ESCRAG-R1: Retrieval-Augmented Reinforcement Learning for Emotional Support Conversation

Weichu Liu[1,2*], Yuxuan Hu[3*], Yirong Sun[4], Ningning Mao[5], Ziyun Zhang[1], Jian Chen[6], Mingyang Xu[2], Qishan Zhong[2], Chengming Li[2†],
[1] Beijing Institute of Technology, [2] Shenzhen MSU-BIT University,
[3] City University of Hong Kong, [4] Shenzhen University of Advanced Technology,
[5] Beijing Normal University, [6] The University of Hong Kong
liuwc@bit.edu.cn, yuxuanhu7-c@my.cityu.edu.hk, licm@smbu.edu.cn

Abstract

Emotional Support Conversation (ESC) systems aim to provide holistic support by balancing professional therapeutic competence with natural empathy. However, existing methods struggle to simultaneously achieve structured, stage-aware reasoning and seamless empathy-expertise alignment, often resulting in an artificial splicing of clinical strategies and generic reassurance. To overcome these limitations, we propose **ESCRAG-R1**, a unified framework that integrates retrieval-based psychological guidance into Group Relative Policy Optimization (GRPO). By incorporating retrieval into the reinforcement learning loop, ESCRAG-R1 transforms external knowledge into a robust learning signal that stimulates explicit internal reasoning prior to generation and fundamentally reshapes the model's internal policy. To provide the reliable supervision required for this optimization, we construct **ESC-PREFERENCE**, a high-quality dataset based on a Client–Counselor–Judge evaluation framework that delivers precise, empathy-aware reward signals. Extensive experiments demonstrate that ESCRAG-R1 significantly outperforms existing baselines by mitigating superficial splicing and realizing a natural integration of professional guidance and empathetic expression. Code and datasets are released at https://github.com/Matcha-Liu/ESCRAG-R1.

1 Introduction

Driven by growing psychological needs and LLM advancements (Yi et al., 2025; Li et al., 2025), Emotional Support Conversation (ESC) systems have become vital resources for users seeking comfort and guidance (DeepSeek-AI, 2024; OpenAI, 2025; Comanici et al., 2025). In real-world scenarios, users often share their experiences and emotional states with ESC systems, expecting empathetic and constructive responses (Rashkin et al., 2019; Lin et al., 2019). The development of ESC systems is advancing toward deeper Emotional Intelligence, aiming to provide users with comprehensive emotional support that balances professional competence and empathy (Zhong et al., 2021; Wang et al., 2023; Chen et al., 2025; Hu et al., 2026).

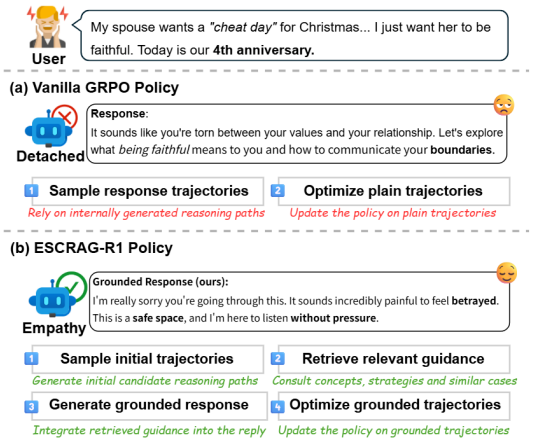

Figure 1: Comparison of GRPO and ESCRAG-R1. By integrating retrieved guidance into GRPO, ESCRAG-R1 produces more grounded and empathetic responses.

To achieve deep emotional intelligence, an effective ESC system must possess two core capabilities. The first is *structured, stage-aware reasoning*. Because emotional support is inherently process-oriented, generating a high-quality response requires inferring the client's counseling stage, assessing their psychological state, and selecting appropriate interventions (Beck, 2020; Hayes et al., 2006; Elliott, 2002). This explicit reasoning ensures that empathy and guidance emerge from a coherent framework rather than isolated, reactive utterances. The second capability is *empathy-expertise alignment*. Psychologically grounded guidance must be delivered in a supportive, non-didactic manner. Responses that are emotionally warm but clinically ungrounded offer only superfi-

*Equal contribution.
†Corresponding author.

cial reassurance, while technically accurate yet affectively detached replies risk alienating vulnerable users (Nienhuis et al., 2018). Ultimately, effective support hinges on seamlessly integrating rigorous psychological reasoning with carefully calibrated empathetic language.

Existing ESC research generally follows two paradigms, both struggling to simultaneously fulfill these ideal capabilities. (i) Parameter-based approaches use supervised fine-tuning (Chen et al., 2023; Zhang et al., 2024a) or reinforcement learning (Zhao et al., 2025; Wang et al., 2025a) to mimic empathetic patterns. However, they critically lack explicit guidance for structured, stage-aware reasoning, relying on surface-level imitation rather than dynamically adapting to the client's evolving psychological state. (ii) Inference-augmented approaches incorporate agents (Zhou et al., 2025; Zhang et al., 2024b) or retrieval-augmented generation (Hu et al., 2024; Guo et al., 2024) to inject psychological knowledge. Yet, this knowledge typically serves merely as external auxiliary guidance at inference time, rather than a definitive learning signal reshaping the model's internal policy. Consequently, avoiding this artificial splicing requires a unified framework that integrates explicit reasoning guidance directly into the model's internal policy.

To overcome these limitations, we propose **ESCRAG-R1**, a unified framework that integrates retrieval-based psychological guidance into Group Relative Policy Optimization (GRPO), as shown in Figure 1. ESCRAG-R1 leverages retrieved counseling strategies to stimulate explicit, stage-aware reasoning prior to generation, ensuring the model conducts an internal psychological assessment rather than surface-level imitation. Furthermore, by incorporating this retrieval process into the GRPO loop, external knowledge acts as a robust learning signal that reshapes the model's internal policy, seamlessly integrating professional therapeutic strategies with empathetic expression. To provide reliable reward supervision for this optimization, we construct a Client–Counselor–Judge evaluation framework and build **ESC-PREFERENCE**. This dataset delivers precise reward signals for GRPO, while its preferred responses serve as practice-grounded exemplars in the retrieval corpus for both training and inference.

Our contributions are summarized as:

- We propose ESCRAG-R1, a retrieval-augmented reinforcement learning framework that incorporates psychological guidance into GRPO for stage-aware emotional support.
- We construct ESC-PREFERENCE based on a three-dimensional Client–Counselor–Judge evaluation framework, supporting both reward modeling and retrieval grounding.
- Extensive experiments show that ESCRAG-R1 improves empathy-expertise alignment and generates more grounded, supportive, and therapeutically appropriate responses.

2 Related Work

Parameter-based ESC Systems. Parameter-based approaches internalize supportive behaviors through optimization on specialized datasets. As the foundation for training, studies propose psychological counseling dialogue datasets with diverse topics and strategies to support subsequent model training (Zhang et al., 2024a; Liu et al., 2021; Zheng et al., 2023). Many works adopt supervised fine-tuning (SFT) on counseling dialogues, enabling models to learn therapeutic patterns from the data and thereby deliver better services (Chen et al., 2023; Zhou et al., 2025). In reinforcement learning, some studies construct preference pair data and employ Direct Preference Optimization (DPO) (Rafailov et al., 2023) to align models with human preferences (Zhao et al., 2025). With the emergence of Group Relative Policy Optimization (GRPO), several works have conducted preliminary explorations, though these attempts rely entirely on LLMs to explore strategies autonomously, lacking external guidance or structured intervention (Wang et al., 2025b; Yang et al., 2025; Wang et al., 2025a). However, these parameter-based methods mainly learn supportive patterns from fixed training data, without explicitly incorporating practice-grounded counseling knowledge into policy optimization. As a result, they may lack guidance for stage-aware reasoning and context-sensitive intervention, motivating our retrieval-augmented reinforcement learning framework.

Inference-augmented ESC Systems. Inference-augmented approaches enhance ESC systems by introducing external knowledge or structured reasoning during inference without modifying model parameters. For example, some studies develop dialogue agent frameworks to construct ESC systems that are more empathetic and helpful (Zhang et al.,

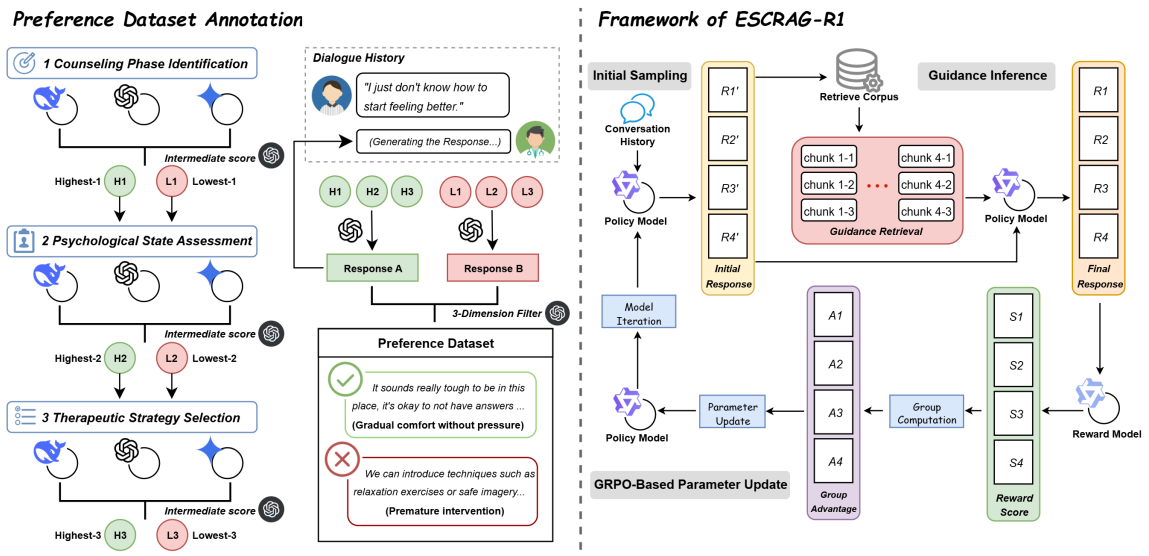

Figure 2: The left panel shows the stage-aware annotation pipeline for ESC-Preference, and the right panel illustrates how retrieval-augmented guidance is integrated into GRPO to optimize grounded, empathetic support responses.

2024b, 2025; Li et al., 2024a). Beyond reasoning-focused approaches, a growing body of research draws upon established psychological theories to ground system behavior in clinically validated principles (Xiao et al., 2024; Xu et al., 2025). Additionally, recent studies adopt retrieval-augmented generation (RAG) frameworks to incorporate external knowledge or relevant dialogue exemplars during inference (Xiong et al., 2024; Li et al., 2024b; Liu et al., 2025). Despite their strengths, the augmented information in these approaches often remains peripheral to the model's core decision-making, serving as auxiliary input rather than actively shaping the underlying intervention policy. This motivates us to integrate retrieval-based psychological guidance into reinforcement learning, so that external counseling knowledge can participate in policy optimization rather than only assisting inference.

3 Method

The overall framework of ESCRAG-R1 consists of two main components: **ESC-Preference Construction** and **Retrieval-Guided Policy Optimization**, as shown in Figure 2. ESC-Preference is constructed to provide both preference supervision for reward modeling and high-quality counseling exemplars for retrieval grounding. Based on this dataset, we first train a multi-perspective reward model, and then optimize the ESC policy through supervised fine-tuning and retrieval-augmented GRPO.

3.1 Task Definition

Emotional Support Conversation (ESC) is formulated as a turn-level counselor response generation task. A dialogue consists of alternating user and counselor utterances, denoted as $\mathcal{D} = \{u_1, c_1, u_2, c_2, \ldots, u_T, c_T\}$, where u_i is the i-th user utterance, c_i is the corresponding counselor response, and T is the number of counselor turns. At the i-th counselor turn, the model observes the dialogue history $\mathcal{H}_i = \{u_1, c_1, \ldots, u_{i-1}, c_{i-1}, u_i\}$ and generates a response c_i according to the counselor policy $\pi_\theta(\cdot \mid \mathcal{H}_i)$, where θ denotes the policy parameters.

The goal is to learn a counselor policy π_θ that maximizes the expected reward of generated responses over observed dialogue histories:

$$\pi_\theta^* = \arg\max_{\pi_\theta} \mathbb{E}_{\mathcal{H}_i, c_i \sim \pi_\theta(\cdot|\mathcal{H}_i)} \left[R(\mathcal{H}_i, c_i) \right], \quad (1)$$

where $R(\mathcal{H}_i, c_i)$ denotes the reward function that evaluates the emotional support quality of response c_i under dialogue history $\mathcal{H}_i$.

3.2 ESC-Preference Construction

To obtain reward signals that balance professional competence and empathetic quality, as well as practice-grounded guidance for both training and inference, we construct the ESC-Preference dataset, which provides multi-perspective preference pairs and high-quality counseling exemplars. Based on ESConv (Liu et al., 2021), we simulate multi-turn client–counselor interactions to construct counseling dialogues. As shown in Figure 2, candidate

models perform three-stage counseling inference to build contrasting reasoning paths, from which reliable preference pairs are obtained through stage-wise selection and final preference filtering. The prompts are provided in Appendix A.

Stage-wise Response Reasoning. Given a dialogue history, we construct candidate responses through a three-stage counseling reasoning framework. This design follows the progressive nature of counseling interactions, where the counselor first understands the counseling context, then assesses the client's psychological state, and finally selects an intervention strategy. This process grounds response generation in counseling-oriented reasoning rather than relying only on surface dialogue context. To obtain contrasting response candidates, three candidate models, including GPT-4o (OpenAI, 2024), Gemini-2.5-Pro (Comanici et al., 2025), and DeepSeek-V3 (DeepSeek-AI, 2024), independently perform the three-stage inference. At each stage, intermediate selection is applied to retain both high- and low-quality reasoning states. These selected states are propagated through the subsequent stages, producing two contrasting reasoning trajectories, which are then used by GPT-4o to generate paired counselor responses.

Multi-perspective Preference Evaluation. To evaluate the paired responses beyond a single overall score, we use GPT-5-chat (OpenAI, 2025) as the evaluator and adopt a multi-perspective evaluation framework covering the client, counselor, and judge viewpoints. The client perspective focuses on perceived emotional support quality, the counselor perspective evaluates therapeutic coherence and professional grounding, and the judge perspective assesses whether the response fits the client's state and targets the core problem. Based on these evaluation results, we apply a dominance-based filtering criterion: the preferred response must achieve non-negative improvement across all metrics and strictly outperform the alternative response in at least three dimensions. Only pairs satisfying these conditions are retained. Through this process, we obtain 3,667 high-confidence preference pairs, containing 7,334 responses in total.

3.3 Retrieval-Guided Policy Optimization

To move beyond inference-only guidance and enable retrieved counseling knowledge to shape the model policy, we use ESC-Preference to guide policy optimization. The preference pairs are used to train a reward model, and the preferred responses are organized as a retrieval corpus to provide counseling guidance. We first initialize the policy with a supervised reasoning-response pattern, and then integrate retrieval augmentation into GRPO. In this way, retrieved counseling exemplars are used during both policy optimization and inference, rather than serving only as inference-time guidance.

Preference-Guided Reward Learning. To provide reward signals for policy optimization, we train a reward model on the preference pairs in ESC-Preference. Each preference pair is represented as $(\mathcal{H}_i, c_i^+, c_i^-)$, where c_i^+ and c_i^- denote the preferred and rejected responses under the same dialogue state $\mathcal{H}_i$, respectively. The reward model estimates the relative quality of candidate responses by mapping each dialogue state and response to a scalar reward score. Specifically, it consists of a pretrained backbone $f_\theta(\cdot)$ and a linear scoring head $g_\psi(\cdot)$, and the reward score r_i^{RM} for a candidate response is computed as:

$$r_i^{\mathrm{RM}} = g_\psi(f_\theta(\mathcal{H}_i, c_i)) \,. \tag{2}$$

During training, the backbone parameters θ are frozen and only the scoring head parameters ψ are optimized. For each preference pair, the reward model produces scores $r_i^{RM+} = R_\phi(\mathcal{H}_i, c_i^+)$ and $r_i^{RM-} = R_\phi(\mathcal{H}_i, c_i^-)$ for the preferred and rejected responses respectively. The loss function of the reward model is defined as:

$$\mathcal{L}_{RM} = \frac{1}{N}\sum_{i=1}^{N}\Big[-\log\sigma(r_i^{RM+} - r_i^{RM-}) + \lambda(r_i^{RM+} + r_i^{RM-})^2\Big], \tag{3}$$

where $\sigma(\cdot)$ denotes the sigmoid function and λ is a regularization coefficient. The first term encourages the reward model to assign higher scores to preferred responses, while the second term constrains the reward scale to prevent score drifting during training.

During policy optimization, we further introduce an auxiliary *format reward* to encourage the policy model to follow the structured output format. Instead of directly generating the final counselor response, the model is expected to first organize its counseling reasoning and then provide the final supportive response. The format reward r_i^{format} is defined as:

$$r_i^{format} = \begin{cases} 1, & \text{if format satisfied;} \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The final reward r_i used during reinforcement learning is defined as:

$$r_i = r_i^{RM} + r_i^{format}. \tag{5}$$

Supervised Pattern Initialization. Before reinforcement learning, we initialize the policy model using a supervised fine-tuning (SFT) stage to establish a structured reasoning pattern for emotional support responses. Specifically, we generate 500 cold-start demonstrations using GPT-4o (OpenAI, 2024). Following the prompt design in Appendix B.2, each demonstration consists of a reasoning segment and a final response, formatted as `<think></think>` and `<response></response>`.

Given a dialogue state $\mathcal{H}_i$, let y_i denote the target structured output in the cold-start demonstration, which contains both the reasoning segment and the final response. The policy model π_θ is trained to generate y_i conditioned on $\mathcal{H}_i$. The supervised fine-tuning objective is defined as:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(\mathcal{H}_i, y_i)} \left[\log \pi_\theta(y_i \mid \mathcal{H}_i) \right]. \tag{6}$$

This stage enables the model to learn the reasoning–response output format and provides a stable initialization for subsequent reinforcement learning with GRPO.

Retrieval-Augmented GRPO. After supervised pattern initialization, we further optimize the counselor policy using GRPO with retrieval augmentation. Different from inference-only RAG, our goal is to expose the policy to retrieved counseling exemplars during the rollout process, so that retrieval-based guidance can influence both response generation and policy updates. Given a dialogue state $\mathcal{H}_i$, the policy first generates an initial response $\hat{c}_i$:

$$\hat{c}_i \sim \pi_\theta(\cdot \mid \mathcal{H}_i). \tag{7}$$

The initial response is used as a query to retrieve relevant counseling exemplars from the retrieval corpus constructed during dataset generation. Let $\mathcal{E}_i = \{e_1, \ldots, e_k\}$ denote the retrieved exemplars. The policy then generates the final response conditioned on the dialogue state, the initial response, and the retrieved exemplars:

$$c_i \sim \pi_\theta(\cdot \mid \mathcal{H}_i, \hat{c}_i, \mathcal{E}_i). \tag{8}$$

During training, for each dialogue state $\mathcal{H}_i$, the current policy samples a group of N retrieval-augmented candidate responses $\{c_i^j\}_{j=1}^N$, and each response is evaluated using the reward signal defined above. The group-relative advantage A_i^j for each sampled response is computed as:

$$A_i^j = \frac{r_i^j - \frac{1}{N}\sum_{k=1}^N r_i^k}{\sqrt{\frac{1}{N}\sum_{k=1}^N \left(r_i^k - \frac{1}{N}\sum_{l=1}^N r_i^l\right)^2 + \epsilon}}, \tag{9}$$

where r_i^j denotes the reward of the j-th sampled response, while r_i^k is used as the summation term over all responses in the same group. The constant ϵ is used for numerical stability.

We define the importance sampling ratio ρ_i^j between the current policy and the old policy for the j-th sampled response as:

$$\rho_i^j = \frac{\pi_\theta(c_i^j \mid \mathcal{H}_i, \hat{c}_i, \mathcal{E}_i)}{\pi_{\theta_{\text{old}}}(c_i^j \mid \mathcal{H}_i, \hat{c}_i, \mathcal{E}_i)}, \tag{10}$$

where ρ_i^j measures how the likelihood of the sampled response changes under the current policy π_θ compared with the old policy $\pi_{\theta_{\text{old}}}$.

Based on the importance sampling ratio ρ_i^j and the group-relative advantage A_i^j, the policy is optimized with the following clipped GRPO objective:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{\mathcal{H}_i, \{c_i^j\}_{j=1}^N} \Bigg[\frac{1}{N} \sum_{j=1}^{N} \min\left(\rho_i^j A_i^j,\ \text{clip}(\rho_i^j, 1-\delta, 1+\delta) A_i^j \right) \Bigg], \tag{11}$$

where $\text{clip}(\rho_i^j, 1-\delta, 1+\delta)$ constrains the policy ratio within $[1-\delta, 1+\delta]$, and δ is a clipping coefficient that limits excessively large policy updates. The overall procedure of retrieval-augmented GRPO and RAG inference is summarized in Appendix D.

4 Experiment

4.1 Experiment Settings

Dataset. We conduct experiments on ESConv (Liu et al., 2021) using its default training and test split. The training set contains 910 dialogues with 10,939 turns, which are used for reward model training, retrieval corpus construction, and GRPO-based policy optimization. During evaluation, we assess model performance on the test split, which contains 195 dialogues with a total of 2,505 turns.

Model	Client			Counselor		Judge		Avg
	EI	NE	TA	SC	TH	SF	PT	
General-Purpose Models (Raw)								
GPT-4o	8.19	8.61	8.07	8.87	7.90	8.40	7.56	8.23
GPT-4o *with RAG*	**8.79**	8.96	8.56	<u>8.95</u>	**8.13**	**8.76**	**7.96**	<u>8.59</u>
Llama-3.1-8B-Instruct	8.20	8.48	8.06	8.71	7.78	8.27	<u>7.65</u>	8.16
Llama-3.1-8B-Instruct *with RAG*	8.73	8.50	<u>8.59</u>	8.09	7.71	8.29	7.57	8.21
Qwen-2.5-7B-Instruct	7.62	7.63	7.51	8.49	7.36	7.63	6.92	7.59
Qwen-2.5-7B-Instruct *with RAG*	8.31	8.50	8.12	8.22	7.49	8.00	7.18	7.97
Specialized ESC models								
ChatCounselor-7B	6.22	5.59	6.25	7.06	6.07	6.09	5.56	6.12
PsyLLM-8B	8.59	9.18	8.34	8.68	7.87	8.55	<u>7.65</u>	8.41
ESCRAG-R1								
ESCRAG-R1-3B *with RAG*	8.68	<u>9.31</u>	8.55	8.80	8.02	8.58	7.56	8.50
ESCRAG-R1-7B *with RAG*	<u>8.74</u>	**9.60**	**8.68**	**8.98**	<u>8.09</u>	<u>8.66</u>	<u>7.65</u>	**8.63**

Table 1: Evaluation results from three perspectives. Client: EI (Emotional Impact), NE (Non-Lecturing Empathy), TA (Therapeutic Alliance); Counselor: SC (Strategic Coherence), TH (Theoretical Application); Judge: SF (Strategy-Client Fit), PT (Problem Targeting). The best and second-best results are **bolded** and <u>underlined</u>, respectively.

Evaluation Metrics. We evaluate model performance from three complementary perspectives: Client, Counselor, and Judge. For automatic evaluation, we use GPT-5-chat as the evaluator under a unified scoring prompt. Detailed descriptions of the metrics are provided in Appendix A.2.

Implementation Details. For reward model, we adopt Qwen-2.5-7B (Yang et al., 2024; Team, 2024) as the backbone. For policy optimization, we use Qwen-2.5-3B-Instruct and Qwen-2.5-7B-Instruct as the backbone policy models. For the retrieval component, we use bge-m3 (Chen et al., 2024) to encode dialogue chunks and construct the vector database. For fairness, all RAG-based experiments use the same retrieval number, with $N = 3$ retrieved exemplars. Detailed hyperparameter settings are provided in Appendix C.

Baselines. We compare our method against two categories of baselines: general-purpose large language models and specialized emotional support dialogue models. The general-purpose models include GPT-4o (OpenAI, 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen-2.5-7B-Instruct(Yang et al., 2024; Team, 2024), while the specialized models include ChatCounselor-7B (Liu et al., 2023) and PsyLLM-8B (Hu et al., 2025). Considering the different architectures and usage settings of the models, we additionally evaluate the RAG setting only for general-purpose models.

4.2 Main Results

Table 1 presents the overall comparison results from the Client, Counselor, and Judge perspectives. The first key finding is that ESCRAG-R1 achieves the strongest overall performance among all compared models. Specifically, ESCRAG-R1-7B with RAG obtains the highest average score of 8.63, while ESCRAG-R1-3B with RAG also achieves a competitive score of 8.50. These results demonstrate that the proposed framework is effective across different model scales and consistently improves the overall quality of ESC responses.

The second key finding is that ESCRAG-R1 achieves a better balance between empathetic expression and professional grounding. Compared with GPT-4o with RAG, ESCRAG-R1-7B obtains a substantially higher score on Non-Lecturing Empathy (9.60 vs. 8.96) and a higher score on Therapeutic Alliance (8.68 vs. 8.56), indicating that our model produces responses that are perceived as more supportive and less didactic even when compared with a strong retrieval-augmented general-purpose model. Meanwhile, ESCRAG-R1-7B also maintains strong Counselor-side performance, with 8.98 on Strategic Coherence and 8.09 on Theoretical Application. These results show that ESCRAG-R1 improves empathetic expression while preserving professional therapeutic reasoning, thereby better supporting empathy-expertise alignment.

The third key finding is that specialized ESC

Model	Client			Counselor		Judge		Avg
	EI	NE	TA	SC	TH	SF	PT	
Ablation on Framework Components								
Qwen-2.5-3B-Instruct	7.18	7.20	7.18	7.97	6.92	7.08	6.31	7.12
Qwen-2.5-3B-Instruct *with RAG*	7.93	8.16	7.81	7.18	6.90	7.00	6.12	7.30 ↑ 0.18
Cold-Start SFT	8.14	8.46	8.00	8.88	7.88	8.21	7.43	8.14 ↑ 1.02
+ *Vanilla-GRPO*	8.19	9.06	8.13	9.00	8.02	8.52	7.55	8.35 ↑ 1.23
+ *Vanilla-GRPO* + *RAG Inference*	8.49	9.30	8.35	8.96	8.03	8.46	7.52	8.44 ↑ 1.32
ESCRAG-R1-3B + *RAG Inference*	8.68	9.31	8.55	8.80	8.02	8.58	7.56	8.50 ↑ 1.38

Table 2: Ablation on framework components. Vanilla-GRPO uses standard GRPO training, RAG Inference applies retrieval augmentation only at test time, and ESCRAG-R1-3B uses our full joint training framework with retrieval augmentation during both training and test time.

models still show limitations in jointly maintaining empathy and professional grounding. Although PsyLLM-8B achieves a strong average score of 8.41 and performs well on empathy-related metrics, ESCRAG-R1-7B further improves all three Client-side metrics and achieves higher Counselor-side scores, increasing Strategic Coherence from 8.68 to 8.98 and Theoretical Application from 7.87 to 8.09. Meanwhile, ChatCounselor-7B obtains a much lower average score of 6.12, with weak performance across both Client and Judge dimensions. These results suggest that counseling-oriented training alone may be insufficient to consistently produce responses that are both empathetic and strategically appropriate, highlighting the importance of integrating structured reasoning and preference-based policy optimization.

4.3 Ablation Analysis

Table 2 reports the ablation study on the core components of the framework, examining how each module contributes to the overall performance. Figure 3 further investigates how varying the number of retrieved exemplars affects model performance.

Ablation on Framework Components. Table 2 shows that each component of ESCRAG-R1 contributes to the overall performance. Starting from the backbone, Qwen-2.5-3B-Instruct achieves an average score of 7.12, and adding RAG brings a modest improvement to 7.30, suggesting that retrieved exemplars provide useful supportive cues but are insufficient when used only as external inference-time context. After Cold-Start SFT establishes a structured reasoning-and-response pattern, *Vanilla-GRPO* further improves the average score from 8.14 to 8.35, showing that reinforcement learning can enhance the general quality of emotional support generation. Adding *RAG Inference* on top of Vanilla-GRPO further raises the score to 8.44, indicating that retrieved examples can provide additional guidance during generation. The full ESCRAG-R1-3B framework achieves the best overall score of 8.50, outperforming both *Vanilla-GRPO* and *Vanilla-GRPO + RAG Inference*. Notably, compared with *Vanilla-GRPO + RAG Inference*, ESCRAG-R1 further improves the Client-side metrics such as Emotional Impact and Therapeutic Alliance, as well as the Judge-side metrics of Strategy–Client Fit and Problem Targeting. These improvements suggest that integrating retrieval into policy optimization helps the model better internalize practice-grounded counseling strategies, leading to more supportive responses that are better aligned with the client's state and core problem.

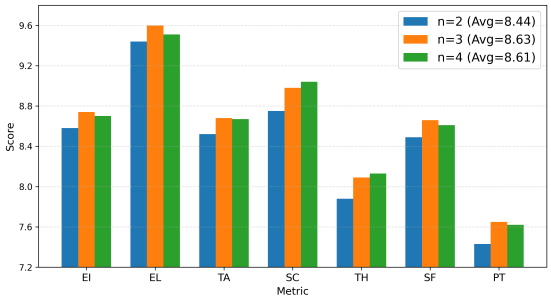

Figure 3: Performance of ESCRAG-R1-7B with different retrieval numbers across evaluation dimensions.

Ablation on Retrieval Number. Figure 3 illustrates the effect of different retrieval numbers on ESCRAG-R1-7B. Overall, increasing the retrieval number from $n = 2$ to $n = 3$ leads to consistent improvements across all evaluation dimensions, raising the average score from 8.44 to 8.63. This

Client: I feel like my friend group judges me for simply being myself. Although they are probably joking, I sometimes take their words personally, and it makes me upset. **Counselor:** It sounds like your friends' jokes can feel judgmental to you, and even if they do not mean harm, their words still end up hurting you. **Client:** Should I still hang out with them if they keep making me feel this way?		
Response	**Initial Response (Vanilla-GRPO)** It sounds like this situation is really tough for you. It's understandable that you feel upset when people you care about are not being supportive. Maybe we can explore together what being in this situation feels like for you and what boundaries, if any, you might want to set. Would you be comfortable discussing how you want to approach your friend group moving forward?	**Final Response (ESCRAG-R1)** It sounds like this situation is really affecting you, and I want to acknowledge how difficult that must be. It makes sense that you feel upset when people you care about say things that hurt, even if they are joking. Would you like to talk more about how this is affecting you? We can take our time to understand your feelings and what matters most to you. Remember, you're not alone in this, and I'm here to support you as you explore your options.
Retrieved Chunks	**[Chunk 1]** *History:* The client is hurt and confused after a friend said she hated them while they were also dealing with family problems. *Response:* The counselor validates the client's pain and confusion, encourages them to process feelings at their own pace, and reassures them that they do not have to go through this alone. **[Chunk 2]** *History:* The client feels rejected because friends are cold and unwilling to meet, then asks what they should do. *Response:* The counselor acknowledges the hurt, avoids direct advice, invites the client to explore how the coldness affects them, and supports taking things one step at a time.	

Table 3: Case study of ESCRAG-R1. Retrieved chunks guide the model to produce a more supportive final response, with highlighted text showing improved counseling behaviors. During training, the reward of such retrieval-guided responses is used to update the policy, allowing external guidance to shape model parameters.

suggests that using too few retrieved exemplars may provide insufficient counseling guidance for response refinement. When the retrieval number further increases to $n = 4$, the overall performance remains competitive but slightly decreases to 8.61, indicating that additional retrieved examples do not necessarily bring further gains and may introduce redundant or less focused guidance.

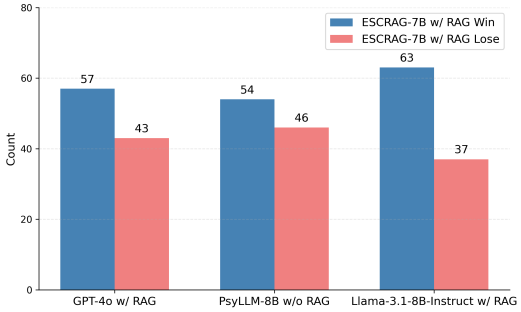

Figure 4: Human evaluation results on 100 sampled dialogue contexts from the ESConv test set.

4.4 Human Evaluation

We further conduct human evaluation on 100 randomly sampled dialogue contexts from the ESConv test set. As shown in Figure 4, ESCRAG-R1-7B with RAG consistently achieves more wins than losses against all three baselines, obtaining 57 wins against GPT-4o with RAG, 54 wins against PsyLLM-8B without RAG, and 63 wins against Llama-3.1-8B-Instruct with RAG. Interestingly, although GPT-4o with RAG performs strongly in LLM-as-Judge evaluation, human annotators still prefer ESCRAG-R1 in more cases, and the comparison with PsyLLM-8B is relatively close. This suggests that models whose policies are updated with psychological counseling data may produce responses that better match human expectations.

4.5 Case Study

Table 3 illustrates how retrieval augmentation helps ESCRAG-R1 refine its response. Compared with the initial response, the final response provides warmer emotional validation, avoids premature advice, and maintains a more supportive stance. The retrieved chunks guide the model to acknowledge the client's hurt and encourage further exploration rather than directly prescribing solutions. Since retrieval-guided responses are evaluated during GRPO training, such guidance can further shape the policy beyond inference-time generation.

5 Conclusion

In this paper, we propose **ESCRAG-R1**, a retrieval-augmented reinforcement learning frame-

work for emotional support conversation. We construct **ESC-PREFERENCE** with a multi-perspective evaluation framework to support reward modeling and retrieval grounding. Experimental results show that ESCRAG-R1 improves emotional support quality across model scales and achieves stronger empathy-expertise alignment than general-purpose and specialized ESC models.

6 Limitations

Despite the effectiveness of ESCRAG-R1, this work still has several limitations. First, due to computational resource constraints, our experiments are mainly conducted on 3B and 7B policy models. Although the results demonstrate consistent improvements across different model scales, we have not fully investigated how the proposed framework performs on larger-scale LLMs. Second, this work focuses on text-only Emotional Support Conversation. However, real-world emotional support often involves multimodal signals, such as facial expressions, speech tone, and other behavioral cues, which may provide important information about the user's emotional state.

7 Ethical Considerations

This work aims to improve the performance of Emotional Support Conversation (ESC) models and is intended for research purposes only. All data and models used in this work are based on publicly available and open-source resources, and our experiments do not involve private user data or personally identifiable information.

During dataset construction, we carefully design the annotation and generation pipeline to reduce the risk of producing harmful or inappropriate content. The prompts used for client simulation, counselor reasoning, and response evaluation are designed to guide the models toward supportive, non-harmful, and therapeutically appropriate interactions, as detailed in Appendix A. Nevertheless, we acknowledge that automatically generated emotional support responses may still carry potential risks if used in real-world settings without proper safeguards.

Therefore, ESCRAG-R1 and the constructed dataset are intended only for academic research. They should not be directly used as a substitute for professional psychological counselors or mental health services. Any future practical deployment should include strict safety mechanisms, human oversight, and clear instructions for users to seek professional help in high-risk situations.

8 Acknowledgements

This work was supported in part by Guangdong Province Science and Technology Foundation (No. 2026A1515011806, No. 2024TQ08X559), in part by Innovation Team Project of Guangdong Province (No. 2024KCXTD017), in part by Shenzhen Science and Technology Foundation (No. JCYJ20240813145816022).

References

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

Jian Chen, Yuxuan Hu, Haifeng Lu, Wei Wang, Min Yang, Chengming Li, and Xiping Hu. 2025. Mghft: Multi-granularity hierarchical fusion transformer for cross-modal sticker emotion recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5794–5803.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

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. *arXiv preprint arXiv:2311.00273*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Robert Elliott. 2002. The effectiveness of humanistic therapies: A meta-analysis.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qiming Guo, Jinwen Tang, Wenbo Sun, Haoteng Tang, Yi Shang, and Wenlu Wang. 2024. Soullmate: An application enhancing diverse mental health support with adaptive llms, prompt engineering, and rag techniques. *arXiv preprint arXiv:2410.16322*.

Steven C Hayes, Jason B Luoma, Frank W Bond, Akihiko Masuda, and Jason Lillis. 2006. Acceptance and commitment therapy: Model, processes and outcomes. *Behaviour research and therapy*, 44(1):1–25.

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

Yuxuan Hu, Jian Chen, Yuhao Wang, Zixuan Li, Jing Xiong, Pengyue Jia, Wei Wang, Chengming Li, and Xiangyu Zhao. 2026. Emotion and intention guided multi-modal learning for sticker response selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 14883–14891.

Yuxuan Hu, Minghuan Tan, Chenwei Zhang, Zixuan Li, Xiaodan Liang, Min Yang, Chengming Li, and Xiping Hu. 2024. Aptness: Incorporating appraisal theory and emotion support strategies for empathetic response generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 900–909.

Junlin Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024a. Be helpful but don't talk too much-enhancing helpfulness in conversations through relevance in multi-turn emotional support. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1976–1988.

Zixuan Li, Binzong Geng, Jing Xiong, Yong He, Yuxuan Hu, Jian Chen, Dingwei Chen, Xiyu Chang, Ngai Wong, Liang Zhang, and 1 others. 2025. Ctr-sink: Attention sink for language models in click-through rate prediction. *arXiv preprint arXiv:2508.03668*.

Zixuan Li, Jing Xiong, Fanghua Ye, Chuanyang Zheng, Xun Wu, Jianqiao Lu, Zhongwei Wan, Xiaodan Liang, Chengming Li, Zhenan Sun, and 1 others. 2024b. Uncertaintyrag: Span-level uncertainty enhanced long-context modeling for retrieval-augmented generation. *arXiv preprint arXiv:2410.02719*.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132.

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

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pages 3469–3483.

Weichu Liu, Jing Xiong, Yuxuan Hu, Zixuan Li, Minghuan Tan, Ningning Mao, Hui Shen, Wendong Xu, Chaofan Tao, Min Yang, and 1 others. 2025. Longemotion: Measuring emotional intelligence of large language models in long-context interaction. *arXiv preprint arXiv:2509.07403*.

Jacob B Nienhuis, Jesse Owen, Jeffrey C Valentine, Stephanie Winkeljohn Black, Tyler C Halford, Stephanie E Parazak, Stephanie Budge, and Mark Hilsenroth. 2018. Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: A meta-analytic review. *Psychotherapy Research*, 28(4):593–605.

OpenAI. 2024. OpenAI: Hello GPT-4o. https://openai.com/zh-Hans-CN/index/hello-gpt-4o/. Accessed: 2025-07-24.

OpenAI. 2025. Gpt-5 system card.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5370–5381.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, and 1 others. 2025a. Rlver: Reinforcement learning with verifiable emotion rewards for empathetic agents. *arXiv preprint arXiv:2507.03112*.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Yunxiao Wang, Meng Liu, Wenqi Liu, Kaiyu Jiang, Bin Wen, Fan Yang, Tingting Gao, Guorui Zhou, and Liqiang Nie. 2025b. Compeer: Controllable empathetic reinforcement reasoning for emotional support conversation. *arXiv preprint arXiv:2508.09521*.

Mengxi Xiao, Qianqian Xie, Ziyan Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. Healme: Harnessing cognitive reframing in large language models for psychotherapy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725.

Jing Xiong, Zixuan Li, Chuanyang Zheng, Zhijiang Guo, Yichun Yin, Enze Xie, Zhicheng Yang, Qingxing Cao, Haiming Wang, Xiongwei Han, and 1 others. 2024. Dq-lore: Dual queries with low rank approximation re-ranking for in-context learning. In *International Conference on Learning Representations*, volume 2024, pages 41179–41203.

Ancheng Xu, Di Yang, Renhao Li, Jingwei Zhu, Minghuan Tan, Min Yang, Wanxin Qiu, Mingchen Ma, Haihong Wu, Bingyu Li, and 1 others. 2025. Autocbt: An autonomous multi-agent framework for cognitive behavioral therapy in psychological counseling. *arXiv preprint arXiv:2501.09426*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Ting Yang, Li Chen, and Huimin Wang. 2025. Towards open-ended emotional support conversations in llms via reinforcement learning with future-oriented rewards. *arXiv preprint arXiv:2508.12935*.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2025. A survey on recent advances in llm-based multi-turn dialogue systems. *ACM Computing Surveys*, 58(6):1–38.

Chenhao Zhang, Renhao Li, Minghuan Tan, Min Yang, Jingwei Zhu, Di Yang, Jiahao Zhao, Guancheng Ye, Chengming Li, and Xiping Hu. 2024a. Cpsycoun: A report-based multi-turn dialogue reconstruction and evaluation framework for chinese psychological counseling. *arXiv preprint arXiv:2405.16433*.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024b. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412.

Xinjie Zhang, Wenxuan Wang, and Qin Jin. 2025. IntentionESC: An intention-centered framework for enhancing emotional support in dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26494–26516, Vienna, Austria. Association for Computational Linguistics.

Weixiang Zhao, Xingyu Sui, Xinyang Han, Yang Deng, Yulin Hu, Jiahe Guo, Libo Qin, Qianyun Du, Shijin Wang, Yanyan Zhao, Bing Qin, and Ting Liu. 2025. Chain of strategy optimization makes large language models better emotional supporter. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15361–15381, Suzhou, China. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021. Care: commonsense-aware emotional response generation with latent concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14577–14585.

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

A Prompt Design for ESC-Preference Construction

A.1 Client Simulator Prompt

We design a role-playing prompt to simulate the client as a patient seeking emotional support. The prompt enforces the client to remain consistent with the given emotional and situational context, respond naturally to the therapist, and generate only one short and succinct sentence for the next turn without additional explanation. The prompt design can be seen in Figure 5.

Client Simulation Prompt

Now enter the role-playing mode. In the following conversation, you will play as a patient in a counselling conversation with a therapist.

You are the patient who is looking for the help from the therapist, because you have the emotional issue about {emotion_type} regarding {experience_type}. Please reply with only one short and succinct sentence. You are currently facing problem {problem_type} and are in situation {situation}.

Here is the existing dialogue between you and the therapist:
{dialogue}

Based on the above information, please provide your next turn of response.

Figure 5: Client role-playing prompt.

A.2 Evaluator Simulator Prompt

We design an evaluator prompt to evaluate candidate responses from multiple perspectives and provide reliable preference signals. The evaluation considers three complementary roles: client, counselor, and judge, ensuring both empathetic quality and professional appropriateness. During automatic evaluation, we also adopt the same multi-perspective criteria in the LLM-as-a-Judge setting. The prompt design can be seen in Figure 6.

A.3 Counselor Simulator Prompt

We design a stage-aware prompting framework that decomposes the counselor's reasoning into three stages, each implemented with a dedicated prompt.

Counseling stage identification. The model infers the client's current counseling phase based on the dialogue context. The prompt design can be seen in Figure 7.

```
## Evaluation Prompt

You are a psychotherapy process evaluator. Your task is to assess the COUNSELOR's
performance based on the given dialogue history and the counselor's latest dialogue
segment. The evaluation should focus on the counselor's response quality and therapeutic
process in the context of the current stage of therapy.

[Dialogue History]:
{dialogue_history}

[Counselor's Latest Dialogue Segment]:
{latest_dialogue_segment}

[Evaluation Dimensions & Scoring]:
Provide a comprehensive evaluation from the following three perspectives. Please provide a
score for each sub-item using the following scale:
Scoring Scale: 0-10 (where 10 = Excellent/Perfect, 5 = Average, 0 = Poor/Missing or
Completely Inadequate)

Perspective 1: Client's Perspective
Emotional Impact (Score: ): Did the response help the client feel better, validated, and
understood?
Empathy vs. Lecturing (Score: ): Did the counselor demonstrate empathy and attunement,
or resort to lecturing or premature advice?
Therapeutic Alliance (Score: ): Did the response reduce power distance and foster
collaboration and trust?

Perspective 2: Counselor's Perspective
Strategic Coherence (Score: ): Was there a clear logical connection between the counselor's
strategy, reasoning, and final response?
Theoretical Application (Score: ): Was the application of psychological theory accurate and
appropriate?

Perspective 3: Judge's Perspective
Strategy-Client Fit (Score: ): Does the chosen strategy match the client's issues and current
stage of therapy?
Problem Targeting (Score: ): Did the response address the client's most pressing problem at
this point?

[Output Format]:
Return the evaluation in JSON format:
```json
{{
   "Emotional Impact": [score],
   "Empathy vs. Lecturing": [score],
   "Therapeutic Alliance": [score],
   "Strategic Coherence": [score],
   "Theoretical Application": [score],
   "Strategy-Client Fit": [score],
   "Problem Targeting": [score]
}}
```

After scoring, please provide a brief overall summary of the counselor's strengths and areas
for improvement.
```

Figure 6: Evaluator role-playing prompt.

```
## Stage 1: Counseling Phase Identification

You are an experienced psychological counselor.

Your task: Determine which counseling process stage the client is currently in, based on the
dialogue.

Possible stages:
1. Relationship-building and goal assessment (Early stage)
2. Problem exploration and clarification – Conceptualization (Early middle stage)
3. Cognitive intervention and action promotion (Late middle stage)
4. Consolidation and termination (Late stage)

Use the conversation below to infer the current counseling stage.

[Conversation History]
{conversation_history}

Output strictly in the following format:

Stage number: [1–4]
Stage name: [corresponding name]
Reasoning: [A short paragraph explaining your reasoning, including the client's focus,
communication tone, and therapeutic process stage]
```

Figure 7: Counselor role-playing prompt for counseling stage identification.

Psychological state assessment. The model evaluates the client's psychological state in terms of distress, resilience, and motivation. The prompt design can be seen in Figure 8.

```
## Stage 2: Psychological State Assessment

You are an experienced psychological counselor.

You have already determined the client's current counseling stage from Stage 1.

[Stage 1 Result]
{stage1_output}

[Conversation History]
{conversation_history}

Now, based on the conversation and the above stage assessment, determine the client's
psychological state using the categories below:

1. Low distress, low resilience, low motivation
2. High distress, low resilience, high motivation
3. High distress, low resilience, low motivation
4. High distress, high resilience, low motivation
5. Low distress, low resilience, high motivation
6. High distress, high resilience, high motivation
7. Low distress, high resilience, high motivation
8. Low distress, high resilience, low motivation

Output strictly in the following format:

Category number: [1–8]
Category description: [corresponding description]
Reasoning: [Explain emotional distress level, resilience, and motivation to change, and why
this matches the category]
```

Figure 8: Counselor role-playing prompt for psychological state assessment.

Therapeutic strategy selection. The model selects appropriate counseling strategies conditioned on the inferred stage and psychological state. The prompt can be seen in Figure 9 and Figure 10.

```
## Stage 3: Therapeutic Strategy Selection

You are an experienced psychological counselor.

You have already determined the client's current counseling stage from Stage 1.
[Stage 1 Result]
{stage1_output}

You have determined the client's current psychological state based on Stage 2.
[Stage 2 Result]
{stage2_output}

[Conversation History]
{conversation_history}

Your task: Select the most appropriate counseling strategies from the following table based
on the client's state.

Stage 3: Strategy Selection
Choose the appropriate strategy based on the client's state:
1.Low distress, low resilience, low motivation
Fragile state maintenance - Supportive accompaniment and resource activation
Possible strategies:
Supportive empathy: Emotional reflection + Positive attention
Resource activation: Support system reinforcement + Strengths review
Psychological education: Normalization of the problem + Resilience building

2.High distress, low resilience, high motivation
Urgent help-seeking in crisis - Stabilization and structuring
Possible strategies:
Emotional stabilization: Relaxation training + Emotional labeling + Safe imagery
Structured sessions: Agenda setting + Psychological education + Goal breakdown
Skill training: Stress tolerance techniques + Emotional regulation exercises

3.High distress, low resilience, low motivation
Resistance in crisis state - Safety establishment and motivation cultivation
Possible strategies:
Emotional containment: Accepting listening + Non-invasive accompaniment
Motivation enhancement: Encouraging small changes + Respecting autonomy
Relationship building: Trust reinforcement + Strengthening cooperation alliance

4.High distress, high resilience, low motivation
Resources available but lacking motivation - Motivation awakening and value connection
Possible strategies:
Motivational interviewing: Exploring ambivalence about change + Decision balance analysis
(pros and cons of change)
Value clarification: Value prioritization + Behavior-value connection
Motivation activation: Stress relief + Intrinsic motivation guidance
```

Figure 9: Counselor role-playing prompt for therapeutic strategy selection (part 1).

5.Low distress, low resilience, high motivation
Fragile but eager for change - Skill development and confidence building
Possible strategies:
Reality testing: Adjusting expectations + Accumulating small successes
Behavioral activation: Assigning progressive tasks + Encouraging behavioral experiments
Self-efficacy enhancement: Identifying strengths + Specific feedback and reinforcement of behaviors

6.High distress, high resilience, high motivation
Ideal working state - In-depth intervention and core work
Possible strategies:
Experiential exploration: Role-playing + Emotion-focused + Body awareness
Cognitive restructuring: Belief testing + Thought recording
Relationship pattern analysis: Transference exploration + Schema identification

7.Low distress, high resilience, high motivation
Growth-oriented counseling - Optimization and potential development
Possible strategies:
Cognitive flexibility training: Guiding multi-angle thinking + Challenging cognitive rigidity
Behavioral optimization: Habit reshaping plan + Effectiveness improvement strategies
Meaning exploration: Value clarification exercises + Life narrative reconstruction

8.Low distress, high resilience, low motivation
Staying in the comfort zone - Developmental catalysis and meaning stimulation
Possible strategies:
Existential issues exploration: Meaning exploration + Responsibility and freedom discussion
Future-oriented intervention: Vision building + Staircase goal setting
Challenging support: Gentle confrontation + Subtle potential stimulation

Now, considering the client's emotional state, resilience, and motivation, choose the strategies that best suit this client's needs.

Output strictly in the following format:

Selected strategies: [List the primary strategy or combined approach chosen]
Reasoning: [Briefly explain why these are appropriate given the client's state and counseling process stage]

Figure 10: Counselor role-playing prompt for therapeutic strategy selection (part 2).

B Prompt Design for ESCRAG-R1 Training Framework

B.1 Reward Model Prompt

The reward model is trained to evaluate whether a counselor response effectively addresses the client's emotional problem under a given conversation history.

Reward Model Prompt

Instruction: You are a helpful assistant tasked with evaluating whether a patient's emotional problem has been effectively addressed following a conversation with a therapist.
Conversation History:
{conversation_history}
Candidate Response:
{candidate_response}

B.2 Policy Model Prompt

The policy model is prompted to generate responses in a structured reasoning–response format. During retrieval-augmented generation, the model receives the dialogue history, an initial response, and retrieved dialogue exemplars from the external corpus, and then generates the final counselor response.

Policy Model Prompt

System Prompt:
Now enter the role-playing mode. In the following conversation, you will play as a therapist in a counselling conversation with a patient. Your goal is to help the patient reduce their emotional distress and support them working through their challenges. You first think about the reasoning process in the mind and then provide the patient with the response. The reasoning process and response are enclosed within `<think> </think>` and `<response> </response>` tags, respectively, i.e., `<think> reasoning process here </think> <response> supportive response here </response>`.

User Prompt:
You are an empathetic and helpful assistant. Given the following dialogue context, your initial response and response guidelines retrieved from an external library, your task is to generate a final response to the user. This response should incorporate empathy and understanding, provide helpful guidance or suggestions, and be conversational and natural.
[Dialogue History]:
{dialogue_history}
[Initial Response]:
{initial_response}
[Retrieved Chunks from External Library]:
{retrieved_dialogue}

C Parameter Setting

Table 4 summarizes the main parameter settings used for reward model training and Retrieval-Augmented GRPO. The reward model is built on Qwen-2.5-7B and trained using two A800 GPUs. For policy optimization, ESCRAG-R1-3B is trained on two A800 GPUs, while ESCRAG-R1-7B is trained on four A800 GPUs.

Parameter	Value
Reward Model Training	
Epochs	5
Per-device batch size	1
Gradient accumulation steps	16
Learning rate	5×10^{-5}
Reward centering coefficient	0.01
Retrieval-Augmented GRPO	
Batch size	128
Epochs	2
Learning rate	1×10^{-6}
KL coefficient	0.05
Number of sample generation	4
Number of retrieved chunks	3
Temperature	0.9
Top-p	0.95

Table 4: Hyperparameter settings for reward model training and Retrieval-Augmented GRPO.

D Retrieval-Augmented GRPO and RAG Inference

Algorithm 1 summarizes how retrieval is used in ESCRAG-R1 during both training and inference. In retrieval-augmented GRPO, retrieved counseling exemplars are incorporated into the rollout process, and the reward of retrieval-guided responses is used to update the policy. In RAG inference, the optimized policy uses the same retrieval-augmented generation process to produce the final response, but no parameter update is performed. Thus, retrieval serves as a learning signal during training and as generation guidance during inference.

Algorithm 1 Retrieval-Augmented GRPO and RAG Inference of ESCRAG-R1

Require: Training set $\mathcal{D}_{\text{train}} = \{\mathcal{H}_i\}_{i=1}^{M}$, test set $\mathcal{D}_{\text{test}} = \{\mathcal{H}_i\}_{i=1}^{M'}$, policy model π_θ, old policy $\pi_{\theta_{\text{old}}}$, reward model R_ϕ, retriever $\mathcal{R}$, group size N

Ensure: Optimized policy parameters θ and final responses $\{c_i^{\text{final}}\}_{i=1}^{M'}$

Retrieval-Augmented GRPO

for each training step **do**
  Sample a dialogue state $\mathcal{H}_i$ from $\mathcal{D}_{\text{train}}$
  for $j = 1$ to N **do**
    Generate an initial response $\hat{c}_i^j \sim \pi_\theta(\cdot \mid \mathcal{H}_i)$
    Retrieve relevant exemplars $\mathcal{E}_i^j \leftarrow \mathcal{R}(\hat{c}_i^j)$
    Generate a retrieval-augmented response $c_i^j \sim \pi_\theta(\cdot \mid \mathcal{H}_i, \hat{c}_i^j, \mathcal{E}_i^j)$
    Compute reward $r_i^j \leftarrow R_\phi(\mathcal{H}_i, c_i^j)$
  end for
  Compute group-relative advantages $\{A_i^j\}_{j=1}^{N}$ from rewards $\{r_i^j\}_{j=1}^{N}$
  Compute importance sampling ratios $\{\rho_i^j\}_{j=1}^{N}$ with respect to $\pi_{\theta_{\text{old}}}$
  Compute the clipped GRPO loss $\mathcal{L}_{\text{GRPO}}(\theta)$
  Update policy parameters θ by optimizing $\mathcal{L}_{\text{GRPO}}(\theta)$
end for
Obtain the optimized policy π_θ

RAG Inference

for each dialogue state $\mathcal{H}_i$ in $\mathcal{D}_{\text{test}}$ **do**
  Generate an initial response $\hat{c}_i \sim \pi_\theta(\cdot \mid \mathcal{H}_i)$
  Retrieve relevent exemplars $\mathcal{E}_i \leftarrow \mathcal{R}(\hat{c}_i)$
  Generate the final response $c_i^{\text{final}} \sim \pi_\theta(\cdot \mid \mathcal{H}_i, \hat{c}_i, \mathcal{E}_i)$
end for
return θ and $\{c_i^{\text{final}}\}_{i=1}^{M'}$

E Human Evaluation

To complement automatic evaluation, we conduct a human evaluation to assess the quality of generated responses. We recruit three students with backgrounds in psychology as human annotators and provide them with training before the annotation process. During training, annotators are introduced to the evaluation criteria and the scoring prompt used in our study, ensuring that they understand the meaning of each evaluation dimension and apply the criteria consistently.

For each dialogue context, annotators are presented with a blind pair of responses generated by different models, where the model identities are hidden to reduce potential bias. Following the evaluation prompt, annotators score each response from the Client–Counselor–Judge perspectives, covering emotional support quality, therapeutic coherence, and contextual appropriateness. After annotation, we compare the scores assigned to each response in the blind pair and determine model preference according to the higher overall score.