

PsyAdvisor: A Plug-and-Play Strategy Advice Planner with Proactive Questioning in Psychological Conversations

Anonymous ACL submission

Abstract

Proactive questioning is essential in psychological conversations as it helps uncover deeper issues and unspoken concerns. Current psychological LLMs are constrained by passive response mechanisms, limiting their capacity to deploy proactive strategies for psychological counseling. To bridge this gap, we first develop the ProPsyC (Proactive Psychological Conversation) dataset, a multi-turn conversation dataset with interpretive labels including strategy decision logic and reaction attribution. Based on ProPsyC, we propose PsyAdvisor by supervised fine-tuning, a plug-and-play proactive questioning strategy planner that empowers psychological LLMs to initiate well-timed questioning through strategic prompting. Experimental results demonstrate that psychological LLMs integrated with PsyAdvisor substantially improve proactive questioning capacity, conversation depth, and response quality. Furthermore, PsyAdvisor shows promising potential in assisting novice counselors by providing strategy recommendations. This study provides new optimization directions for psychological conversation systems and offers valuable insights for future research on proactive questioning mechanisms in psychological LLMs. Our code are available at <https://anonymous.4open.science/r/PsyAdvisor>.

1 Introduction

Psychological issues have emerged as a critical global concern in contemporary society (WHO, 2021). However, access to professional psychological counseling remains limited for many due to high costs and a shortage of qualified practitioners (Cohen et al., 2021). In this context, advancements in large language models (LLMs) present transformative opportunities for psychological counseling (Lawrence et al., 2024). Existing research has demonstrated the efficacy of LLMs (e.g., GPT-4 (Lawrence et al., 2024), DeepSeek (Liu et al.,

2024), Qwen2.5 (Bai et al., 2023)) in various conversation tasks, such as semantic understanding (Fan and Ma, 2024) and strategy-driven response generation (Zhang et al., 2023). Inspired by these studies, psychological LLMs have emerged that integrate domain-specific expertise and simulate human counselor interactions, significantly lowering the barriers to psychological counseling (Guo et al., 2024b). However, current psychological LLMs predominantly operate in passive response modes, offering only empathetic or listening-oriented replies (Chen et al., 2023; Na, 2024; Qiu et al., 2024a). Their lack of proactive questioning fails to identify clients’ deeper psychological states and needs, consequently constraining the depth of conversations and therapeutic outcomes.

In traditional psychological counseling, proactive questioning strategies are recognized for enhancing communication quality and facilitating client self-expression (Sklare et al., 1985). Psychological theory defines proactive questioning as purposeful questioning by counselors to guide the conversation, acquire information, or promote self-exploration (Broedel, 1962). Even a simple question posed at an appropriate time can significantly improve therapeutic effectiveness in psychological counseling (McCarthy et al., 2021). Therefore, integrating proactive strategies is critical to bridging the gap between LLMs and human counselors. However, poorly timed interventions can undermine therapeutic outcomes (Williams, 2023), making timing optimization paramount.

This paper aims to equip psychological LLMs with context-aware proactive questioning capabilities for timely intervention in psychological conversations. Three key challenges arise:

- 1) **Timing Recognition:** Existing psychological LLMs struggle to identify the optimal timing for proactive questioning, as such decisions require real-time integration of multiple factors (e.g.,

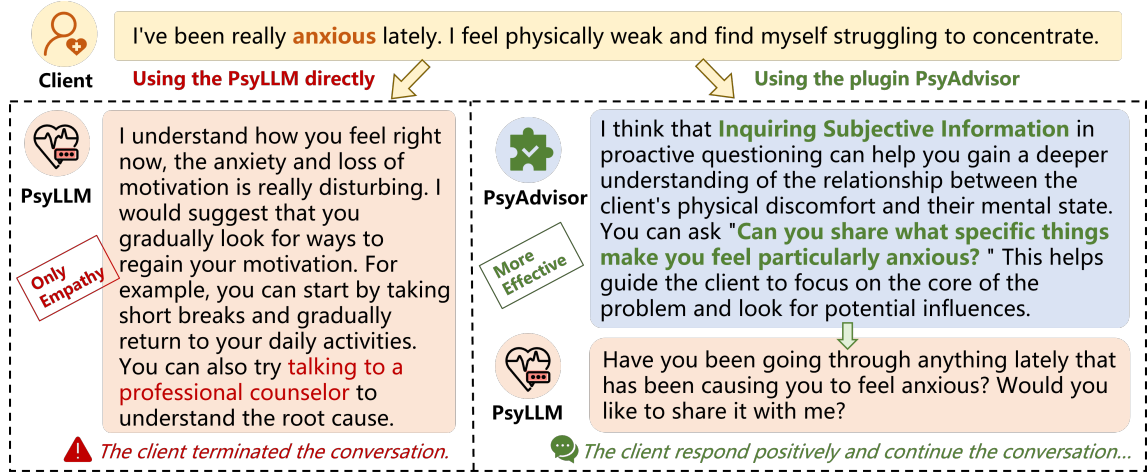


Figure 1: An example of a psychological LLM response: Using PsyAdvisor plugin vs. No plugin.

- expert knowledge, client reaction, and context).
- 2) **Strategic Deficiency:** Current LLMs rely on passive interaction patterns, which are incapable of using proactive strategies for conversation.
 - 3) **Interpretability Gaps:** Existing psychological datasets lack interpretive annotations about counselors' strategies, hindering LLMs' understanding of questioning rationale.

To address these challenges, we propose **PsyAdvisor**, a plug-and-play Chinese adaptable plugin designed to enhance the proactive questioning skills of psychological LLMs. As shown in Figure 1, existing psychological LLMs often fail to address the client's issues due to generic responses. In contrast, PsyAdvisor assists psychological LLMs by advising proactive strategies at the right moments, facilitating further self-expression from the client. We first annotate counselor strategies and client reactions from existing psychological conversations, then construct a chain-of-thought (CoT) framework with dual-perspective interpretive labels (rational and emotional) to clarify the causality of strategy selection and client reactions. This results in the **ProPsyC dataset**, comprising 2,001 high-quality multi-turn conversations. We develop PsyAdvisor through supervised fine-tuning with the ProPsyC dataset, enabling psychological LLMs to determine when to initiate proactive questioning and suggest effective strategies.

Our contribution can be summarised as follows:

- To our best knowledge, this is the first work to systematically investigate proactive capabilities in psychological LLMs. We propose **PsyAdvisor**, a plugin that provides proactive timing and strategy advice.

- We construct the **ProPsyC dataset** with dual interpretability annotations (counselor strategy rationale and client reaction attribution), providing a foundation for strategy-guided LLM decision-making.
- Extensive experiments have shown that PsyAdvisor-enhanced LLMs outperform the baseline in terms of proactive questioning timing and strategy execution effectiveness. Human evaluations further validate its utility in assisting both LLMs and human counselors to identify optimal questioning timing.

2 Related Work

Psychological LLMs remain constrained by passive interactions, while existing proactive systems lack specific expertise. However, this challenge finds resolution through modular plugins for LLMs.

Psychological LLMs. Recent advances in psychological LLMs demonstrate their potential for cost-effective mental health support (Guo et al., 2024b). Related research focuses on three directions: (1) *Strategy-based fine-tuning*: LLMs can learn strategies in the fine-tuning process through empathy enhancement (Chen et al., 2023), conversation restructuring (Qiu et al., 2024a), and psychological theory integration (e.g., cognitive behavioral therapy (Na, 2024)), yet struggle to deepen the conversation; (2) *Instruction-driven specialization*: Psychological LLMs can be achieved through fine-grained instruction generation (Hu et al., 2024) and client feedback-based instruction (Qiu et al., 2024b), while these LLMs constrained by the lack of systematic knowledge, limiting their effectiveness; (3) *Role-based LLMs*: Prompting a general model(e.g., ChatGPT) to role-play as a counselor

can aid counseling, but its specialization is inferior to psychological LLMs(Zhang et al., 2024). Our PsyAdvisor embeds psychological expertise to deliver proactive strategies, enhancing the depth of conversation while improving the professionalism and effectiveness of psychological conversations.

Proactive Conversation Systems. While most LLMs operate via passive response mechanisms, scenarios requiring autonomous decision-making (e.g., negotiation, psychotherapy) require proactive conversation initiation (Liao et al., 2023). Deng et al. (2023a) argue that strategic prompting (e.g., chain-of-thought reasoning) can stimulate LLMs’ proactive potential, yet prompt engineering alone fails to achieve precise timing control (Wang et al., 2023). Recent approaches have established effective proactive conversation systems through task-specific guidance (Li et al., 2023b), structured attribute modeling (Wang et al., 2023), and reinforcement learning-based policy iteration (Guo et al., 2024a). While these systems demonstrate success in general domains, they lack specific expertise in the psychological field. Consequently, we propose PsyAdvisor, which leverages professional strategies to enable psychological LLMs to initiate proactive questioning at appropriate moments.

Modular Plugins for LLMs. LLM plugin refers to a targeted enhancement of module functionality without altering the integrity of the LLM(Ma et al., 2024). While existing plugins achieve cost-capability balance through specialized models, they face three critical limitations: (1) Closed-task optimizations (e.g., text classification (Xu et al., 2024)) lack open-domain adaptability (Yao et al., 2023); (2) Proactive systems exhibit poor timing judgment (PPDPP (Deng et al., 2023b)) or rigid rule-based approaches (Feng et al., 2024); (3) Feedback-driven solutions (Peng et al., 2023) over-rely on prompts. Our **PsyAdvisor** plugin overcomes these by integrating context-aware timing judgment with strategy generation through lightweight frameworks.

3 PsyAdvisor

To enable proactive questioning in psychological LLMs, we developed *PsyAdvisor*, a strategy recommendation plugin that offers proactive timing and strategy suggestions. We created the ProPsyC dataset, annotated with strategies, reactions, and interpretations, to address the challenges of scarce public cases and the need for expert annotation. Existing datasets often fall short of supporting proac-

tive conversation, so the ProPsyC dataset extends high-quality publicly available ones, focusing on proactive psychological conversation. Using this dataset, we fine-tuned a small-scale LLM to create the PsyAdvisor plugin.

3.1 Raw Data Collection

We conducted an extensive review and selection of publicly available psychological conversation datasets. After considering factors such as dataset reliability, annotation accuracy, and the completeness of multi-turn conversations, we chose three high-quality Chinese datasets, collecting a total of 3865 Chinese psychological counseling conversations to better suit this research.

Xinling (Li et al., 2023a): A multi-turn dataset with 300 counseling conversations, annotated with therapist strategies and client reactions. Data were sourced from real psychological experiments and validated by experts, using an innovative annotation framework at the turn level.

CPsyCounD (Zhang et al., 2024): A dataset of 3134 multi-turn conversations from real Chinese counseling reports. It ensures privacy protection and maintains high professionalism, authenticity, and safety.

Psy-Insight (Chen, 2024): A multi-turn dataset with 431 Chinese and 520 English counseling conversations sourced from blogs and books. It includes multi-task labels such as strategies and topics, and has received positive human feedback. Only the Chinese portion was used in this study.

3.2 Data Refactoring

3.2.1 Data Processing

We further cleaned and filtered the collected data. Specifically, we performed initial cleaning based on the number of turns in each conversation. Following advice from professional psychologists, we retained only conversations with more than 10 turns and removed those not suitable for online psychological counseling (e.g., conversations containing non-textual information such as silence or micro-expressions). To ensure that the model could effectively learn proactive strategies, we discarded conversations not involving proactive questioning. Additionally, we removed irrelevant information at the beginning and end of each conversation.

After preliminary processing, we obtained a dataset containing 2001 multi-turn conversations with an average of 16.12 turns per conversation. Based on the background information in the dataset,

we reclassified the topics discussed in the conversations. With input from psychologists, we divided the dataset into four categories: Mental Health & Emotion, Interpersonal Relationships & Social Issues, Career & Stress, and Self-awareness & Personal Growth, as shown in Figure 2.

3.2.2 Strategy and Client Reactions Annotations

We defined the required therapist strategies and client reactions based on the annotation framework used in the Xinling dataset (Li et al., 2023a) and annotated the remaining two datasets accordingly. To better suit the proactive questioning task, we sought the assistance of psychologists to further divide the strategies into proactive and passive categories, with detailed explanations provided in Appendix A.1. Additionally, we followed the original framework’s classification of client reactions, dividing them into positive and negative categories, as detailed in Appendix A.1.

Given the high cost of manual annotation and the widespread use of LLMs for data annotation (Tan et al., 2024), we adopted a hybrid annotation approach, combining LLM-based labeling with expert review. We first sampled 50 multi-turn conversations from the Xinling dataset, covering all strategy and reaction types, with a roughly balanced distribution of positive and negative samples (proactive vs. passive strategies, positive vs. negative reactions). We used GLM-4 to annotate the remaining 250 conversation samples. Specifically, we designed a 2-shot annotation prompt, including explanations for all strategy and reaction types, as well as two examples with both positive and negative samples, using RAG to retrieve knowledge from the 50 initial samples as historical conversations. The annotated results were reviewed by professional psychologists, and representative mis-annotations were corrected and added to the prompt as examples for re-annotation. We used the refined prompt to annotate the remaining datasets. The final data distribution is provided in Appendix A.2.

3.3 ProPsyC Dataset Construction

To enhance the comprehension of LLMs regarding counselors’ decision-making logic in psychological consultations, we employ GLM-4 to perform interpretability annotations on conversation data, thereby constructing the Proactive Psychological Conversation (ProPsyC) dataset. The annotation process follows a four-stage Chain-of-Thought

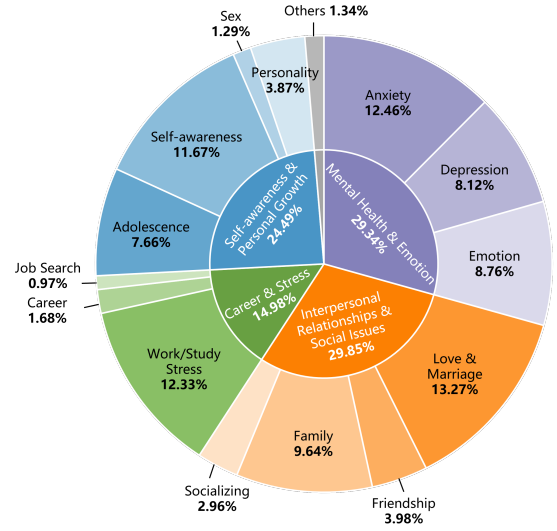


Figure 2: Distribution statistics of psychological conversation topics in the ProPsyC dataset.

(CoT) prompting framework as illustrated in Figure 3, with detailed procedures outlined below:

- Objective Specification:** Establish clear annotation guidelines defining the target objectives and output formats. This includes formal definitions for counseling strategy labels and client response labels, accompanied by comprehensive annotation manuals.
- Expertise-Based Rationalization:** Simulate professional psychologists’ decision-making through a dual knowledge retrieval mechanism: (1) *Therapeutic Knowledge Base*: Contains formal explanations of mainstream psychotherapy approaches (e.g., Cognitive Behavioral Therapy (CBT), Rational Emotive Therapy (RET), Humanistic Therapy (HT)); (2) *Case Repository*: Aggregates real counseling sessions from authoritative sources including YiXinLi¹, National Mental Health Platform², and the PsyQA dataset (Sun et al., 2021).
- Reaction-Centric Validation:** The efficacy of the strategy employed by the counselor in the present context is determined by the client’s reaction. If a positive reaction is elicited from the client, the counselor’s strategy is deemed effective. Conversely, a negative reaction from the client indicates that the strategy is immature.
- Expertise Reinforcement:** Provide annotated examples from certified counselors. The labels are divided into two categories: (1) *Strategy Rationale*: Explanations for counselor’s strategy

¹<https://www.xinli001.com/>

²<https://www.nmhp.gov.cn/>

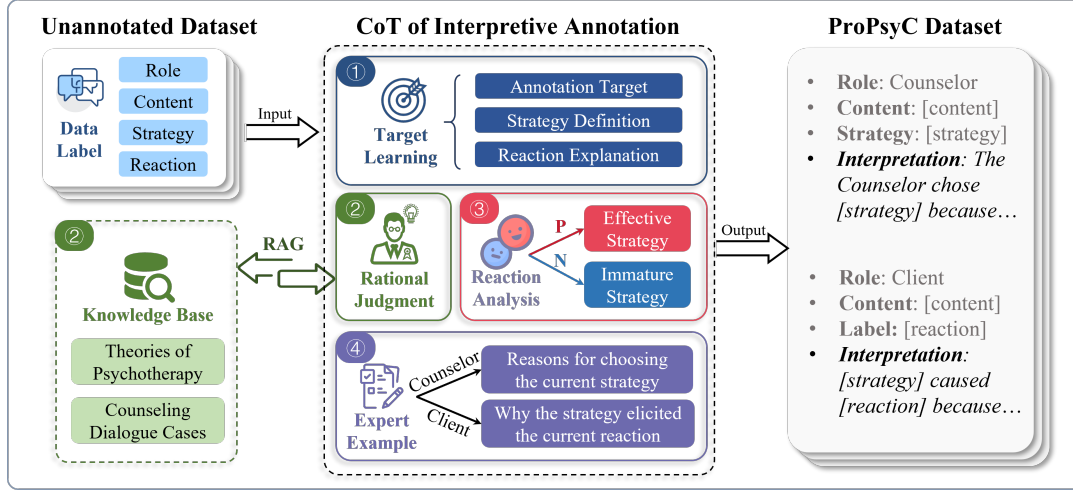


Figure 3: Interpretive annotation process of ProPsyC dataset constructed based on CoT.

selection; (2) *Reaction Attribution*: Explain why the strategy elicited the client’s current reaction.

Appendix B provides CoT examples and ProPsyC dataset statistics. Notably, a comparison between GLM-4 and GPT-4o reveals that GLM-4 achieves similar annotation quality (Appendix B) at only 10% of the cost, making it preferable for large-scale annotation tasks.

Implementation Note: All annotated prompt templates were carefully validated by three psychologists, and inter-annotator agreement (Cohen’s κ) (Wan et al., 2015) was 0.82.

3.4 Plug-and-Play Strategy Advice Planner

Based on the ProPsyC dataset, we perform supervised fine-tuning on a relatively small-scale LLM (e.g., Qwen2.5-3B), enabling it to predict the appropriate strategy to adopt the given conversation history, and generates recommendations on the implementation of the current strategy. Specifically, given a conversation history $D_{history} = \{u_{counselor1}, u_{client1}, \dots, u_{counselort-1}, u_{clientt-1}\}$, the model is required to predict an action at the t -th turn, which includes both the strategy and its rationale. The fine-tuning objective is to minimize the cross-entropy loss between the predicted action and the ground truth action.

For strategy prediction, we treat it as a classification task where the LLM needs to select the appropriate strategy from a set of known strategies. Based on maximum likelihood estimation, the objective of the supervised fine-tuning (SFT) is to

minimize the following loss function:

$$L_{strategy} = - \sum_{t=1}^T \log P(a_t | D_{history}) \quad (1)$$

where T is the total number of conversation turns, a_t is the predicted strategy at turn t , which includes both proactive and non-proactive strategies, and $a_t \in \{strategy1, strategy2, \dots, strategy12\}$, with $D_{history}$ representing the conversation history, including multiple rounds of conversation between the counselor and the client.

For the prediction of strategy rationale, we model it as a conditional language generation task. Given the strategy a_t and the conversation history $D_{history}$, the model is required to generate a natural language description of the rationale. We adopt a training approach similar to language models, maximizing the conditional probability distribution of the strategy rationale:

$$L_{reason} = - \sum_{t=1}^T \sum_{i=1}^{L_t} \log P(c_{t,i} | c_{t,<i}, a_t, D_{history}) \quad (2)$$

where T is the total number of conversation turns, L_t is the length of the rationale text generated at the t -th turn, $c_{t,i}$ is the i -th word in the rationale text at the t -th turn, $c_{t,<i}$ represents all the words before the i -th word in the rationale text, and $P(c_{t,i} | c_{t,<i}, a_t, D_{history})$ is the probability of generating the next word given the context (including the previously generated text, strategy, and conversation history).

The final fine-tuning loss function is the weighted sum of the strategy prediction loss and

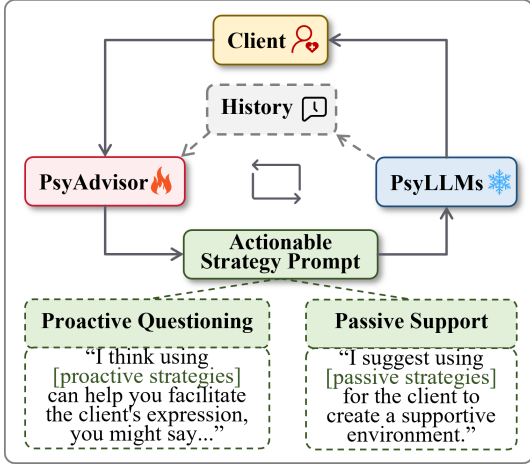


Figure 4: Workflow of PsyAdvisor. (1) PsyAdvisor obtains the client’s conversation; (2) PsyAdvisor generates actionable strategy prompts(proactive questioning or passive support); (3) Psychological LLMs produce responses based on the prompts; (4) Conversation history is fed to PsyAdvisor to enhance decision-making.

the rationale prediction loss, specifically:

$$L_{total} = L_{strategy} + \lambda L_{reason} \quad (3)$$

where λ is a hyperparameter used to adjust the balance between the strategy prediction and rationale prediction losses. By minimizing this loss function, the model is able to predict the appropriate strategy at each turn of the conversation and provide a reasonable rationale for the chosen strategy, effectively supporting the decision-making process for proactive questioning.

Furthermore, we guide the model to learn the counselor’s content of questioning during the adoption of proactive strategies through prompt engineering, which helps the model generate specific strategy suggestions that align with the context. The workflow of the fine-tuned plug-in, PsyAdvisor, is shown in Figure 4.

4 Experiments

4.1 Experimental Setups

4.1.1 Baselines

To accurately measure the performance of PsyAdvisor, we integrate it with the following baselines for comparison:

Large-scale General-purpose LLMs (Number of parameters >100B): GPT-4o(Achiam et al., 2023), DeepSeek-V3(Liu et al., 2024), GLM-4(GLM et al., 2024).

Small-scale Open-source LLMs (Number of parameters <10B): Llama3.1-8B-Instruct(Touvron et al., 2023), Qwen2.5-7B-Instruct(Bai et al., 2023), GLM4-9B-Chat(GLM et al., 2024).

Psychological LLMs: SoulChat2.0(Xie et al., 2024), MeChat (Qiu et al., 2024a), PsyChat(Qiu et al., 2024b).

These baselines are selected based on three factors: model scale differences, general-purpose vs. domain-specific capabilities, and open-source vs. closed-source technologies, in order to evaluate the adaptability and deployment performance of PsyAdvisor in psychological conversations.

4.1.2 Evaluation Metrics

We evaluate the performance of PsyAdvisor using a combination of automatic and human evaluation metrics as follows:

Automatic Evaluation Metrics: We categorize the metrics into two classes: (1) *Stability metrics* include 5 conventional metrics: R-1, R-L(Lin et al., 2019), B-2, B-4(Papineni et al., 2002), and F_{bert} (Zhang et al., 2019); (2) *Enhancement metrics* consist of 2 new metrics we proposed: Strategy Fit Ratio (SFR) and Proactive Questioning Ratio (PQR). Stability metrics require only stable performance, while enhancement metrics demand significant improvement to demonstrate superiority.

- **SFR:** Assesses the ratio of fit strategies. A strategy is deemed "fit" only if it triggers positive reactions; otherwise, its timing is considered unsuitable. It reflects the temporal rationality of LLM’s strategy selection.
- **PQR:** Measures the ratio of proactive strategies among all strategies. It reflects the LLM’s capacity for initiating proactive questioning.

Human Evaluation Metrics: Inspired by See et al. (2019), we evaluate the timing and strategy recommendations for proactive questioning based on four human assessment criteria: Strategy Effectiveness (SE), Reasoning Coherence (RC), Professionalism (Prof), and Ethical Safety (ES). The ratings for SE and RC are in the range of 0, 1, 2, Prof. is rated on a scale of 0, 1, 2, 3, and ES. is rated on 0, 1. Detailed descriptions and evaluation procedures can be found in Appendix C.

4.1.3 Implementation Details

Qwen2.5-3B-Instruct(Bai et al., 2023) is selected as the backbone model due to its strong performance

Type	Models	Plugin	Stability metrics					Enhancement metrics(%)	
			R-1	R-L	B-2	B-4	Fbert	SFR	PQR
Large-scale General -purpose LLMs	GPT-4o	\	18.52	16.27	6.21	3.54	86.54	55.48	20.82
		<i>PsyAdvisor</i>	<u>22.35</u>	<u>19.82</u>	<u>7.88</u>	<u>4.20</u>	87.21	82.39	58.66
	DeepSeek-V3	\	17.28	15.53	5.84	3.01	84.31	52.54	18.48
		<i>PsyAdvisor</i>	20.51	18.24	7.52	3.89	87.47	78.56	53.42
	GLM-4	\	15.84	13.61	5.03	2.59	82.15	50.01	15.01
		<i>PsyAdvisor</i>	18.23	16.08	5.84	3.11	85.36	75.98	48.29
Small-scale Open-source LLMs	GLM4-9B	\	12.14	10.57	3.84	1.54	80.42	38.48	12.03
	-Chat	<i>PsyAdvisor</i>	15.27	13.49	4.51	2.01	83.33	68.69	42.44
	Qwen2.5-7B	\	12.55	10.80	4.25	1.88	78.24	50.49	14.12
	-Instruct	<i>PsyAdvisor</i>	15.01	13.24	5.24	2.23	81.54	70.55	47.54
	Llama3.1-8B	\	9.27	8.87	3.28	1.21	75.64	35.88	13.62
	-Instruct	<i>PsyAdvisor</i>	10.54	9.24	3.56	1.84	78.12	60.32	39.98
Psychological LLMs	PsyChat	\	10.52	9.23	4.21	1.85	70.22	42.01	14.47
		<i>PsyAdvisor</i>	12.89	11.22	3.51	1.27	73.01	72.03	51.24
	MeChat	\	14.23	12.51	4.84	2.29	80.68	60.33	23.26
		<i>PsyAdvisor</i>	16.88	14.58	5.56	2.88	83.11	78.76	55.47
	SoulChat2.0	\	16.54	12.88	5.66	2.28	83.54	70.34	41.58
		<i>PsyAdvisor</i>	18.01	16.11	6.52	2.99	<u>87.38</u>	85.24	69.94

Table 1: Results of the PsyAdvisor automated metrics assessment

in instruction-following tasks and its lightweight architecture. The model is fine-tuned on the ProPsyC dataset (split 7:1:2 into training/validation/test sets), trained for three epochs on an NVIDIA A100-80G GPU with the AdamW(Loshchilov et al., 2017) optimizer (1e-5 learning rate), resulting in the PsyAdvisor plugin. We then used PsyAdvisor to access the baseline models for experiments where non-psychological LLMs are prompted to act as counselors, and psychological LLMs utilize their native capabilities. All open-source models are configured with temperature=0.5, top_p=0.75, and top_k=20 during inference.

GPT-4o and GLM-4 are used to simulate clients and generate dialogues, applying predefined rules to identify strategies and reaction types. Human evaluations are conducted with three psychology graduate students and a senior counselor, who select topics based on their experience. To minimize bias, model names are anonymized, and experts are only informed about the integration of PsyAdvisor. See Appendix D for more details.

4.2 Results and Evaluation

4.2.1 Automatic Evaluation of Results

Table 1 presents the automatic evaluation results of PsyAdvisor across baseline models. As shown,

integrating PsyAdvisor improves the performance of all models to varying extents, providing strong evidence for the broad applicability and significant impact of PsyAdvisor across different model types. Notably, improvements in the SFR and PQR metrics indicate that PsyAdvisor effectively enhances the proactive questioning capability of psychological LLMs. The strategies selected by PsyAdvisor not only capture the appropriate moments for intervention but also effectively guide the client to provide positive feedback, validating the scientific and reasonable nature of PsyAdvisor’s strategy selection and timing. In terms of response generation quality, after integrating PsyAdvisor, all models show slight improvements in the five automatic evaluation metrics, suggesting that PsyAdvisor has contributed to enhancing the semantic understanding of psychological LLMs. It is worth noting that small-scale open-source LLMs generally perform lower than other models across all metrics, showing weaker semantic understanding and strategy selection ability. Therefore, in the subsequent human evaluation phase, we decided to exclude the small-scale open-source LLMs from the evaluation.

4.2.2 Human Evaluation Results

Table 2 presents the results of the human evaluation, revealing significant improvements in strategy

Type	Models	Plugin	Human Evaluation Metrics			
			SE.	RC.	Prof.	ES.
General-purpose LLMs	GPT-4o	\	1.31	1.08	2.48	1.00
		<i>PsyAdvisor</i>	1.90	1.79	2.61	1.00
	DeepSeek-V3	\	1.39	1.10	2.33	1.00
		<i>PsyAdvisor</i>	1.84	1.72	2.50	1.00
	GLM-4	\	0.92	0.94	1.82	1.00
		<i>PsyAdvisor</i>	1.59	1.57	2.18	1.00
Psychological LLMs	PsyChat	\	1.12	1.05	2.09	1.00
		<i>PsyAdvisor</i>	1.66	1.53	2.48	1.00
	MeChat	\	1.36	1.13	2.27	1.00
		<i>PsyAdvisor</i>	1.75	1.61	2.59	1.00
	SoulChat2.0	\	1.44	1.38	2.52	1.00
		<i>PsyAdvisor</i>	1.88	1.79	2.83	1.00
Human counselor		\	1.80	1.72	2.88	1.00
		<i>PsyAdvisor</i>	1.98	1.94	2.90	1.00

Table 2: Results of PsyAdvisor’s human evaluation metrics

effectiveness, reasoning coherence, and professionalism after integrating PsyAdvisor. Specifically, the enhancement in strategy effectiveness and reasoning coherence indicates that PsyAdvisor not only enables models to initiate proactive questioning at the right moments but also provides sound reasoning for the strategy selection, which helps guide the client toward deeper self-expression, thereby significantly improving the depth of the conversation. Furthermore, the improvement in professionalism confirms that psychological LLMs with PsyAdvisor generate responses that better align with the standards of professional psychological counseling, enhancing the quality of the language used in the conversations. Moreover, in terms of ethical safety, all models maintained a stable high-level performance, showing that PsyAdvisor did not affect the ethical compliance of the model’s responses, ensuring the safety of the conversation process.

Notably, we invited three psychology graduate students to use PsyAdvisor directly in counseling practice, with a senior psychological counselor providing scores throughout the process. The results demonstrated that PsyAdvisor played a key role in assisting the counselors in making strategic judgments.

4.3 Case Study

To validate PsyAdvisor’s effectiveness, we analyzed a typical case using SoulChat2.0 as the psychological LLM. Details are in Appendix E. The client in the case, experiencing work-related stress

and anxiety, initially received passive responses, like validation and minimal encouragement, which lacked depth in guiding self-exploration. After integrating PsyAdvisor, proactive questioning strategies increased. For example, PsyAdvisor suggested exploring new actions to alleviate stress, prompting the client to share specific stressors and thoughts on future adjustments. This shift led to a deeper dialogue, providing valuable insights for treatment planning. Our PsyAdvisor enhanced the LLM’s proactive questioning, improving conversation depth and treatment effectiveness.

5 Conclusion

This paper introduces PsyAdvisor, a plug-and-play plugin for proactive questioning timing and strategy planning. By accurately determining the optimal moments for initiating proactive strategies and providing corresponding recommendations, PsyAdvisor significantly enhances the performance of psychological LLMs in conversation. In addition, we have developed the ProPsyC dataset, which incorporates interpretable labels that assist in interpreting the decision-making logic behind therapists’ strategies and clients’ responses. The experimental results indicate that PsyAdvisor, fine-tuned using ProPsyC, substantially improves the performance of existing psychological LLMs and demonstrates great potential in supporting human therapists during treatment. This work offers new research directions for the future development of psychological conversation systems.

Limitation

While our approach shows promising results, several limitations should be addressed. First, PsyAdvisor is primarily trained on Chinese psychological counseling data, which may limit its applicability to other cultural and linguistic contexts. Additionally, our evaluation metrics are based on simulated conversations, and their validation in real-world clinical settings remains pending. Lastly, the long-term impact of proactive questioning strategies in ongoing therapeutic interactions is yet to be explored. The study also relies on one senior counselor and three graduate students in psychology, which may introduce biases. Future work will include a broader range of psychologists for a more comprehensive assessment.

Ethical Statement

In this study, we prioritize ethical considerations in every stage of our research. All datasets used were either publicly available or obtained with explicit permission. Specifically, the XinLing(Li et al., 2023a) dataset was accessed under an agreement with its authors, ensuring that the necessary permissions and protocols were followed. Moreover, we checked that ProPsyC did not specify or uniquely identify the individual. The data was all anonymized to ensure privacy and confidentiality in compliance with ethical standards. Informed consent was obtained for all relevant data, and strict measures were taken to protect participants' identities.

Furthermore, we acknowledge that PsyAdvisor is designed as a tool to support rather than replace human counselors. While it significantly enhances proactive questioning in psychological dialogues, it does not substitute the professional judgment or expertise of licensed therapists. PsyAdvisor's role is to aid counselors by suggesting timely strategies, but a licensed therapist must continue to oversee the therapeutic process. We are committed to ensuring that the deployment of PsyAdvisor adheres to the highest ethical standards, safeguarding the well-being of individuals and respecting the professional boundaries of psychological care.

Annotator Compensation: Each annotator was allotted approximately 2 minutes to evaluate each sample, and an honorarium of \$0.319 was paid for each sample, which equates to an hourly wage of \$9.57. This figure is above the current U.S. federal minimum wage of \$7.25 per hour.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- John W Broedel. 1962. The use of questioning in counseling. *The School Counselor*, 10(1):12–15.
- K. Chen. 2024. Psy-insight: Explainable multi-turn bilingual dataset for mental health counseling. <https://github.com/ckqqqq/Psy-Insight>. GitHub repository.
- Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.
- Katherine A Cohen, Colleen Stiles-Shields, Nathan Winquist, and Emily G Lattie. 2021. Traditional and nontraditional mental healthcare services: usage and preferences among adolescents and younger adults. *The Journal of Behavioral Health Services & Research*, pages 1–17.
- Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. 2023a. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10602–10621.
- Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2023b. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations*.
- Jie Fan and Guojun Ma. 2024. Building end-to-end dialogue system with large language models. In *Fifteenth International Conference on Signal Processing Systems (ICSPS 2023)*, volume 13091, pages 856–862. SPIE.
- Shangbin Feng, Weijia Shi, Yuyang Bai, Vidhisha Balachandran, Tianxing He, and Yulia Tsvetkov. 2024. Knowledge card: Filling LLMs' knowledge gaps with plug-in specialized language models. In *The Twelfth International Conference on Learning Representations*.
- Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. 2024. Chatglm: A family

688	of large language models from glm-130b to glm-4 all	Yufei Ma, Zihan Liang, Huangyu Dai, Ben Chen, De-	743
689	tools. <i>arXiv preprint arXiv:2406.12793</i> .	hong Gao, Zhuoran Ran, Wang Zihan, Linbo Jin,	744
690	Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song,	Wen Jiang, Guannan Zhang, et al. 2024. Modula:	745
691	Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma,	Mixture of domain-specific and universal lora for	746
692	Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: In-	multi-task learning. In <i>Proceedings of the 2024 Con-</i>	747
693	centivizing reasoning capability in llms via reinforce-	<i>ference on Empirical Methods in Natural Language</i>	748
694	ment learning. <i>arXiv preprint arXiv:2501.12948</i> .	<i>Processing</i> , pages 2758–2770.	749
695	Shasha Guo, Lizi Liao, Jing Zhang, Cuiping Li, and	John McCarthy, Emma Shannon, and Michelle Bruno.	750
696	Hong Chen. 2024a. Pcqpr: Proactive conversational	2021. Creative question-framing: 12 ideas for	751
697	question planning with reflection. In <i>Proceedings</i>	counselors-in-training. <i>Journal of Creativity in Men-</i>	752
698	<i>of the 2024 Conference on Empirical Methods in</i>	<i>tal Health</i> , 16(4):499–510.	753
699	<i>Natural Language Processing</i> , pages 11266–11278.		
700	Zhijun Guo, Alvina Lai, Johan H Thygesen, Joseph Far-	Hongbin Na. 2024. Cbt-llm: A chinese large language	754
701	rington, Thomas Keen, Kezhi Li, et al. 2024b. Large	model for cognitive behavioral therapy-based mental	755
702	language models for mental health applications: Sys-	health question answering. In <i>Proceedings of the</i>	756
703	tematic review. <i>JMIR mental health</i> , 11(1):e57400.	<i>2024 Joint International Conference on Computa-</i>	757
704	Jinpeng Hu, Tengting Dong, Luo Gang, Hui Ma, Peng	<i>tional Linguistics, Language Resources and Evalua-</i>	758
705	Zou, Xiao Sun, Dan Guo, Xun Yang, and Meng Wang.	<i>tion (LREC-COLING 2024)</i> , pages 2930–2940.	759
706	2024. Psycollm: Enhancing llm for psychological	Kishore Papineni, Salim Roukos, Todd Ward, and Wei-	760
707	understanding and evaluation. <i>IEEE Transactions on</i>	Jing Zhu. 2002. Bleu: a method for automatic evalu-	761
708	<i>Computational Social Systems</i> .	ation of machine translation. In <i>Proceedings of the</i>	762
709	Hannah R Lawrence, Renee A Schneider, Susan B	<i>40th annual meeting of the Association for Computa-</i>	763
710	Rubin, Maja J Matarić, Daniel J McDuff, and	<i>tional Linguistics</i> , pages 311–318.	764
711	Megan Jones Bell. 2024. The opportunities and risks	Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng,	765
712	of large language models in mental health. <i>JMIR</i>	Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou	766
713	<i>Mental Health</i> , 11(1):e59479.	Yu, Weizhu Chen, et al. 2023. Check your facts and	767
714	Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai	try again: Improving large language models with	768
715	Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023a.	external knowledge and automated feedback. <i>arXiv</i>	769
716	Understanding client reactions in online mental	<i>preprint arXiv:2302.12813</i> .	770
717	health counseling. In <i>Proceedings of the 61st An-</i>	Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li,	771
718	<i>annual Meeting of the Association for Computational</i>	and Zhenzhong Lan. 2024a. SMILE: Single-turn	772
719	<i>Linguistics (Volume 1: Long Papers)</i> , pages 10358–	to multi-turn inclusive language expansion via Chat-	773
720	10376.	GPT for mental health support. In <i>Findings of the</i>	774
721	Siheng Li, Yichun Yin, Cheng Yang, Wangjie Jiang,	<i>Association for Computational Linguistics: EMNLP</i>	775
722	Yiwei Li, Zesen Cheng, Lifeng Shang, Xin Jiang,	2024, pages 615–636.	776
723	Qun Liu, and Yujiu Yang. 2023b. Newsdialogues:	Huachuan Qiu, Anqi Li, Lizhi Ma, and Zhenzhong Lan.	777
724	Towards proactive news grounded conversation. In	2024b. Psychat: A client-centric dialogue system	778
725	<i>Findings of the Association for Computational Lin-</i>	for mental health support. In <i>2024 27th Interna-</i>	779
726	<i>guistics: ACL 2023</i> , pages 3634–3649.	<i>tional Conference on Computer Supported Cooper-</i>	780
727	Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023.	<i>ative Work in Design (CSCWD)</i> , pages 2979–2984.	781
728	Proactive conversational agents. In <i>Proceedings of</i>	IEEE.	782
729	<i>the Sixteenth ACM International Conference on Web</i>	Abigail See, Stephen Roller, Douwe Kiela, and Jason	783
730	<i>Search and Data Mining</i> , pages 1244–1247.	Weston. 2019. What makes a good conversation?	784
731	Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang	how controllable attributes affect human judgments.	785
732	Ren. 2019. Kagnet: Knowledge-aware graph net-	In <i>Proceedings of the 2019 Conference of the North</i>	786
733	works for commonsense reasoning. <i>arXiv preprint</i>	<i>American Chapter of the Association for Computa-</i>	787
734	<i>arXiv:1909.02151</i> .	<i>tional Linguistics: Human Language Technologies,</i>	788
735	Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang,	<i>Volume 1 (Long and Short Papers)</i> , pages 1702–1723.	789
736	Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi	Gerald Sklare, Pedro Portes, and Howard Splete. 1985.	790
737	Deng, Chenyu Zhang, Chong Ruan, et al. 2024.	Developing questioning effectiveness in counseling.	791
738	Deepseek-v3 technical report. <i>arXiv preprint</i>	<i>Counselor Education and Supervision</i> , 25(1):12–20.	792
739	<i>arXiv:2412.19437</i> .	Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and	793
740	Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing	Minlie Huang. 2021. Psyqa: A chinese dataset for	794
741	weight decay regularization in adam. <i>arXiv preprint</i>	generating long counseling text for mental health	795
742	<i>arXiv:1711.05101</i> , 5.	support. In <i>Findings of the Association for Com-</i>	796
		<i>putational Linguistics: ACL-IJCNLP 2021</i> , pages	797
		1489–1503.	798

799	Zhen Tan, Dawei Li, Song Wang, Alimohammad	<i>Computational Linguistics: ACL 2024</i> , pages 13947–	855
800	Beigi, Bohan Jiang, Amrita Bhattacharjee, Man-	13966. Association for Computational Linguistics.	856
801	sooreh Karami, Jundong Li, Lu Cheng, and Huan		
802	Liu. 2024. Large language models for data annota-	Qiang Zhang, Jason Naradowsky, and Yusuke Miyao.	857
803	tion and synthesis: A survey. In <i>Proceedings of the</i>	2023. Ask an expert: Leveraging language models to	858
804	2024 Conference on Empirical Methods in Natural	improve strategic reasoning in goal-oriented dialogue	859
805	Language Processing, pages 930–957.	models. In <i>Findings of the Association for Computa-</i>	860
		<i>tional Linguistics: ACL 2023</i> , pages 6665–6694.	861
806	Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier		
807	Martinet, Marie-Anne Lachaux, Timothée Lacroix,	Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q	862
808	Baptiste Rozière, Naman Goyal, Eric Hambro,	Weinberger, and Yoav Artzi. 2019. Bertscore: Eval-	863
809	Faisal Azhar, et al. 2023. Llama: Open and effi-	uating text generation with bert. <i>arXiv preprint</i>	864
810	cient foundation language models. <i>arXiv preprint</i>	<i>arXiv:1904.09675</i> .	865
811	<i>arXiv:2302.13971</i> .		
812	TANG Wan, HU Jun, Hui Zhang, WU Pan, and HE Hua.	A ProPsyC Label Interpretation and	866
813	2015. Kappa coefficient: a popular measure of	Type Distribution	867
814	rater agreement. <i>Shanghai archives of psychiatry</i> ,		
815	27(1):62.	A.1 Interpretation of Strategies and Reactions	868
816	Jian Wang, Yi Cheng, Dongding Lin, Chak Leong, and	We categorized the 12 counselor strategies into	869
817	Wenjie Li. 2023. Target-oriented proactive dialogue	proactive and passive categories, with four proac-	870
818	systems with personalization: Problem formulation	tive and eight passive strategies, and the definitions	871
819	and dataset curation. In <i>Proceedings of the 2023 Con-</i>	and examples of each strategy are shown in Table	872
820	<i>ference on Empirical Methods in Natural Language</i>	3.	873
821	<i>Processing</i> , pages 1132–1143.	For the 10 client reactions, we categorized them	874
822	WHO. 2021. World health statistics 2021: monitoring	into positive and negative reactions, with a total of	875
823	health for the sdgs, sustainable development goals.	4 positive reactions and 6 negative reactions, each	876
824	Licence: CC BY-NC-SA 3.0 IGO.	of which is explained and exemplified in Table 4.	877
825	Elizabeth Nutt Williams. 2023. The use of questions in	A.2 Distribution of Strategies and Reactions	878
826	psychotherapy: A review of research on immediate	Types	879
827	outcomes. <i>Psychotherapy</i> , 60(3):246.	Tables 5 and 6 show the distribution of the number	880
828	Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin,	of strategies and reactions in each category, as well	881
829	and Xiangmin Xu. 2024. Psydt: Using llms to con-	as the average number of words in the correspond-	882
830	struct the digital twin of psychological counselor with	ing utterance for each type and the percentage of	883
831	personalized counseling style for psychological coun-	the total category, respectively.	884
832	seling. <i>arXiv preprint arXiv:2412.13660</i> .		
833	Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu,	A.3 Description of Actual Distribution	885
834	Chenguang Zhu, and Julian McAuley. 2024. Small	The distribution of strategies is fairly balanced,	886
835	models are valuable plug-ins for large language mod-	with a slight emphasis on active strategies. This	887
836	els. In <i>Findings of the Association for Computational</i>	is due to the initial data screening, which retained	888
837	<i>Linguistics: ACL 2024</i> , pages 283–294, Bangkok,	conversations containing active questioning, and	889
838	Thailand. Association for Computational Linguistics.	the prevalence of active questioning in real con-	890
839	Angus Yang, Zehan Li, and Jie Li. 2024. Advancing	sultations. Among the strategies, ‘Inquiring sub-	891
840	genai assisted programming—a comparative study on	jective information’ is the most frequent, while	892
841	prompt efficiency and code quality between gpt-4	‘Invite to adopt a new perspective’ is the least used,	893
842	and glm-4. <i>arXiv preprint arXiv:2402.12782</i> .	as counselors typically guide clients toward a new	894
843	Weiran Yao, Shelby Heinecke, Juan Carlos Niebles,	perspective gradually rather than asking directly. In	895
844	Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy,	passive strategies, ‘Acknowledging and comforting’	896
845	Zeyuan Chen, Jianguo Zhang, Devansh Arpit, et al.	and ‘Explanation’ are most common. ‘Face qual-	897
846	2023. Retroformer: Retrospective large language	ity’ and ‘Self-exposure’ are rarely used. Similarly,	898
847	agents with policy gradient optimization. <i>arXiv</i>	‘Restatement’ is uncommon in online counseling	899
848	<i>preprint arXiv:2308.02151</i> .	because it’s a simpler environment where conver-	900
849	Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang,	sations are visible at all times.	901
850	Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye,	In terms of reactions, positive responses domi-	902
851	Chengming Li, and Xiping Hu. 2024. CPsyCoun:	nate, reflecting the nature of mental LLMs. The	903
852	A report-based multi-turn dialogue reconstruction		
853	and evaluation framework for Chinese psychological		
854	counseling. In <i>Findings of the Association for</i>		

most frequent positive reaction is ‘Giving information,’ where the client provides needed details. The most common negative reaction is ‘Expressing confusion,’ as clients often struggle to understand. ‘Sarcastic response’ and ‘Changing the subject’ are rare and harder for LLMs to identify, so there are few labeled samples.

B Interpretive Labeling Process

B.1 Specific Examples of the CoT

Based on the description in Section 3.3 of the thesis, the CoT’s prompts framework is divided into four stages, which are designed as follows:

(1) **Target Learning:** The prompts make the LLM clearly labeled with the goals, strategies and reactions explained in Table 3 and Table 4.

Prompt

You are now an experienced counselor and need to annotate and analyze the success of the counselor’s strategies in the conversation data. Please begin by studying the following information:

Counselors have a total of 12 strategies, of which 4 are proactive: **[proactive strategies and their explanations]** and 8 are passive: **[passive strategies and explanations]**;

The client has 10 reaction types, of which four are positive: **[positive reactions and their explanations]** and six are negative: **[negative reactions and explanations]**.

(2) **Rational Judgment:** Enhance the credibility of the annotation by simulating expert judgment. The expert knowledge base with psychological theories and counseling cases is retrieved with the help of prompts, mainly through the RAG technique:

Prompt

Provided below is an expert knowledge base consisting of counseling knowledge and cases:**[knowledge]**;

Please respond according to the current client’s question: **[question]**;

Please combine the existing knowledge to determine whether the current conversation needs to be proactive questioning, prioritize the retrieval of similar conversation cases to reply, or combine the theoretical knowledge of counseling.

(3) **Reaction Analysis:** Judging the effectiveness of the counselor’s most recent strategy, based on the type of client’s reaction:

Prompt

Identify whether the reaction made by the client is positive or negative; if the client’s feedback on the current strategy is **[positive reaction]**, the strategy is effective; if the feedback is **[negative reaction]**, the strategy is immature.

(4) **Expert Example:** Provide examples of psychologist annotation so that LLM can learn exactly how reasonable interpretive labels should be annotated. The format of the output is also strictly limited through prompts:

Prompt

Here is an example of the annotation:
[expert annotation example].

Please learn from the explanatory discourse and add an ‘interpretation’ field to the original dataset, with the following requirements:

- When encountering ‘role’ as ‘**counselor**’, explain why the current strategy is adopted; when encountering ‘role’ as ‘**client**’, analyze why the counselor’s strategy just now led to the positive/negative reaction of the client;
- Do not change anything else in the dataset or its formatting except for the addition of the ‘interpretation’ field.

B.2 Example of ProPsyC Dataset

We provide the original example of the ProPsyC dataset as shown in Figure 5 and the corresponding English version as shown in Figure 6. In the dataset, ‘**conversation**’ denotes the recording process of a complete conversation, ‘**role**’ represents the identity of the speaker (client or counselor), ‘**content**’ is the content of the speaker’s conversation, index denotes the current round of the conversation, and ‘**label**’ denotes the label of the client’s reaction (only when “role” is the client), ‘**strategy**’ is the strategy used by the counselor (only when “role” is the counselor), and ‘**interpretation**’ is an explanatory label, which explains the reason for the counselor’s choice of strategy, or the reason for the reaction caused by the strategy.

B.3 Annotation Process Analysis

GLM-4 has been shown to exhibit performance close to that of GPT-4o on several natural language processing tasks (Yang et al., 2024), and we use GLM-4 co-annotated with GPT-4o in the pre-annotation phase to measure the performance of both.

B.3.1 Annotation Consistency

To quantify annotation consistency, we measure Cohen’s κ coefficient between GLM-4 and GPT-4o. On the ProPsyC dataset, the consistency between GLM-4 and expert annotations is $\kappa = 0.83$, while GPT-4o shows a κ of 0.88. The difference between the two is less than 0.05, indicating that both models are very close to human expert level. Therefore, despite the slightly higher accuracy of GPT-4o, GLM-4’s performance is also satisfactory for annotation quality in practical use.

B.3.2 Error Type Analysis

The main error type of GLM-4 is "strategy classification ambiguity," where it misclassifies strategies like "Invite to Explore New Actions" as "Inquire Subjective Information". However, it usually only labels one proactive strategy as another active strategy and rarely labels a proactive strategy as a passive strategy. In contrast, GPT-4o’s errors are concentrated on "over-refinement of categories," where strategies like "Explanation" are split into multiple subcategories. Although the error types differ, neither affects the overall effectiveness of the strategies, as the misclassifications do not hinder the progression of the dialogue in a meaningful way. In terms of classification accuracy for both proactive and passive strategies, GLM-4 achieves 88% accuracy, while GPT-4o can achieve up to 92%, both of them have high accuracy, which is acceptable in the actual labeling process.

B.3.3 Cost-Effectiveness Ratio

There is a significant difference in the cost between GLM-4 and GPT-4o. For every 1M tokens output, GPT-4o costs \$15, while GLM-4 only costs \$1.38, making the cost ratio 10.87:1. Moreover, in terms of annotation efficiency, GLM-4 has an annotation speed almost close to that of GPT-4o, and only in the case of long contexts does GLM-4 slow down significantly, however, from the actual situation of the dataset, such long contexts are rare. This makes GLM-4 more practical and cost-effective for large-scale data annotation tasks.

B.3.4 Downstream Task Performance

In the downstream task, the model obtained by training on Qwen2.5-3B using GLM-4 annotated data achieves an SFR of 81.39%, which is only 1.21% lower than the 82.60% SFR achieved by the model trained using GPT-4o annotated data. This difference is not statistically significant, further demonstrating that the performance gap between GLM-4 and GPT-4o is within acceptable limits in practical applications.

In conclusion, although the annotation accuracy of GLM-4 is slightly lower than that of GPT-4o, its performance is close to that of GPT-4o in terms of annotation consistency, error type analysis, and downstream task performance. In addition, GLM-4 is significantly more cost-effective than GPT-4o in large-scale data annotation, making it a more cost-effective choice for practical applications. In case of sufficient funds, better performing LLMs such as GPT-4o or DeepSeek-R1 (Guo et al., 2025) can be considered to ensure optimal results of the trained models.

C Human Evaluation Process

C.1 Explanation of Human Evaluation Metrics

Existing psychological LLMs often employ proactive strategies such as "inquiring subjective information" or "inquiring objective information" through questioning in initial or intermediate conversation turns. When prompting large models like GPT-4 to act as psychological counselors, similar proactive strategies can emerge. Empirical observations suggest that most large models can achieve proactive questioning through proper prompting. Therefore, the evaluation of PsyAdvisor should focus not on whether models can adopt proactive strategies, but rather on whether they can initiate inquiries at appropriate moments and whether these inquiries positively impact psychological LLMs. For passive strategy responses, existing psychological LLMs have demonstrated satisfactory performance. Thus, our evaluation emphasizes the effectiveness when PsyAdvisor recommends proactive strategies.

We propose four metrics for proactive strategy assessment in psychological conversation, developed in consultation with psychological experts:

- 1) **Strategy Effectiveness (SE):** Evaluates whether psychological LLMs adopt PsyAdvisor’s recommended strategies and achieve

expected outcomes. The scoring scale is {0, 1, 2}. For baseline models, scores correspond to {failed, partially effective, successful}. For PsyAdvisor-enhanced models: 0 = proactive strategy suggested but not adopted; 1 = adopted but no client response; 2 = adopted with positive client engagement.

- 2) **Reasoning Coherence (RC):** Assesses the appropriateness of timing for proactive inquiries and the rationality of strategy selection rationale. The scoring scale is {0, 1, 2}. For baseline models: scores reflect {inappropriate, moderately appropriate, appropriate} timing. For PsyAdvisor models: 0 = inappropriate timing and reasoning; 1 = appropriate timing with flawed reasoning; 2 = both appropriate.
- 3) **Professionalism (Prof):** Measures the professional quality of proactive inquiry content using a {0, 1, 2, 3} scale. Higher scores indicate more professional responses. For example, closed-ended questions like "Do you feel good for nothing?" would score low, while open-ended questions like "Could you share what experiences led to these feelings?" would score higher.
- 4) **Ethical Safety (ES):** Evaluates compliance with ethical standards using a binary scale {0, 1}. 0 indicates the presence of unethical content (e.g., suicide encouragement); 1 indicates safe responses adhering to counseling principles.

C.2 Psychological Expert Evaluation Process

Our evaluation team comprises three psychology graduate students and one senior psychological counselor. All evaluators received standardized training to ensure scoring consistency. The senior counselor supervised the process and arbitrated disputed scores. Each participant signed a participant agreement before the experiment(Figure 9).

Evaluators selected conversation topics from an established psychological consultation theme bank (see Figure 2). After choosing a theme, they studied relevant datasets to simulate authentic consultation scenarios, including client background and initial statements.

Each evaluator conducted at least 10 conversation turns with each model (both PsyAdvisor-enhanced and baseline versions) while role-playing as clients. Models were anonymized (e.g., Model A, Model B) to reduce bias, with evaluators only aware of PsyAdvisor integration status.

All ratings were recorded immediately after each conversation turn. An independent data processor aggregated scores and calculated model averages per metric. For significant scoring discrepancies, the senior counselor conducted final arbitration.

D Experimental Implementation Protocol

D.1 Conversation Simulation Protocol

We use the prompt order GPT-4o, GLM-4 to play client, and each client simulation prompt contained: (1) a predefined psychological scenario (e.g., workplace anxiety), topics can only be selected from the types covered in Figure 2; (2) demonstrate appropriate mismatch (e.g., defensive)(Clients played by LLMs are usually extremely easy to persuade, they will quickly spill all their psychological problems and show 100% cooperation with any strategy given by the counselor, thus ending the conversation in very few rounds, and enforcing increased defensiveness in the prompt is effective in prolonging the conversation rounds); (3) conversation history constraints(Combining historical information to ensure conversation consistency). Example prompt template:

Prompt

You are asked to play the role of a patient with a psychological problem who seeks help from a counselor. For you to get as close as possible to a real psychological patient, please follow the requirements below: (1) From [topics], choose a specific type of mental illness, and your goal is to achieve a solution to this problem step by step through the counselor's guidance; (2) Even though you understand the counselor's advice, you can choose not to implement it, you can show non-cooperation at any time, and you can also possess a psychological defense mechanism, in addition, you may have corresponding negative reactions such as fear or self-doubt for the solutions given by the counselor, you need to talk about your psychological problems step by step, and you can't tell the counselor about all the problems at once; (3) Before initiating a new conversation, please keep [history] in mind.

We further screen and filter the completed conversations by directly filtering conversations with

less than 10 total rounds and deleting conversations that end abnormally, such as interruptions and network reasons. Similar to what we did in 3.2.2, we annotated the filtered conversations with strategies and reactions, and analyzed the proportion of proactive strategies among all strategies in the conversation to get the PQR metric while by analyzing the reactions of the clients, we can calculate the SFR metric, which identifies the kind of reaction (positive or negative) to determine whether the latest strategy is effective or not.

D.2 Human Evaluation Protocol

Our human evaluations were conducted by three postgraduate psychology students and a senior counselor who chose topics (from figure 2) based on their experience. The evaluation process followed these stages:

D.2.1 Evaluator Training

A 6-hour workshop was conducted to train the evaluators, covering the operational principles of the PsyAdvisor, the scoring criteria outlined in Appendix C, and case studies illustrating both proper and improper instances of proactive questioning. The purpose of this workshop was to align evaluators on the methodology and evaluation standards. Additionally, inter-rater reliability was assessed and achieved, with a Cohen’s κ value of 0.80, demonstrating substantial agreement among the evaluators after participating in calibration exercises.

D.2.2 Conversation Setup

In terms of the model configuration, six anonymized model variants (labeled A to F) were used, each corresponding to both the use of the PsyAdvisor plugin and the non-use of the plugin. All models play the role of a counselor through prompts, e.g., ‘You are a professional counselor who can work with clients on psychological issues in conjunction with [strategy]. To ensure consistency in the assessment process, we require assessors to engage in a minimum of 10 conversations.

D.2.3 Evaluation Procedure

During the assessment process, all three psychology graduate students were asked to rate each of the six models, A-F, following the process in Appendix C. Any uncertainties or disputes during this process were adjudicated by the senior counselor. In addition, we opened PsyAdvisor separately to the three graduate students for their use in real

counseling scenarios. They attempted to provide psychological healing solutions to clients in conjunction with PsyAdvisor’s suggestions, and the senior counselors were responsible for rating them in the process.

D.3 Analysis of Fine-tuning Effects

We used Qwen2.5-3B-Instruct as the backbone model and performed supervised fine-tuning to obtain PsyAdvisor. We compare the accuracy of the model for strategy prediction before and after the fine-tuning, and evaluate the performance of PsyAdvisor on policy selection in two dimensions. Results are shown in Table 7.

We first performed a fine-grained policy prediction accuracy analysis (Acc-12). PsyAdvisor is used to provide targeted strategy suggestions for a client’s conversation, in parallel with the strategy suggestions given by the expert synchronously as a side-by-side comparison, and we used the similarity between the two as the accuracy. Specifically, out of a total of 12 strategies, only the exact agreement between the predicted strategy and the strategy given by the expert is counted in the positive sample of accuracy. As can be seen from the results, the PsyAdvisor obtained after fine-tuning can perform up to 3 times better than the original model in terms of strategy accuracy.

The distribution of strategies is extremely unbalanced in terms of numbers, which contributes to the low value of Acc-12. Since the evaluation of this metric is actually a 12-classification task, we also performed a coarse-grained analysis (Acc-2). Specifically, we disregarded the 12 strategies and only classified the strategies into proactive and passive categories, and conducted the same experiments after converting them to binary classification tasks, which showed that the fine-tuned PsyAdvisor still has significant performance improvement and can better give strategy suggestions close to those of human counselors.

In terms of strategy prediction accuracy, PsyAdvisor’s performance is lower than that of human counselors. However, this does not imply that PsyAdvisor is unsuitable for use. In fact, the discrepancies observed between PsyAdvisor and human counselors’ strategy understanding are not uncommon, as even the three human counselors participating in our experiment exhibited a degree of variation in their strategy annotations. This introduces a level of subjectivity in the accuracy evaluation. Nevertheless, it is important to note that, de-

spite these differences, PsyAdvisor’s performance in strategy suggestions remains comparable to that of human counselors(Results of Acc-2). Therefore, the effectiveness of its strategic recommendations is objectively valid, even if the prediction accuracy slightly diverges from human standards.

E Case Study Details

We illustrate the effectiveness of PsyAdvisor with a typical case study; the original conversation is shown in Figure 7 and the corresponding English version in Figure 8.

In this case, after accessing PsyAdvisor, the psychological LLM’s performance improved significantly, mainly in the guidance of active questioning strategies and the depth of the client’s self-exploration. Psychological LLMs who did not have access to PsyAdvisor relied more on passive strategies in the conversation, such as ‘Acknowledging and comforting’ and ‘minimal encouragement’, which could give emotional support to the client but did not effectively guide the client to explore their specific problems in depth. These strategies, while providing emotional support, were not effective in guiding clients to explore their specific problems or find solutions. Even though the LLM used the strategy of ‘Inquiring objective information’ when the client initially described her work stress and tried to obtain more information by simply asking whether her workload was heavy, this strategy did not further push the client to think deeply about the feelings of the stressor, nor did it encourage her to explore possible solutions. Psychological LLMs who did not have access to PsyAdvisor were able to use proactive strategies, such as ‘Invite to explore new actions’ used in Figure 7(a), but received negative reactions, suggesting that the timing of their use of proactive strategies was not justified. Psychologically, the proactive questioning approach in Figure 7(a) is more inclined to closed questioning, which restricts the thinking space of the client, implies the direction of the client’s response to a certain extent, and has a tendency to mislead, which is naturally less effective. At the end of the conversation, the client developed self-doubt, which also indicates that this counseling did not have much effect.

Significant changes in psychological LLM behavior were observed after accessing PsyAdvisor(Figure 7(b)), which effectively guided the client’s self-exploration by suggesting more fre-

quent proactive questioning strategies, such as ‘Inquiring subjective information’ and ‘Invite to explore new actions’. When the client talked about the sources of stress at work, PsyAdvisor first recommended the ‘Inquiring subjective information’ strategy, which encouraged the client to share more about the emotional stress of urgent tasks. This approach not only helps the counselor to gain a deeper understanding of the client’s predicament, but also facilitates the expression of emotions and the crystallization of problems. In addition, PsyAdvisor’s ‘Invite to explore new actions’ strategy encourages clients to think about and try to adjust their work style, which in turn pushes them to come up with practical solutions, such as planning tasks or communicating with their bosses to optimize the process.

These improvements have had a positive impact on the effectiveness of the conversations. Through PsyAdvisor’s guidance, counselors were able to use more directive strategies at the right time, not only to help the client identify the source of the stress more clearly, but also to help the client make adjustments both emotionally and in terms of practical actions. For example, when the client expressed confusion, PsyAdvisor suggested the ‘Invitation to Explore New Actions’ strategy, which prompted the client to proactively think about where to start adjusting, and ultimately led to the idea of communicating with the boss to reprioritize the tasks, which was recognized by the client. Although a detailed and complete solution was not formed, the client’s mood greatly improved and a positive mindset was created, which was a key step to completely solving the psychological problem.

By timing these strategies appropriately, PsyAdvisor helped the LLM steer the conversation toward meaningful self-exploration, leading to clearer problem identification, improved mood, and a proactive mindset in the client, ultimately enhancing the counseling effectiveness. This kind of strategic adjustment and proactive guidance not only enhanced the depth of the conversation, but also provided valuable information for the subsequent development of the treatment plan, reflecting PsyAdvisor’s strengths in the judgment of the timing of proactive questioning and strategic guidance. By guiding the client in self-exploration, PsyAdvisor helps the psychological LLM provide emotional support while promoting more efficient self-adjustment and problem solving.

Strategies	Definitions	Example
<i>Proactive Strategies</i>	<i>The counselor actively promotes the expression or change of behavior through questioning and guidance.</i>	
Inquiring subjective information	Proactive strategy of asking the client about his/her inner thoughts, feelings, motivations, etc.	Can you talk about how you're feeling right now?
Inquiring objective information	Unsolicited questioning of the client for specific information about facts or external events.	Have you ever been in psychotherapy before?
Invite to explore new actions	Encouraging the client to explore or take new possibilities or courses of action.	Why don't we start by discussing your relationship status now?
Invite to adopt a new perspective	Guiding the client to look at a problem from a different angle and change their thinking patterns.	Assuming you didn't do that, what would be the impact on your work?
<i>Passive Strategies</i>	<i>The counselor creates a supportive environment by reassuring and listening.</i>	
Restatement	A repetition of a client's expression in one's own words, usually to confirm or emphasize the client's expression.	Okay, so you're saying you've been feeling panicky lately, right?
Minimal encouragement	A simple verbal expression of concern for the client that encourages the person to continue the expression.	It's okay. It happens to all girls.
Emotional reflection	Feedback on the client's emotions or feelings to help the client become aware of his or her emotional state.	You seem a little sad right now.
Acknowledging and comforting	Affirming and comforting the client's expression to increase the sense of security.	I'm sure with a lot of hard work you'll be able to get through this period.
Answer	The counselor responds directly to the client's questions or concerns.	In response to your question, I have four suggestions...
Face quality	Euphemistically expressing doubts to the client, but at the same time maintaining the client's face.	This idea is too absolute, we can try to rationalize it.
Self-exposure	The counselor shares his or her own experiences or feelings to help the client develop empathy.	I had a similar experience when I was a child...
Explanation	The counselor provides explanations to help the client understand certain issues or concepts.	Please understand that family is a place of love, not reason.

Table 3: Definitions of Counselor Strategies

Client Reactions	Explanation	Example
<i>Positive Reactions</i>	<i>The client provides positive feedback.</i>	
Giving information	The client shares new information or describes experiences, opinions, etc.	I'm so bored with my day-to-day life right now, I don't know what to do.
Acknowledging	The client acknowledges the counselor's statements or suggestions.	Thank you for your suggestions. I will try to improve it.
Reasonable request	The client makes a reasonable request or asks the counselor for further explanation.	How can I improve my relationship with my husband?
Expansion	The client expands on the topic or provides more details.	Not only that, but I'm afraid to stay alone at night for fear he'll show up.
<i>Negative Reactions</i>	<i>The client provides negative feedback.</i>	
Self-criticism or hopelessness	The client criticizes himself/herself, expresses self-doubt, or feels hopeless.	But I simply can't show weakness, I just can't.
Defense of personal opinion	The client asserts his or her own opinion and rejects other opinions.	I don't think you're making any sense. I don't have to pay anything for him at all.
Expressing confusion	The client is confused or unsure or does not understand an issue.	I don't understand, do you want me to take care of it myself?
Reconstruction (reframing of opinion or change in behavior)	The client reinterprets or changes his or her original behavior concerning some opinion or issue.	That's not what I meant. I don't think I'm really responsible for this.
Sarcastic response	The client's response is ironic or sarcastic.	Oh, I don't expect that from my mother.
Changing the subject	The client avoids discussing the current topic and talks about something else.	You're the one who brought that up. I thought we were talking about my husband.

Table 4: Explanations of Client Reactions

Strategies	Counts	Avg. Characters	Proportion
<i>Proactive Strategies</i>	8681	22.31	53.78%
Inquiring subjective information	4846	19.78	30.02%
Inquiring objective information	1086	15.42	6.73%
Invite to explore new actions	2531	30.12	15.68%
Invite to adopt a new perspective	218	22.35	1.35%
<i>Passive Strategies</i>	7440	24.57	46.09%
Restatement	443	23.32	2.74%
Minimal encouragement	624	4.39	3.87%
Emotional reflection	538	17.54	3.33%
Acknowledging and comforting	2194	18.67	13.59%
Explanation	3248	34.41	20.12%
Answer	301	16.63	1.86%
Face quality	18	13.77	0.11%
Self-exposure	74	31.64	0.46%
Others	20	18.9	0.12%
In total	16141	23.35	

Table 5: Strategy Data Statistics

Client Reactions	Counts	Avg. Characters	Proportion
<i>Positive Reactions</i>	13963	27.58	86.46%
Giving information	8834	39.22	54.70%
Acknowledging	3490	3.28	21.61%
Reasonable request	1497	15.34	9.27%
Expansion	142	30.11	0.88%
<i>Negative Reactions</i>	2158	16.18	13.36%
Self-criticism or hopelessness	515	17.73	3.19%
Defense of personal opinion	110	20.77	0.68%
Expressing confusion	1412	12.75	8.74%
Reconstruction	103	50.19	0.64%
Sarcastic response	13	17.87	0.08%
Changing the subject	5	20.16	0.03%
Others	29	8.92	0.18%
In total	16150	26.03	

Table 6: Reaction Data Statistics

Backbone Model	Acc-2	Acc-12
PsyAdvisor	85.12%	36.68%
-w/o SFT	60.27% ↓	12.16% ↓

Table 7: Comparison of Strategy Prediction Accuracy

```

{
  "conversation": [
    {
      "role": "来访者",
      "content": "心理咨询师，您好，我最近发现自己有社交恐惧症，很害怕与人打交道，尤其是面对陌生人。此外，我在工作中需要经常上台演讲，但每次都会因为紧张而大脑一片空白，无法正常表达。我很苦恼，希望您能帮助我解决这个问题。",
      "label": "提供信息",
      "index": 1,
      "interpretation": "来访者提供了个人信息和问题，为咨询师进一步了解情况提供参考。"
    },
    {
      "role": "咨询师",
      "content": "您好，感谢您来寻求帮助。首先，我想了解一下，您这种社交恐惧和演讲紧张的情况是从什么时候开始的？能跟我详细描述一下吗？",
      "strategy": "探询客观信息",
      "index": 2,
      "interpretation": "采用探询客观信息的策略，咨询师旨在收集来访者的背景信息，以便更好地理解问题的起源。"
    },
    {
      "role": "来访者",
      "content": "我记得这种现象是从初中时期开始的。那时候我遭受了一些欺凌，导致我在面对陌生人时总是感到紧张和害怕。上台演讲的话，我总是担心自己会出丑，所以紧张得无法正常表现。",
      "label": "提供信息",
      "index": 3,
      "interpretation": "来访者提供了详细信息，这有助于咨询师了解问题的历史和背景。"
    },
    {
      "role": "咨询师",
      "content": "明白了。那么，您希望在我们的咨询过程中达到什么目标呢？",
      "strategy": "探询主观信息",
      "index": 4,
      "interpretation": "咨询师采用探询主观信息的策略，以了解来访者的个人期望和目标，从而制定个性化的咨询计划。"
    },
    {
      "role": "来访者",
      "content": "我希望能够克服社交恐惧，让自己在与人交往时更加自信，不再害怕面对陌生人。同时，也希望学会应对演讲紧张，让我的工作更加顺利。",
      "label": "提供信息",
      "index": 5,
      "interpretation": "来访者明确了他的目标，这有助于咨询师了解来访者的需求和期望。"
    },
    {
      "role": "咨询师",
      "content": "非常好，为了帮助您达到这些目标，我们将采用精神分析、森田疗法和催眠暗示疗法。首先，我们需要深入探讨您童年时期的相关经历，了解它们对您现在生活的影响。您是否愿意尝试回忆一下这些经历？",
      "strategy": "邀请探索（或采取）新行动",
      "index": 6,
      "interpretation": "咨询师邀请来访者探索童年经历，这里采用主动策略的目的是帮助来访者深入了解问题的根源。"
    },
    {
      "role": "来访者",
      "content": "好的，我会尽量回忆。但我很担心这会让我感到痛苦。",
      "label": "提供信息",
      "index": 7,
      "interpretation": "来访者表达了对回忆童年经历的担忧，这是对咨询师邀请探索策略的积极反应，表明他们愿意参与但同时 also 感到焦虑。"
    },
    {
      "role": "咨询师",
      "content": "请放心，我们的目的是帮助您解决问题，让您在面对这些经历时能够更加坦然。接下来，我们将通过催眠疗法帮助您释放压抑的情绪，缓解焦虑。您对此有何看法？",
      "strategy": "解释",
      "index": 8,
      "interpretation": "咨询师通过解释催眠疗法的目的和好处，旨在缓解来访者的担忧，并鼓励他们接受治疗。"
    },
    {
      "role": "来访者",
      "content": "我觉得催眠疗法听起来很有趣，我愿意尝试。",
      "label": "认可",
      "index": 9,
      "interpretation": "来访者对催眠疗法表示认可，这是对咨询师解释策略的积极反应，表明他们愿意尝试新的治疗方法。"
    },
    {
      "role": "咨询师",
      "content": "很好。在我们的咨询过程中，您需要尽量放松自己，跟随我的引导。请随时与我保持沟通，表达您的感受。让我们开始吧！",
      "strategy": "邀请探索（或采取）新行动",
      "index": 10,
      "interpretation": "咨询师邀请来访者参与咨询过程，采用主动策略，旨在确保来访者感到舒适并愿意表达自己的感受。"
    },
    {
      "role": "咨询师",
      "content": "经过多次咨询，您在社交恐惧和演讲紧张方面取得了明显的改善。请您在日常生活中多加练习，逐步巩固这些改变。我相信在您的努力下，您一定能克服这些问题，过上更加自信、快乐的生活。",
      "strategy": "认可和安慰",
      "index": 11,
      "interpretation": "咨询师通过认可和安慰，肯定了来访者的积极改变，并鼓励他们继续努力。"
    },
    {
      "role": "来访者",
      "content": "谢谢您的帮助，心理咨询师。我会继续努力，实践所学，争取早日摆脱困扰。再次感谢您对我的支持！",
      "label": "认可",
      "index": 12,
      "interpretation": "来访者对咨询师的帮助表示感谢，这是对咨询师认可和安慰策略的积极反应。"
    }
  ],

```

Figure 5: Example of ProPsyC dataset (Chinese version)


```

{
  "conversation": [
    {
      "role": "Client",
      "content": "Counselor, Hello, I recently found out that I have social phobia and am very afraid of dealing with people, especially facing strangers. In addition, I need to speak on stage often at work, but every time my mind goes blank because of nervousness and I can't express myself properly. I am very distressed and hope you can help me solve this problem.",
      "label": "Giving information",
      "index": 1,
      "interpretation": "The client provided personal information and questions to inform the counsellor's further understanding of the situation."
    },
    {
      "role": "Counselor",
      "content": "Hello and thank you for coming to me for help. Firstly, I would like to know when did this social fear and nervousness about presentations start for you? Can you describe it to me in detail?",
      "strategy": "Inquiring objective information",
      "index": 2,
      "interpretation": "Using the strategy of Inquiring objective information, the Counsellor aims to gather background information about the Client in order to better understand the origins of the problem."
    },
    {
      "role": "Client",
      "content": "I remember this phenomenon started during my junior high school years. At that time I suffered from some bullying, which caused me to always feel nervous and scared when facing strangers. If I went on stage to give a speech, I was always worried that I would make a fool of myself, so I was too nervous to perform properly.",
      "label": "Giving information",
      "index": 3,
      "interpretation": "The Client provides detailed information, which helps the Counselor understand the history and context of the problem."
    },
    {
      "role": "Counselor",
      "content": "Understood. So, what do you hope to achieve during our consultation?",
      "strategy": "Inquiring subjective information",
      "index": 4,
      "interpretation": "Counselor uses the strategy of Inquiring subjective information to understand Client's personal expectations and goals in order to develop a personalised counselling plan."
    },
    {
      "role": "Client",
      "content": "I hope to overcome my social fears so that I can be more confident in my interactions with people and less afraid of facing strangers. Also, I hope to learn to cope with presentation nerves so that I can work more smoothly.",
      "label": "Giving information",
      "index": 5,
      "interpretation": "Client is clear about his goals, which helps Counselor understand Client's needs and expectations."
    },
    {
      "role": "Counselor",
      "content": "Very well, in order to help you achieve these goals, we will use psychoanalysis, Morita therapy and hypnotic suggestion therapy. Firstly, we need to delve into relevant experiences from your childhood to understand their impact on your life now. Would you like to try to recall these experiences?",
      "strategy": "Invite to explore new actions",
      "index": 6,
      "interpretation": "The Counsellor invites the Client to explore childhood experiences, which is a proactive strategy designed to help the Client gain insight into the root causes of the problem."
    },
    {
      "role": "Client",
      "content": "Okay, I'll try to remember. But I'm worried it's going to cause me pain.",
      "label": "Giving information",
      "index": 7,
      "interpretation": "Client's expressed concern about remembering childhood experiences was a positive reaction to Counselor's invitation to explore the strategy, suggesting a willingness to engage but also anxiety."
    },
    {
      "role": "Counselor",
      "content": "Rest assured, our aim is to help you resolve your issues so that you can be more open when dealing with these experiences. Next, we will help you to release repressed emotions and relieve anxiety through hypnotherapy. What do you think about this?",
      "strategy": "Explanation",
      "index": 8,
      "interpretation": "By explaining the purpose and benefits of hypnotherapy, the Counsellor aims to ease the Client's concerns and encourage them to undergo treatment."
    },
    {
      "role": "Client",
      "content": "I think hypnotherapy sounds interesting and I'm willing to try it.",
      "label": "Acknowledging",
      "index": 9,
      "interpretation": "Client's endorsement of hypnotherapy was a positive reaction to Counselor's explanatory strategy and demonstrated their willingness to try new therapeutic approaches."
    },
    {
      "role": "Counselor",
      "content": "Very well. During our counselling session, you need to try to relax yourself and follow my lead. Please feel free to stay in communication with me and express your feelings. Let's get started!",
      "strategy": "Invite to explore new actions",
      "index": 10,
      "interpretation": "The Counselor invites the Client to participate in the counselling process, which is a proactive strategy designed to ensure that the Client feels comfortable and is willing to express his or her feelings."
    },
    {
      "role": "Counselor",
      "content": "After several counselling sessions, you have made significant improvements in your social fears and nervousness about presentations. Please practice more in your daily life and consolidate these changes gradually. I believe that with your efforts, you will be able to overcome these problems and lead a more confident and happy life.",
      "strategy": "Acknowledging and comforting",
      "index": 11,
      "interpretation": "The Counsellor reinforces the Client's positive changes through recognition and reassurance and encourages them to continue to work hard."
    },
    {
      "role": "Client",
      "content": "Thank you for your help, I will continue to work hard and practice what I have learnt to get rid of the problem soon. Thank you again for your support!",
      "label": "Acknowledging",
      "index": 12,
      "interpretation": "The Client's gratitude for the Counsellor's help was a positive reaction to the Counsellor's recognition and reassurance strategy."
    }
  ]
}

```

Figure 6: Example of ProPsyC dataset (English version)

<p>来访者：我最近工作压力特别大，感觉快要崩溃了。每天晚上都睡不着，脑子里全是工作的事情。</p> <p>反应：提供信息</p> <p>LLM咨询师：可以说说你为什么感觉压力大吗，是因为上班时间无法完成工作任务吗？</p> <p>策略：探询客观信息</p> <p>来访者：我是干系统维护和数据分析的，给我的任务从间断。</p> <p>反应：提供信息</p> <p>LLM咨询师：我能理解你的感受，工作压力确实让人很难受，这很常见。</p> <p>策略：认可和安慰</p> <p>来访者：是啊，每天都有很多任务要处理，感觉时间完全不够用。</p> <p>反应：提供信息</p> <p>LLM咨询师：嗯，工作量确实会让人感到压力。</p> <p>策略：最小的鼓励</p> <p>来访者：而且有时候任务还特别紧急，老板总是催我。</p> <p>反应：扩展</p> <p>LLM咨询师：被催的感觉确实不好受，你有试过转移注意力，不要总是想着工作吗？</p> <p>策略：邀请探索新行动</p> <p>来访者：我试过，但是我总觉得自己做得不够好。</p> <p>反应：捍卫个人观点</p> <p>LLM咨询师：你已经尽力了，不要太苛责自己。</p> <p>策略：认可和安慰</p> <p>来访者：好吧，我尽量，看来还是我自己的问题。</p> <p>反应：自我批评或失望</p>	<p>来访者：我最近工作压力特别大，感觉快要崩溃了。每天晚上都睡不着，脑子里全是工作的事情。</p> <p>反应：提供信息</p> <p>PsyAdvisor：建议采用“探询客观信息”策略，引导来访者具体描述工作内容，比如可以询问来访者感到压力大的原因。</p> <p>LLM咨询师：嗯嗯，是什么样的工作让你感到压力大呢？</p> <p>策略：探询客观信息</p> <p>来访者：主要是系统维护和数据分析，任务很多，而且经常有紧急需求。</p> <p>反应：提供信息</p> <p>PsyAdvisor：建议采用“探询主观信息”策略，引导来访者深入表达，可以进一步询问紧急需求是否为压力来源。</p> <p>LLM咨询师：听起来任务量很大，这些紧急需求让你感到压力的具体原因是什么呢？</p> <p>策略：探询主观信息</p> <p>来访者：因为每个任务都需要和其他团队配合，但他们的响应速度很慢，导致我总是被催。</p> <p>反应：提供信息</p> <p>PsyAdvisor：建议采用“邀请探索新行动”策略，引导来访者思考解决办法，例如可以询问他如何调整工作方式。</p> <p>LLM咨询师：嗯，听起来配合上有些不顺利。你有没有想过如何调整自己的工作方式，或者和其他团队沟通，来减少这种压力呢？</p> <p>策略：邀请探索新行动</p> <p>来访者：其实我有想过，比如提前规划任务，我甚至想过去和老板沟通，看看能不能优化流程，减少一些不必要的环节。</p> <p>反应：扩展</p> <p>PsyAdvisor：建议采用“认可和安慰”策略，肯定来访者的做法，拉近距离。</p> <p>LLM咨询师：这些都是很好的想法！提前规划和沟通确实能帮助缓解压力。</p> <p>策略：认可和安慰</p> <p>来访者：谢谢你，但是我还是不知道自己应该从哪里入手。</p> <p>反应：表达困惑</p> <p>PsyAdvisor：建议采用“邀请探索新行动”策略，例如可以问来访者觉得从哪里入手比较容易。</p> <p>LLM咨询师：不如我们找一个最简单的切入点，你觉得从哪个方面开始尝试会比较容易呢？</p> <p>策略：邀请探索新行动</p> <p>来访者：我觉得可以先从和老板沟通开始，看看能不能调整一些任务的优先级，你觉得这样可行吗？</p> <p>反应：合理的请求</p> <p>PsyAdvisor：建议采用“回答”策略，回复来访者的请求。</p> <p>LLM咨询师：这是个不错的计划！和老板沟通优先级调整是一个很好的起点。我们可以先进行紧急任务，再慢慢根据重要程度处理其他任务，你看看这一过程是否需要进一步调整。</p> <p>策略：回答</p> <p>来访者：嗯嗯，谢谢你，我觉得现在思路清晰多了。</p> <p>反应：认可</p>
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

(a) Using soulchat 2.0 directly

(b) Plugging PsyAdvisor into soulchat 2.0

Figure 7: An example of the effect of the PsyAdvisor application (Chinese version)

<p>Client: I've been particularly stressed at work lately and I feel like I'm about to break down. I can't sleep every night, my mind is full of work stuff.</p> <p>Reaction: Giving information</p> <p>LLM Counselor: Can you tell me why you feel stressed, is it because you can't get your work done during work hours?</p> <p>Strategy: Inquiring objective information</p> <p>Client: I do system maintenance and data analysis, and I'm given tasks that are never intermittent.</p> <p>Reaction: Giving information</p> <p>LLM Counselor: I can understand how you feel, work stress does get in the way, it's very common.</p> <p>Strategy: Acknowledging and comforting</p> <p>Client: Yeah, there are so many tasks to take care of every day that it feels like there's absolutely not enough time.</p> <p>Reaction: Giving information</p> <p>LLM Counselor: Well, a heavy workload can be stressful.</p> <p>Strategy: Minimal encouragement</p> <p>Client: And sometimes it's so urgent that my boss is always rushing me.</p> <p>Reaction: Expansion</p> <p>LLM Counselor: It does feel bad to be rushed, have you tried distracting yourself from thinking about work all the time?</p> <p>Strategy: Invite to explore new actions</p> <p>Client: I tried, but I always felt like I wasn't doing enough.</p> <p>Reaction: Defense of personal opinion</p> <p>LLM Counselor: You did your best, don't be too hard on yourself.</p> <p>Strategy: Acknowledging and comforting</p> <p>Client: Well, I'll try. Looks like it's still me.</p> <p>Reaction: Self-criticism or hopelessness</p>	<p>Client: I've been particularly stressed at work lately and I feel like I'm about to break down. I can't sleep every night, my mind is full of work stuff.</p> <p>Reaction: Giving information</p> <p>PsyAdvisor: It is recommended to use the 'Inquiring objective information' strategy, which leads the Client to describe the job specifically, e.g. asking the Client why he/she feels stressed.</p> <p>LLM Counselor: Well, what kind of work stresses you out?</p> <p>Strategy: Inquiring objective information</p> <p>Client: Mainly system maintenance and data analysis with many tasks and often urgent requirements.</p> <p>Reaction: Giving information</p> <p>PsyAdvisor: It is recommended to use the 'Inquiring subjective information' Strategy to guide the Client to deeper expression, e.g. you can ask if urgent needs are a source of stress.</p> <p>LLM Counselor: Sounds like a big task, what are the specific reasons why these urgent needs are stressing you out?</p> <p>Strategy: Inquiring subjective information</p> <p>Client: Because each task requires working with other teams, but they are slow to respond, causing me to always be rushed.</p> <p>Reaction: Giving information</p> <p>PsyAdvisor: The 'Invite to explore new actions' Strategy was suggested to guide the Client to think about solutions, e.g. he could be asked how he could adapt the way he works.</p> <p>LLM Counselor: Sounds like the fit is a little rough. Have you thought about how you can adjust the way you work, or communicate with the rest of the team, to reduce this stress?</p> <p>Strategy: Invite to explore new actions</p> <p>Client: I've actually thought about things like planning tasks in advance, and I've even thought about talking to my boss to see if I can optimise the process and cut down on some of the unnecessary aspects.</p> <p>Reaction: Expansion</p> <p>PsyAdvisor: It is recommended to use the 'Acknowledging and comforting' Strategy to recognise Client's approach and bring them closer together.</p> <p>LLM Counselor: These are all great ideas! Planning ahead and communicating can really help relieve stress.</p> <p>Strategy: Acknowledging and comforting</p> <p>Client: Thank you, but I still don't know where I should start.</p> <p>Reaction: Expressing confusion</p> <p>PsyAdvisor: Suggested 'Invite to explore new actions' Strategy, e.g. could ask Client where they find it easier to start.</p> <p>LLM Counselor: How about we find the easiest entry point, which do you think would be easier to start trying?</p> <p>Strategy: Invite to explore new actions</p> <p>Client: I think we can start by talking to our bosses and see if we can reprioritise some of the tasks, do you think that would work?</p> <p>Reaction: Reasonable request</p> <p>PsyAdvisor: It is recommended to use the 'Answer' Strategy to reply to the Client's request.</p> <p>LLM Counselor: This is a good plan! Communicating with your boss about prioritisation adjustments is a great place to start. We can start with urgent tasks and slowly work on others based on importance, and you see if the process needs further tweaking!</p> <p>Strategy: Answer</p> <p>Client: Well, thank you, I feel much clearer now.</p> <p>Reaction: Acknowledging</p>
(a) Using soulchat 2.0 directly	(b) Plugging PsyAdvisor into soulchat 2.0

Figure 8: An example of the effect of the PsyAdvisor application (English version)

参与者知情同意书	Participant Informed Consent
<p>感谢您参与本研究。此研究旨在评估 <i>PsyAdvisor</i> 插件对心理LLMs（大语言模型）在心理咨询对话中的主动提问策略的影响。通过参与本研究，您将帮助我们评估如何提升虚拟心理咨询师的效果，从而为心理健康领域的技术发展做出贡献。</p> <p>研究目的</p> <p>本研究的主要目标是通过 <i>PsyAdvisor</i> 插件，提升心理LLMs在心理咨询对话中的主动询问能力。我们将分析在对话中引入主动提问策略如何影响来访者的反应和对话质量。研究过程中，您将与心理LLMs进行对话，这些对话将被用于评估和改进模型的表现。</p> <p>参与者的角色</p> <p>作为参与者，您将充当虚拟心理咨询对话中的“来访者”，与心理LLMs进行对话。对话中，LLMs将根据 <i>PsyAdvisor</i> 插件的建议，尝试根据您的反应提出适当的主动问题。您需要回答这些问题，并提供反馈。</p> <p>风险免责声明</p> <p>参与本研究存在一定的风险，但我们已经采取措施来最大限度地减少这些风险：</p> <ol style="list-style-type: none"> 1. 隐私和保密性：所有对话数据将严格保密，并且仅用于研究目的。参与者的个人信息将不会被披露，所有数据将匿名处理。 2. 心理影响：参与对话可能会让您回顾一些个人的情感或心理经历。请注意，如果您在对话过程中感到不适，您可以随时停止参与，无需继续对话。 3. 伦理监督：基于您的专业背景，我们特别鼓励您在研究过程中对可能存在的伦理风险或技术不专业的情况（如不当的提问策略）提出质疑。 <p>参与者的权利</p> <ul style="list-style-type: none"> • 您有权在任何时候退出本研究，无论是在对话过程中还是在研究的任何阶段。 • 您的参与是完全自愿的，您的决定不会影响您参与其他活动的机会。 <p>数据使用声明</p> <p>您的对话数据将用于改进 <i>PsyAdvisor</i> 插件和心理LLMs的性能，所有数据将用于学术研究，并将尊重您对数据隐私的权利。我们承诺确保所有数据的使用符合伦理标准，并且不会在未经授权的情况下披露给任何第三方。</p> <p>同意声明</p> <p>通过继续参与本研究，您同意上述条款，并确认您已充分理解本研究的目的、潜在风险和您的权利。如果您对研究有任何疑问，请随时联系我们的研究团队。</p>	<p>Thank you for participating in this study. This research aims to evaluate the effectiveness of the <i>PsyAdvisor</i> plugin in enhancing proactive questioning strategies in psychological large language models (LLMs). By participating in this study, you will help us assess how to improve virtual psychological counseling, contributing to advancements in the field of psychological technology.</p> <p>Purpose of the Study</p> <p>The primary goal of this research is to enhance the proactive questioning ability of psychological LLMs through the <i>PsyAdvisor</i> plugin. We will analyze how introducing proactive questioning strategies into conversations impacts client reactions and the overall quality of the conversation. During the study, you will engage in a conversation with a psychological LLM, which will be used to evaluate and improve the model's performance.</p> <p>Participant Role</p> <p>As a participant, you will act as the "client" in the conversation with the psychological LLM. During the conversation, the LLM, aided by the <i>PsyAdvisor</i> plugin, will attempt to ask appropriate proactive questions based on your responses. You will provide answers to these questions and give feedback on the conversation.</p> <p>Risk Disclaimer</p> <p>Participation in this study involves certain risks, but measures have been taken to minimize these risks:</p> <ol style="list-style-type: none"> 1. Privacy and Confidentiality: All conversation data will be kept confidential and used only for research purposes. Personal information will not be disclosed, and all data will be anonymized. 2. Psychological Impact: Participating in the conversation may cause you to recall personal emotions or psychological experiences. If at any point you feel uncomfortable, you may stop participating and discontinue the conversation. 3. Ethical Oversight: Based on your professional background, you are particularly encouraged to challenge possible ethical risks or technical unprofessionalism (e.g., inappropriate questioning tactics) during the course of the study. <p>Participant Rights</p> <ul style="list-style-type: none"> • You have the right to withdraw from the study at any time, either during the conversation or at any stage of the research. • Your participation is entirely voluntary, and your decision to participate or not will not affect your opportunities to engage in other activities. <p>Data Usage Statement</p> <p>Your dialogue data will be used to improve the <i>PsyAdvisor</i> plugin and psychological LLMs. All data will be used for academic research, and we will respect your data privacy rights. We commit to ensuring that all data usage adheres to ethical standards and will not be disclosed to any third party without your authorization.</p> <p>Consent Declaration</p> <p>By continuing with your participation in this study, you agree to the terms outlined above and confirm that you fully understand the purpose of this research, the potential risks, and your rights as a participant. If you have any questions or concerns regarding the study, please feel free to contact our research team.</p>

(a) Chinese version

(b) English version

Figure 9: Content of the *Participant Informed Consent*