimental results show that MentalGuide surpasses baselines in automated and human expert evaluations, demonstrating the closest alignment with real-world multi-turn counseling dynamics.

Introduction

In recent years, the global prevalence of mental health disorders has continued to rise, while the number of qualified clinicians remains critically insufficient, presenting unprecedented challenges (McGorry et al. 2024; Pinchuk et al. 2024). The advent of large language models (LLMs) offers new opportunities for AI mental health care.

Current approaches fall into two categories: direct application of LLMs and fine-tuned models (Liu et al. 2023; Lai et al. 2023; Qiu et al. 2024). The former typically provides immediate advice but lacks support for multi-turn conversations. The latter improves the therapeutic tone, yet fails to organize comfort, advice, and analysis across conversational rounds. These limitations stem from two issues (Sun et al. 2021; Li et al. 2023): (1) a disparity between single-turn Q&A and multi-turn psychotherapy—whereas the latter is guidance-oriented and demands careful pacing; and (2) LLMs lack the knowledge of real-world therapeutic practice, especially the counseling strategies.

To address the above issues, we annotate client states and counseling strategies (Hill et al. 1992; Ribeiro et al. 2013; Hill 1999; Chamberlain et al. 1984) from real-world dialogues, and find while the specific content is variable, the progression of strategies aligns with psychological principles: (1) specific client state often corresponds to specific strategies usage; (2) therapists need to analyze the situation on a case-by-case basis, ensuring the strategies remain flexible; (3) therapists need to integrate both approaches to determine the most suitable strategy. We propose the idea of introducing the expert statistics of real-world counseling and combining them with real-time inferences from LLMs.

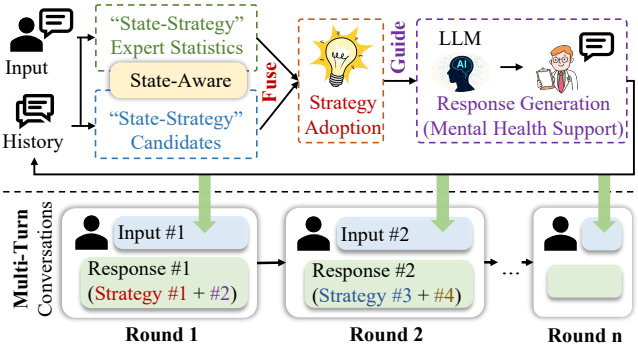


Figure 1: MentalGuide determines the counseling strategies for each round by integrating (1) prior strategy distributions from expert and (2) LLM-based strategy ranking, which ensures clinically appropriate and adaptive strategies.

As shown in Figure 1, to achieve our idea, we propose **MentalGuide**, which is a MULTI-TURN, STATE-AWARE AND STRATEGY-DRIVEN CONVERSATION FRAMEWORK for mental health support, consisting of four key modules: (1) **State-Aware Strategy Mapping**: Using real expert data, we derive statistical patterns between client states and strategies. LLM assesses the client state and computes prior probabilities for strategy selection. (2) **Adaptive Strategy Ranking**: Based on the ongoing conversation, LLM dynamically ranks the top-3 most appropriate strategies. (3) **Integrated Decision Fusion**: The prior probabilities and LLM ranking are combined to determine the best strategies. (4) **Strategy-Driven Response Generation**: The determined strategies guide the LLM in generating clinically aligned responses.

Experimental results demonstrate that our approach is closest to the real-world strategies usage in multi-turn mental health counseling, achieving state-of-the-art performance in both automatic and expert evaluations.

*Corresponding author

Overall, our contributions are as follows:

1. **New Paradigm**: To the best of our knowledge, we are the first to focus on the reasoning for states and strategies in multi-turn mental health conversations with LLMs.
2. **Novel Method**: We propose MentalGuide, which combines expert prior statistics with LLMs' reasoning, enabling principled and adaptive strategy determination.
3. **Effective Performance**: Our approach most accurately approximates real-world strategy usage and achieves superior results in automatic and human expert evaluations.

Preliminary Research

Clients' States and Counselors' Strategies

States	Definitions
Statement (Sta.)	Client provides information based on the counselor's specific request.
Query (Que.)	Client seeks clarification, understanding, information, advice, or opinions from the counselor.
Discussion (Discu.)	Client agrees with the counselor's intervention and engages in discussion.
Revolt (Rev.)	Client expresses confusion, or insists on their own perspective.
Self-denial (Se-d.)	Client falls into self-criticism or a significantly low emotional state.
Disconnect (Disco.)	Instead of addressing the counselor's intervention, the client shifts to another topic or focuses on their own concerns.
Common Talk (C&T)	Conversation without clear characteristics of the above categories.

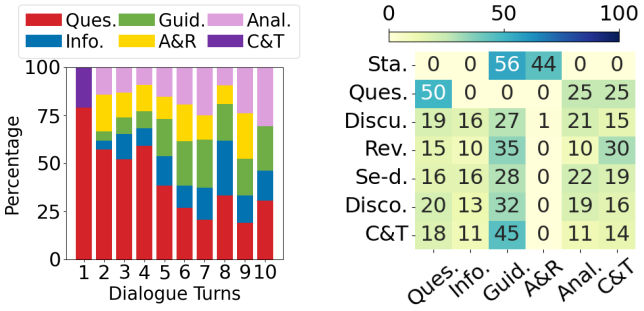
Table 1: Definitions of our chosen clients' states.

Extensive research on counselor-client interactions has demonstrated that therapists dynamically adjust strategies based on clients' evolving states (Sun et al. 2021; Li et al. 2023). This evidence underscores the critical relationship between client states and therapeutic strategy selection.

The definition of client states and counseling strategies is a fundamental part of our study. Through collaboration with licensed clinical psychologists, we conduct a systematic review and synthesis of existing classification frameworks, including the Client Behavior System (Hill et al. 1992), the Therapeutic Collaboration Coding Scheme (Ribeiro et al. 2013), the Helping Skills Framework (Hill 1999), and the Client Resistance Coding System (Chamberlain et al. 1984). Summarizing these theoretical foundations, we develop a new taxonomy consisting of: **seven client states**, including Statement, Query, Discussion, Revolt, Self-denial, Disconnect, and Common Talk, and **six counseling strategies**, including Open Question, Provide Information, Direct Guidance, Approval and Reassurance, Analysis, and Common Talk. Tables 1 and 2 present the definitions of seven client states and six counseling strategies.

Strategies	Definitions
Ask Question (Ques.)	Elicit key information through heuristic questioning (overview of the event, emotional detail, etc).
Provide Information (Info.)	Deliver effective information by offering factual evidence, professional insights, etc.
Direct Guidance (Guid.)	Provide clear behavioral suggestions and actionable solutions to guide the client toward practical steps.
Approval and Reassurance (A&R)	Use empathy techniques to offer emotional comfort and enhance the client's psychological strength.
Analysis (Anal.)	Help the clients understand themselves through multidimensional analysis.
Common Talk (C&T)	Conversation without clear characteristics of the above categories.

Table 2: Definitions of our chosen counselors' strategies.



(a) Strategy distribution across the first 10 rounds of the mental health conversation. Different colors indicate different counseling strategies. The X-axis shows conversation rounds, and the y-axis displays the percentage of each strategy.

(b) Client state-counselor strategy correspondence matrix. Columns represent the strategies, rows indicate the client states, and cell values show the strategy distribution percentage. Color intensity reflects association magnitude.

Figure 2: Statistics of multi-turn counseling.

Statistics of Real-World Mental Counseling

We invite licensed clinical psychologists to annotate a collection of real counseling conversations, categorizing client states and counselor strategies for each conversation round. To better explore the strategy patterns, we analyze the first 10 rounds. As shown in Figure 2a, the counselor strategies change dynamically during the process of counseling, for instance, questions decrease while suggestions increase.

Through heatmap visualization (Figure 2b), we quantitatively characterize the state-strategy correspondence. The

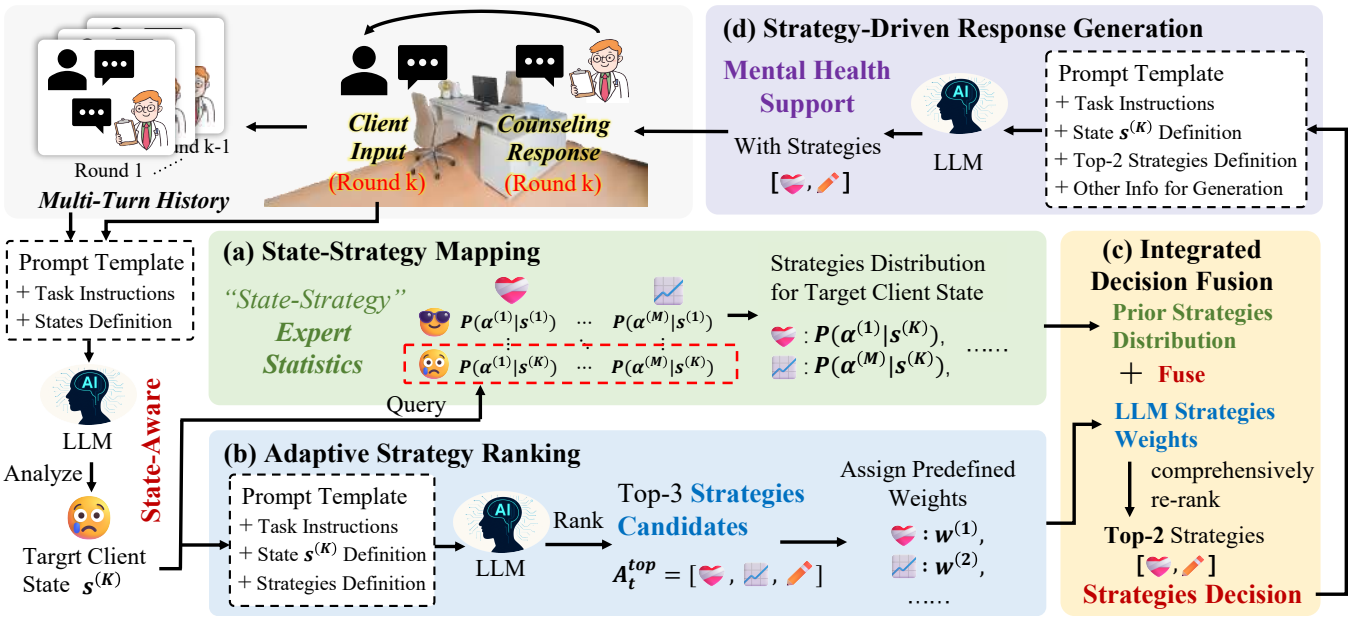


Figure 3: Our MentalGuide framework adopts a four-stage strategy determination approach: (a) State-Aware Strategy Mapping, (b) Adaptive Strategy Ranking, (c) Integrated Decision Fusion and (d) Strategy-Driven Response Generation. Our method unifies the strengths of LLMs' flexible reasoning capabilities with real-world expert prior knowledge.

distribution demonstrates statistically significant associations ($p < 0.01$) between specific client states and preferred counselor strategies. These empirical results confirm that professional counselors adaptively modulate their strategic approach in response to clients' evolving states.

MentalGuide

Overall Structure

We propose the MentalGuide framework, as shown in Figure 3. It is based on the strategic patterns observed in real-world counseling. The core idea of this approach is to simultaneously incorporate prior strategy distributions from experts and the reasoning of LLMs to determine optimal strategies: (a) **State-Aware Strategy Mapping**: LLMs assess client states and compute prior strategy probabilities using statistical patterns from actual expert counseling sessions. (b) **Adaptive Strategy Ranking**: LLMs dynamically rank and select the top-3 most appropriate strategies based on conversation history. (c) **Integrated Decision Fusion**: The prior probabilities and LLMs rankings are combined to determine the optimal strategies. (d) **Strategy-Driven Response Generation**: The determined strategies guide LLMs in generating mental health counseling responses.

State-Aware Strategy Mapping

Our statistical analysis of real-world expert counseling conversations yields the state-strategy mapping, whose implementation results are presented in Figure 2b. Let $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^{N}$ denote the counseling dataset where $s_i \in \mathcal{S}$ represents client states and $a_i \in \mathcal{A}$ represents applied strategies. The prior probability distribution is given by:

$$P(a \mid s; \mathcal{D}) = \frac{\text{count}_{\mathcal{D}}(a, s)}{\sum_{a' \in \mathcal{A}} \left(\text{count}_{\mathcal{D}}(a', s)\right)} \tag{1}$$

This constructs a state-strategy probability matrix:

$$\mathbf{M}_{\mathcal{D}} = \begin{bmatrix} P(a_1 \mid s_1; \mathcal{D}) & \cdots & P(a_M \mid s_1; \mathcal{D}) \\ \vdots & \ddots & \vdots \\ P(a_1 \mid s_K; \mathcal{D}) & \cdots & P(a_M \mid s_K; \mathcal{D}) \end{bmatrix} \tag{2}$$

Let H_t represents the conversation history until round t. θ_{LLM} is LLMs parameters. $\mathcal{P}$ is LLMs prompt. Client state is classified using LLMs: $s_t = \arg\max_{s \in \mathcal{S}} P(s \mid H_t, \mathcal{P}; \theta_{\text{LLM}})$.

Let $\pi_t(a_j) = P(a_j \mid s_t; \mathcal{D})$ represents the prior probability of strategy a_j given state s_t. The strategy distribution for s_t is retrieved from the precomputed probability matrix:

$$\boldsymbol{\pi}_t = \mathbf{M}_{\mathcal{D}}[s_t, :] = \left[\pi_t(a_1), \pi_t(a_2), \ldots, \pi_t(a_M)\right] \tag{3}$$

Adaptive Strategy Ranking

Relying solely on prior probabilities cannot flexibly respond to actual situations, so we introduce the real-time inference of LLMs. We use LLMs to rank the top-3 most applicable strategies based on the client state and conversation history: $\mathcal{A}_t^{\text{top}} = \text{LLM}_{\text{rank}}\left(s_t, H_t, \mathcal{P}_{\text{rank}}; \theta_{\text{LLM}}\right)$, where $\mathcal{A}_t^{\text{top}} = [a^{(1)}, a^{(2)}, a^{(3)}]$ represents the set of top-3 strategies. $\mathcal{P}_{\text{rank}}$ represents LLMs prompt. Then we assign predefined weights $w^{(1)}, w^{(2)}, w^{(3)}$ to $a^{(1)}, a^{(2)}, a^{(3)}$:

$$\phi_t(a) = \begin{cases} w^{(i)} & \text{if } a = a^{(i)}, i = 1, 2, 3 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\phi_t(a)$ represents weight for strategy a.

Method	Correctness (%) (↑)	Single-turn Evaluation (↑)				Multi-turn Evaluation (↑)	
		Emotional.	Cognitive.	State.	Safety.	Naturalness.	Strategic.
SoulChat	23.58	1.46	1.28	1.65	1.00	2.24	1.27
MeChat	32.87	1.12	1.09	1.54	1.00	2.39	1.37
RolePlay (Deepseek-V3)	33.33	2.31	2.42	2.47	1.00	2.44	1.55
RolePlay (Qwen-3-Plus)	36.13	2.25	2.37	2.33	1.00	2.58	1.87
Ours PK-Free (Deepseek-V3)	38.68	2.34	2.42	2.46	1.00	2.76	2.35
Ours PK-Free (Qwen-3-Plus)	41.15	2.42	2.41	2.42	1.00	2.75	2.51
Ours (Deepseek-V3)	**52.45**	**2.52**	**2.50**	**2.58**	**1.00**	**2.88**	**2.52**
Ours (Qwen-3-Plus)	**48.25**	**2.58**	**2.60**	**2.64**	**1.00**	**2.76**	**2.69**

Table 3: Experimental results for automatic evaluation, including correctness of strategies, evaluation for single-turn responses and multi-turn conversations. Ours PK-Free means our method without the prior knowledge which is an ablation approach.

Integrated Decision Fusion

We integrate the prior probabilities $\pi_t(a)$ and LLMs ranking weights $\phi_t(a)$ through direct linear combination: $\text{Score}_t(a) = \pi_t(a) * 100\% + \phi_t(a)$. In order to reflect real-world counseling practices where more than one strategies are applied in each round, we select the k highest-scoring strategies as the final decision, we set $k = 2$:

$$\{a_1^*, a_2^*\} = \arg\operatorname*{top2}_{a \in \mathcal{A}} \text{Score}_t(a) \tag{5}$$

Strategy-Driven Response Generation

We employ LLMs to generate responses r_t based on the determined counseling strategies $\{a_1^*, a_2^*\}$. $\mathcal{P}_{\text{gen}}$ represents LLMs prompts. We incorporate detailed explanations and examples of the specific strategy, along with general requirements in it. Then, r_t is presented to the client, history H_t is updated and a new round of conversation starts.

$$r_t = \text{LLM}_{\text{gen}}\big(\{a_1^*, a_2^*\}, s_t, H_t, \mathcal{P}_{\text{gen}}; \theta_{\text{LLM}}\big) \tag{6}$$

Experiments and Discussions

Settings and Baselines

Given the Chinese-language origin of the data we used, we employ two state-of-the-art Chinese-optimized LLMs as foundation LLMs: Deepseek-V3 (Liu et al. 2024) and Qwen-3-Plus (Yang et al. 2025), both accessed through their official APIs. This selection excludes specialized reasoning LLMs (e.g., Deepseek-R1 (Guo et al. 2025)) due to their prohibitive latency - generation delays frequently exceeding 60 seconds, which violates the real-time requirements of mental health conversations. Our framework maintains an average 5-second response latency per turn, well within the acceptable threshold for conversational flow.

For comparative methods, we select: (1) Fine-tuned baselines: SoulChat (Chen et al. 2023) and MeChat (Qiu et al. 2023); (2) The RolePlay method (Shanahan, McDonell, and Reynolds 2023; Tseng et al. 2024), a established approach for domain-specialized LLM applications, which shares our foundation models; (3) An ablated variant of our method (Ours-PKF, where "PKF" denotes Prior Knowledge-Free) that eliminates expert-derived priors and relies exclusively on LLM reasoning. Regarding test data, to facilitate comparison with real-world conditions, we collected real-world counseling dialogues (total of 176 dialogues, avg. 15 turns, covering 10+ counseling topics) exclusively for experimental evaluation, independent of method development. [1]

Automatic Evaluation

Correctness of Strategies. Using real-world multi-turn mental health counseling conversation data as ground truth, we use each method to generate new responses based on existing history for each round of the conversations. The generated responses are then annotated with the defined strategies using LLMs. Each response is labeled with one or two strategies. Then we compare the strategies used by each method with the real strategies and calculate the accuracy. We apply a stringent criterion: a method's response strategies are considered correct only when all strategies used in the real data are included in the method's response.

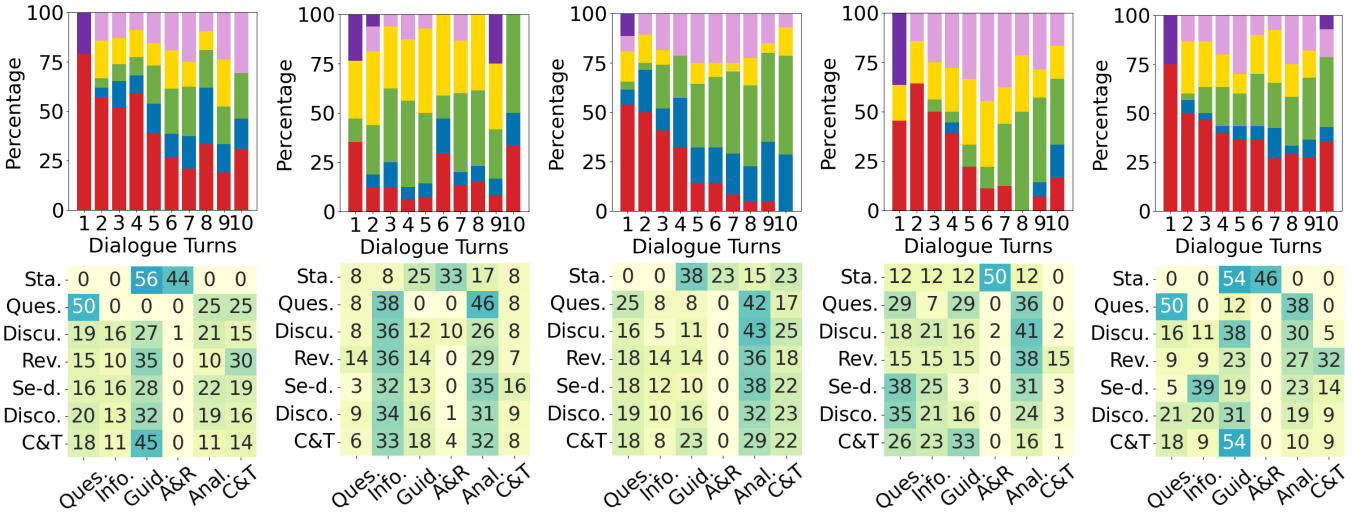
As shown in the second column of Table 3, our method achieves the highest accuracy (52.45% and 48.25%) in aligning with the real psychological counseling strategies. This indicates that MentalGuide more accurately replicates real therapeutic decision-making patterns.

Evaluation for Single-turn Response. We evaluate single-turn response quality using the same dataset from the previous section. Following established metrics from previous study (Chen et al. 2023; Xie et al. 2024), we assess four dimensions: (1) Emotional Empathy, (2) Cognitive Empathy, (3) State and Attitude (all 0-3 scales, the higher the better), and (4) Safety (binary, 1 indicating safety).

We use three different LLMs (including Deepseek-V3, Deepseek-R1 and Qwen-3-Plus) as evaluators to score each method's performance on above metrics respectively.

Table 3 shows the average scores. All methods achieve safe responses (score=1). However, concerning the other three metrics, fine-tuned LLMs show limited capability

[1] In response to the reviewers' comments, the relevant content will be made publicly available in an appropriate format after undergoing necessary ethical and privacy reviews. Please follow our subsequent updates in https://phi-ai.buaa.edu.cn.



(a) Human Expert Data. (b) SoulChat. (c) RolePlay. (d) Ours PK-Free. (e) **Ours.**

Figure 4: Statistical results for the usage of counseling strategies in each method.

Metrics	SoulChat	RolePlay	Ours PK-Free	**Ours**	Ours $\mathcal{S}_A$	Ours $\mathcal{S}_B$	Ours $\mathcal{S}_C$	Ours $\mathcal{S}_D$
$\mathcal{M}_r$ (↓)	14.00	8.81	8.46	**6.02**	8.29	9.23	10.91	7.45
$\mathcal{M}_s$ (↓)	13.80	9.72	11.62	**5.11**	7.90	9.90	10.48	7.80

Table 4: Quantitative evaluation for the statistical results of the usage of counseling strategies in Figure 4 and Figure 5.

(scores 1.0-1.7) potentially due to smaller model sizes. RolePlay and Ours PK-Free perform better (scores 2.2-2.5), but still need improvement, proving that relying solely on LLMs is not enough for the task. Our method achieves the highest scores across all evaluation metrics (scores 2.5-2.7), demonstrating that although designed for multi-turn conversations, it remains fully compatible with previously proposed evaluation criteria for single-turn responses.

Evaluation for Multi-turn Conversation. In this section, we evaluate the quality of the complete multi-turn mental health counseling conversations generated by each method. To enable automated evaluation, we use Deepseek-V3 (Liu et al. 2024) to simulate client interactions during counseling sessions with each tested approach (Han et al. 2024).

We introduce two evaluation metrics: Naturalness of Conversation and Strategic Execution Capability. Naturalness of Conversation assesses the overall fluency and coherence of the counseling process. Strategic Execution Capability evaluates the model's ability to effectively apply diverse counseling strategies throughout multi-turn mental health support conversations. Using the same LLM evaluator panel from last experiment, we automatically score each method's performance on these metrics (0-3 scale).

The experimental results, as shown in the rightmost columns in Table 3, are the average scores. Our method achieves the best performance across both metrics (Naturalness: 2.88 and 2.76; Strategic: 2.52 and 2.69), especially the strategic execution capability. These findings validate our framework's effectiveness in maintaining high-quality, strategy-aware counseling conversations.

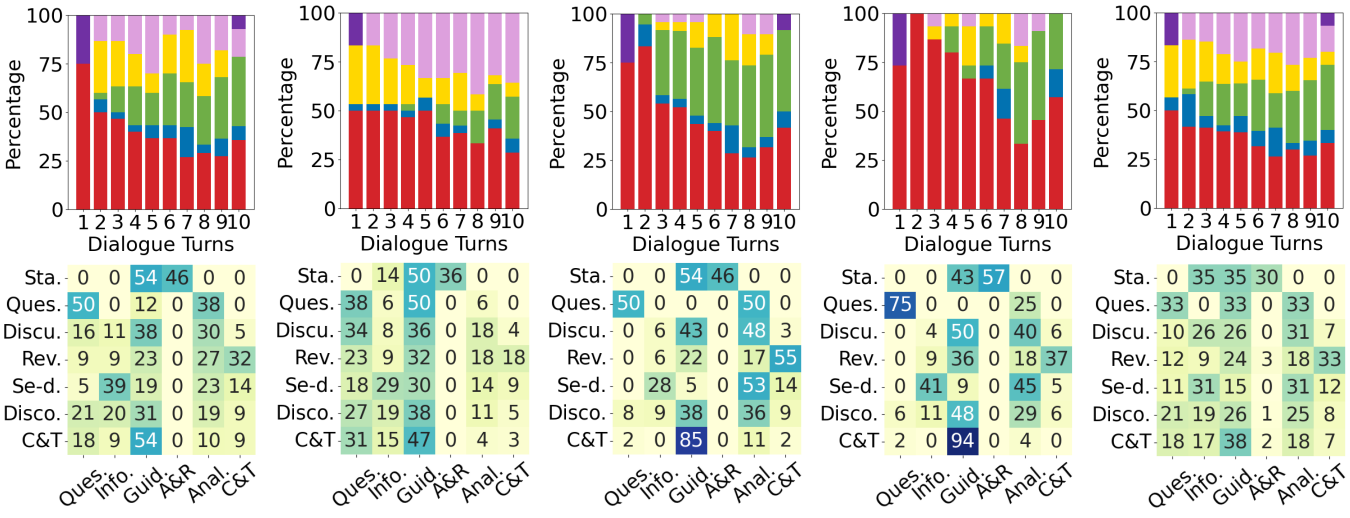
Statistical Results

Settings. Based on real-world multi-turn mental health counseling conversation, we employ each method to generate new responses. Subsequently, we annotate client states and response strategies, comparing them against ground truth to study their statistical patterns. We use two types of statistical graphs in Figure 4, the definitions of which are the same as in Figure 2. Due to space limitations, we only show the results of SoulChat, RolePlay, Ours PK-Free, and Ours, the last three are based on Deepseek-V3 (Liu et al. 2024). To quantitatively evaluate the results, we define $\mathcal{M}_r$ and $\mathcal{M}_s$. $\mathcal{M}_r$ evaluates the first-row graph by calculating the difference between the strategy distribution and ground truth in each round, while $\mathcal{M}_s$ evaluates the second-row graph by computing the strategic difference per client state:

Given the true strategy distribution $\{p_i\}$ and model's strategy distribution $\{q_i\}$, r denotes conversation round, s denotes client state, i denotes strategy type and n is the number of strategy types, m is the number of client states:

$$\mathcal{M}_r = \frac{1}{10n} \sum_{r=1}^{10} \sum_{i=1}^{n} \left| p_i^{(r)} - q_i^{(r)} \right| \tag{7}$$

$$\mathcal{M}_s = \frac{1}{mn} \sum_{s=1}^{m} \sum_{i=1}^{n} \left| p_i^{(s)} - q_i^{(s)} \right| \tag{8}$$



(a) Best setting: ranking weights $w^{(1)}$, $w^{(2)}$, $w^{(3)}$ = **72, 64, 36** and final strategies number k =**2**

(b) Setting A ($\mathcal{S}_A$): weights $w^{(1)}$, $w^{(2)}$, $w^{(3)}$ = **300, 200, 100** and strategies number k = **2**

(c) Setting B ($\mathcal{S}_B$): weights $w^{(1)}$, $w^{(2)}$, $w^{(3)}$ = **30, 20, 10** and strategies number k = **2**.

(d) Setting C ($\mathcal{S}_C$): weights $w^{(1)}$, $w^{(2)}$, $w^{(3)}$ = **72, 64, 36** and strategies number k = **1**.

(e) Setting D ($\mathcal{S}_D$): weights $w^{(1)}$, $w^{(2)}$, $w^{(3)}$ = **72, 64, 36** and strategies number k = **3**.

Figure 5: Statistical results for the usage of counseling strategies under different parameters settings in our method: (1) top three LLM-ranked strategies weights $w^{(1)}$, $w^{(2)}$, $w^{(3)}$, and (2) the number of final strategies k.

Overall Results. The qualitative results are shown in Figure 4. Our MentalGuide demonstrates the closest alignment with the ground truth in terms of strategy distribution (top-row graphs) across the first 10 rounds, with statistically significant effectiveness. In contrast, SoulChat exhibits a pronounced bias toward employing comforting and advising strategies, while underutilizing questioning and analytical strategies. RolePlay method shows partial variability consistent with real-world dynamics. However, it over-relies on guidance strategies. Ours PK-Free method shows a significant increase in the use of analytical strategies, while questioning and suggestion strategies exhibit more extreme variations. These differences indicate that LLMs cannot directly make accurate strategy determinations.

From the second row of Figure 4, it can be observed that all comparative methods exhibit significant deviations from the ground truth when responding to client states, while only our method shows a notably closer alignment.

As shown in Table 4, our method achieves the best quantitative results on both metrics (6.02 for $\mathcal{M}_r$ and 5.11 for $\mathcal{M}_s$), which is most closely aligned with the strategy distribution of real-world counseling approaches. RolePlay and Ours PK-Free methods demonstrate suboptimal performance for $\mathcal{M}_s$ when addressing specific client states, while SoulChat performs poorly on both metrics. All above results strongly suggest that combining the prior probabilities of multi-turn mental health counseling with LLMs reasoning can better determine the suitable strategies.

Discussion for Weight Settings. Our framework incorporates two preset parameters: (1) weights assigned to the top three LLM-generated strategies in Adaptive Strategy Ranking $w^{(1)}$, $w^{(2)}$, $w^{(3)}$, and (2) the number of final strategies k in Integrated Decision Fusion. Their values may influence method performance, which we discuss in this section.

As shown in Figures 5a, our final setting uses $w^{(1)} = 72$, $w^{(2)} = 64$, $w^{(3)} = 36$, with the number of finalized strategies $k = 2$. Figures 5b and 5c demonstrate that excessive strategy weights lead to over-reliance on LLMs strategies ranking, resulting in averaged outcomes. Conversely, insufficient weights amplify the influence of prior probabilities, causing the method to exclusively select the top two high-prior strategies and produce polarized results. Similarly, Figures 5d and 5e reveal that when limited to a single strategy, questioning is typically prioritized, leading to its predominant selection and consequently extreme strategy distribution. In contrast, selecting three strategies reduces strategic specificity, producing more averaged outcomes.

These findings indicate that parameter selection could affect the performance of MentalGuide. Extreme configurations induce extreme preferences, potentially leading to polarized results—an observation that aligns with intuitive expectations. As shown in Table 4, the more balanced results ($\mathcal{S}_A$ and $\mathcal{S}_D$) still exhibit smaller strategy deviations than other comparative methods. While the more extreme results ($\mathcal{S}_B$ and $\mathcal{S}_C$) perform slightly worse than RolePlay method, they remain comparable to Ours PK-Free and outperform SoulChat on the metircs. This demonstrates that our method maintains competitive performance even under extreme parameter settings, highlighting its robustness.

Human Expert Evaluation

In addition to automatic evaluation, we conduct subjective evaluation by human experts. We compare three approaches: (1) SoulChat (Chen et al. 2023) (fine-tuning-

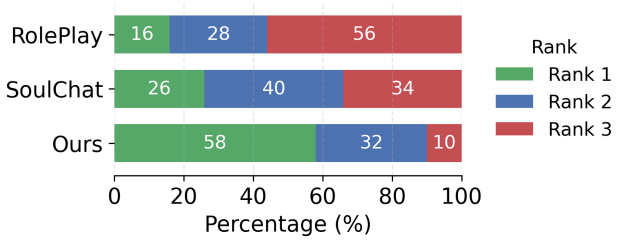


Figure 6: Subjective evaluation by human experts. Different colors represent different rankings, and the figure shows the proportion of different rankings obtained by three methods.

based), (2) RolePlay (Shanahan, McDonell, and Reynolds 2023) (prompt-based), and (3) our proposed MentalGuide. For evaluation, we select 10 counseling scenarios covering different topics from the data in the multi-turn conversation evaluation section. Each scenario includes conversations generated by the three methods, resulting in a total of 30 multi-turn mental health dialogues. Ten licensed clinical psychologists conducted blind evaluations of the conversations using established therapeutic competency criteria. Each independently ranked the performance of the methods (1-3 scale, where 1=best), with no knowledge of the generating approach. The results are shown in Figure 6.

Our method significantly outperforms the other two approaches, achieving the highest ranking (Rank 1) in 58% of cases, with only 10% for Rank 3, and the inter-rater agreement remains high with Kendall's W = 0.91. This demonstrates that our approach has been recognized by professional psychological counselors for its effectiveness, highlighting its practical potential. Moreover, it is worth noting that in expert evaluation and multi-turn conversation evaluation, SoulChat outperforms RolePlay, whereas the opposite holds true in single-turn evaluation. This indicates that SoulChat and RolePlay methods have their own strengths in different situations, while our approach consistently achieves superior results.

Related Works

Large Language Models In recent years, the groundbreaking development of large language models (LLMs) has marked a significant leap forward in the field of artificial intelligence (Vaswani et al. 2017; Touvron et al. 2023; Achiam et al. 2023; Zeng et al. 2022; Jiang et al. 2023). Trained on massive datasets and enhanced by unprecedented parameter scaling, these models demonstrate remarkable capabilities in language comprehension and generation. They not only process complex textual interactions with fluency but also accomplish higher-order cognitive tasks such as logical reasoning and knowledge association (Wei et al. 2022; Wang et al. 2022; Yao et al. 2024b; Sun et al. 2023). With ongoing technological iterations, LLMs are rapidly permeating vertical domains such as finance (Zhang and Yang 2023), healthcare (Zhang et al. 2023; Xiong et al. 2023), law (Cui et al. 2023; Yao et al. 2024a), etc. Key enabling technologies include domain-specific fine-tuning (Chen et al. 2023; Liu et al. 2023), agents (Shen et al. 2024; Zhang et al. 2024b), and prompt engineering (Besta et al. 2024; Yao et al. 2024b).

Mental Health Support with LLMs While LLMs have strong linguistic competence that suggests promise for mental health applications, their effective deployment faces unique challenges. The implicit nature of mental health counseling expertise—requiring professional knowledge and skill application, creates significant adaptation barriers. A common approach to adapting LLMs for vertical domains involves role-playing methods (Shanahan, McDonell, and Reynolds 2023; Tseng et al. 2024; Han et al. 2024; Qiu et al. 2024), yet these fail to address LLMs' inherent lack of clinical experience. Consequently, model outputs often remain isolated, single-turn responses incapable of sustaining the guided, multi-turn conversations.

Currently, influential works applying LLMs to mental health counseling, such as SoulChat (Chen et al. 2023), MeChat (Qiu et al. 2023), CPsyCoun (Zhang et al. 2024a) primarily rely on generating large-scale synthetic datasets (either single-turn or multi-turn) to train or fine-tune LLMs (Sun et al. 2021; Xie et al. 2024; Lai et al. 2023). While they have achieved impressive results in enhancing empathetic tone and emotional resonance for smaller-scale LLMs, their multi-turn conversational performance often suffers from limited response diversity due to data and model constraints. Consequently, they struggle to strategically plan interactions with client across consecutive rounds.

Conclusion

For mental health counseling, we propose **MentalGuide**, a MULTI-TURN, STATE-AWARE AND STRATEGY-DRIVEN CONVERSATION FRAMEWORK, which integrates the prior expert knowledge of client states to counseling strategies with LLMs' real-time strategies ranking, balancing professionalism and flexibility. Our approach addresses the limitations of existing methods in sustaining multi-turn conversations and implementing therapeutic counseling strategies. Experiments show that our proposed MentalGuide achieves superior performance across both automatic and human expert evaluation. Our approach is closest to the strategic patterns of real mental health counseling, establishing a new paradigm for AI mental health support.

Although we establishes validation through automated and expert evaluations, several future directions remain: (1) collecting larger scale real-world counseling data to enhance generalizability, (2) conducting controlled trials with human participants to assess therapeutic outcomes under ethical supervision, and (3) a deeper investigation of cultural adaptation needs. These steps will further address current constraints in data diversity and human evaluations. Moreover, when this psychological dialogue technology becomes mature in the future, it can be applied to wearable devices to perceive human emotions in real time and provide support, thereby contributing to human-AI collaboration paradigms such as cobodied intelligence (Lu and Zhao 2026).

Ethical Statement

Data Private We collect real-world multi-turn mental health counseling conversations through both online and offline sources. To protect privacy, these original data will not be made publicly available. In our github repository, we only provide experimental results obtained from a limited set of publicly available online counseling data for reference purposes. These results undergo rigorous data cleaning and human verification processes to ensure the complete removal of any sensitive or private information.

Potential Risks of the Model During the expert evaluation phase, we limit the human assessment to the ranking of generated conversation content, with no direct human-model interaction involved in our experiments. We explicitly emphasize that while our study enhances strategy implementation for LLM-based multi-turn mental health counseling and strives to approximate real counseling scenarios, significant gaps still remain compared to professional human counselors. Importantly, we cannot guarantee complete avoidance of harm. The application of our work in actual psychotherapeutic practice is strictly prohibited.

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*.

Besta, M.; Blach, N.; Kubicek, A.; Gerstenberger, R.; Podstawski, M.; Gianinazzi, L.; Gajda, J.; Lehmann, T.; Niewiadomski, H.; Nyczyk, P.; et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17682–17690.

Chamberlain, P.; Patterson, G.; Reid, J.; Kavanagh, K.; and Forgatch, M. 1984. Observation of client resistance. *Behavior therapy*, 15(2): 144–155.

Chen, Y.; Xing, X.; Lin, J.; Zheng, H.; Wang, Z.; Liu, Q.; and Xu, X. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. *arXiv preprint arXiv:2311.00273*.

Cui, J.; Li, Z.; Yan, Y.; Chen, B.; and Yuan, L. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

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*.

Han, J.-E.; Koh, J.-S.; Seo, H.-T.; Chang, D.-S.; and Sohn, K.-A. 2024. PSYDIAL: Personality-based synthetic dialogue generation using large language models. *arXiv preprint arXiv:2404.00930*.

Hill, C. E. 1999. Helping skills: Facilitating exploration, insight, and action. *American Psychological Association*.

Hill, C. E.; Corbett, M. M.; Kanitz, B.; Rios, P.; Lightsey, R.; and Gomez, M. 1992. Client behavior in counseling and therapy sessions: Development of a pantheoretical measure. *Journal of Counseling psychology*, 39(4): 539.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Lai, T.; Shi, Y.; Du, Z.; Wu, J.; Fu, K.; Dou, Y.; and Wang, Z. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Li, A.; Ma, L.; Mei, Y.; He, H.; Zhang, S.; Qiu, H.; and Lan, Z. 2023. Understanding client reactions in online mental health counseling. *arXiv preprint arXiv:2306.15334*.

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*.

Liu, J. M.; Li, D.; Cao, H.; Ren, T.; Liao, Z.; and Wu, J. 2023. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Lu, F.; and Zhao, Q. 2026. Towards cobodied/symbodied AI: concept and eight scientific and technical problems. *SCIENCE CHINA Information Sciences*, 69(1): 116101:1–116101:3.

McGorry, P. D.; Mei, C.; Dalal, N.; Alvarez-Jimenez, M.; Blakemore, S.-J.; Browne, V.; Dooley, B.; Hickie, I. B.; Jones, P. B.; McDaid, D.; et al. 2024. The Lancet Psychiatry Commission on youth mental health. *The Lancet Psychiatry*, 11(9): 731–774.

Pinchuk, I.; Leventhal, B. L.; Ladyk-Bryzghalova, A.; Lien, L.; Yachnik, Y.; Dias, M. C.; Virchenko, V.; Szatmari, P.; Protsenko, O.; Chaimowitz, G. A.; et al. 2024. The Lancet Psychiatry Commission on mental health in Ukraine. *The Lancet Psychiatry*, 11(11): 910–933.

Qiu, H.; He, H.; Zhang, S.; Li, A.; and Lan, Z. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Qiu, H.; Li, A.; Ma, L.; and Lan, Z. 2024. Psychat: A client-centric dialogue system for mental health support. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2979–2984. IEEE.

Ribeiro, E.; Ribeiro, A. P.; Gonçalves, M. M.; Horvath, A. O.; and Stiles, W. B. 2013. How collaboration in therapy becomes therapeutic: The therapeutic collaboration coding system. *Psychology and Psychotherapy: Theory, Research and Practice*, 86(3): 294–314.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

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 preprint arXiv:2106.01702*.

Sun, H.; Xu, W.; Liu, W.; Luan, J.; Wang, B.; Shang, S.; Wen, J.-R.; and Yan, R. 2023. From Indeterminacy to Determinacy: Augmenting Logical Reasoning Capabilities with Large Language Models. *arXiv preprint arXiv:2310.18659*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tseng, Y.-M.; Huang, Y.-C.; Hsiao, T.-Y.; Chen, W.-L.; Huang, C.-W.; Meng, Y.; and Chen, Y.-N. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Xie, H.; Chen, Y.; Xing, X.; Lin, J.; and Xu, X. 2024. PsyDT: Using LLMs to Construct the Digital Twin of Psychological Counselor with Personalized Counseling Style for Psychological Counseling. *arXiv preprint arXiv:2412.13660*.

Xiong, H.; Wang, S.; Zhu, Y.; Zhao, Z.; Liu, Y.; Huang, L.; Wang, Q.; and Shen, D. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yao, S.; Ke, Q.; Wang, Q.; Li, K.; and Hu, J. 2024a. Lawyer GPT: A legal large language model with enhanced domain knowledge and reasoning capabilities. In *Proceedings of the 2024 3rd International Symposium on Robotics, Artificial Intelligence and Information Engineering*, 108–112.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2024b. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. 2022. GLM-130B: An Open Bilingual Pre-trained Model. In *The Eleventh International Conference on Learning Representations*.

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, H.; Chen, J.; Jiang, F.; Yu, F.; Chen, Z.; Chen, G. H.; Li, J.; Wu, X.; Zhiyi, Z.; Xiao, Q.; et al. 2023. HuatuoGPT, Towards Taming Language Model to Be a Doctor. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhang, S.; Fu, D.; Liang, W.; Zhang, Z.; Yu, B.; Cai, P.; and Yao, B. 2024b. Trafficgpt: Viewing, processing and interacting with traffic foundation models. *Transport Policy*, 150: 95–105.

Zhang, X.; and Yang, Q. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4435–4439.