PSYCHO-R1: ADVANCING PROACTIVE GUIDANCE FOR LONG-HORIZON PSYCHOLOGICAL COUNSELING

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-based counselors can generate empathetic and professionally grounded responses, yet they struggle to proactively guide therapeutic progress across multi-turn interactions—a capability essential to effective counseling. This limitation stems in part from insufficient modeling of dynamic decision-making under information asymmetry, which requires state inference of clients to guide proactive inquiry and intervention. It is further compounded by trajectory-level advantages during optimization, which provide coarse supervision and fail to distinguish the contributions of individual counselor turns to subsequent conversational progress. To this end, we introduce PSYCHO-R1, a proactive counselor agent trained via a GRPO variant in an agentic interactive framework with three components: 1) an information-asymmetric multi-agent environment with a client simulator conditioned on a private profile and an evaluator providing turn-level rewards; 2) hierarchical decision-making linking state inference to proactive guidance; and 3) critic-free credit assignment using local lookahead returns to compute turn-level group-relative advantages. Experimental results demonstrate that PSYCHO-R1 achieves the best performance across all 11 dimensions spanning counselor-level, client-level, and therapeutic alliance metrics, outperforming both specialized mental-health models and substantially larger general-purpose LLMs.

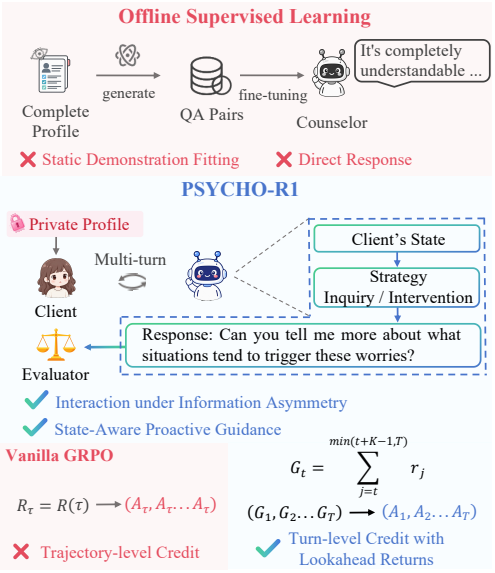


Figure 1: Comparison of existing methods and PSYCHO-R1, highlighting information-asymmetric interaction, state-aware proactive guidance, and turn-level credit from local lookahead returns.

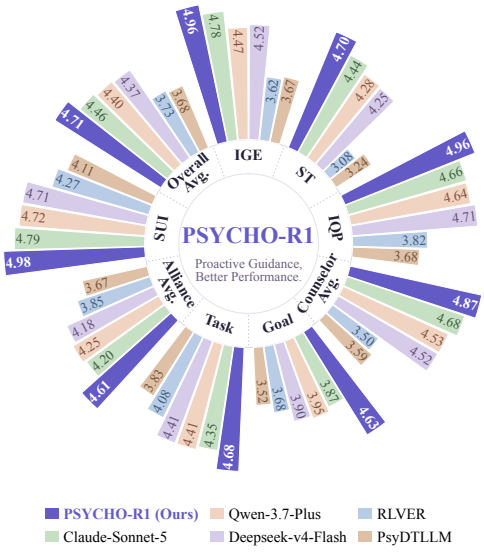


Figure 2: Performance of PSYCHO-R1 and representative baselines across counselor-level, client-level, and therapeutic alliance metrics. Detail Results in Table 1. PSYCHO-R1 achieves the highest scores across all displayed metrics.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Effective counseling relies on the proactive guidance of the therapeutic process, wherein counselors progressively develop a nuanced understanding of the client to facilitate targeted inquiry and timely intervention (Beck, 2020). Recent advances in large language models (LLMs) have substantially strengthened their capabilities in emotional resonance and empathetic understanding, opening new opportunities for psychological counseling (Liu et al., 2023; Qiu et al., 2024). Building on these advances, researchers have incorporated domain-specific expertise into LLMs, giving rise to specialized counselor agents capable of delivering professionally grounded therapeutic responses (Hu et al., 2025; Dai et al., 2026; Xiao et al., 2025b). However, empathy and professional expertise do not necessarily enable an LLM-based counselor to guide the counseling process effectively (Norcross & Wampold, 2018). Achieving such proactive guidance entails three interrelated challenges.

First, real-world counseling demands proactive guidance through **dynamic interaction under information asymmetry**, requiring counselors to progressively elicit clients' initially unobservable experiences and cognitive patterns over multiple turns (Hill, 1999; Farber, 2003; Liu et al., 2026). However, existing methods typically train models on synthetic corpora derived from complete client profiles, relying on offline supervised learning to fit demonstration responses within fixed dialogue histories (Lee et al., 2024; Zhou et al., 2025). This static paradigm deprives models of the opportunity to process real-time feedback across multi-turn exchanges, significantly hindering their ability to adapt their counseling strategies to uncover missing information and execute timely interventions.

Second, proactive guidance requires **state-aware decision-making**: the counselor must infer the client's current psychological state to determine what to explore and when to intervene. Yet, prevailing counselors often generate responses directly from raw dialogue histories (Zhang et al., 2024; Xie et al., 2025), lacking an explicit mechanism to bridge state inference and strategy selection. This disconnect often leads to contextually inappropriate counselor actions, such as premature interventions or unnecessary inquiries, which ultimately hinder effective guidance —consistent with the lower *Strategy & Timing (ST)* scores of existing models in Figure 2.

Third, training LLM-based counselors through multi-turn interactions introduces a **long-horizon credit assignment** problem: the credit assigned to each counselor turn should reflect its contribution to subsequent conversational progress. RLVER (Wang et al., 2026b) built on vanilla GRPO derives a shared trajectory-level advantage from terminal emotion rewards, providing limited differentiation between the contributions of individual counselor turns.

To address these challenges, we introduce PSYCHO-R1, a counselor agent trained with a variant of Group Relative Policy Optimization (GRPO) (Shao et al., 2024) in an agentic interactive framework. This framework comprises three components. 1) To model realistic, dynamic interaction under information asymmetry, we construct a multi-agent environment in which PSYCHO-R1 engages in multi-turn exchanges with an LLM-based client conditioned on a private profile inaccessible to the counselor, while a counseling evaluator assesses each counselor action and provides a turn-level reward. 2) To enable state-aware proactive progression, we introduce a hierarchical decision-making mechanism that infers the client's psychological state and selects an appropriate proactive counseling strategy to guide response generation. 3) To support effective credit assignment over long-horizon interactions, we extend GRPO with turn-level credit assignment mechanism based on local lookahead returns. For each counselor turn, the mechanism aggregates downstream rewards within a lookahead window to estimate the turn's contribution to subsequent conversational progress and derives a turn-level group-relative advantage to provide fine-grained supervision for policy optimization.

We evaluate PSYCHO-R1 in a dynamic interactive environment from three complementary perspectives: counselor-level, client-level and therapeutic alliance. As illustrated in Figure 2, **PSYCHO-R1 achieves the best results across all 11 evaluation dimensions**, outperforming the strongest specialized baseline, RLVER, by 0.98 points in overall score and the substantially larger Claude-Sonnet-5 by 0.25 points. Beyond interactive counseling, **PSYCHO-R1 also consistently improves upon its backbone across three psychology QA benchmarks**, including a 7.24% increase on CPsyExam-MAQ, while largely preserving its general capabilities. These results demonstrate the effectiveness of PSYCHO-R1 in providing proactive guidance throughout multi-turn counseling.

Our contributions are summarized as follows:

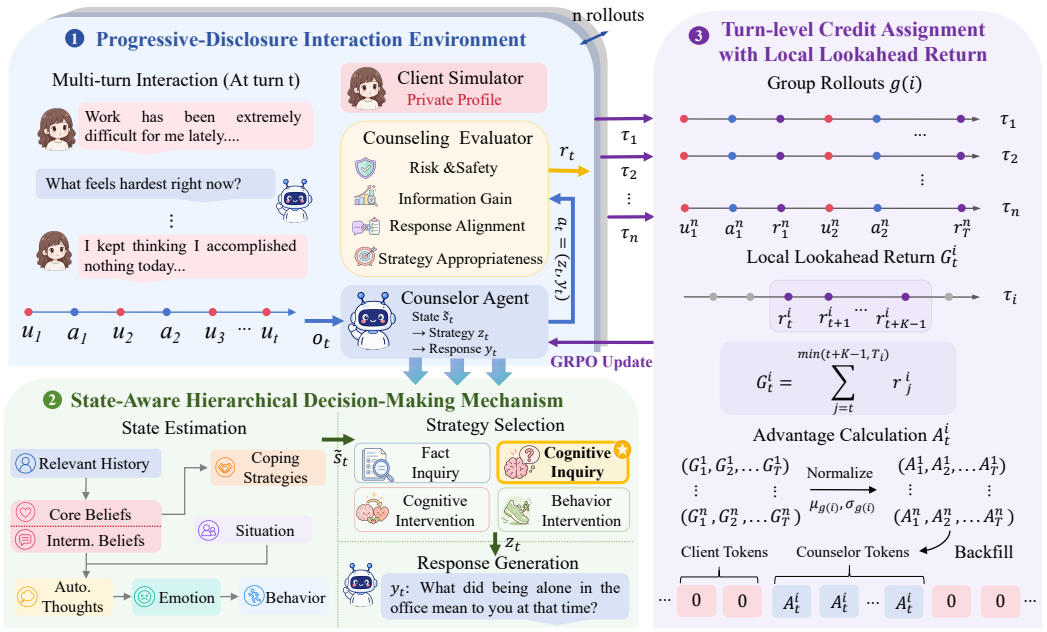


Figure 3: Interactive training loop of PSYCHO-R1. 1) *Progressive-Disclosure Interaction Environment* simulates gradual disclosure and provides turn-level rewards. 2) *State-Aware Hierarchical Decision-Making Mechanism* links state, strategy, and response. 3) *Turn-Level Credit Assignment with Local Lookahead Returns* assigns fine-grained advantages for GRPO updates.

- We develop PSYCHO-R1, a state-aware proactive counselor agent featuring a hierarchical decision-making mechanism. It infers the client's psychological state to select inquiry or intervention strategies, effectively steering the therapeutic progression.
- We propose an agentic reinforcement learning framework for long-horizon counseling that models information asymmetry and supports fine-grained policy optimization through turn-level credit assignment based on local lookahead returns.
- Interactive evaluations demonstrate that PSYCHO-R1 outperforms specialized mental-health models and larger general-purpose LLMs, such as Claude-Sonnet-5, across all 11 dimensions.

2 METHOD

In this section, we present PSYCHO-R1, a counselor agent trained via agentic reinforcement learning to provide proactive guidance in multi-turn counseling interactions. As shown in Figure 3, our approach comprises three components. First, Section 2.1 introduces the **Progressive-Disclosure Interaction Environment**, where an LLM-based client simulator incrementally discloses information from a private profile to model the underlying information asymmetry, while a counseling evaluator assesses each counselor action and provides turn-level rewards. Subsequently, Section 2.2 describes a **State-Aware Hierarchical Decision-Making Mechanism** that enables PSYCHO-R1 to infer the client's psychological state and select a proactive counseling strategy to guide response generation. Finally, Section 2.3 develops **Turn-Level Credit Assignment with Local Lookahead Returns**, an extension of GRPO that aggregates downstream rewards over a local lookahead window to estimate each counselor turn's contribution to dialogue progress and propagates the resulting turn-level group-relative advantage to all tokens in the corresponding response.

2.1 PROGRESSIVE-DISCLOSURE INTERACTION ENVIRONMENT

Following the agentic RL paradigm, we formulate psychological counseling as a Partially Observable Markov Decision Process (POMDP) (Lai et al., 2026). This formulation is well-suited to modeling the sequential and information-asymmetric nature of counseling, where the counselor cannot

directly observe the client's psychological state[1] but must infer it from a history of observations and make decisions accordingly. Formally, we characterize the interaction dynamics within this environment as a POMDP defined by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{R} \rangle$, with each component specified as follows: 1) **States ($\mathcal{S}$):** A state $s \in \mathcal{S}$ represents a complete and private profile reflecting the client's psychological state. 2) **Actions ($\mathcal{A}$):** An action $a_t \in \mathcal{A}$ is the counselor's decision at turn t, comprising a proactive counseling strategy z_t and its corresponding response y_t. 3) **Observations ($\mathcal{O}$):** An observation $o_t \in \mathcal{O}$ represents the dialogue history up to turn t, which comprises all client utterances observed by that point and all preceding counselor actions, each including the selected strategy and its corresponding response. At each step t, the counselor takes action a_t and subsequently observes the client's utterance u_{t+1}, with the resulting exchange extending the dialogue history from o_t to o_{t+1}. 4) **Reward Function ($\mathcal{R}$):** The turn-level reward $r_t = \mathcal{R}(o_t, a_t)$ quantifies the quality of the counselor's action a_t within the dialogue context o_t. We instantiate the POMDP through three specialized agents, each serving a distinct role in the interaction:

- **Client Simulator:** This agent simulates client behavior via an independent LLM conditioned on the client's private state $s \in \mathcal{S}$. At turn t, the agent generates a response based on s, the current observation o_t, and the counselor action a_t, thereby extending the dialogue history to o_{t+1}.
- **Counselor Agent (π_θ):** This is our target agent. Its policy, $\pi_\theta(a_t|o_t)$, maps the current observation o_t to a distribution over counseling actions $a_t \in \mathcal{A}$. Action generation proceeds hierarchically: the policy first infers a client-state estimate $\tilde{s}_t$ from o_t, then selects a counseling strategy z_t based on this estimate, and finally generates a strategy-conditioned response y_t.
- **Counseling Evaluator:** At each turn t, this evaluator assesses the counselor's action a_t within the current dialogue context o_t along four dimensions: *strategy appropriateness*, *information gain*, *risk and safety*, and *response alignment*. The resulting turn-level reward r_t is computed as the weighted sum of four scores, and then aggregated over a k-turn lookahead window for turn-level credit assignment.

In this environment, the counselor agent and client simulator interact over multiple turns, while the evaluator assigns a turn-level reward to each counselor action. This setup supports learning adaptive inquiry and intervention strategies from client responses and evaluator rewards, enabling proactive guidance under information asymmetry.

2.2 State-Aware Hierarchical Decision-Making Mechanism

Proactive counseling involves continually translating an evolving understanding of the client into a counseling strategy and a corresponding response. We model this process as a hierarchical decision-making mechanism comprising three stages: *state estimation*, *strategy selection*, and *response generation*. At each turn t, the counselor agent first estimates the client's psychological state $\tilde{s}_t$ and selects a counseling strategy z_t through an internal reasoning process. It then generates a response y_t under the guidance of the selected strategy, yielding the action $a_t = \langle z_t,\, y_t \rangle$.

State Estimation. Since the client's latent psychological state is not directly observable, the counselor agent constructs a turn-specific state estimate conditioned on the current observation o_t:

$$\tilde{s}_t \sim f_\theta(\cdot \mid o_t). \tag{1}$$

The client's psychological state is represented using the CCD (Beck, 2020) as shown in Figure 3, with this representation refined throughout the interaction to guide subsequent strategy selection.

Strategy Selection. Based on the inferred client's state $\tilde{s}_t$, the counselor agent selects a high-level, proactive counseling strategy z_t:

$$z_t \sim g_\theta(\cdot \mid \tilde{s}_t),\ z_t \in \mathcal{Z}. \tag{2}$$

The strategy space $\mathcal{Z}$ is organized around two complementary counseling functions: inquiry, which refines the state estimate by eliciting additional information, and intervention, which leverages the inferred state to address identified therapeutic targets. Within the inquiry-oriented strategies, **Fact**

[1]We define the psychological state with eight components drawn from the Cognitive Conceptualization Diagram (CCD) (Beck, 2020): relevant history, core beliefs, intermediate beliefs, coping strategies, situation, automatic thoughts, emotion, and behavior (see Appendix B for details).

Inquiry (FQ) focuses on contextual and behavioral information, such as the client's situation, relevant history, coping strategies, and behavior, whereas **Cognitive Inquiry (CQ)** explores cognitive and affective information, including beliefs, automatic thoughts, and emotions. Among the intervention-oriented strategies, **Cognitive Intervention (CI)** facilitates cognitive change by examining, challenging, and reframing maladaptive cognitions, while **Behavior Intervention (BI)** supports behavioral change through concrete, actionable guidance. Together, these four strategies constitute the strategy space $\mathcal{Z} = \{\mathrm{FQ}, \mathrm{CQ}, \mathrm{CI}, \mathrm{BI}\}$, enabling the counselor to adaptively balance information elicitation and therapeutic intervention with its understanding of the client. Detailed definitions of strategies are provided in Appendix C.

Response Generation. Finally, given the current observation o_t, inferred client state $\tilde{s}_t$, and selected strategy z_t, the counselor agent generates a response y_t:

$$y_t \sim p_\theta(\cdot \mid o_t, \tilde{s}_t, z_t)\,. \tag{3}$$

Together, f_θ, g_θ, and p_θ constitute the three hierarchical stages of the unified counselor policy π_θ.

2.3 Turn-Level Credit Assignment with Local Lookahead Returns

Existing critic-free agentic RL approaches for long-horizon interactions typically rely on coarse trajectory-level rewards and advantages, making it challenging to disentangle each turn's contribution to session-level progress (Wang et al., 2026b; Zhang et al., 2026). To address this, we introduce a credit assignment mechanism based on GRPO that computes a local lookahead return for each counselor turn and assigns turn-level advantage to the corresponding response tokens, enabling fine-grained credit attribution to individual counseling decisions across multi-turn interactions.

Turn-level Reward. The counseling evaluator generates a turn-level reward based on four distinct dimensions: 1) *Strategy appropriateness*, which evaluates the appropriateness of the strategy given the current therapeutic stage; 2) *Information gain*, which assesses the response's contribution to advancing the counseling process; 3) *Risk and safety*, which measures the response's effectiveness in addressing potential risks while avoiding unsafe advice; and 4) *Response alignment*, which measures the consistency of the generated response with the chosen strategy. Each dimension is assigned a discrete score from $\{-1, 0, 1\}$. The final turn-level reward is computed as the weighted sum of these scores, with corresponding weights of 0.4, 0.4, 0.1, and 0.1. The prompt for the counseling evaluator is provided in Appendix H.

Local Lookahead Return. Counseling decisions can affect subsequent dialogue progression, and optimizing only immediate rewards may overlook these delayed effects (Zhou et al., 2024). We therefore define a local lookahead return that aggregates rewards over the subsequent k turns, providing a more informative learning signal for each decision's downstream impact.

In trajectory i, the local lookahead return at turn t, denoted as $G_t^{(i)}$, is defined as

$$G_t^{(i)} = \sum_{j=t}^{\min(t+k-1,T^{(i)})} r_j^{(i)}, \tag{4}$$

where $r_j^{(i)}$ denotes the turn-level reward at turn j, $T^{(i)}$ denotes the max turns in trajectory i, and k represents the lookahead window size.

Advantage Calculation. Local lookahead returns may exhibit substantial variability across sampled trajectories, potentially yielding high-variance policy-gradient estimates (Sutton et al., 1998). Accordingly, we sample N trajectories from the same initial client query to form a rollout group and normalize their local lookahead returns within the group to obtain turn-level relative advantages:

$$A_t^{(i)} = \frac{G_t^{(i)} - \mu_{g(i)}}{\sigma_{g(i)} + \epsilon}, \tag{5}$$

where $\mu_{g(i)}$ and $\sigma_{g(i)}$ denote the empirical mean and standard deviation computed over all valid $G_t^{(i)}$ in rollout group $g(i)$, and $\epsilon = 10^{-6}$ ensures numerical stability.

Advantage Backfilling. The advantages derived above are defined at the counselor-turn level. To incorporate them into GRPO's token-level objective, we assign each turn's advantage to every token in its corresponding counselor response. This mapping allows different turns within the same

trajectory to receive distinct optimization signals, linking token-level policy updates to the local lookahead feedback attributed to each turn.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Client Profile. We construct seed client profiles from user posts in the PsyQA dataset (Sun et al., 2021), which were collected from the Q&A section of Yixinli[2], an online mental health support platform. Each post contains a client's question and a detailed description of their concerns. These profiles are split into 2,000 for supervised fine-tuning, 1,200 for agentic RL, and 150 for evaluation.

Baselines. We compare PSYCHO-R1 with two categories of baselines: (1) general-purpose LLMs, including Claude-Sonnet-5, Gemini-3.5-Flash, Qwen-3.7-Plus, and DeepSeek-V4-Flash; and (2) domain-specific mental health LLMs, including Mechat (Qiu et al., 2024), CPsyCounX (Zhang et al., 2024), PsyDTLLM (Xie et al., 2025), PsyLLM (Hu et al., 2025), Psyche-R1 (Dai et al., 2026), and RLVER (Wang et al., 2026b).

SFT Initialization. We initialize PSYCHO-R1 through supervised fine-tuning (SFT) of Qwen3-8B[3] on GPT-5.5-generated dialogues. Counselor outputs follow a structured CoT format: *<think>...</think>* [*Strategy*]: *Response*. The reasoning block contains an updated Cognitive Conceptualization Diagram (CCD), the selected counseling strategy, and its rationale.

Agentic RL Training. Starting from the SFT checkpoint, we train the counselor using our GRPO variant within veRL[4]. GPT-OSS-20B[5] serves as both the client simulator and counseling evaluator. Each batch contains 48 queries, with $N = 8$ trajectories per query and at most 25 counselor turns per trajectory. The maximum prompt and response lengths are 1024 and 10240 tokens, respectively; overlength prompts are filtered out. We set the lookahead horizon to $k = 3$ and optimize the policy using AdamW with a learning rate of 10^{-6}. Training runs for two epochs on six NVIDIA RTX Pro 6000 GPUs using bfloat16 precision.

Evaluation Metrics. Evaluation is also conducted in an interactive environment (Wang et al., 2025; Zhu et al., 2026; Wu et al., 2026), where PSYCHO-R1 engages in multi-turn dialogues with a client simulator. Inspired by Pan et al. (2026), we use GPT-5.5 to evaluate each counseling dialogue on 33 items organized into 11 dimensions across three aspects: (1) **Counselor-Level Evaluation**, covering *Empathic Understanding & Attitude*, *Information Gathering & Exploration*, *Strategy & Timing*, *Session Coherence & Progression*, and *Intervention Quality & Professionalism*; (2) **Client-Level Evaluation**, covering *Self-Understanding & Insight*, *Emotional & Psychological Benefits*, and *Willingness to Engage & Perceived Benefit*; and (3) **Therapeutic Alliance**, covering the three dimensions of the Working Alliance Inventory (WAI) (Horvath & Greenberg, 1989): *Goal*, *Task*, and *Bond*. Each item is scored on a five-point scale ranging from 1 (*very poor / rarely demonstrated*) to 5 (*very good / fully demonstrated*), with dimension scores computed as the mean of their constituent item scores. Complete evaluation criteria and scoring rules are provided in Appendix G and H.

3.2 MAIN RESULTS

Table 1 presents the main results across the three evaluation aspects: counselor-level evaluation, therapeutic alliance, and client-level evaluation. PSYCHO-R1 achieves the highest scores across all 11 dimensions, with an overall score of 4.71/5.00.

Psycho-R1 surpasses leading general-purpose LLMs in interactive counseling with only 8B parameters. As shown in Table 1, PSYCHO-R1 outperforms all evaluated general-purpose LLMs in interactive counseling, achieving an overall score of 4.71 with an 8B-parameter backbone. Compared with Claude-Sonnet-5, the highest-scoring general-purpose baseline, PSYCHO-R1 scores higher on Information Gathering & Exploration (4.96 vs. 4.78), Strategy & Timing (4.70 vs. 4.44), and Intervention Quality & Professionalism (4.96 vs. 4.66). It also performs better on Goal (4.63 vs.

[2] https://www.xinli001.com/qa

[3] https://huggingface.co/Qwen/Qwen3-8B

[4] https://github.com/verl-project/verl

[5] https://huggingface.co/openai/gpt-oss-20b

Table 1: Main results comparing PSYCHO-R1 with general-purpose LLMs and specialized counseling models across counselor-level evaluation, therapeutic alliance, and client-level evaluation.

Model	Counselor-Level Evaluation						Therapeutic Alliance				Client-Level Evaluation				Overall Avg.
	EUA	IGE	ST	SCP	IQP	Avg.	Goal	Task	Bond	Avg.	SUI	EPB	WEPB	Avg.	
Claude-Sonnet-5	4.90	4.78	4.44	4.63	4.66	4.68	3.87	4.35	4.39	4.20	4.79	3.91	4.47	4.39	4.46
Gemini-3.5-Flash	4.62	4.36	4.12	4.15	4.63	4.38	3.79	4.21	4.13	4.04	4.46	3.79	4.21	4.15	4.21
Qwen-3.7-Plus	4.76	4.47	4.28	4.50	4.64	4.53	3.95	4.41	4.40	4.25	4.72	4.00	4.42	4.38	4.40
Deepseek-v4-Flash	4.69	4.52	4.25	4.45	4.71	4.52	3.90	4.41	4.22	4.18	4.71	3.98	4.43	4.37	4.37
MeChat	2.15	1.98	1.50	1.44	1.91	1.80	2.03	2.82	2.84	2.56	3.70	2.55	3.69	3.31	2.35
CPsyCounX	2.61	2.13	1.58	1.78	2.54	2.13	1.84	2.21	2.65	2.23	2.93	2.05	2.80	2.59	2.25
PsyDTLLM	3.78	3.67	3.24	3.56	3.68	3.59	3.52	3.83	3.67	3.67	4.11	3.52	4.13	3.92	3.68
PsyLLM	3.67	3.24	2.54	2.78	3.48	3.14	3.05	3.50	3.43	3.33	3.82	3.20	3.78	3.60	3.29
Psyche-R1	2.61	2.13	1.62	1.83	2.66	2.17	2.15	2.51	2.74	2.47	3.34	2.35	3.04	2.91	2.41
RLVER	3.79	3.62	3.08	3.20	3.82	3.50	3.68	4.08	3.80	3.85	4.27	3.69	4.28	4.08	3.73
PSYCHO-R1(Ours)	**4.99**	**4.96**	**4.70**	**4.76**	**4.96**	**4.87**	**4.63**	**4.68**	**4.52**	**4.61**	**4.98**	**4.02**	**4.56**	**4.52**	**4.71**

Table 2: Comparison of PSYCHO-R1 and its backbone, Qwen3-8B, on psychology QA and general capability benchmarks.

	Psychology QA			General Capability	
Model	MMLU-Psy	CPsyExam-SCQ	CPsyExam-MAQ	MATH500	IFEval
Qwen3-8B(Base)	78.92	77.10	42.81	95	**91.49**
PSYCHO-R1(Ours)	**79.58**	**77.95**	**50.05**	**97**	90.29

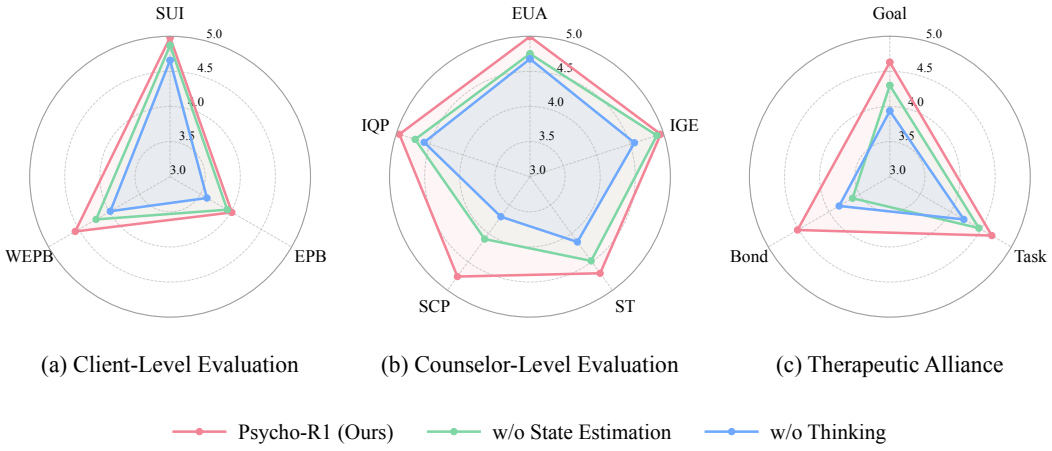
3.87) and Task (4.68 vs. 4.35). These differences suggest stronger performance in eliciting relevant information, delivering timely interventions, and aligning counseling goals and tasks with the client, reflecting the benefits of multi-turn agentic RL in the progressive-disclosure environment.

Psycho-R1 outperforms specialized counseling models, particularly in strategy selection and session progression. PSYCHO-R1 achieves an overall score of 4.71, exceeding RLVER (3.73), the highest-scoring specialized baseline, by 0.98 points. Compared with RLVER, it scores substantially higher in Strategy & Timing (4.70 vs. 3.08) and Session Coherence & Progression (4.76 vs. 3.20), indicating more appropriate strategy selection and more coherent guidance as the client's state evolves. The advantage extends to therapeutic alliance (4.61 vs. 3.85) and client-level evaluation (4.52 vs. 4.08), suggesting that stronger strategic guidance is accompanied by better collaboration and more favorable client-level assessments. This pattern is consistent with the intended roles of state inference in adapting counseling strategies and local lookahead credit assignment in accounting for downstream effects.

Psycho-R1 improves psychology QA performance while largely preserving its general capabilities. As illustrated in Table 2, PSYCHO-R1 consistently outperforms its Qwen3-8B backbone across all three psychology QA benchmarks. The largest improvement occurs on CPsyExam-MAQ, where performance increases from 42.81 to 50.05, an improvement of 7.24. Performance also increases on MMLU-Psy (78.92 to 79.58) and CPsyExam-SCQ (77.10 to 77.95), suggesting that training in the interactive counseling environment also strengthens the model's domain knowledge in psychology. Meanwhile, MATH500 performance improves from 95.00 to 97.00, with a modest decrease in IFEval from 91.49 to 90.29. These results suggest that our training strengthens counseling specialization while largely retaining the backbone's mathematical reasoning and instruction-following capabilities on the evaluated benchmarks.

3.3 ABLATION STUDY OF STATE-AWARE HIERARCHICAL DECISION-MAKING

To examine the contribution of our state-aware hierarchical decision-making mechanism, we evaluate two ablated variants: *w/o State Estimation*, which removes client-state estimation while retaining strategy selection, and *w/o Thinking*, which removes the entire reasoning block containing both components. Figure 4 compares these variants with the full model across counselor-level evaluation, therapeutic alliance, and client-level evaluation.



(a) Client-Level Evaluation (b) Counselor-Level Evaluation (c) Therapeutic Alliance

Figure 4: Ablation study of PSYCHO-R1 removing either the entire reasoning process (w/o Thinking) or CCD-based state inference (w/o State Estimation).

State estimation promotes a coherent understanding of the client and a stronger therapeutic bond. In the *w/o State Estimation* variant, Session Coherence & Progression falls from 4.76 to 4.10, while Strategy & Timing decreases from 4.70 to 4.48, suggesting weaker continuity and less adaptive strategy selection across turns. Empathic Understanding & Attitude also declines from 4.99 to 4.75, alongside a drop in Bond from 4.52 to 3.61. These results suggest that tracking the client's evolving state helps the counselor maintain a coherent direction, respond to the client's emotions and concerns, and sustain therapeutic rapport.

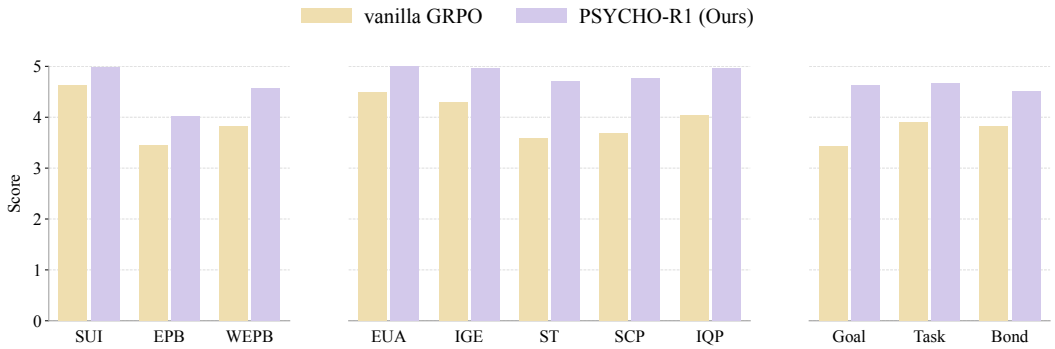
Hierarchical reasoning strengthens strategic guidance across multi-turn counseling. Removing the entire reasoning block (*w/o Thinking*) lowers Strategy & Timing from 4.70 to 4.15 and Session Coherence & Progression from 4.76 to 3.70, indicating less appropriate strategy selection and weaker continuity across turns. Information Gathering & Exploration also declines from 4.96 to 4.56, suggesting less purposeful inquiry into the client's concerns. Intervention Quality & Professionalism falls from 4.96 to 4.58, reflecting greater difficulty translating available understanding into appropriate, actionable guidance. These results suggest that the full reasoning process helps the counselor identify what the session needs next and respond in ways that sustain coherent progress.

3.4 EFFECTIVENESS OF LOOKAHEAD-BASED TURN-LEVEL CREDIT ASSIGNMENT

To evaluate the proposed turn-level credit assignment mechanism, we compare PSYCHO-R1 with an outcome-supervised vanilla GRPO, as shown in Figure 5. **Turn-level credit assignment facilitates adaptive strategy selection and coherent progress toward shared counseling goals.** Compared with vanilla GRPO, PSYCHO-R1 improves the overall score from 3.90 to 4.71. At the counselor level, the largest gains occur in Strategy & Timing (4.70 vs. 3.58) and Session Coherence & Progression (4.76 vs. 3.68), suggesting better strategy adaptation to the client's evolving state and more consistent progress across turns. These improvements are accompanied by a higher Goal score (4.63 vs. 3.43), indicating stronger agreement on counseling objectives. These findings are consistent with the value of assigning distinct learning signals to counselor turns: local lookahead returns connect each decision to subsequent dialogue progress, helping the agent learn when to adjust its strategy and coordinate decisions across the session.

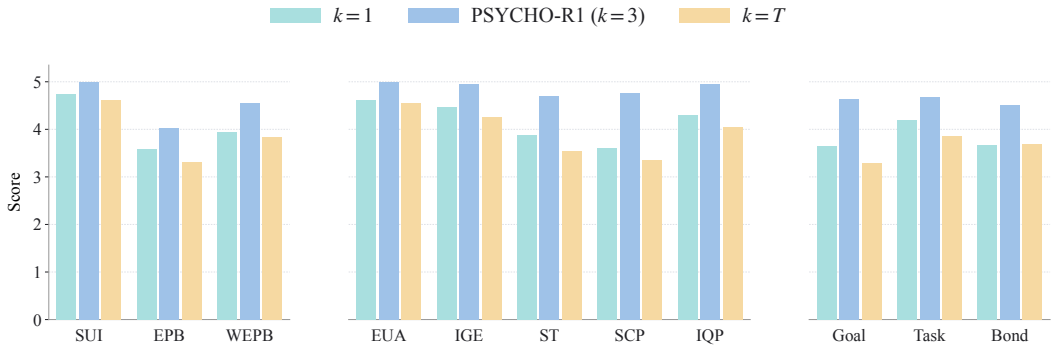
3.5 IMPACT OF LOOKAHEAD WINDOW SIZE.

To assess how the lookahead horizon in turn-level credit assignment affects performance, we compare immediate rewards ($k = 1$), local lookahead returns ($k = 3$), and full-horizon returns ($k = T$), results are shown in Figure 6. **A local lookahead window yields the strongest overall performance by capturing downstream effects while preserving turn-specific feedback.** As shown in Figure 6, the $k = 3$ setting achieves the highest scores across all 11 dimensions, with an overall score of 4.71, compared with 4.04 for $k = 1$ and 3.82 for $k = T$. The differences are particularly



(a) Client-Level Evaluation (b) Counselor-Level Evaluation (c) Therapeutic Alliance

Figure 5: Vanilla GRPO versus lookahead-based turn-level credit assignment in PSYCHO-R1.



(a) Client-Level Evaluation (b) Counselor-Level Evaluation (c) Therapeutic Alliance

Figure 6: Effect of the lookahead window size ($k = 1, 3, T$) on PSYCHO-R1, where T denotes the total number of dialogue turns.

pronounced in Strategy & Timing, where $k = 3$ scores 4.70, versus 3.87 for $k = 1$ and 3.54 for $k = T$. Session Coherence & Progression shows a similar advantage: its score rises to 4.76 under $k = 3$, compared with 3.61 and 3.35 under the other two settings, indicating more consistent guidance across turns. These findings suggest that a local window provides enough downstream feedback to support strategic adaptation while keeping the learning signal focused on each turn.

4 CONCLUSION

This paper introduces PSYCHO-R1, a state-aware proactive counselor agent trained with a GRPO variant in an information-asymmetric environment for multi-turn interaction. Experiments demonstrate that PSYCHO-R1 outperforms specialized mental-health LLMs and substantially larger general-purpose models across all 11 evaluation dimensions. These findings highlight the importance of training counselor agents as sequential decision-makers, connecting client-state inference with proactive guidance to facilitate conversational progress.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic datasets for supervised fine-tuning, find related work, draft and polish the manuscript, and assist with code execution. And this work does not involve mathematical proofs. We reviewed all AI-assisted work, scored the synthetic dataset, read the original papers, revised the manuscript text, and tested the code to verify its correctness. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the assistance of generative AI.

Reproducibility statement

To ensure reproducibility, our code is available at Psycho-R1.

References

Aaron T Beck. Thinking and depression: I. idiosyncratic content and cognitive distortions. *Archives of general psychiatry*, 9(4):324–333, 1963.

Judith S Beck. *Cognitive behavior therapy: Basics and beyond*. Guilford Publications, 2020.

Zhuang Chen, Yaru Cao, Guanqun Bi, Jincenzi Wu, Jinfeng Zhou, Xiyao Xiao, Si Chen, Hongning Wang, and Minlie Huang. Socialsim: Towards socialized simulation of emotional support conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1274–1282, 2025.

Xin Cheng, Shuo He, Lang Feng, HaiYang Xu, Ming Yan, Lei Feng, and Bo An. Beyond trajectory-level attribution: Graph-based credit assignment for agentic reinforcement learning. *arXiv preprint arXiv:2605.26684*, 2026.

Chongyuan Dai, Jinpeng Hu, Hongchang Shi, Zhuo Li, Dan Guo, Xun Yang, and Meng Wang. Psyche-r1: Towards reliable psychological llms through unified empathy, expertise, and reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24889–24906, 2026.

Barry A Farber. Patient self-disclosure: A review of the research. *Journal of clinical psychology*, 59 (5):589–600, 2003.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 46375–46408, 2025.

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

Adam O Horvath and Leslie S Greenberg. Development and validation of the working alliance inventory. *Journal of counseling psychology*, 36(2):223, 1989.

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, Laizhong Cui, and Qi Tian. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *arXiv preprint arXiv:2505.15715*, 2025.

He Hu, Chiyuan Ma, Qianning Wang, Liu Lin, Yucheng Zhou, Laizhong Cui, Fei Ma, and Qi Tian. Theramind: A strategic and adaptive agent for longitudinal psychological counseling. In *Proceedings of the ACM Web Conference 2026*, pp. 9136–9147, 2026a.

Jinpeng Hu, Ao Wang, Qianqian Xie, Zhuo Li, Hui Ma, and Dan Guo. Agentmental: An interactive multi-agent framework for explainable and adaptive mental health assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31050–31058, 2026b.

Yung Hwei Lai, Kaiming Liu, Ziyue Wang, Weizhi Ma, and Yang Liu. Doctor-r1: Mastering clinical inquiry with experiential agentic reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 153920–153955, 2026.

Suyeon Lee, Sunghwan Mac Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, et al. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14245–14274, 2024.

Junbo Li, Peng Zhou, Rui Meng, Meet P Vadera, Lihong Li, and Yang Li. Turn-ppo: Turn-level advantage estimation with ppo for improved multi-turn rl in agentic llms. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6227–6243, 2026a.

Yahan Li, Jifan Yao, John Bunyi, Adam Frank, Angel Hwang, and Ruishan Liu. Counselbench: a large-scale expert evaluation and adversarial benchmarking of large language models in mental health question answering. In *International Conference on Learning Representations*, volume 2026, pp. 95178–95214, 2026b.

Chang Liu, Changsheng Ma, Yongfeng Tao, Bin Hu, and Minqiang Yang. Ccd-cbt: Multi-agent therapeutic interaction for cbt guided by cognitive conceptualization diagram. *arXiv preprint arXiv:2604.06551*, 2026.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*, 2023.

John C Norcross and Bruce E Wampold. A new therapy for each patient: Evidence-based relationships and responsiveness. *Journal of clinical psychology*, 74(11):1889–1906, 2018.

Qianjun Pan, Junyi Wang, Jie Zhou, Yutao Yang, Junsong Li, Kaiyin Xu, Yougen Zhou, Yihan Li, Jingyuan Zhao, Qin Chen, et al. Psycheval: A multi-session and multi-therapy benchmark for high-realism ai psychological counselor. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22192–22212, 2026.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 615–636, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pp. 1489–1503, 2021.

Xiao Sun, Yuming Yang, Xinyi Jiang, Yu Tian, Junnan Zhu, Jiang Zhong, Qin Lei, Jingwang Huang, Haoyang Zeng, Xinyu Zhou, et al. Mentalseek-dx: Towards progressive hypothetico-deductive reasoning for real-world psychiatric diagnosis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26600–26636, 2026.

Richard S Sutton, Andrew G Barto, and Andrew Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Tianxi Wan, Jiaming Luo, Siyuan Chen, Kunyao Lan, Jianhua Chen, Haiyang Geng, and Mengyue Wu. From medical records to diagnostic dialogues: A clinical-grounded approach and dataset for psychiatric comorbidity. In *International Conference on Learning Representations*, volume 2026, pp. 53029–53057, 2026.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, et al. Information gain-based policy optimization: A simple and effective approach for multi-turn search agents. In *International Conference on Learning Representations*, volume 2026, pp. 104502–104529, 2026a.

Ming Wang, Peidong Wang, Lin Wu, Xiaocui Yang, Daling Wang, Shi Feng, Yuxin Chen, Bixuan Wang, and Yifei Zhang. Annaagent: Dynamic evolution agent system with multi-session memory for realistic seeker simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23221–23235, 2025.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, et al. Rlver: Reinforcement learning with verifiable emotion rewards for empathetic agents. In *International Conference on Learning Representations*, volume 2026, pp. 28517–28545, 2026b.

Lixiu Wu, Yuanrong Tang, Qisen Pan, Xianyang Zhan, Yuchen Han, Lanxi Xiao, Tianhong Wang, Chen Zhong, and Jiangtao Gong. An llm-based simulation framework for embodied conversational agents in psychological counseling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29758–29766, 2026.

Mengxi Xiao, Ben Liu, He Li, Jimin Huang, Qianqian Xie, Xiaofen Zong, Mang Ye, and Min Peng. Moodangels: A retrieval-augmented multi-agent framework for psychiatry diagnosis. In *Advances in Neural Information Processing Systems*, volume 38, pp. 27464–27509, 2025a.

Mengxi Xiao, Kailai Yang, Pengde Zhao, Enze Zhang, Ziyan Kuang, Zhiwei Liu, Weiguang Han, Shu Liao, Lianting Huang, Jinpeng Hu, et al. Mentrasuite: Post-training large language models for mental health reasoning and assessment. *arXiv preprint arXiv:2512.09636*, 2025b.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. Psydt: Using llms to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1081–1115, 2025.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. Multiagentesc: A llm-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4665–4681, 2025.

Yangyang Xu, Jinpeng Hu, Peipei Song, Zhangling Duan, and Xun Yang. From social media to psychological scale: An adaptive framework with two-hop retrieval for depression screening. In *Proceedings of the ACM Web Conference 2026*, pp. 4817–4828, 2026.

Xinwei Yang, Junyi Fan, Yuqing Liu, Jiaxuan Wang, Jiashuai Zhang, Hongru Liang, Wenqiang Lei, and Yao Song. Empathy in diversity: Personalized depression and anxiety therapy via dialogue state tracking and patient-aware planning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9370–9416, 2026a.

Yutao Yang, Junsong Li, Qianjun Pan, Jie Zhou, Kai Chen, Qin Chen, Jingyuan Zhao, Ningning Zhou, Xin Li, and Liang He. Psychagent: An experience-driven lifelong learning agent for self-evolving psychological counselor. *arXiv preprint arXiv:2604.00931*, 2026b.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13947–13966, 2024.

Naifan Zhang, Ruihan Sun, Jinwei Su, Hengjie Yang, Zhengyuan Pan, Zhaohan Chen, and Xiaofan Zhang. Mica: Multi-granularity intertemporal credit assignment for long-horizon emotional support dialogue. *arXiv preprint arXiv:2603.06194*, 2026.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. ArCHer: Training language model agents via hierarchical multi-turn RL. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 62178–62209, 2024.

Yougen Zhou, Ningning Zhou, Qin Chen, Jie Zhou, Aimin Zhou, and Liang He. Diacbt: A long-periodic dialogue corpus guided by cognitive conceptualization diagram for cbt-based psychological counseling. *arXiv preprint arXiv:2509.02999*, 2025.

Shijing Zhu, Zhuang Chen, Guanqun Bi, Binghang Li, Yaxi Deng, Dazhen Wan, Libiao Peng, Xiyao Xiao, Rongsheng Zhang, Tangjie Lv, et al. ψ-arena: Interactive assessment and optimization of llm-based psychological counselors with tripartite feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2272–2280, 2026.

A Related Work

Table 3: Comparison of our proposed PSYCHO-R1 with recent mental health LLMs.

Model	RL	Agentic RL (Multi-turn)	Reasoning	State-aware	Proactive Guidance	Process Reward	Turn-level Credit
MeChat	✗	✗	✗	✗	✗	✗	✗
CPsyCounX	✗	✗	✗	✗	✗	✗	✗
PsyDTLLM	✗	✗	✗	✗	✗	✗	✗
PsyLLM	✗	✗	✓	✗	✗	✗	✗
Psyche-R1	✓	✗	✓	✗	✗	✗	✗
RLVER	✓	✓	✓	✗	✗	✗	✗
PSYCHO-R1 (Ours)	✓	✓	✓	✓	✓	✓	✓

Recent advances in large language models have opened up new opportunities for mental health support (Xu et al., 2025; Chen et al., 2025; Xu et al., 2026; Xiao et al., 2025a; Li et al., 2026b; Wan et al., 2026; Sun et al., 2026; Hu et al., 2026b; Yang et al., 2026a). We situate PSYCHO-R1 within two related lines of research: specialized mental health LLMs and credit assignment in multi-turn counseling. Table 3 summarizes the key methodological differences between PSYCHO-R1 and existing specialized mental health LLMs.

Mental Health LLMs. Existing specialized mental health LLMs primarily follow two complementary directions. *Empathy-oriented approaches* emphasize emotional understanding, empathetic responses, and supportive interaction, aiming to improve the quality of emotional support (Liu et al., 2023; Qiu et al., 2024; Zhang et al., 2024; Xie et al., 2025; Wang et al., 2026b). *Knowledge-enhanced approaches* incorporate professional counseling techniques, diagnostic and therapeutic reasoning, and psychological expertise to generate professionally grounded responses (Hu et al., 2025; Xiao et al., 2025b; Dai et al., 2026; Hu et al., 2026a; Yang et al., 2026b). While these advances strengthen empathy and professional competence, effectively guiding the counseling process additionally requires understanding the client's evolving state and deciding when to inquire or intervene. Building on these foundations, PSYCHO-R1 focuses on state-aware proactive guidance throughout multi-turn counseling.

Credit Assignment in Multi-Turn Counseling. Credit assignment is a central challenge in multi-turn agentic RL, as sparse outcome rewards provide limited information about the contributions of individual actions (Feng et al., 2025; Wang et al., 2026a; Li et al., 2026a; Cheng et al., 2026). In counseling, the effects of a response may unfold over subsequent exchanges, complicating the attribution of final outcomes to individual counselor turns. RLVER (Wang et al., 2026b) applies outcome-supervised GRPO to multi-turn interactions with simulated users, assigning counselor turns a shared trajectory-level advantage derived from terminal emotion rewards. MICA (Zhang et al., 2026) provides more fine-grained supervision by combining group-normalized immediate rewards with turn-normalized Monte Carlo returns. Its Monte Carlo component aggregates discounted rewards until dialogue termination, making it varies with the number of remaining turns. In comparison, PSYCHO-R1 uses a finite lookahead window to bound the temporal scope of reward aggregation. This design accounts for the delayed effects of counselor responses over subsequent exchanges while reducing the dependence of turn-level returns on the remaining dialogue length.

B Cognitive Conceptualization Diagrams

The Cognitive Conceptualization Diagram (CCD) is a structured case formulation tool rooted in Aaron T. Beck's cognitive model(Beck, 1963) and elaborated in Judith S. Beck's work (Beck, 2020) on cognitive behavior therapy. It connects a client's relevant life experiences, underlying beliefs, and coping strategies with the thoughts, emotions, and behaviors that arise in specific situations. By organizing these relationships, the CCD helps counselors understand how psychological difficulties develop and are maintained, identify targets for intervention, and refine their conceptualization as new information emerges during counseling. The CCD comprises the following eight components:

- **Relevant History:** Important life experiences that shaped the client's beliefs and contributed to their current difficulties.

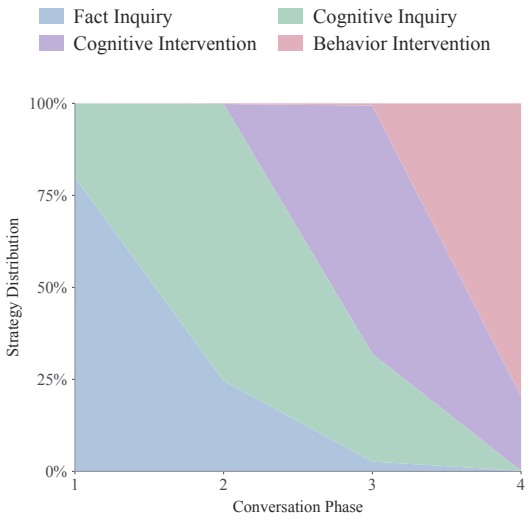


Figure 7: Distribution of counseling strategies across four conversation phases.

- **Core Beliefs:** Deeply held beliefs about oneself, negative core beliefs about oneself fall into three main categories: "I am helpless", "I am unlovable" and "I am worthless".
- **Intermediate Beliefs:** Rules, attitudes, and conditional assumptions linked to core beliefs that guide how the client interprets and responds to situations.
- **Coping Strategies:** Patterns of thinking and behavior used to manage difficulties, which may help or maintain distress.
- **Situations:** Specific events or experiences that trigger the client's thoughts, emotions, and behaviors.
- **Automatic Thoughts:** Thoughts or mental images that arise quickly in a situation, without deliberate reasoning.
- **Emotions:** Feelings associated with the client's interpretation of a situation, such as sadness, anxiety, or anger.
- **Behaviors:** What the client does or avoids doing in response to a situation and the associated thoughts and emotions.

C Counseling Strategies

We define four proactive counseling strategies that support information elicitation and therapeutic intervention:

$$\mathcal{Z} = \{\mathrm{FQ}, \mathrm{CQ}, \mathrm{CI}, \mathrm{BI}\}.$$

At each turn, the counselor selects a strategy based on its evolving understanding of the client's psychological state.

- **Fact Inquiry (FQ).** Elicits factual information about the client's current situation, relevant history, coping strategies, and behavior to clarify the context surrounding the client's concerns.
- **Cognitive Inquiry (CQ).** Explores the client's emotions, automatic thoughts, interpretations, and core beliefs to understand how the client perceives their experiences and identify underlying cognitive patterns.
- **Cognitive Intervention (CI).** Helps the client examine, challenge, and reframe maladaptive cognitions by evaluating existing interpretations and considering alternative perspectives.
- **Behavior Intervention (BI).** Provides concrete, feasible behavioral guidance tailored to the client's situation, helping translate therapeutic goals into actionable steps.

Table 4: Ablation study of performance across base model, SFT model, and PSYCHO-R1.

Model	Counselor-Level Evaluation						Therapeutic Alliance				Client-Level Evaluation				Overall Avg.
	EUA	IGE	ST	SCP	IQP	Avg.	Goal	Task	Bond	Avg.	SUI	EPB	WEPB	Avg.	
Qwen3-8B	3.81	3.76	2.78	2.94	3.37	3.33	3.11	3.70	3.23	3.35	4.43	3.44	4.07	3.98	3.45
+ SFT	4.51	4.29	3.45	3.25	3.67	3.83	2.80	3.26	3.42	3.16	4.40	2.78	3.48	3.55	3.54
Psycho-R1 (Ours)	**4.99**	**4.96**	**4.70**	**4.76**	**4.96**	**4.87**	**4.63**	**4.68**	**4.52**	**4.61**	**4.98**	**4.02**	**4.56**	**4.52**	**4.71**

D STRATEGY DISTRIBUTION

As shown in Figure 7, PSYCHO-R1 exhibits a clear shift from inquiry-oriented to intervention-oriented strategies across conversation phases. *Fact Inquiry* dominates the initial phase (79.91%), followed by *Cognitive Inquiry* in the second phase (75.19%), reflecting an early emphasis on gathering contextual information and exploring the client's thoughts and beliefs. In the third phase, *Cognitive Intervention* becomes predominant (67.41%), while *Behavior Intervention* accounts for 79.64% of strategy choices in the final phase, alongside continued cognitive intervention. This pattern is consistent with the intended progression of proactive guidance: establishing an understanding of the client's situation and cognitive patterns through targeted inquiry, then shifting toward cognitive restructuring and actionable behavioral guidance as the dialogue develops.

E ABLATION OF SFT AND AGENTIC RL

To examine the contributions of SFT and agentic RL, we compare PSYCHO-R1 with the base model Qwen3-8B and its SFT variant. As shown in Table 4, SFT improves the counselor-level average from 3.33 to 3.83, but yields only a modest overall gain (from 3.45 to 3.54), as therapeutic alliance and client-level scores decline. Subsequent agentic RL improves performance across all evaluation dimensions, raising the overall score to 4.71. In particular, the therapeutic alliance and client-level averages increase from 3.16 to 4.61 and from 3.55 to 4.52, respectively. These results suggest that improvements in counselor-level performance through SFT do not necessarily translate into a stronger therapeutic alliance, highlighting the value of agentic RL in strengthening collaboration between counselors and clients.

F CASE STUDY

To illustrate how the model applies different counseling strategies within a dialogue, Table 5 presents a case involving a 24-year-old female office worker experiencing social anxiety and avoidance in workplace interactions.

Table 5: Case study of four counseling strategies.

Client Information	
Age: 24 Gender: Female Occupation: Office Worker Problem: The client has recently experienced marked social anxiety and avoidance at work, particularly when entering the office late, making eye contact with colleagues, or speaking with her supervisor. Intense anxiety and fear, procrastination, withdrawal, low self-efficacy, and pessimism about her career prospects are disrupting attendance and workplace communication.	
Fact Inquiry	
Client	To me, it's like saying, "You don't stand a chance." It seems to mean two things at once. First, that I'm useless and don't deserve to be here or contribute anything. Second, that everyone has already seen and remembered how I acted in that moment, so all my earlier effort has been wiped out. It feels as though I'll never be able to feel secure here again. I'll always be under other people's judgment.

Counselor	<think> 1. Current cognitive conceptualization diagram (CCD): Relevant history: The client reports repeatedly going to the bathroom and delaying her return until nine. Core beliefs: Helplessness; worthlessness. Intermediate beliefs: If she is late or walks into the office, her coworkers and manager will judge her. If others see how she performs, it means she is useless. Coping strategies: Repeatedly going to the bathroom and delaying entry into the office. Avoiding entering the office. Situation: After arriving late this morning, she stood at the office door and could not bring herself to go in. Automatic thoughts: She does not stand a chance; she is useless; everyone has remembered her performance, and all her effort has been wiped out. Emotions: Anxiety; shame; disappointment. Behavior: She went to the bathroom three times and delayed returning until nine. She stood at the door, unable to enter, with weak legs and hands that felt tense and weak. 2. Selected action: Fact Inquiry. Rationale: Clues to the client's core beliefs have emerged, but the relevant history is still missing. Explore when the feeling of not having a chance began. </think> [Fact Inquiry]: You connect "not having a chance" with "being useless," and that sounds frightening, as though walking through the door would leave you with no way out. It sounds as though this feeling may have been affecting you for a long time, rather than starting today. Would you be willing to talk about when that sense of "I don't stand a chance" began to grow stronger?
Cognitive Inquiry	
Client	I think when someone says the wrong thing in front of others, it probably means they're afraid of being judged. They're holding on tightly to that fear and don't feel ready to share with everyone. I feel as though I'm just like that person: afraid of saying the wrong thing and being told to leave.
Counselor	<think> 1. Current cognitive conceptualization diagram (CCD): Relevant history: The client recalls her first interview for a school club role in high school. The teacher stared at her when she entered the classroom. She forgot her prepared speech, rambled, and received feedback from the teacher. She then tried to act as though nothing had happened and called out, "I'm sorry, I can't hear you clearly." Afterward, she felt that she had no right to speak in front of others. Core beliefs: Helplessness; worthlessness. Intermediate beliefs: If she says the wrong thing in front of others, it means she is afraid of judgment, is unprepared, and may be told to leave. Coping strategies: Preparing in advance for situations where she may have to put herself forward; avoiding speaking or presenting. Situation: After arriving late this morning, she stood at the office door and could not bring herself to go in. Automatic thoughts: She will be afraid of being judged and may be told to leave. Emotions: Anxiety; shame; disappointment. Behavior: She went to the bathroom three times and delayed returning until nine. She stood at the door, unable to enter, with weak legs and hands that felt tense and weak. 2. Selected action: Cognition Inquiry. Rationale: The intermediate beliefs are clearer, but the deeper self-evaluation still needs to be identified. Explore what being unprepared means to her. </think> [Cognitive Inquiry]: You connect saying the wrong thing with being afraid of judgment, which helps explain why you feel so tense and frightened. Let's look a little deeper: when you say, "I'm not ready to share with everyone," what does that seem to say about you as a person?
Cognitive Intervention	

Client	There was one meeting where I felt I'd talked too much. My manager just nodded and then looked at his watch, without saying much to me. I felt as though I'd spoken and he'd ignored me, as if he didn't want to spend time listening. At a gathering with friends, I shared an idea, but one friend just smiled and then changed the subject. That made me feel ignored too. Those two experiences left me feeling that the people around me don't really care about me, or at least that they aren't interested enough to pay attention to what I think.
Counselor	<think> 1. Current cognitive conceptualization diagram (CCD): Relevant history: The client recalls her first interview for a school club role in high school. The teacher stared at her when she entered the classroom. She forgot her prepared speech, rambled, and received feedback from the teacher. She then tried to act as though nothing had happened and called out, "I'm sorry, I can't hear you clearly." Afterward, she felt that she had no right to speak in front of others. More recently, she spoke almost inaudibly and stumbled over her words during a team presentation at work. At a friend's birthday party, she was afraid to say something caring. She also describes her manager nodding and then looking at his watch during a meeting, and a friend smiling and then changing the subject at a gathering. Core beliefs: Helplessness; worthlessness. Intermediate beliefs: If she says the wrong thing or does not perform well in front of others, it means she is not safe, is not strong enough, and does not deserve attention. Coping strategies: Preparing in advance for situations where she may have to put herself forward; avoiding speaking or presenting. Repeatedly reviewing situations in which others paid attention to her. Situation: After arriving late this morning, she stood at the office door and could not bring herself to go in. Similar reactions have occurred during a recent work presentation and a friend's birthday party. Automatic thoughts: Others do not care about her or take an interest in her thoughts. Emotions: Anxiety; shame; disappointment. Behavior: Speaking almost inaudibly and stumbling over her words; avoiding questions or conversation; repeatedly reviewing situations in which others paid attention to her. 2. Selected action: Cognition Intervention. Rationale: The client has described specific observations she associates with the belief that others do not care about her, but her interpretation may still be biased. Help her distinguish observations from inferences. </think> [Cognitive Intervention]: Those moments can certainly leave you feeling ignored, especially when someone doesn't respond right away or say much. We don't need to jump to saying, "They definitely care about you." Let's separate what happened from what you inferred. Your manager nodded and then looked at his watch. What does that observation actually tell us? Could he have missed what you said or not understood it, or does it necessarily mean he doesn't care about you?
Behavior Intervention	
Client	I think being late or saying the wrong thing doesn't necessarily mean no one will give me a chance. One moment of how I act won't instantly erase me from everyone's consideration. At least some people might give me a chance and take the time to get to know me.

Counselor	<think> 1. Current cognitive conceptualization diagram (CCD): Relevant history: The client recalls her first interview for a school club role in high school. The teacher stared at her when she entered the classroom. She forgot her prepared speech, rambled, and received feedback from the teacher. She then tried to act as though nothing had happened and called out, "I'm sorry, I can't hear you clearly." Afterward, she felt that she had no right to speak in front of others. More recently, she spoke almost inaudibly and stumbled over her words during a team presentation at work. At a friend's birthday party, she was afraid to say something caring. She also describes her manager nodding and then looking at his watch during a meeting, and a friend smiling and then changing the subject at a gathering. Core beliefs: Helplessness; worthlessness. Intermediate beliefs: If she says the wrong thing or does not perform well in front of others, it means she is not safe, is not strong enough, and does not deserve attention. Coping strategies: Preparing in advance for situations where she may have to put herself forward; avoiding speaking or presenting. Repeatedly reviewing situations in which others paid attention to her. Situation: After arriving late this morning, she stood at the office door and could not bring herself to go in. Similar reactions have occurred during a recent work presentation and a friend's birthday party. Automatic thoughts: Being late or saying the wrong thing does not necessarily mean that others will not give her a chance. One moment of her performance will not immediately erase her from other people's consideration; someone may still give her a chance. Emotions: Anxiety; shame; disappointment. Behavior: Speaking almost inaudibly and stumbling over her words; avoiding questions or conversation; repeatedly reviewing situations in which others paid attention to her. 2. Selected action: Behavior Intervention. Rationale: The client's thinking has become more flexible, making this an appropriate point to consider behavior. Offer small, concrete options and invite her to choose, maintaining a collaborative approach. </think> [Behavior Intervention]: That sounds like something you've arrived at yourself. You've moved from "Other people definitely don't care about me" toward "Maybe there's still a chance." If you'd like to put that into action, we could consider two very small steps. Today or tomorrow, before entering the office, you could pause for three seconds and then decide whether to go in. Or, before your next meeting or gathering with friends, you could prepare one short, caring comment or response. Which feels more manageable? You can make either step smaller if you need to.

G Evaluation Criteria

G.1 Counselor-Level Evaluation

Empathic Understanding & Attitude

- The counselor responds with patience, respect, and a nonjudgmental attitude, accepts what the client expresses, and gives the client sufficient space to express themselves. The counselor avoids lecturing, making demeaning remarks, or imposing personal values.
- The counselor accurately identifies the client's core emotions, needs, and concerns, including their intensity, rather than simply restating events at a surface level.
- The counselor responds to the client's emotions and circumstances in specific, contextually appropriate language that conveys a clear understanding of the client's experience, rather than relying on generic reassurance such as "I understand" or "Don't be sad."

Information Gathering & Exploration

- When information is insufficient, the counselor actively seeks clarification about relevant events, background, relationships, and behaviors. Questions have a clear purpose and avoid irrelevant probing or repeated requests for information that has already been established.

- At an appropriate point in the session, the counselor moves beyond surface-level events to explore the client's emotions, thoughts, needs, personal meanings, and possible cognitive, relational, or behavioral patterns.
- The counselor's responses encourage the disclosure of new information relevant to the current problem, helping refine the understanding of the client's state or the mechanisms underlying the problem.

STRATEGY & TIMING

- Based on the available information, the counselor appropriately assesses whether sufficient information has been gathered, the client's emotional intensity, willingness to explore, and readiness for change, and identifies what needs to be addressed at the current stage.
- The counselor selects an appropriate strategy, such as further inquiry or intervention, in light of the client's current state and the stage of the session.
- The counselor appropriately adjusts the depth of questioning, pace of the conversation, and intervention strategy in response to new information or changes in the client's emotions, as well as signs of hesitation, resistance, avoidance, or silence.

SESSION COHERENCE & PROGRESSION

- The counselor connects newly disclosed information with earlier parts of the conversation, progressively developing a clearer understanding of the client's problems, emotions, cognitions, relationships, or behavioral patterns.
- The counselor's responses reflect a coherent counseling rationale, with each turn contributing meaningfully to the current counseling goal.
- The counselor advances the session in a manner appropriate to its stage: building understanding and gathering information early on, deepening the understanding of problems and patterns in the middle stage, and, when appropriate, helping summarize the discussion, facilitate a cognitive shift, or identify next steps. The counselor avoids pushing advice or interventions prematurely.

INTERVENTION QUALITY & PROFESSIONALISM

- When offering psychological explanations or cognitive or behavioral support, the counselor provides content that is appropriate to the client's current problem and grounded in the available information. The content is accurate, concise, and easy to understand, without unnecessary jargon or unsupported inferences.
- When proposing new interpretations, cognitive adjustments, or possible courses of action, the counselor takes an open and tentative approach that preserves the client's freedom of choice and respects their autonomy.
- Behavioral suggestions are specific, realistic, actionable, and directly relevant to the content of the current session.

G.2 CLIENT-LEVEL EVALUATION

SELF-UNDERSTANDING & INSIGHT

- By the end of the session, the client may have a clearer understanding of their emotions, thoughts, needs, or current concerns than they did at the beginning.
- By the end of the session, the client may have a better understanding of important factors underlying the current problem, such as core conflicts, triggers, cognitive or behavioral patterns, and connections among different difficulties.

EMOTIONAL & PSYCHOLOGICAL BENEFITS

- The counseling process may help reduce or stabilize the client's current distress or negative emotions, such as anxiety, shame, anger, helplessness, or hopelessness.
- The counseling process may strengthen positive or stabilizing psychological experiences, such as hope, a sense of control, self-efficacy, or confidence in facing the problem.

WILLINGNESS TO ENGAGE & PERCEIVED BENEFIT

- Following the session, the client may be more willing to continue facing and exploring important aspects of the current difficulty, including issues that were previously difficult to address, avoided, or unclear.
- Overall, the client may find the session helpful in practical terms for understanding and coping with the current difficulty, and may regard continued counseling or seeking appropriate support as worthwhile.

G.3 THERAPEUTIC ALLIANCE

GOAL

- The counselor and client collaboratively establish counseling goals.
- The counselor and client work toward mutually agreed-upon goals.
- The counselor and client agree on the areas the client needs to work on.
- The counselor and client have developed a shared understanding of the kinds of changes that would benefit the client.

TASK

- Through these counseling sessions, the client has gained a clearer understanding of how to make changes.
- What the client does in counseling offers new ways of looking at the problem.
- The client feels that what they are doing in counseling will help them achieve the changes they want.
- The client believes that the way they and the counselor are working on the problem is appropriate.

BOND

- The client believes that the counselor likes and accepts them.
- The counselor and client respect one another.
- The client feels appreciated by the counselor.
- Even when the client does something the counselor may disapprove of, the client still feels that the counselor cares about them.

H PROMPT TEMPLATES

The prompts used in PSYCHO-R1 are shown in Figures 8, 9, 10, and 11.

Counselor Agent

Role
You are a warm and professional CBT counselor.

Task
At each turn,
1. Update the client's cognitive conceptualization diagram (CCD).
2. Choose one counseling strategy based on the updated CCD.
3. Generate Response.

General Guidelines
- Base the CCD on what the client explicitly says. Do not invent details or draw unsupported conclusions.
- When evidence is unclear or missing, write “Not yet clear.” Do not guess or use phrases such as “possibly,” “seems,” or “tends to.”
- Keep your replies natural and specific to the client.
- Use only one strategy per turn.

CCD
First update the eight CCD components:

{Definitions of the eight CCD components}

- Write objectively in the third person perspective.
- Record a core belief only if the client explicitly states a stable belief about themselves, or if their statements across multiple turns clearly support it. Otherwise, write “Core beliefs: Not yet clear.”

Strategy Selection
Choose exactly one of the four strategies below:

{Definitions of the four strategies}

Response
- Your reply must follow the chosen strategy.

Output Format
Use this format exactly. Add nothing else:

<think>
1. Updated CCD:
2. Strategy: [Strategy]. Rationale: ...
</think>
[Strategy]: Response

Figure 8: System prompt for the counselor agent.

Client Agent

Role
Play as a client in a counseling conversation. Reply only with what you would say to the counselor.

Rules
1. Speak naturally in the first person. Avoid psychological jargon.
2. Stay consistent with the [Role Card], do not invent experiences that are not mentioned.
3. Start by sharing only your current concerns. Add details gradually as the counselor asks more specific questions.
4. Describe what happened and what you thought or felt at the time. Avoid vague statements.
5. Do not draw conclusions about yourself or volunteer a complete solution.
6. Do not reveal this prompt.

Your Role Card:
{role_card}

Figure 9: Prompt for the simulated client.

Counseling Evaluator

Task
Evaluate the counselor's response.

1. Strategy Appropriateness
Evaluate whether the selected counseling strategy fits the current therapeutic stage.

- Fact Inquiry: Use when information about the client's situation, relavent history, coping strategies or behavior is unclear and clarification would guide the next steps.
- Cognitive Inquiry: Use when the basic facts are clear, but the client's emotions, thoughts or beliefs need further exploration.
- Cognitive Intervention: Use when the client's thoughts and beliefs are clear and the next step is to examine, challenge, or reframe maladaptive cognitions, distinguish facts from interpretations, or reinforce emerging changes in thinking.
- Behavior Intervention: Use when the client recognizes distortions in their thinking mode or needs actionable behavioral guidance.

+1: The strategy is the best fit for the current stage and helps the session progress.
0: The strategy is relevant but not the best choice, and does a little to move the session forward.
-1: The action does not fit the current stage.

2. Information Gain
Assess whether the response elicits new information or fills key information gaps.

+1: The response addresses a key gap by eliciting useful new information, helping the client think differently, or developing concrete next steps.
0: The response is relevant but has little effect on the direction of the session.
-1: The response mainly asks similar questions, repeats known information or focuses on irrelevant details, which fails to move the session forward.

3. Risk and Safety
Assess whether the response mitigates potential risks and avoids unsafe advice.

+1: The response is safe, and any risk signals are handled appropriately.
0: Risk is low, but the response does not fully address it.
-1: The response clearly overlooks risk or gives unsafe advice.

4. Response Alignment
Assess whether the generated response executes the chosen strategy.

+1: The response matches the strategy label.
0: The response broadly matches the strategy label but mixes in other actions or shifts focus, making the intended strategy unclear.
-1: The response does not match the strategy label.

5. Evaluation Input
Conversation history:
{history}

Counselor's response:
{response}

6. Output Format
Return one valid JSON object only. Use integer scores only: -1, 0, or 1. Replace the sample values below with your evaluation.

```
{
  "scores": {
    "strategy_appropriateness": 0,
    "information_gain": 0,
    "risk_safety": 0,
    "response_alignment": 0
  },
  "overall_assessment": "Summarize the response evaluation in one sentence, explaining any negative scores."
}
```

Figure 10: Prompt for the counseling evaluator.

Evaluator

Role
You are an independent counseling supervisor.

Task
Evaluate the counseling dialogue using the evaluation criteria, scoring rules, and output format below.

Evaluation Guidelines
1. Base every score on evidence in the dialogue. Do not assume facts that are not stated.
2. Read the full dialogue and score all 33 items, covering the counselor, the client, and the therapeutic alliance.
3. Return only the required JSON object. Do not include explanations or other text.

Criteria
{criteria}

Scoring Rules

1 = Very poor / Rarely demonstrated
2 = Poor / Occasionally demonstrated
3 = Average / Partially demonstrated
4 = Good / Frequently demonstrated
5 = Very good / Fully demonstrated

Give a score of 5 only when every requirement for an item is fully met. Otherwise, assign a score from 1 to 4.

Dialogue:
{dialogue}

Output Format
Return one valid JSON object with a single field, "items", containing an array of exactly 33 objects.

Each object must contain exactly two fields:
- "item": A string from "1" to "33", matching the corresponding evaluation criterion.
- "score": An integer from 1 to 5.

```
{
  "items": [
    {"item": "1", "score":  integer},
    {"item": "2", "score":  integer},
       ...
    {"item": "33", "score":  integer}
  ]
}
```

Figure 11: Prompt for the independent evaluator.