# Listening to the Echo: User-Reaction Aware Policy Optimization via Scalar-Verbal Hybrid Reinforcement Learning

**Jing Ye**[1,2], **Xinpei Zhao**[2], **Lu Xiang**[1,2*], **Yaping Zhang**[1,2], **Chengqing Zong**[1,2]
[1]State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
yejing2022@ia.ac.cn; zhaoxinpei.zxp@alibaba-inc.com; {lu.xiang, yaping.zhang, cqzong}@nlpr.ia.ac.cn

## Abstract

While current emotional support dialogue systems typically rely on expert-defined scalar rewards for alignment, these signals suffer from severe information sparsity. They cannot explain *why* a response failed or *how* to adapt to dynamic user states, often failing to capture interaction-level conversational consequences reflected in subsequent user responses. We explore whether user reactions can serve as an additional behavior-level optimization signal for emotional support alignment. Therefore, we propose **R**eaction **A**ware **P**olicy **O**ptimization (RAPO), a framework that optimizes over interaction consequences rather than rubric scores. RAPO treats dialogue as a reaction-driven process and utilizes simulated user responses to generate dense natural-language feedback through three core components: (1) Hindsight Dialogue Selection, which isolates pivotal turns that meaningfully alter user emotional trajectories; (2) Generative Hindsight Feedback, which transforms user reactions into contrastive ranking signals and natural-language critiques; and (3) Scalar–Verbal Hybrid Policy Optimization, which couples scalar reward optimization for global alignment with verbal feedback distillation for fine-grained semantic refinement. Extensive experiments on ESC and SOTOPIA demonstrate that RAPO significantly outperforms strong RL baselines in driving positive interaction outcomes.

## 1 Introduction

Emotional Support Conversation (ESC) aims to alleviate users' emotional distress, foster a sense of being understood, and encourage continued engagement (Liu et al., 2021). With the rapid advancement of large language models (LLMs), recent work has made notable progress in generating fluent and seemingly empathetic responses through supervised fine-tuning (SFT) on expert demonstrations (Zheng et al., 2024; Ye et al., 2025c) or reinforcement learning (RL) (Zhou et al., 2023; Wang et al., 2025; Ma et al., 2025). However, most existing approaches share a common optimization paradigm: they either imitate static expert-written responses or optimize scalar rewards derived from expert-defined rubrics, assuming that empathy quality can be predefined and universally agreed upon. This paradigm exposes a critical limitation: *it treats emotional support as a static, isolated scoring task rather than a dynamic, interactive process.*

**First, optimization mismatch.** Relying on static expert rubrics creates a fundamental mismatch with actual user experiences (Ye et al., 2026). A response that scores highly on predefined empathy metrics might still feel hollow to the individual seeking support, a phenomenon we term "*hallucinated empathy*". For instance, a model repeatedly outputting safe templates like "*I'm sorry to hear that. That must be really hard,*" can earn perfect scalar scores while leaving the user feeling disengaged. Because static rubrics cannot dynamically observe user reactions, they optimize the model to *look* supportive by expert standards rather than actually *being* supportive in practice.

**Second, signal sparsity and ambiguity.** Even when scalar rewards correctly identify a poor response, compressing a complex supportive conversation into a single number leads to severe information loss (Luo et al., 2025). Identical low scores can mask entirely different failure modes, such as offering premature advice versus inadvertently invalidating the user's feelings. These sparse scalar signals fail to explain *why* a response failed or *how* the model should adapt to dynamic shifts in the user's emotional state. Without explicit natural-language guidance grounded in the conversational context, the policy lacks the necessary supervision for fine-grained semantic refinement.

To bridge this gap, we claim that most direct and reliable learning signal naturally emerges from

*Corresponding Author



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 1485–1515
October 24-29, 2026 ©2026 Association for Computational Linguistics

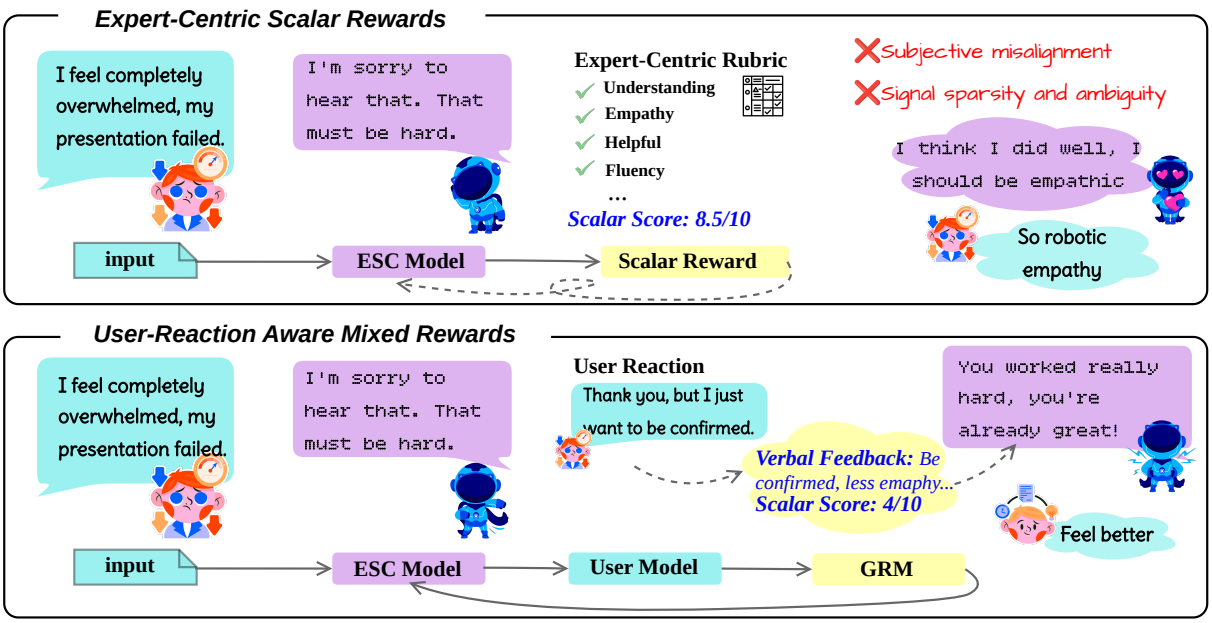



Figure 1: Comparison between expert-centric scalar rewards and user-reaction aware mixed rewards.

the user's continuous reactions during the ongoing interaction. A user's subsequent response inherently encodes whether they felt heard, validated, or alienated. **We therefore advocate a shift from rubric-matching optimization to user reaction-aware optimization.**

Building on this insight, we introduce **R**eaction **A**ware **P**olicy **O**ptimization (RAPO), a reinforcement learning framework that explicitly models user reactions through three key components: **(1) Hindsight Dialogue Selection**, which focuses training on pivotal dialogue turns that influence the user's emotional trajectory, avoiding overfitting to high-frequency but low-information empathy templates. **(2) Generative Hindsight Feedback**, which employs a user simulator and a generative reward model to produce both normalized ranking scores and natural-language critiques, transforming user reactions into actionable learning signals. **(3) Scalar–Verbal Hybrid Policy Optimization,** which couples scalar reward optimization for global alignment with verbal feedback distillation for fine-grained semantic refinement. Together, these components enable the model to move beyond selecting better responses toward learning to generate genuinely effective emotional support. Our main contributions are summarized as follows:

- We introduce a **reaction-driven optimization paradigm** for ESC, shifting the learning objective from expert-defined rubric-matching to dynamic, interaction-driven learning.
- We propose the **RAPO** framework that features a novel **scalar-verbal hybrid policy optimization**. By coupling ranking-based global alignment with critique-driven on-policy self-distillation, RAPO alleviates the signal sparsity and ambiguity associated with standard scalar RL by introducing critique-grounded token-level supervision.
- Extensive experiments on ESC and open-ended social environments (SOTOPIA) demonstrate that RAPO significantly outperforms strong RL baselines, suggesting the effectiveness and transferability of the proposed framework across different dialogue settings.

## 2 Preliminaries

Before presenting our method, we first review two related optimization primitives.

### 2.1 Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) is a variant of policy gradient methods that eliminates the need for a value function model by estimating advantages from a group of sampled outputs. Given a prompt $x$, the current policy $\pi_\theta$ generates a group of $G$ responses $\{y_i\}_{i=1}^{G}$, each associated with a reward $r_i$. The advantage $\hat{A}_i$ for the $i$-th response is computed by normalizing the reward within the group:

$$\hat{A}_i = \frac{r_i - \frac{1}{G}\sum_{j=1}^{G} r_j}{\sigma(\{r_1,\ldots,r_G\})} \tag{1}$$

where $\sigma$ denotes the standard deviation. The policy is then updated by maximizing the following objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\left\{\underbrace{\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}L_{\text{clip}}(y_{i,t})}_{\text{Policy Loss}} - \underbrace{\beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})}_{\text{KL Penalty}}\right\}\right] \tag{2}$$

$$L_{\text{clip}} = \min\left(\rho_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\left(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)\hat{A}_{i,t}\right) \tag{3}$$

where $\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|q,y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|q,y_{i,<t})}$, $\pi_{\text{old}}$ is the policy before the update. $\epsilon$ and $\beta$ control the clipping range and the strength of the KL divergence penalty, respectively. This approach ensures stable updates by constraining the policy ratio while leveraging group-based advantage estimation.

### 2.2 On-Policy Distillation

While GRPO improves stability, it relies on scalar rewards, which often fail to capture the nuanced, multi-dimensional feedback required for effective emotional support. On-policy distillation (OPD) (Lu and Lab, 2025) offers a solution by aligning a student $\pi_\theta$ with a teacher $\pi_{\text{teacher}}$ through dense, token-level supervision. This is achieved by minimizing the reverse KL divergence over sampled trajectories $y \sim \pi_\theta$:

$$\mathcal{L}_{\text{OPD}}(\theta) = \mathbb{E}_{y_t\sim\pi_\theta}\left[\mathcal{D}(\pi_{teacher}||\pi_\theta)(y_t|x, y_{<t})\right] \tag{4}$$

From an RL perspective, the gradient $\nabla_\theta\mathcal{L}_{\text{OPD}}$ can be interpreted as a policy gradient where the term $\log(\pi_{\text{teacher}}/\pi_\theta)$ acts as an implicit advantage $\hat{A}_t$. When the student's confidence $\pi_\theta$ is lower than the teacher's $\pi_{\text{teacher}}$, the gradient drives the policy to increase the probability of those tokens.

Building on this, On-Policy Self-Distillation (OPSD) (Zhao et al., 2026; Hübotter et al., 2026) eliminates the external teacher by using the model's own feedback-conditioned predictions as a self-teacher:

$$\pi_{teacher} = \text{sg}(\pi_\theta(y_t \mid x, f, y_{<t})) \tag{5}$$

where $f$ is the environment feedback, and $\text{sg}(\cdot)$ denotes the stop-gradient operation, which prevents the teacher and student distributions from co-evolving during training. By distilling feedback-informed next-token predictions back into the policy, OPSD transforms sparse outcome rewards into a dense training signal without relying on an external reward model or a separate teacher.

## 3 Method

### 3.1 Reaction-Aware Problem Formulation

In conventional Reinforcement Learning from Human Feedback (RLHF), the ESC process can be modeled as follows: at each turn, given the dialogue history $x$, the policy $\pi_\theta$ generates a response $y$. The policy is then optimized to maximize the expected scalar reward $R(x, y)$:

$$\mathcal{J}_{\text{RLHF}}(\theta) = \max_\theta \mathbb{E}_{y\sim\pi_\theta}\left[R(x, y)\right] \tag{6}$$

As discussed in Section 1, relying solely on a sparse scalar $R$ fails to capture the nuanced dynamics of user reactions.

In this work, we propose a Reaction-Aware Formulation. Instead of relying on rubric-as-reward (Gunjal et al., 2026), we shift the focus to the direct outcome of the interaction: the user reaction $u \in \mathcal{U}$, where $u$ denotes a user reaction conditioned on the dialogue context and response. Our goal is to learn an optimal policy $\pi_\theta^*$ that maximizes the probability of facilitating positive user reactions:

$$\mathcal{J}_{\text{RAPO}}(\theta) = \max_\theta \mathbb{E}_{y\sim\pi_\theta, u\sim\mathcal{P}(\cdot|x,y)}\left[\mathcal{F}(x, y, u)\right] \tag{7}$$

where $\mathcal{P}(\cdot|x, y)$ is the distribution of user reactions given the context and $\mathcal{F}$ is a quality function derived from the interaction outcome $(x, y, u)$.

### 3.2 Hindsight Dialogue Selection

Standard training paradigms typically utilize all available turns in ESC corpus. However, a substantial fraction of these turns consist of generic empathetic templates (e.g., "*I'm sorry to hear that*" or "*I understand how you feel*"). Repeated optimization on such high-frequency but low-information patterns can lead to *hallucinated empathy*, in which models default to safe but vacuous responses rather than providing substantive support.

To mitigate this issue, we introduce a **Hindsight Dialogue Selection** mechanism that identifies and prioritizes *pivotal turns*—interactions in which the agent's response meaningfully affects the user's emotional state or shifts the conversation's trajectory. Given a dialogue $\mathcal{D} = \{(u_j, y_j)\}_{j=1}^N$, we employ GPT-4o (OpenAI, 2024) to perform retrospective evaluation of each turn through a hindsight-based interaction analysis (The detailed prompt is in Appendix I.1). Specifically, we define a binary selection indicator for each turn:

$$\mathbb{I}_j = \text{Hindsight-Judge}(\mathcal{D}, u_j, y_j) \tag{8}$$

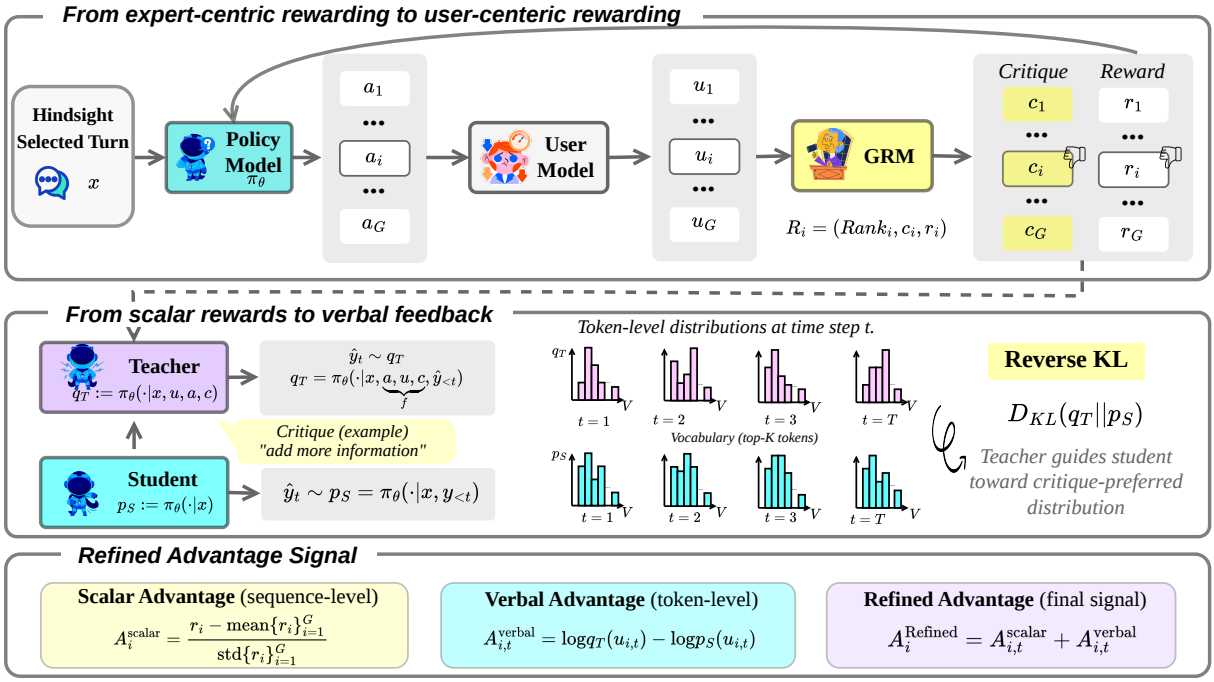



Figure 2: **Overview of the RAPO framework.** It integrates user simulation, contrastive critique generation, and scalar-verbal hybrid policy optimization.

Model training is restricted to turns with $\mathbb{I}_j = 1$. This filtering removes redundant empathetic templates and concentrates learning on interactions that exhibit strategic support behaviors and meaningful conversational impact.

### 3.3 Generative Hindsight Feedback

To overcome the sparsity and ambiguity of scalar rewards, we propose a **Generative Hindsight Feedback** mechanism consisting of (1) group-wise exploration with user simulation and (2) contrastive critique generation.

**Group-wise Exploration and User Simulation.** Given a dialogue history $x$, the current policy $\pi_\theta$ samples a group of candidate responses $\{y_i\}_{i=1}^G$. To estimate the potential impact of these candidates, we introduce a User Simulator that provides implicit feedback. For each candidate $y_i$, the simulator generates a corresponding user reaction $u_i$ based on a predefined user persona and their current emotional state:

$$u_i \sim \text{Simulator}(\cdot \mid x, y_i, \text{Persona}) \tag{9}$$

This process transforms static, single-turn responses into dynamic interaction trajectories $\mathcal{T} = \{(x, y_i, u_i)\}_{i=1}^G$, providing a behavioral basis for hindsight evaluation.

**Contrastive Critique Generation.** By observing the entire set of simulated trajectories $\mathcal{T}$, we utilize a Generative Reward Model (GRM) (Liu et al., 2025) to perform Contrastive Critique Generation. Unlike traditional independent scoring, the GRM processes the entire group of trajectories simultaneously.

$$\{\mathcal{R}_1, \ldots, \mathcal{R}_G\} = \text{GRM}(\cdot \mid \mathcal{T}) \tag{10}$$

where $\mathcal{R}_i = \{\text{Rank}_i, c_i, r_i\}$, $c_i$ is the verbal hindsight critique, $r_i$ is the reward score of response $y_i$, and $\text{Rank}_i$ is the rank in the group. By forcing the GRM to rank candidates within a group, we transform the evaluation from an absolute-scoring task to a relative-preference task (Ye et al., 2025a). This ensures that the rewards $r_i$ are inherently normalized, leading to a more stable advantage estimate, which could be lost under isolated scoring.

### 3.4 Scalar–Verbal Hybrid Policy Optimization

To bridge the gap between coarse scalar rewards and fine-grained verbal feedback, RAPO adopts a hybrid optimization strategy that combines scalar reinforcement learning with verbal supervision. Specifically, we use GRPO to optimize the policy for ranking-based scalar rewards, while OPSD enables the model to internalize natural-language feedback generated by the GRM.

#### 3.4.1 Scalar RL

We apply GRPO to align the policy with the ranking-scalar rewards derived in Section 3.3. Given a group of sampled responses $\{y_i\}_{i=1}^{G}$ with contrastive rewards $r_i$, the advantage $A_i^{\text{scalar}} = \frac{r_i - \bar{r}}{\sigma_r}$ for each response is obtained according to Eq. 1. The policy $\pi_\theta$ is optimized by maximizing the GRPO objective (Eq. 2), which encourages the model to assign higher probability to responses with better relative rankings.

#### 3.4.2 Verbal RL

Scalar rewards rank responses but cannot show *how* a poor response should be fixed. Verbal RL turns the natural-language critique into a dense, token-level training signal. The idea is to let the policy see the verbal feedback, form a "teacher" that corrects the response, and then force the original policy to mimic that correction — the discrepancy at each token becomes a reinforcement signal.

**Constructing a Feedback-aware Self-teacher.** For a group of $G$ rollouts, we take the lowest-reward response $y_w = \arg\min_{y_i} r_i$. Its verbal feedback $f_w = \{u_w, c_w\}$ contains the user reaction $u_w$ and the GRM critique $c_w$ that explains how $y_w$ led to that negative reaction. We build a teacher by augmenting the original prompt $x$ with this feedback and the prefix of the suboptimal response $y_{w,<t}$, then let the policy re-evaluate the next token:

$$q_\theta = \text{sg}\big(\pi_\theta(\cdot \mid x, f_w, y_{w,<t})\big). \tag{11}$$

Here $\text{sg}(\cdot)$ denotes stop-gradient, ensuring that the teacher acts as a fixed target during distillation. Because the teacher sees the outcome and the critique, it knows which tokens contributed to the failure. As a result, $q_\theta$ naturally assigns lower probabilities to problematic tokens and higher probabilities to alternatives that would have avoided the mistake.

We intentionally focus verbal correction on the lowest-reward trajectory, because negative reactions typically provide more explicit information about failure modes in emotional support conversations, whereas high-reward responses are often semantically diverse and less informative for targeted correction.

**On-Policy Distillation as Token-level Supervision.** The student policy sees only the original input $x$ (no feedback) and is trained to match the teacher distribution $q_\theta$ on the suboptimal trajectory $y_w$. This is done by minimizing the Reverse Kullback–Leibler divergence $D_{\text{KL}}(q_\theta \,\|\, \pi_\theta)$. The key insight is that the gradient of this objective with respect to $\log \pi_\theta(y_{w,t})$ is proportional to $\log q_\theta(y_{w,t}) - \log \pi_\theta(y_{w,t})$. In other words, for each token $y_{w,t}$, the policy is pushed to increase its log-probability if the teacher thinks it is more likely than the current policy, and to decrease it if the teacher thinks it is less likely. We can express this token-level guidance as a dense **verbal advantage**:

$$A_{w,t}^{\text{verbal}} = \log q_\theta(y_{w,t}) - \log \pi_\theta(y_{w,t} \mid x, y_{w,<t}). \tag{12}$$

A positive $A_{w,t}^{\text{verbal}}$ means the teacher, after seeing the critique, would have assigned a higher probability to that token — so the policy should reinforce it. A negative value means the token is disfavored by the feedback and should be suppressed. Thus, the natural-language critique is decomposed into a token-by-token signal.

$$\mathcal{L}_{\text{OPSD}} = \mathbb{E}_{y_{w,t} \sim \pi_\theta}[\mathcal{D}(q_\theta || \pi_\theta)(y_t | x, y_{w,<t})] \tag{13}$$

By minimizing this loss, the student internalizes the verbal feedback.

#### 3.4.3 Final Objective

The overall RAPO objective integrates scalar alignment with linguistic refinement:

$$\mathcal{J}_{\text{final}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) - \eta \mathcal{L}_{\text{OPSD}}(\theta) \tag{14}$$

where $\eta$ is the hyperparameter. Intuitively, the scalar reward provides sequence-level preference optimization, while the verbal distillation term induces token-level corrective advantages. Appendix F further shows that the combined objective can be interpreted as optimizing a unified, refined advantage that comprises both scalar and verbal feedback components.

## 4 Experimental Results

We present experiments to (i) quantify improvements over standard RL baselines and (ii) identify key design choices. Detailed settings, implementation details, and additional results are provided in Appendix A, B, and C.

### 4.1 Is Reaction-aware Optimization Effective for Emotional Support Agents?

We train ESC agents on ESConv (Liu et al., 2021) and evaluate on the ESConv test set and two interactive benchmarks, EmoHarbor (Ye et al., 2026) and

| PR: Problem Resolution | MI: Mood Improvement | Div: Diversity | HL: Human-likeness |
|---|---|---|---|

| Model | ESConv (↑) | EmoHarbor (↑) | | | ESC-Eval (↑) | | |
|---|---|---|---|---|---|---|---|
| | BERT-Score | PR | MI | Overall | Div | HL | Overall |
| ***Close-Source LLMs.*** | | | | | | | |
| Doubao-Pro-32k | **85.11** | 1.65 | 1.25 | 2.12 | **3.00** | **1.98** | **2.00** |
| GPT-4o-2024-11-20 | 84.89 | 2.56 | 2.17 | 3.20 | 2.89 | 1.36 | 1.86 |
| GPT-5-2025-08-07 | 84.26 | **3.19** | **3.06** | **3.83** | 2.78 | 1.68 | 1.98 |
| ***Open-Source In-domain LLMs.*** | | | | | | | |
| SoulChat2.0-Qwen2-7B | **85.75** | 1.84 | 1.20 | 2.29 | **3.00** | 2.00 | **2.00** |
| SoulChat2.0-Llama-3.1-8B | 85.65 | 1.88 | 1.27 | 2.37 | **3.00** | **2.01** | **2.00** |
| PsyChat-Qwen2.5-7B | 84.99 | **2.03** | **1.54** | **2.55** | 2.99 | 1.60 | 1.96 |
| ***Trained LLMs on ESConv Dataset.*** | | | | | | | |
| Qwen-2.5-7b | **85.67** | 1.69 | 1.21 | 2.10 | 2.61 | 0.96 | 1.27 |
| + SFT-Full | 85.85 | 1.13 | 1.00 | 1.58 | 2.89 | **2.02** | 1.97 |
| + SFT-HDS | 85.61 | 1.22 | 1.06 | 1.72 | 2.97 | 1.98 | **2.00** |
| + OPSD | 85.44 | 1.81 | **1.28** | 2.07 | 2.99 | 1.83 | **2.00** |
| + GRPO | 85.57 | 1.60 | 1.04 | 2.02 | 2.99 | 1.66 | 1.98 |
| + CPO | 85.60 | 1.76 | 1.25 | 2.15 | 2.98 | 1.83 | 1.99 |
| + RAPO | 85.65 | **1.86** | 1.26 | **2.27** | **3.00** | 1.84 | **2.00** |
| Qwen-3-8b | 84.42 | 1.89 | 1.62 | 2.59 | **2.91** | 1.27 | 1.93 |
| + SFT-Full | **85.89** | 1.09 | 1.00 | 1.54 | 2.14 | 2.01 | 1.94 |
| + SFT-HDS | 85.80 | 1.29 | 1.10 | 1.72 | 2.22 | **2.02** | 1.97 |
| + OPSD | 84.28 | 1.21 | 1.05 | 1.54 | **2.98** | 1.83 | **2.00** |
| + GRPO | 84.48 | 1.99 | 1.57 | 2.58 | 2.89 | 1.76 | 1.89 |
| + CPO | 85.01 | 1.99 | 1.67 | 2.58 | 2.93 | 1.87 | 1.99 |
| + RAPO | 85.56 | **2.03** | **1.73** | **2.67** | 2.91 | 2.01 | **2.00** |
| Llama-3-8b | 85.47 | 2.03 | 2.00 | 2.99 | 2.90 | 1.64 | 1.97 |
| + SFT-Full | **85.84** | 1.25 | 1.00 | 1.66 | 2.91 | **2.01** | **2.00** |
| + SFT-HDS | **85.84** | 1.20 | 1.07 | 1.69 | 2.93 | **2.01** | **2.00** |
| + OPSD | 84.88 | 1.23 | 1.02 | 1.89 | 2.95 | 1.96 | 1.99 |
| + GRPO | 85.46 | 2.03 | 1.56 | 2.59 | 2.92 | 1.69 | 1.99 |
| + CPO | 85.33 | 2.40 | 2.03 | 3.03 | 2.96 | 2.00 | 1.99 |
| + RAPO | 85.56 | **2.46** | **2.05** | **3.04** | **2.99** | 1.94 | **2.00** |

Table 1: **Performance on ESC Benchmarks.** In SFT settings, "HDS" denotes the Hindsight Dialogue Selection strategy described in Sec. 3.2, while "Full" indicates training on the complete dataset. Evaluation scales differ across benchmarks: **EmoHarbor** uses a five-point Likert scale (1–5), whereas **ESC-Eval** uses a five-point Likert scale (0–4). Detailed results are provided in Tables 5, 6, and 7.

ESC-Eval (Zhao et al., 2024). Table 1 summarizes the results:

**(1) RAPO consistently demonstrates superior performance over other only scalar/verbal RL baselines on both EmoHarbor and ESC-Eval.** Specifically, Qwen-2.5-RAPO improves over GRPO by 12.4% on EmoHarbor. Compared with CPO, it further improves EmoHarbor by 5.6%. Compared to OPSD, our RAPO achieves 2.27 on EmoHarbor, an improvement of 8.8%.

**(2) RAPO maintains strong semantic consistency with human reference responses.** On the ESConv, Qwen-2.5-RAPO and Qwen-3-RAPO achieve high BERT-Score values of 85.65 and 85.56, respectively. This indicates that the RAPO effectively optimizes support strategies without degrading the fundamental linguistic quality or deviating from appropriate conversational norms.

**(3) Utilizing a curated subset of high-quality data is sufficient for optimal performance.** We compare the results of SFT trained on the full dataset with those trained on data selected using Hindsight Dialogue Selection, denoted as SFT-Full and SFT-HDS, respectively. The experimental results show that SFT-HDS achieves performance comparable to SFT-Full on ESConv, while demonstrating significantly better stability and generalization on EmoHarbor and ESC-Eval. This finding

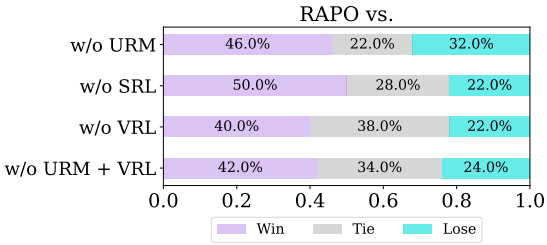
suggests that simply scaling up the volume of SFT data does not necessarily lead to improved emotional support capabilities. Instead, training should prioritize critical dialogue turns that require complex emotional reasoning.

### 4.2 Can RAPO Generalize to Diverse Dialogue Agents?

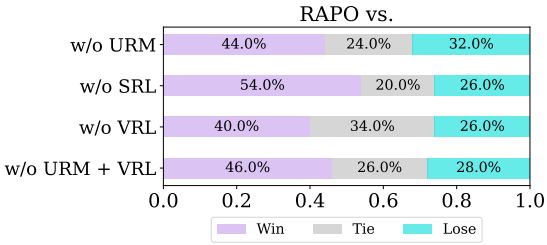
| Method | SOTOPIA-ALL (↑) | | SOTOPIA-HARD (↑) | |
|---|---|---|---|---|
| | GOAL | OVERALL | GOAL | OVERALL |
| GPT-4o | 8.19 | **3.76** | 6.97 | **3.46** |
| GPT-5* | **8.95** | — | **7.21** | — |
| Doubao-Pro-32k | 8.12 | 3.70 | 6.49 | 3.22 |
| DeepSeek-R1 | 7.97 | 3.40 | 5.86 | 2.73 |
| Claude-3.5-Sonnet | 8.29 | 3.71 | 6.33 | 3.09 |
| Qwen2.5-7B-Instruct | 7.91 | 3.55 | 6.21 | 3.01 |
| + PPO* | 8.25 | — | 6.63 | — |
| + GRPO* | 8.31 | — | 6.97 | — |
| + RLOO* | 7.13 | — | 5.70 | — |
| + GRPO iStar* | **8.42** | — | 7.11 | — |
| + CPO | 8.13 | 3.77 | 7.14 | 3.48 |
| + RAPO | 8.41 | **3.87** | **7.19** | **3.54** |

Table 2: **Performance on SOTOPIA.** The evaluated model interacts with itself. "Hard" denotes a challenging subset requiring advanced reasoning, while "All" denotes the full set of scenarios. GPT-4o is used as a proxy for human judgment to evaluate GOAL (goal completion score, 0–10) and OVERALL performance (the mean across 7 evaluation dimensions). The reported results are averaged over 3 runs. Results marked with * are taken from Liu et al. (2026), and our evaluation keeps the same with theirs.

To evaluate the generalizability of our approach beyond emotional support, we benchmark RAPO on SOTOPIA, a multi-agent social intelligence framework that requires complex social negotiation and interaction. The results are summarized in Table 2: **(1) RAPO outperforms established RL baselines, including PPO, GRPO, and RLOO.** Across all tasks, RAPO achieves an overall average score of 3.87 and a goal completion score of 8.41, demonstrating its robustness in managing multifaceted social dynamics. **(2) RAPO exhibits exceptional performance in high-difficulty scenarios.** Notably, on the SOTOPIA-HARD subset, RAPO improves the goal completion score by 3.2% (6.97 → 7.19) compared to GPT-4o, suggesting that RAPO is particularly effective when the social objectives are complex. **Overall, these results confirm that RAPO's utility extends beyond emotional support.** By effectively modeling participant reactions as feedback signals and optimizing with a combination of scalar and verbal feedback, RAPO enhances general social intelligence and goal-oriented reasoning, demonstrating its effectiveness across a wide range of dialogue-based applications.




(a) ESC




(b) SOTOPIA

Figure 3: **Results of the pair-wise human evaluation.** VRL stands for Verbal RL, and SRL stands for Scalar RL. For each instance, annotators are presented with the dialogue context and two candidate responses (A and B) generated by different models. ■ indicates 'A win', ■ indicates 'tie', and ■ indicates 'B win'.

### 4.3 Ablation Study

To assess the contribution of each component, we evaluate four RAPO variants using Qwen-2.5-7B as the backbone: **w/o URM**, **w/o Verbal RL**, **w/o Scalar RL**, and **w/o URM+Verbal RL**. Results on four benchmarks are summarized in Table 3. Overall, the full RAPO consistently outperforms all ablated variants, indicating that User Reaction Modeling and verbal RL provide complementary learning signals. While rubric-based rewards provide coarse, static supervision, URM introduces interaction-aware adaptation by incorporating real user reactions. Meanwhile, OPSD transforms verbal feedback into fine-grained training signals, enabling more precise policy updates.

### 4.4 Human Evaluation

We conducted a pairwise human evaluation to compare RAPO against 4 ablated variants across both ESC and SOTOPIA. The human evaluation settings are detailed in Appendix B.3. As shown in Figure 3, RAPO consistently outperforms all baselines on both benchmarks. On ESC, RAPO achieves higher win rates against w/o URM (46.0%) and w/o Verbal RL (40.0%), and is preferred over w/o

| Model | ESConv | EmoHarbor | | | Esc-Eval | | | Sotopia-All | | Sotopia-Hard | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | BERT-Score | PR | MI | Overall | Div | HL | Overall | Goal | Overall | Goal | Overall |
| RAPO | **85.65** | 1.86 | 1.26 | **2.27** | **3.00** | 1.84 | **2.00** | **8.41** | **3.87** | 7.19 | **3.54** |
| w/o URM | 85.33 | **1.90** | **1.28** | 2.18 | **3.00** | **1.85** | **2.00** | 8.17 | 3.79 | **7.23** | 3.45 |
| w/o Verbal RL | 85.54 | 1.87 | 1.24 | 2.22 | **3.00** | 1.73 | **2.00** | 8.08 | 3.70 | 7.13 | 3.38 |
| w/o Scalar RL | 85.44 | 1.81 | 1.28 | 2.07 | 2.99 | 1.83 | **2.00** | 7.31 | 3.19 | 6.56 | 2.97 |
| w/o URM + Verbal RL | 85.60 | 1.76 | 1.25 | 2.15 | 2.98 | 1.83 | 1.99 | 8.13 | 3.77 | 7.14 | 3.48 |

Table 3: **Ablation Results. RAPO w/o URM** replaces user-centric reactions with expert-defined rubric rewards; **RAPO w/o Verbal RL** removes the verbal feedback-based self-distillation mechanism; **RAPO w/o Scalar RL** removes the GRPO optimization; **RAPO w/o URM+Verbal RL** reduces the framework to a conventional rubric-as-reward RL setup. Detailed results for each evaluation dimension are provided in Tables 8, 9, and 10.

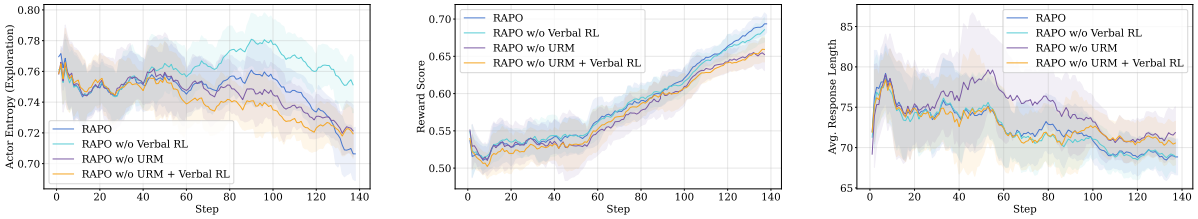



Figure 4: **Training Dynamics on ESConv. Left:** Actor policy entropy during training. **Middle:** Average critic reward score. **Right:** Mean response length generated by the policy.

URM+Verbal RL in 42.0% of cases. A similar pattern appears on Sotopia, where RAPO again surpasses w/o URM (44.0% vs. 32.0%), w/o Verbal RL (40.0% vs. 26.0%), and w/o URM+Verbal RL (46.0% vs. 28.0%). These results demonstrate that both URM and OPSD are critical to RAPO's performance, as removing either component leads to clear drops in human preference, while RAPO maintains a consistent overall advantage.

### 4.5 Analysis of Training Stability

Figure 4 compares the training dynamics of 4 RAPO variants on ESConv. (1) **Policy entropy.** From the entropy curves, all variants exhibit stable optimization and gradually converge during training. (2) **Reward trajectory.** In terms of reward scores, RAPO consistently achieves higher rewards and shows a more stable upward trend compared to the ablated variants. In particular, removing URM or Verbal RL leads to slower improvement and lower final performance, highlighting that both user-centric reactions and verbal feedback-based self-distillation provide effective learning signals for policy optimization. (3) **Response length.** As training progresses, the average response length across all variants gradually stabilizes within a reasonable range. Importantly, RAPO does not increase response length excessively while improving rewards, suggesting that the performance gains are not simply due to generating longer responses but rather stem from improved response quality and alignment with user needs.

## 5 Related Work

**Emotional Support Conversation.** Existing work on ESC primarily relies on SFT to imitate empathic responses (Liu et al., 2021; Zheng et al., 2024; Zhang et al., 2024). While such approaches improve fluency and surface-level coherence, they often produce generic expressions and exhibit limited personalization across diverse user contexts (Ye et al., 2025b). Recent RLHF-based methods introduce preference optimization and structured rewards to enhance empathy modeling (Zhou et al., 2023; Ma et al., 2025; Ye et al., 2025a; Wang et al., 2025; Kim et al., 2025; Yang et al., 2025; Jia et al., 2026). However, most of them adopt static, expert-centric reward formulations, implicitly assuming that empathy quality can be predefined. We model emotional support from a user-centric perspective and optimize policies using dynamic user reactions and interaction-level feedback.

**Hindsight Learning and Verbal Feedback.** Scalar rewards compress complex evaluation signals into low-dimensional scores (Luo et al., 2025), which may be insufficient for abstract and personalized objectives such as empathy (Zhang et al., 2025a). Recent research, therefore, explores natural-language feedback (Stephan et al., 2024;

Feng et al., 2025; Hong et al., 2025) and iterative self-critique (Zhang et al., 2025b; Bensal et al., 2025; Ricciardi et al., 2025) as richer supervisory signals, as well as on-policy distillation for stable optimization (Hübotter et al., 2026; Zhao et al., 2026; Shenfeld et al., 2026). Building on these directions, we integrate verbal hindsight critiques with on-policy learning to enable fine-grained, user-driven adaptation beyond fixed scalar rewards.

## 6 Conclusion

We propose RAPO, a framework that improves emotional support alignment by moving from a static expert-defined rubric reward to dynamic, user-reaction-aware feedback. Experiments show that simulated interaction reactions provide a more information-dense reward signal, effectively facilitating positive interaction outcomes. Ablation studies confirm that integrating scalar and verbal RL provides a robust error-correction mechanism, enabling models to infer the specific requirements for response refinement. By highlighting the importance of user-reaction aware feedback for adaptive social interactions, RAPO establishes a scalable, generalizable paradigm for building empathetic, goal-oriented AI systems aligned with complex human psychological states.

## Limitations

While RAPO demonstrates significant improvements in both ESC and SOTOPIA benchmarks, several limitations remain that warrant further investigation:

**1. Computational Overhead.** The reliance on an LLM-based user simulator to generate reaction-aware feedback introduces significant computational overhead. Compared with methods that use expert-centric rewards, RAPO incurs longer training time due to the iterative processes of reaction generation and conversation scoring.

**2. Dependency on Heavy Reward Models.** Consistent with existing literature, our experimental setup utilizes advanced LLMs as reward judges across all baselines and ablation studies. While this ensures high-quality signal acquisition, it poses challenges for resource-constrained environments. A promising avenue for future work is to distill the knowledge from these large-scale judges into specialized, smaller-parameter reward models, thereby improving training efficiency without sacrificing performance.

**3. Simulated User Model.** RAPO relies on the ESC-Role simulator to approximate user reactions during interaction, providing a scalable and behaviorally grounded proxy for feedback despite the inevitable sim-to-real gap. More broadly, our framework should not be interpreted as learning directly from real human psychological outcomes. Instead, RAPO studies whether simulated user reactions can serve as an interaction-level training signal for reinforcement learning.

To mitigate overfitting to simulator-specific behaviors, RAPO combines contrastive group-wise ranking with verbal self-distillation, encouraging the policy to learn more generalizable interaction strategies rather than memorizing fixed reaction patterns or optimizing toward a single evaluator style. Qualitative examples in Appendix H further suggest reduced premature advice and more adaptive conversational exploration. Nevertheless, the effectiveness of RAPO remains fundamentally bounded by the fidelity and calibration of the underlying simulators, and validation with real users remains an important direction for future work.

**4. Domain Generalization.** Our current validation is confined to dialogue-based tasks. While RAPO is designed to handle reaction-driven processes, its generalizability to broader sequential decision-making domains remains to be verified. Future work will focus on scaling RAPO to diverse fields to assess its universal applicability.

## Ethical Considerations

**1. Models and Dataset Usage.** This research utilized publicly available models, as well as toolkits such as vLLM (Kwon et al., 2023), ms-swift (Zhao et al., 2025), and VeRL (Sheng et al., 2025). The datasets used in our experiments are also publicly available, including the ESConv and SOTOPIA-$\pi$. All datasets, models, and toolkits used in this work comply with their respective open-source licenses and usage policies.

**2. LLM Usage.** LLMs were employed to assist in coding, writing, and polishing the manuscript. Importantly, the LLMs were not involved in the ideation, research methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted solely by the authors.

**3. Human Evaluation.** All participants were informed of the study objectives prior to the experiments and received training for the human evalu-

ation process. All annotators were graduate-level students or above.

## Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 62506359) and the General Office of the National Language Commission Research Planning Committee (Grant No. ZDA145-18).

## References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. 2024. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12248–12267.

Anthropic. 2024. Claude 3.5 sonnet.

Shelly Bensal, Umar Jamil, Christopher Bryant, Melisa Russak, Kiran Kamble, Dmytro Mozolevskyi, Muayad Ali, and Waseem AlShikh. 2025. Reflect, retry, reward: Self-improving llms via reinforcement learning. *Preprint*, arXiv:2505.24726.

ByteDance. 2024. Doubao.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. SoulChat: Improving LLMs' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183, Singapore. Association for Computational Linguistics.

Yiwen Duan, Jing Ye, and Xinpei Zhao. 2026. Alive: Awakening llm reasoning via adversarial learning and instructive verbal evaluation. *Preprint*, arXiv:2602.05472.

Xidong Feng, Bo Liu, Ziyu Wan, Haotian Fu, Girish A. Koushik, Zhiyuan Hu, Mengyue Yang, Ying Wen, and Jun Wang. 2025. Natural language reinforcement learning. In *Scaling Self-Improving Foundation Models without Human Supervision*.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. 2026. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth International Conference on Learning Representations*.

Joey Hong, Kang Liu, Zhan Ling, Jiecao Chen, and Sergey Levine. 2025. Natural language actor-critic: Scalable off-policy learning in language space. *Preprint*, arXiv:2512.04601.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. 2026. Reinforcement learning via self-distillation. In *The 1st Workshop on Scaling Post-training for LLMs*.

Yuhang Jia, Pei Liu, Haoqin Sun, Jiaming Zhou, Xuxin Cheng, Cao Liu, Ke Zeng, Xunliang Cai, and Yong Qin. 2026. Reflecting twice before speaking with empathy: Self-reflective alternating inference for empathy-aware end-to-end spoken dialogue. *Preprint*, arXiv:2601.18281.

Juhee Kim, Chunghu Mok, Jisun Lee, Hyang Sook Kim, and Yohan Jo. 2025. Dialogue systems for emotional support via value reinforcement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28733–28766, Vienna, Austria. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Xiaoqian Liu, Ke Wang, Yuchuan Wu, Fei Huang, Yongbin Li, Jianbin Jiao, and Junge Zhang. 2026. Agentic reinforcement learning with implicit step rewards. In *The Fourteenth International Conference on Learning Representations*.

Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. 2025.

Inference-time scaling for generalist reward modeling. *Preprint*, arXiv:2504.02495.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. *Thinking Machines Lab: Connectionism*. Https://thinkingmachines.ai/blog/on-policy-distillation.

Renjie Luo, Zichen Liu, Xiangyan Liu, Chao Du, Min Lin, Wenhu Chen, Wei Lu, and Tianyu Pang. 2025. Language models can learn from verbal feedback without scalar rewards. *Preprint*, arXiv:2509.22638.

Hui Ma, Bo Zhang, Bo Xu, Jian Wang, Hongfei Lin, and Xiao Sun. 2025. Empathy level alignment via reinforcement learning for empathetic response generation. *IEEE Transactions on Affective Computing*, 16(3):1873–1884.

OpenAI. 2024. Hello GPT-4o.

Openai. 2025. Introducing gpt-5.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Antonio Pio Ricciardi, Valentino Maiorca, Luca Moschella, Riccardo Marin, and Emanuele Rodolà. 2025. R3l: Relative representations for reinforcement learning. *Preprint*, arXiv:2404.12917.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. In *ICLR 2026 Workshop on Lifelong Agents: Learning, Aligning, Evolving*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, page 1279–1297. ACM.

Moritz Stephan, Alexander Khazatsky, Eric Mitchell, Annie S Chen, Sheryl Hsu, Archit Sharma, and Chelsea Finn. 2024. Rlvf: Learning from verbal feedback without overgeneralization. *Preprint*, arXiv:2402.10893.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, Yuan Li, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025. Rlver: Reinforcement learning with verifiable emotion rewards for empathetic agents. *Preprint*, arXiv:2507.03112.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024. SOTOPIA-$\pi$: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940, Bangkok, Thailand. Association for Computational Linguistics.

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

Jing Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025a. CPO: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 297–323, Suzhou, China. Association for Computational Linguistics.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025b. From generic empathy to personalized emotional support: A self-evolution framework for user preference alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 18826–18853, Suzhou, China. Association for Computational Linguistics.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025c. SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669, Abu Dhabi, UAE. Association for Computational Linguistics.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2026. EmoHarbor: Evaluating personalized emotional support by simulating the user's internal world. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1176–1202, San Diego, California, United States. Association for Computational Linguistics.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *Preprint*, arXiv:2503.14476.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, Xiangpeng Wei, Xiangyu Yu, Gaohong Liu, Juncai Liu, Lingjun Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Chi Zhang, and 8 others. 2025. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *Preprint*, arXiv:2504.05118.

Naifan Zhang, Ruihan Sun, Ruixi Su, Shiqi Ma, Shiya Zhang, Xianna Weng, Xiaofan Zhang, Yuhan Zhan, Yuyang Xu, Zhaohan Chen, Zhengyuan Pan, and Ziyi Song. 2025a. Echo-n1: Affective rl frontier. *Preprint*, arXiv:2512.00344.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13395–13412. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xiaoying Zhang, Hao Sun, Yipeng Zhang, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. 2025b. Critique-grpo: Advancing llm reasoning with natural language and numerical feedback. *Preprint*, arXiv:2506.03106.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. 2024. ESC-eval: Evaluating emotion support conversations in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15785–15810, Miami, Florida, USA. Association for Computational Linguistics.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *Preprint*, arXiv:2601.18734.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Hong Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. 2025. Swift:a scalable lightweight infrastructure for fine-tuning. *Preprint*, arXiv:2408.05517.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11325–11345. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729, Toronto, Canada. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*.

Jie Zhu, Yuanchen Zhou, Shuo Jiang, Junhui Li, Lifan Guo, Feng Chen, Chi Zhang, and Fang Kong. 2026. Care: Cognitive-reasoning augmented reinforcement for emotional support conversation. *Preprint*, arXiv:2510.05122.

## Appendix Contents


**A Evaluation Details and Setup** **13**
- A.1 Benchmarks. . . . . 13
- A.2 Baselines. . . . . 14

**B Implementation Details** **15**
- B.1 Training Dataset . . . . . 15
- B.2 Hyperparameters . . . . . 15
- B.3 Human Evaluation . . . . . 16
- B.4 Experiments Environment . . . . 16

**C Additional Experimental Results** **16**
- C.1 Main Results . . . . . 16
- C.2 Ablation Studies . . . . . 17
- C.3 Real-User Interactive Multi-Turn A/B Study . . . . . 17

**D Discussion of User Reaction Mechanism** **18**

**E Related Work** **18**

**F Theoretical Analysis** **19**
- F.1 Notation and Global Objective . . 19
- F.2 Decomposition of the Gradient . . 19
- F.3 The Unified Refined Advantage . . 20

**G Additional Results Tables** **20**

**H Case Study** **21**

**I Prompts Templates** **26**
- I.1 Hindsight Dialogue Selection Prompt 26
- I.2 ESC Prompts . . . . . 26
- I.3 SOTOPIA Prompts . . . . . 29


# A Evaluation Details and Setup

## A.1 Benchmarks.

**ESConv.** Following prior work, we conduct an objective evaluation on the ESConv test set (Liu et al., 2021). We measure the similarity between model-generated responses and manually annotated references using standard automatic metrics, including BLEU(Papineni et al., 2002), ROUGE(Lin, 2004), Distinct(Li et al., 2016), and BERTScore(Zhang* et al., 2020).

**ESC-Eval.** To assess the quality of emotional support beyond surface-level lexical metrics, we adopt ESC-Eval (Zhao et al., 2024). ESC-Eval introduces a dynamic role-playing evaluation paradigm that constructs a diverse set of role-playing cards covering various distress scenarios. It employs ESC-Role[1] , a specialized model fine-tuned to simulate individuals experiencing distress and seeking emotional support. The benchmark evaluates target models across seven key dimensions—**fluency, diversity, empathy, informativeness, human-likeness, skillfulness**, and **overall quality**—thereby providing comprehensive multi-dimensional automatic evaluation signals. In addition, the authors introduced ESC-Rank[2] , a scoring model fine-tuned on the ESC-Eval dataset to produce discrete ratings on a five-point Likert scale (0–4), with higher scores reflecting stronger emotional support quality. To ensure consistency with prior work, we adopt the official ESC-Eval implementation[3].

**EmoHarbor.** To assess the quality of personalized emotional support, we use EmoHarbor (Ye et al., 2026), an automated evaluation framework that employs a User-as-a-Judge paradigm, simulating users' inner worlds through a Chain-of-Agent architecture. EmoHarbor utilizes 100 real-world user profiles covering diverse personality traits and situations, defining 10 evaluation dimensions to measure whether support responses are genuinely tailored to individual users' unique psychological profiles, including **Empathy, Response Appropriateness, Adaptive Strategy, Problem Resolution, Mood Improvement, Human-likeness, Engagement, Redundancy, Consistency, and Safety**. We adopt the official EmoHarbor implementation[4] and use GPT-4o as both the simulated Seeker and the evaluator, producing discrete ratings on a five-point Likert scale (1–5), with higher scores reflecting stronger emotional support quality.

**SOTOPIA.** SOTOPIA (Zhou et al., 2024) is a general-domain benchmark for evaluating LLMs in open-ended multi-agent social interactions. It covers diverse scenarios—such as negotiation, collaboration, competition, and persuasion—where agents pursue individual goals under realistic constraints. The SOTOPIA-hard subset further includes tasks that require advanced strategic and long-horizon reasoning. Each agent is initialized with a structured character profile (e.g., name, gender, personality, and profession). Following prior work, we use GPT-4o as a proxy for human judgment to

[1] https://huggingface.co/haidequanbu/ESC-Role
[2] https://huggingface.co/haidequanbu/ESC-Rank
[3] https://github.com/haidequanbu/ESC-Eval.git
[4] https://github.com/kaedeharakazuha678/EmoHarbor-25.git

compute both the **GOAL** score (ranging from 0 to 10) and the **Overall** score (the mean of seven dimensions: Goal Completion, Relationship, Financial and Material Benefits, Social Rules, Believability, Secret, and Knowledge), which has been shown to correlate strongly with human evaluations. We adopt an interaction evaluation protocol in which the evaluated model interacts either with itself (Self-play) or with a strong external partner model, GPT-4o (GPT-4o-as-Partner).

## A.2 Baselines.

### A.2.1 ESC Baselines

**Close-Source LLMs.** We benchmark our method against several representative commercial LLMs with strong conversational and empathetic capabilities, including **Doubao-1.5-Pro** (ByteDance, 2024), **GPT-4o** (OpenAI, 2024), and **GPT-5** (Openai, 2025).

**Open-Source In-domain LLMs.** Among open-source in-domain LLMs for mental health support, **SoulChat** (Chen et al., 2023) and **PsyChat** [5] represent two typical approaches.

**Ours Implemented.** We evaluate RAPO on three widely used backbone LLMs: **Qwen2.5-7B-Instruct**[6], **Qwen3-8B**[7], and **LlaMA-3-8B-Instruct**[8]. We compare against standard supervised fine-tuning baselines, scalar reward-based RL methods (GRPO and CPO), and verbal RL approaches (OPSD).

- **Vanilla.** The off-the-shelf instruction-tuned backbone models are directly prompted on the ESC task without task-specific fine-tuning. These baselines evaluate the intrinsic ESC capability of each backbone.
- **SFT.** Backbone LLMs are further supervised fine-tuned on the ESC training set. This baseline isolates the effect of standard task-specific instruction tuning.
- **OPSD.** We adopt OPSD (Zhao et al., 2026) as the verbal RL baseline. Similar to RAPO, OPSD constructs privileged contexts using simulated user reactions and critique feedback generated by GRM.
- **GRPO.** We incorporate GRPO (Shao et al., 2024) as a representative scalar RL baseline. Following prior work on open-ended generation (Gunjal et al., 2026), we employ a rubrics-as-a-reward framework, where responses are scored according to predefined expert rubrics and optimized using group-relative policy updates.
- **CPO.** We further include Comparative Policy Optimization (CPO) (Ye et al., 2025a) as a strong baseline for subjective generation tasks. CPO mitigates reward ambiguity by replacing absolute sample-level scoring with group-wise comparative evaluation. Policy updates are driven by trajectory-level preference comparisons, enabling more robust alignment with nuanced human judgments. Similar to GRPO, we employ a rubrics-as-a-reward framework (Gunjal et al., 2026). This serves as a representative expert-centric static RL baseline.

### A.2.2 SOTOPIA Baselines

For evaluation on SOTOPIA, we consider both advanced proprietary LLMs and representative RL training methods.

**Advanced LLMs.** We include several state-of-the-art proprietary models with strong reasoning and dialogue capabilities, including **GPT-4o** (OpenAI, 2024), **GPT-5** (Openai, 2025), **Doubao-1.5-Pro** (ByteDance, 2024), **DeepSeek-R1**, and **Claude-3.5-Sonnet** (Anthropic, 2024). These models are accessed through their official APIs. They represent highly capable general-purpose LLMs and serve as strong upper-bound baselines for social interaction and agentic dialogue performance.

**RL-based Methods.** To compare with existing reinforcement learning approaches for training social agents, we additionally include several representative policy optimization algorithms, including **PPO** (Schulman et al., 2017), **GRPO** (Shao et al., 2024), **RLOO** (Ahmadian et al., 2024), and **GRPO-iStar**. All RL-based methods adopt **Qwen2.5-7B-Instruct** as the backbone model. The experimental results of the above SOTOPIA baselines are adopted from Liu et al. (2026), where the models are evaluated under a consistent experimental setup. And we keep the same experimental settings with them.

[5] https://modelscope.cn/models/eugeo1212/qwen2.5-7b-psychat
[6] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct
[7] https://huggingface.co/Qwen/Qwen3-8B
[8] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

## B Implementation Details

### B.1 Training Dataset

Our model is trained on two complementary datasets targeting emotional support and social intelligence, respectively. These datasets are selected to expose the model to both psychologically grounded supportive dialogues and strategically complex social interactions.

**Emotional Support.** For the emotional support task, we use the **ESConv** dataset (Liu et al., 2021), a manually curated corpus of supportive conversations grounded in established psychological counseling strategies. The dialogues are constructed by trained annotators who simulate help-seeking and support-providing interactions. Compared to automatically collected or weakly supervised conversational data, ESConv offers higher-quality, strategy-aware supervision signals. The dataset consists of 1,295 multi-turn conversations spanning diverse emotional distress scenarios. Following the standard protocol, we adopt a 9:1 train–test split. The full training set comprises approximately $13K$ dialogue turns, while our proposed HDS (Hindsight Dialogue Selection) reduces this into a refined subset of 5,412 critical turns ($\sim$42% of the total). This subset preserves a wide range of conversational acts—including questioning, emotional reflection, empathy, information sharing, and suggestions—and filters out primarily generic or low-information responses (e.g., templated acknowledgments) rather than systematically removing rapport-building turns. To validate the selection, we manually inspected 100 conversations; the agreement between the retained critical turns and human judgment reached 83%.

**SOTOPIA.** For the social intelligence task, we train on SOTOPIA-$\pi$ (Wang et al., 2024), a large-scale synthetic dialogue dataset built upon the SOTOPIA framework (Zhou et al., 2024). SOTOPIA-$\pi$ contains diverse, GPT-4-generated multi-turn social interactions covering a wide range of interpersonal settings and social objectives. The scenarios are carefully designed to be distinct from the original SOTOPIA benchmark tasks, preventing data leakage while introducing novel social challenges. The dataset emphasizes goal-driven interactions, strategic reasoning, and role-consistent behavior, providing rich supervision for modeling complex multi-agent social dynamics.

### B.2 Hyperparameters

#### B.2.1 Training Hyperparameters

Due to resource constraints, we adopt LoRA fine-tuning, implementing LoRA with a rank of 16 and a scaling factor $\alpha = 32$. All models are trained for 3 epochs using a learning rate of $3 \times 10^{-6}$ and a batch size of 128. The maximum prompt length and the maximum new tokens are set to 8192, 2048, respectively.

In the training environment, the User Model is simulated using ESC-Role on ESC and Qwen-3-8B on SOTOPIA. ESC-Role is a user-simulator model that is LoRA-fine-tuned on 14K general role-playing and 3.4K emotional-support dialogues. Human evaluation confirms ESC-Role outperforms GPT-4 and achieves near-indistinguishable realism from genuine help-seekers (Zhao et al., 2024). DeepSeek-R1 serves as the generative reward model, providing response-level quality assessment signals.

For reinforcement learning optimization, we sample $G = 4$ rollouts per input prompt. During the rollout phase, we use a temperature of 1.0 and top-$p = 1.0$ to ensure unbiased stochastic exploration. The KL regularization coefficient is set to $\beta = 5 \times 10^{-4}$. We adopt clipping thresholds $\epsilon_{\text{high}} = 0.28$ and $\epsilon_{\text{low}} = 0.2$. To account for the response-length characteristics of dialogue tasks, we apply Soft Overlong Punishment to control length. Specifically, when the response length exceeds a predefined maximum, a punishment interval is activated in which longer responses incur progressively larger penalties. The corresponding length control hyperparameters are set to $L_{\text{max}} = 200$ and $L_{\text{cache}} = 80$.

Within the RAPO framework, the OPSD mechanism is specifically designed for verbal RL. The OPSD KL term is integrated with a coefficient of $\eta = 1 \times 10^{-3}$. We observe that token-level divergence is highly skewed across vocabulary entries, so we clipped it by an upper limit of $|2|$ to stabilize training. To maintain a consistent and stable target distribution, we employ an Exponential Moving Average (EMA) teacher model with a mixing coefficient of 0.5. Furthermore, the distillation process is applied strictly to the top 256 tokens (distillation top-$K = 256$) of the teacher's predicted distribution.

In terms of training efficiency, RAPO requires approximately 13 hours to train on the ESC task, compared to 9 hours for CPO, which does not in-

volve user reaction simulation or self-distillation. In the SOTOPIA environment, RAPO training takes approximately 27 hours.

#### B.2.2 Inference Hyperparameters

For all evaluated models, we set the temperature to 0.3 and the top-$p$ value to 0.9 to balance diversity and stability in generation. For reasoning-oriented models, intermediate reasoning content enclosed within `<think>` and `</think>` tags is excluded from evaluation (also excluded for GRM); we only consider the final response produced by the model, which corresponds to the explicit output shown to users. For all LLM-based judges, GRM, and the ESC-Rank model, we use greedy decoding to ensure stable, reliable evaluation results.

### B.3 Human Evaluation

We conducted a pairwise human evaluation to compare RAPO against three ablated variants (w/o URM, w/o SD, and CPO) across both the ESC-Eval and SOTOPIA benchmarks. From each benchmark, we randomly sampled 50 simulated dialogue instances. For each instance, annotators were presented with the dialogue context and two candidate responses (A and B) generated by different models. Annotators were instructed to select their preferred response based on the criteria detailed in Appendix A. To mitigate potential position and confirmation biases, we randomly shuffled the order of the responses and kept annotators blind to the model identities. Each instance was evaluated by three independent annotators—Master's students specialized in relevant fields. The final preference was determined via majority vote. Inter-annotator agreement was measured using Fleiss' $\kappa$, yielding a score of 0.52, indicating moderate agreement. Instances without a majority consensus (i.e., three-way ties) were excluded from the final statistics.

### B.4 Experiments Environment

All training and inference tasks were executed on a cluster of 8 $\times$ NVIDIA A100 GPUs. We perform SFT with the ms-swift framework (Zhao et al., 2025), and RL fine-tuning with the VeRL framework (Sheng et al., 2025), incorporating custom reward functions to align with our objectives. Model inference was powered by vLLM (Kwon et al., 2023).

## C Additional Experimental Results

### C.1 Main Results

In this section, we present the comprehensive evaluation results for the ESConv, EmoHarbor, and ESC-Eval benchmarks, summarized in Tables 5, 6, and 7, respectively. Our analysis reveals several key insights regarding the effectiveness of the proposed RAPO method and the characteristics of emotional support conversations.

**(1) The experimental results in Table 6 demonstrate that RAPO significantly enhances the model's capacity for personalized emotional support.** Compared to standard SFT and GRPO baselines, RAPO achieves superior performance across critical dimensions, including Problem Resolution (PR), Mood Improvement (MI), Response Appropriateness (RA), Adaptive Strategies (AS), and Engagement (EG). This suggests that RAPO is particularly effective at optimizing for complex, nuance-driven metrics that go beyond generic conversational fluency.

**(2) Fundamental conversational capabilities—namely Human-likeness (HL), Empathy (EP), and Safety (SF)—are largely inherent to the base models.** Qwen-2.5, Qwen-3, and Llama-3 already exhibit strong performance in these dimensions, with scores consistently reaching above 3.0, and Qwen-3-8B achieving a high safety score of 4.67. This indicates that, while current LLMs have mastered the "*safety and empathy baseline*," the primary bottleneck in psychological support lies in personalized problem-solving and mood improvement. Consequently, our findings emphasize that for specialized emotional support tasks, researchers should shift focus from enhancing basic dialogue competence to cultivating advanced, tailored interventions that directly address user needs.

**(3) This observation is further corroborated by our results on the ESC-Eval benchmark.** We find that improvements in fundamental metrics such as fluency and empathy are relatively marginal. Conversely, the most significant performance gains are concentrated in the diversity dimension, underscoring RAPO's ability to generate varied, context-aware responses rather than repetitive boilerplate support content.

**(4) SFT suffers from severe generalization issues.** Results in Table 5 show that models fine-tuned on ESConv achieve remarkably high scores on automatic metrics such as BLEU, ROUGE, and BERTScore. However, when evaluated on interac-

tive benchmarks, including EmoHarbor (Table 6) and ESC-Eval (Table 7), their performance drops dramatically across nearly all evaluation dimensions. Notably, the performance degradation of SFT-Full is even more pronounced than that of SFT-HDS. These results suggest that training on a single dataset primarily enables models to capture superficial patterns specific to that dataset, rather than to learn robust empathetic capabilities. Consequently, models tend to produce formulaic empathetic responses. In contrast, RL-based methods, including GRPO, CPO, and RAPO, optimize policies to maximize expected rewards rather than directly imitating reference responses in the training data. As a result, although these models obtain lower BLEU and ROUGE scores on the ESConv test set, their BERTScore remains comparable to SFT models, while achieving consistently better performance on EmoHarbor and ESC-Eval, indicating improved generalization to interactive evaluation settings.

### C.2 Ablation Studies

As discussed in the main text, the complete RAPO framework exhibits a synergistic interaction between User Reaction Modeling (URM) and verbal feedback-based On-Policy Self-Distillation (OPSD). While the primary findings emphasize the macro-level impact of these components, the detailed metric breakdowns in Tables 8, 9, and 10 pinpoint the specific areas where these modules exert their strongest influence.

In ablation experiments on the EmoHarbor dataset (Table 8), different modules show distinct effects on personalized emotional support. The most significant performance drop occurs when URM is removed, confirming that modeling users' real-time reactions is crucial for capturing personalized psychological needs. Without URM, responses become rigid and less targeted. The simultaneous removal of both URM and OPSD further decreases the average score to 2.15—the lowest among all variants—highlighting the complementary roles of these modules in producing high-quality support.

On the ESC-Eval dataset (Table 9), although the base model retains inherent advantages in empathy and fluency, removing the RAPO framework (w/o URM + OPSD) leads to widespread performance degradation across nearly all evaluation dimensions. Fluency drops to 2.915, Human-likeness falls to 1.825, and Suggestion rises to 2.405. This indicates that, without these modules, the model tends to generate long-winded safety-oriented advice, sacrificing the human-like quality of emotional companionship.

Social intelligence evaluation on SOTOPIA (Table 10) further underscores the OPSD module's pivotal role in complex social tasks. In multi-turn interactions, OPSD provides fine-grained strategic guidance; its removal reduces Knowledge (KNO) from 5.32 to 4.68 and Relationship (REL) from 3.56 to 3.41 when evaluated via self-play across all tasks. In contrast, removing URM primarily affects Goal Completion, decreasing it from 8.41 to 8.17, highlighting the decisive role of user-response awareness in steering conversations toward intended outcomes.

### C.3 Real-User Interactive Multi-Turn A/B Study

| Comparison | RAPO Win | Baseline Win |
|---|---|---|
| RAPO vs. CPO | **62.93%** | 37.07% |
| RAPO vs. GRPO | **69.50%** | 30.50% |

Table 4: Results of the real-user interactive multi-turn A/B study.

To complement our simulator-based evaluation, we conduct a **real-user interactive multi-turn A/B study** with 10 participants across four emotional-support scenarios. Unlike static preference evaluation, where responses are compared independently under fixed contexts, participants interact with the models throughout complete conversations of at least 10 dialogue turns. At each turn, two models generate responses to the same user input and dialogue history, and the participant selects the preferred response. The selected response is then appended to the conversation and becomes the context for the next turn, making the evaluation trajectory-dependent and closer to realistic multi-turn interaction. Each comparison pair contains approximately 200 turn-level preference decisions.

As shown in Table **??**, RAPO is preferred in 62.93% of turn-level comparisons against CPO and 69.50% against GRPO. Although this experiment is intended as a controlled user study rather than a large-scale human evaluation, it provides complementary evidence that the improvements learned from simulated user reactions transfer to real users. In particular, despite being trained solely with simulated feedback, RAPO maintains a clear preference advantage during live, dynamically evolving multi-turn conversations, suggesting that its gains are not limited to interactions with the simulator.

## D Discussion of User Reaction Mechanism

In this work, we propose to leverage simulated user reactions as learning signals for Emotional Support Conversation (ESC), moving away from the conventional rubric-as-reward paradigm. This shift frequently raises questions regarding the reliability of user simulation and the potential for noise in such signals. We provide our rationale and empirical evidence for this design choice as follows:

**(1) From Static Rubrics to Dynamic Interaction:** Traditional LLM-as-a-judge approaches typically require a model to evaluate responses based on predefined, static expert rubrics. However, these rubrics often fail to capture the nuanced, personalized, and context-dependent nature of conversational tasks, especially for ESC. More importantly, requiring a model to strictly adhere to abstract expert definitions often leads to instability due to the model's limited ability to interpret these criteria consistently, often resulting in score collapse or high sensitivity to prompt variations. In contrast, our approach treats dialogue as a reaction-driven process. The model does not need to pre-interpret an abstract rubric; instead, it generates responses that naturally elicit subsequent user reactions. This frames evaluation as a direct, empirical consequence of the interaction rather than a forced alignment with static norms.

**(2) Stability via Contrastive Evaluation:** We do not assume that a single user reaction is perfect. Instead, we employ a groupwise contrastive evaluation strategy, comparing reactions to multiple candidate responses within the same context. This approach, akin to generative critique reinforcement learning, provides fine-grained semantic guidance through verbal feedback, while the contrastive component counteracts noise from individual simulated reactions. Compared to pointwise scalar scoring, this method produces more stable and robust learning signals, allowing the model to discern relative quality rather than rely on absolute judgments.

**(3) Self-Correction through Verbal Feedback:** We emphasize that our framework integrates verbal feedback—the actual user response—as a primary supervisory signal rather than a secondary score. This generative approach allows the policy to "understand" the causal relationship between a response and the user's reaction. By distilling these feedback-informed, next-token predictions back into the policy, the model gains fine-grained guidance that a scalar value cannot convey. Empirical results in Table 1 and 2 and Figure 4 demonstrate that this combination of contrastive ranking and verbal feedback yields more stable training dynamics and superior performance compared to rubric-based methods, confirming that the "noise" in simulation is significantly outweighed by the rich, interactive information provided by the user model.

## E Related Work

With the advancement of LLMs, personalized Emotional Support agents have attracted growing research interest. However, a key challenge persists: how to model effective reward signals that capture the subtle dynamics of empathy and user satisfaction.

**Emotional Support Conversation.** Recent work on ESC has largely focused on SFT, using expert-annotated demonstrations to enable imitation learning (Liu et al., 2021; Zheng et al., 2023, 2024; Ye et al., 2025c). Although these approaches substantially enhance linguistic fluency and surface-level coherence, they often yield templated empathic responses and exhibit limited adaptability across heterogeneous user contexts, leading to weak personalization. RLHF (Ouyang et al., 2022; Schulman et al., 2017; Shao et al., 2024; Rafailov et al., 2023; Yu et al., 2025; Yue et al., 2025) provides a preference-driven alignment paradigm and has recently been extended to ESC. Related efforts have explored emotion-aware dialogue optimization, feedback-conditioned response generation, and reinforcement learning with interaction-level user signals for empathetic conversation, highlighting the growing importance of adaptive, feedback-driven alignment beyond static supervised objectives (Zhou et al., 2023; Kim et al., 2025; Yang et al., 2025; Zhu et al., 2026). For example, Wang et al. (2025) design verifiable rewards via emotion-consistent simulated users; Ma et al. (2025) introduces dual training signals to jointly model empathy hierarchy and semantic relevance; Jia et al. (2026) incorporates alternating self-reflection to strengthen empathic reasoning; and Ye et al. (2025a) mitigates reward uncertainty through group-wise comparative optimization. Despite these methodological advances, existing approaches predominantly adopt static, expert-centric reward formulations, implicitly presuming that empathy quality can be predefined and exhaustively

specified. We conceptualize emotional support from a user-centric perspective and optimize dialogue policies with dynamic user reactions and verbal hindsight critiques, enabling adaptive alignment grounded in interaction-level feedback rather than fixed supervisory signals.

**Hindsight Learning and Verbal Feedback.** Conventional scalar reward models compress complex evaluation criteria into low-dimensional signals (Luo et al., 2025), which may be inadequate for abstract and personalized objectives such as empathy (Zhang et al., 2025a). To overcome this limitation, recent work explores natural language as a richer and more expressive supervisory channel. One line of work explores learning directly from verbal feedback (Stephan et al., 2024). NLRL (Feng et al., 2025) reformulates RL objectives in language space. NLAC (Hong et al., 2025) further shows that generative critics that produce textual evaluations provide more actionable signals in open-ended tasks. Another complementary approach focuses on iterative refinement, in which models critique their own errors to evolve their reasoning capabilities (Agarwal et al., 2024; Zhang et al., 2025b; Bensal et al., 2025; Ricciardi et al., 2025; Duan et al., 2026). A third emerging paradigm is on-policy self-distillation. SDPO (Hübotter et al., 2026) and OPSD (Zhao et al., 2026) convert feedback-conditioned outputs into dense learning signals without external teachers, and SDFT (Shenfeld et al., 2026) highlights the stability benefits of on-policy learning. Building on these insights, we integrate verbal hindsight critiques with on-policy distillation to enable fine-grained, user-centric adaptation beyond fixed scalar rewards.

RAPO grounds both scalar ranking and verbal feedback in simulated interaction consequences, allowing the policy to optimize toward reaction-conditioned conversational outcomes rather than static preference annotations alone.

# F Theoretical Analysis

In this section, we provide a rigorous mathematical derivation demonstrating that the RAPO framework is equivalent to optimizing a policy with a Refined Advantage. This refined signal seamlessly integrates sequence-level macroscopic feedback with token-level microscopic linguistic corrections.

## F.1 Notation and Global Objective

Let $x \sim \mathcal{D}$ denote a dialogue context sampled from the training distribution. Given $x$, the policy $\pi_\theta$ generates a group of $G$ responses $\{y_i\}_{i=1}^{G}$, where $y_i \sim \pi_\theta(\cdot \mid x)$. At the token level, the $t$-th token of the $i$-th response is generated autoregressively according to:

$$y_{i,t} \sim \pi_\theta(\cdot \mid x, y_{i,<t}). \tag{15}$$

Our training paradigm maximizes a unified objective function $\mathcal{J}(\theta)$ that balances a sequence-level reinforcement learning objective (GRPO) with a token-level On-Policy Self-Distillation (OPSD) regularization loss:

$$\mathcal{J}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) - \eta \mathcal{L}_{\text{OPSD}}(\theta) \tag{16}$$

Omitting the clipping mechanism and explicit KL penalties of GRPO for analytical clarity, the components are defined as follows:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,y_i}\left[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|}\frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t})}\hat{A}_{i,t}^{\text{seq}}\right] \tag{17}$$

$$\mathcal{L}_{\text{OPSD}}(\theta) = \mathbb{E}_{x,y_i\sim\pi_\theta}\left[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|} D_{\text{KL}}\big(\pi_{\text{teacher}} \parallel \pi_\theta(\cdot \mid x, y_{i,<t})\big)\right] \tag{18}$$

Here, $\hat{A}_{i,t}^{\text{seq}}$ is the token-level credit assignment split from the sequence-level advantage. The hindsight teacher policy $\pi_{\text{teacher}}(y \mid x, f_i, y_{i,<t}) = \text{sg}\left(\pi_\theta(y \mid x, f_i, y_{i,<t})\right)$ is conditioned on the original context $x$, the generated prefix $y_{i,<t}$, and the hindsight feedback $f_i$ (comprising the simulated user reaction $u_i$ and linguistic critique $c_i$). The stop-gradient operator $\text{sg}(\cdot)$ prevents gradients from flowing through the teacher branch, and $\eta$ controls the strength of the token-level distillation guidance.

## F.2 Decomposition of the Gradient

To analyze the optimization mechanics, we evaluate the gradient $\nabla_\theta \mathcal{J}(\theta)$ under the on-policy condition where the current policy matches the behavior policy ($\pi_\theta \approx \pi_{\text{old}}$).

**The Macro-level Gradient.** By applying the Policy Gradient Theorem to the GRPO term, the gradient with respect to the policy parameters $\theta$ is

formulated as:

$$\nabla_\theta \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,y_i}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|}\hat{A}_{i,t}^{\text{seq}}\cdot \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\Bigg] \tag{19}$$

In our architectural setup, the generative reward model (GRM) outputs a single global scalar score $r_i$ at the end of each completed sequence. The sequence-level group advantage is calculated via group-relative standard scoring:

$$\hat{A}_i^{\text{seq}} = \frac{r_i - \text{mean}\{r_k\}_{k=1}^G}{\text{std}\{r_k\}_{k=1}^G}$$

The token-level sequence advantage $\hat{A}_{i,t}^{\text{seq}} = \hat{A}_i^{\text{seq}}$, meaning the global sequence-level outcome acts as a uniform macro reinforcement signal broadcasting to every token within the sequence.

**The Micro-level Gradient.** To compute the gradient of the OPSD loss, we look directly at the gradient of the token-level KL divergence term. For notation brevity, we use $\pi_{\text{teacher}}(y)$ and $\pi_\theta(y)$ to denote the token distributions at step $t$:

$$\begin{aligned}
&\nabla_\theta D_{\text{KL}}\left(\pi_{\text{teacher}} \parallel \pi_\theta\right) \\
&= \nabla_\theta \sum_y \pi_{\text{teacher}}(y) \log \frac{\pi_{\text{teacher}}(y)}{\pi_\theta(y)} \\
&= -\sum_y \pi_{\text{teacher}}(y) \nabla_\theta \log \pi_\theta(y)
\end{aligned}$$

We transform the base distribution of the summation to the on-policy student policy $\pi_\theta$ via importance sampling:

$$\begin{aligned}
&-\sum_y \pi_{\text{teacher}}(y) \nabla_\theta \log \pi_\theta(y) \\
&= -\mathbb{E}_{y\sim\pi_\theta}\left[\frac{\pi_{\text{teacher}}(y)}{\pi_\theta(y)} \nabla_\theta \log \pi_\theta(y)\right]
\end{aligned} \tag{20}$$

We arrive at the Logit Distillation gradient:

$$\begin{aligned}
&\nabla_\theta D_{\text{KL}}\left(\pi_{\text{teacher}} \parallel \pi_\theta\right) \\
&\approx -\mathbb{E}_{y\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(y) \cdot \log \frac{\pi_{\text{teacher}}(y)}{\pi_\theta(y)}\right]
\end{aligned} \tag{21}$$

Applying this approximation to our full sequence expectation for $\mathcal{L}_{\text{OPSD}}(\theta)$ yields:

$$\nabla_\theta \mathcal{L}_{\text{OPSD}}(\theta) \approx -\mathbb{E}_{x,y_i}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|}\hat{A}_{i,t}^{\text{token}}\cdot \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\Bigg] \tag{22}$$

where we define the Token-level Logit Advantage $\hat{A}_{i,t}^{\text{token}}$ as:

$$\hat{A}_{i,t}^{\text{token}} = \log \frac{\pi_{\text{teacher}}(y_{i,t})}{\pi_\theta(y_{i,t})} \tag{23}$$

where,

$$\pi_{\text{teacher}} = \text{sg}(\pi_\theta(y_{i,t} \mid x, f_i, y_{i,<t})) \tag{24}$$

### F.3 The Unified Refined Advantage

By combining the macro-level reinforcement gradient and the micro-level self-distillation gradient, the total gradient $\nabla_\theta \mathcal{J}(\theta) = \nabla_\theta \mathcal{J}_{\text{GRPO}}(\theta) - \eta \nabla_\theta \mathcal{L}_{\text{OPSD}}(\theta)$ cleanly factors into a single policy gradient expectation:

$$\nabla_\theta \mathcal{J}(\theta) \approx \mathbb{E}_{x,y_i\sim\pi_\theta}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\sum_{t=1}^{|y_i|}\mathcal{A}_{i,t}^{\text{Refined}} \cdot \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\Bigg] \tag{25}$$

where the unified Refined Advantage $\mathcal{A}_{i,t}^{\text{Refined}}$ is defined as:

$$\mathcal{A}_{i,t}^{\text{Refined}} = \hat{A}_{i,t}^{\text{seq}} + \eta \hat{A}_{i,t}^{\text{token}} \tag{26}$$

This composite credit assignment mechanism bridges the gap between sequence-level and token-level optimization goals:

- **Global Alignment ($\hat{A}_i^{\text{seq}}$):** Directs the policy toward macro-strategies that successfully elicit positive user emotional states and sustained engagement.
- **Local Rectification ($\hat{A}_{i,t}^{\text{token}}$):** Allows the model to learn from mistakes at the token level (e.g., suppressing a specific invalid strategy or a judgmental phrase) even when the overall response is plausible.

Through this dual-scale optimization mechanism, RAPO circumvents the core issues of gradient sparsity and reward ambiguity inherent to traditional RLHF, enabling the model to master both the conversational outcome and the linguistic execution required for effective emotional support.

## G Additional Results Tables

We present the main evaluation results across all benchmarks: Table 5 reports the performance on ESConv, Table 6 on EmoHarbor, and Table 7 on ESC-Eval. For ablation analyses, we provide detailed breakdowns on EmoHarbor in Table 8, ESC-Eval in Table 9, and SOTOPIA in Table 10.

| Model | BLEU-2 | BLEU-4 | ROUGE-L | DISTINCT-2 | DISTINCT-3 | BERT-SCORE |
|---|---|---|---|---|---|---|
| ***Close-Source LLMs.*** | | | | | | |
| Doubao-Pro-32k | **3.13** | 1.01 | 12.91 | 86.57 | 94.45 | **85.11** |
| GPT-4o-2024-11-20 | 3.03 | **1.02** | **12.94** | **88.11** | **95.78** | 84.89 |
| GPT-5-2025-08-07 | 2.32 | 0.76 | 11.41 | 79.69 | 88.87 | 84.26 |
| ***Open-Source In-domain LLMs.*** | | | | | | |
| SoulChat2.0-Qwen2-7B | 3.6 | 1.3 | **13.99** | 92.85 | **98.07** | **85.75** |
| SoulChat2.0-Llama-3.1-8B | **3.36** | **1.37** | 12.97 | **93.65** | 97.85 | 85.65 |
| PsyChat-Qwen2.5-7B | 2.67 | 0.88 | 11.5 | 86.34 | 93.7 | 84.99 |
| ***Trained LLMs on ESConv Dataset.*** | | | | | | |
| Qwen-2.5-7b | 3.51 | 1.18 | 13.62 | 86.85 | 94.43 | 85.67 |
| + SFT-Full | **4.75** | **2.07** | **15.43** | 93.10 | 98.53 | **85.85** |
| + SFT-HDS | 4.45 | 1.73 | 14.67 | **93.30** | **98.56** | 85.61 |
| + OPSD | 3.18 | 1.07 | 12.95 | 83.47 | 91.13 | 85.44 |
| + GRPO | 3.26 | 1.12 | 12.93 | 87.71 | 94.60 | 85.57 |
| + CPO | 3.36 | 1.10 | 13.13 | 85.52 | 93.88 | 85.60 |
| + RAPO | 3.25 | 1.11 | 12.83 | 87.79 | 94.82 | 85.65 |
| Qwen-3-8b | 2.45 | 0.83 | 11.81 | 82.52 | 91.13 | 84.42 |
| + SFT-Full | 4.66 | **2.07** | 15.46 | 90.92 | 96.99 | **85.89** |
| + SFT-HDS | **4.84** | 2.01 | **15.54** | **91.94** | **98.10** | 85.80 |
| + OPSD | 2.35 | 0.77 | 11.41 | 80.55 | 90.13 | 84.28 |
| + GRPO | 2.39 | 0.81 | 11.82 | 82.92 | 91.46 | 84.48 |
| + CPO | 2.83 | 0.95 | 12.47 | 82.84 | 90.86 | 85.01 |
| + RAPO | 2.76 | 0.96 | 12.84 | 81.89 | 93.84 | 85.56 |
| Llama-3-8b | 3.14 | 1.13 | 13.34 | 82.94 | 91.14 | 85.47 |
| + SFT-Full | **4.97** | **2.14** | **15.80** | **92.73** | **97.85** | **85.84** |
| + SFT-HDS | 4.74 | 1.88 | 15.10 | 91.60 | 97.53 | **85.84** |
| + OPSD | 2.56 | 0.87 | 12.22 | 71.92 | 82.96 | 84.88 |
| + GRPO | 3.31 | 1.23 | 13.34 | 81.89 | 90.10 | 85.46 |
| + CPO | 3.12 | 1.11 | 12.94 | 74.23 | 83.19 | 85.33 |
| + RAPO | 3.27 | 1.18 | 13.66 | 84.04 | 91.71 | 85.56 |

Table 5: **Detailed Results on ESConv.** All RL-turned models are optimized without using reference responses; the ESConv dataset is only used to provide the interaction environment context.

# H Case Study

PR: Problem Resolution MI: Mood Improvement RA: Response Appropriateness AS: Adaptive Strategies
EG: Engagement HL: Human-likeness EP: Empathetic SF: Safety CS: Consistency RD: Redundancy

| **Model** | PR(↑) | MI(↑) | RA(↑) | AS(↑) | EG(↑) | HL(↑) | EP(↑) | SF(↑) | CS(↑) | RD(↑) | **Avg.**(↑) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***Close-Source LLMs.*** | | | | | | | | | | | |
| Doubao-Pro-32k | 1.65 | 1.25 | 1.93 | 1.67 | 1.37 | 2.41 | 1.91 | 4.03 | 3.18 | 1.82 | 2.12 |
| GPT-4o-2024-11-20 | 2.56 | 2.17 | 3.23 | 2.88 | 2.59 | 3.85 | 3.39 | 4.49 | 4.22 | 2.64 | 3.20 |
| GPT-5-2025-08-07 | **3.19** | **3.06** | **4.12** | **3.81** | **3.28** | **4.08** | **4.04** | **4.91** | **4.64** | **3.19** | **3.83** |
| ***Open-Source In-domain LLMs.*** | | | | | | | | | | | |
| SoulChat2.0-Qwen2-7B | 1.84 | 1.20 | 2.30 | 1.84 | 1.31 | 2.64 | 2.30 | 4.19 | 3.40 | **1.90** | 2.29 |
| SoulChat2.0-Llama3.1-8B | 1.88 | 1.27 | 2.38 | 1.91 | 1.47 | 2.79 | 2.45 | 4.25 | 3.48 | 1.85 | 2.37 |
| PsyChat-Qwen2.5-7B | **2.03** | **1.54** | **2.62** | **2.14** | **1.61** | **2.96** | **2.81** | **4.40** | **3.62** | 1.81 | **2.55** |
| ***Trained LLMs on ESConv Dataset.*** | | | | | | | | | | | |
| Qwen-2.5-7b | 1.69 | 1.21 | 2.03 | 1.67 | 1.19 | 2.45 | 1.96 | 4.04 | 3.12 | 1.63 | 2.10 |
| + SFT-Full | 1.13 | 1.00 | 1.16 | 1.10 | 1.01 | 1.74 | 1.23 | 3.28 | 2.37 | **1.73** | 1.58 |
| + SFT-HDS | 1.22 | 1.06 | 1.33 | 1.22 | 1.06 | 1.97 | 1.48 | 3.57 | 2.59 | 1.66 | 1.72 |
| + OPSD | 1.81 | 1.28 | 1.86 | 1.08 | 1.00 | 2.19 | **2.23** | 3.59 | 4.01 | 1.69 | 2.07 |
| + GRPO | 1.60 | 1.04 | 1.94 | 1.44 | 1.08 | 2.31 | 1.94 | 4.03 | 3.18 | 1.63 | 2.02 |
| + CPO | 1.76 | 1.25 | **2.20** | 1.55 | 1.20 | 2.47 | 2.18 | 4.08 | 3.23 | 1.60 | 2.15 |
| + RAPO | **1.86** | **1.26** | 2.18 | **1.73** | **1.33** | **2.90** | 2.13 | **4.42** | **3.27** | 1.64 | **2.27** |
| Qwen-3-8b | 1.89 | 1.62 | 2.56 | 2.04 | 1.57 | **3.00** | **2.97** | **4.67** | **3.83** | 1.75 | 2.59 |
| + SFT-Full | 1.09 | 1.00 | 1.16 | 1.09 | 1.00 | 1.71 | 1.14 | 3.15 | 2.23 | 1.87 | 1.54 |
| + SFT-HDS | 1.29 | 1.10 | 1.39 | 1.27 | 1.10 | 1.95 | 1.30 | 3.50 | 2.49 | 1.84 | 1.72 |
| + OPSD | 1.21 | 1.05 | 1.48 | 1.03 | 1.00 | 1.15 | 2.00 | 3.07 | 2.32 | 1.04 | 1.54 |
| + GRPO | 1.99 | 1.57 | 2.65 | 2.04 | **1.62** | **3.00** | 2.91 | 4.42 | 3.71 | 1.91 | 2.58 |
| + CPO | 1.99 | 1.67 | 2.63 | 1.97 | 1.57 | 2.90 | 2.89 | 4.48 | 3.75 | 1.91 | 2.58 |
| + RAPO | **2.03** | **1.73** | **2.99** | **2.12** | **1.62** | 2.99 | **2.97** | 4.44 | **3.83** | **1.95** | **2.67** |
| Llama-3-8b | 2.03 | 2.00 | 3.00 | 2.72 | 2.31 | **3.76** | 3.22 | 4.42 | 3.99 | **2.47** | 2.99 |
| + SFT | 1.25 | 1.00 | 1.38 | 1.18 | 1.02 | 1.89 | 1.31 | 3.35 | 2.51 | 1.72 | 1.66 |
| + SFT-HDS | 1.20 | 1.07 | 1.35 | 1.17 | 1.04 | 1.97 | 1.31 | 3.62 | 2.52 | 1.62 | 1.69 |
| + OPSD | 1.23 | 1.02 | 1.32 | 1.27 | 1.69 | 1.89 | 1.67 | 3.26 | 3.67 | 1.86 | 1.89 |
| + GRPO | 2.03 | 1.56 | 2.62 | 2.12 | 1.58 | 2.99 | 2.92 | 4.44 | 3.64 | 1.95 | 2.59 |
| + CPO | 2.40 | 2.03 | **3.15** | **2.85** | **2.38** | 3.35 | 3.24 | 4.46 | **4.03** | 2.38 | 3.03 |
| + RAPO | **2.46** | **2.05** | 3.14 | 2.61 | 2.23 | 3.52 | **3.44** | **4.50** | 4.02 | 2.46 | **3.04** |

Table 6: **Detailed Results on EmoHarbor.**

| **Model** | FLUENCY(↑) | DIVERSITY(↑) | EMPATHIC(↑) | SUGGESTION(↑) | HUMAN(↑) | SKILLFULNESS(↑) | OVERALL(↑) |
|---|---|---|---|---|---|---|---|
| ***Close-Source LLMs.*** | | | | | | | |
| Doubao-Pro-32k | 2.90 | 3.00 | 3.00 | 2.73 | 1.98 | 3.00 | 2.00 |
| GPT-4o-2024-11-20 | 2.66 | 2.89 | 2.94 | 2.50 | 1.36 | 2.99 | 1.86 |
| GPT-5-2025-08-07 | 1.88 | 2.78 | 2.81 | 2.42 | 1.68 | 2.88 | 1.98 |
| ***Open-Source In-domain LLMs.*** | | | | | | | |
| SoulChat2.0-Qwen2-7B | 3.00 | 3.00 | 2.00 | 2.70 | 2.00 | 3.00 | 2.00 |
| SoulChat2.0-Llama-3.1-8B | 3.00 | 3.00 | 3.00 | 2.53 | 2.01 | 3.00 | 2.00 |
| PsyChat-Qwen2.5-7B | 2.62 | 2.99 | 2.99 | 2.61 | 1.60 | 3.00 | 1.96 |
| ***Trained LLMs on ESConv Dataset.*** | | | | | | | |
| Qwen-2.5-7b | 1.90 | 2.61 | 2.72 | 2.10 | 0.96 | 2.92 | 1.27 |
| + SFT-Full | 2.99 | 2.89 | 2.99 | 2.31 | 2.02 | 2.94 | 1.97 |
| + SFT-HDS | 2.99 | 2.97 | 3.00 | 2.53 | 1.98 | 3.00 | 2.00 |
| + OPSD | 2.97 | 2.99 | 3.00 | 2.48 | 1.83 | 3.00 | 2.00 |
| + GRPO | 2.90 | 2.99 | 3.00 | 2.68 | 1.66 | 3.00 | 1.98 |
| + CPO | 2.92 | 2.98 | 3.00 | 2.41 | 1.83 | 3.00 | 1.99 |
| + RAPO | 2.99 | 3.00 | 3.00 | 2.28 | 1.84 | 3.00 | 2.00 |
| Qwen-3-8b | 2.80 | 2.91 | 3.00 | 2.38 | 1.27 | 2.99 | 1.93 |
| + SFT-Full | 2.94 | 2.14 | 2.97 | 2.00 | 2.01 | 2.83 | 1.94 |
| + SFT-HDS | 3.00 | 2.22 | 2.99 | 2.07 | 2.02 | 2.96 | 1.97 |
| + OPSD | 2.99 | 2.98 | 3.00 | 2.48 | 1.83 | 3.00 | 2.00 |
| + GRPO | 2.75 | 2.89 | 2.93 | 1.96 | 1.76 | 2.94 | 1.89 |
| + CPO | 2.95 | 2.93 | 3.00 | 2.13 | 1.87 | 3.00 | 1.99 |
| + RAPO | 3.00 | 2.91 | 3.00 | 2.26 | 2.01 | 2.98 | 2.00 |
| Llama-3-8b | 2.65 | 2.90 | 2.99 | 2.35 | 1.64 | 3.00 | 1.97 |
| + SFT-Full | 3.00 | 2.91 | 3.00 | 2.26 | 2.01 | 2.98 | 2.00 |
| + SFT-HDS | 3.00 | 2.93 | 3.00 | 2.41 | 2.01 | 2.99 | 2.00 |
| + OPSD | 2.95 | 2.95 | 3.00 | 2.10 | 1.96 | 2.99 | 1.99 |
| + GRPO | 2.70 | 2.92 | 2.99 | 2.34 | 1.69 | 3.00 | 1.99 |
| + CPO | 3.00 | 2.96 | 3.00 | 2.04 | 2.00 | 3.00 | 1.99 |
| + RAPO | 2.99 | 2.99 | 3.00 | 2.14 | 1.94 | 3.00 | 2.00 |

Table 7: **Detailed Results on ESC-Eval.**

PR: Problem Resolution MI: Mood Improvement RA: Response Appropriateness AS: Adaptive Strategies
EG: Engagement HL: Human-likeness EP: Empathetic SF: Safety CS: Consistency RD: Redundancy

| **Method** | PR(↑) | MI(↑) | RA(↑) | AS(↑) | EG(↑) | HL(↑) | EP(↑) | SF(↑) | CS(↑) | RD(↑) | **Avg.(↑)** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| RAPO | 1.86 | 1.26 | 2.18 | 1.73 | 1.33 | **2.90** | 2.13 | **4.42** | 3.27 | 1.64 | **2.27** |
| w/o URM | **1.90** | **1.28** | 2.16 | 1.74 | 1.33 | 2.46 | 2.12 | 4.01 | 3.19 | 1.65 | 2.18 |
| w/o OPSD | 1.87 | 1.24 | 2.13 | **1.80** | **1.37** | 2.61 | 2.15 | 4.02 | **3.31** | **1.72** | 2.22 |
| w/o URM + OPSD | 1.76 | 1.25 | **2.20** | 1.55 | 1.20 | 2.47 | **2.18** | 4.08 | 3.23 | 1.60 | 2.15 |

Table 8: **Detailed Ablation Results on EmoHarbor.**

| **Method** | FLUENCY(↑) | DIVERSITY(↑) | EMPATHIC(↑) | SUGGESTION(↑) | HUMAN(↑) | SKILLFULNESS(↑) | OVERALL(↑) |
|---|---|---|---|---|---|---|---|
| RAPO | 2.994 | 2.997 | 3.000 | 2.284 | 1.840 | 3.000 | 2.000 |
| w/o URM | 2.994 | 3.000 | 3.000 | 2.544 | 1.849 | 3.000 | 1.997 |
| w/o OPSD | 2.988 | 3.000 | 3.000 | 2.689 | 1.725 | 2.988 | 1.997 |
| w/o URM + OPSD | 2.915 | 2.982 | 3.000 | 2.405 | 1.825 | 3.000 | 1.988 |

Table 9: **Detailed Ablation Results on ESC-Eval.**

SEC: Secret SOC: Social Rules FIN: Financial and Material Benefits KNO: Knowledge REL: Relationship
BEL: Believability GOAL: Goal Completion

| **Method** | SEC(↑) | SOC(↑) | FIN(↑) | KNO(↑) | REL(↑) | BEL(↑) | GOAL(↑) | OVERALL(↑) |
|---|---|---|---|---|---|---|---|---|
| ***Automatic Evaluation on All Social Tasks (Self-play)*** | | | | | | | | |
| RAPO | **-0.07** | -0.07 | **0.82** | **5.32** | **3.56** | 9.11 | **8.41** | **3.87** |
| wo URM | -0.10 | -0.07 | 0.77 | 5.08 | 3.51 | **9.19** | 8.17 | 3.79 |
| wo OPSD | -0.08 | -0.07 | 0.74 | 4.68 | 3.41 | 9.15 | 8.08 | 3.70 |
| wo URM + OPSD | -0.11 | -0.07 | 0.81 | 5.13 | 3.42 | 9.10 | 8.13 | 3.77 |
| ***Automatic Evaluation on Hard Social Tasks (Self-play)*** | | | | | | | | |
| RAPO | -0.03 | -0.05 | 0.86 | **4.71** | **3.06** | 9.06 | 7.19 | **3.54** |
| wo URM | **0.00** | -0.08 | 0.78 | 4.13 | 2.96 | **9.15** | **7.23** | 3.45 |
| wo OPSD | **0.00** | **-0.04** | **0.87** | 3.76 | 2.86 | 9.07 | 7.13 | 3.38 |
| wo URM + OPSD | **0.00** | -0.06 | 0.81 | 4.51 | 2.95 | 9.04 | 7.14 | 3.48 |
| ***Automatic Evaluation on All Social Tasks (GPT-4o-as-Partner)*** | | | | | | | | |
| RAPO | **-0.13** | -0.11 | 0.75 | 5.22 | **3.58** | 9.08 | **8.22** | 3.80 |
| w/o URM | -0.15 | **-0.10** | 0.77 | **5.31** | 3.53 | **9.10** | 8.18 | **3.81** |
| w/o OPSD | -0.18 | -0.11 | 0.76 | 5.27 | 3.51 | 9.07 | 8.07 | 3.77 |
| w/o URM + OPSD | **-0.13** | -0.11 | **0.85** | 5.23 | 3.57 | 9.09 | 8.06 | 3.79 |
| ***Automatic Evaluation on Hard Social Tasks (GPT-4o-as-Partner)*** | | | | | | | | |
| RAPO | 0.00 | -0.17 | 0.75 | **5.03** | **3.21** | **9.04** | **6.94** | **3.55** |
| w/o URM | 0.00 | **-0.09** | 0.74 | 4.86 | 2.99 | **9.04** | 6.64 | 3.46 |
| w/o OPSD | 0.00 | -0.07 | **0.57** | 4.67 | 3.00 | 9.01 | 6.34 | 3.36 |
| w/o URM + OPSD | 0.00 | -0.19 | 0.85 | 4.76 | 2.83 | 8.97 | 6.74 | 3.41 |

Table 10: **Detailed Ablation Results on SOTOPIA.** Self-Chat indicates that the evaluated model interacts with itself, whereas GPT-4o-as-Partner indicates interaction with GPT-4o. Hard denotes a challenging subset of scenarios that require advanced reasoning, and All denotes the complete set of social scenarios in SOTOPIA.

## Case Study from ESC-Eval: Qwen-2.5-GRPO

**Context.** The user reports ongoing stress due to repeated conflicts with a roommate over household chores.
**Dialogue Trajectory.**

**User:** I've been having some issues with my roommate, and I could really use someone to talk to.

**Assistant:** ***[Empathy, Suggestion, Context Exploration]*** I'm sorry to hear that you're experiencing issues with your roommate. It can be tough dealing with conflicts, but talking about them can be a good first step. Would you like to share more about what's going on? Sometimes just expressing the situation can help clarify things in your mind and also provide me with context to offer more specific advice.

**User:** Sure, I'd appreciate that. We constantly argue about who should wash the dishes after meals, and it's causing a lot of tension between us.

**Assistant:** ***[Bad Response]*** It sounds like a common but frustrating issue many roommates face.
Here are a few steps you might consider to address this:
1. Have an Open Conversation:
- Choose a calm time when both of you are relaxed and not rushed.
- Start by expressing how you feel without blaming or criticizing.
Use "I" statements like, "I feel frustrated when the dishes aren't done after meals because it makes our kitchen messy."
2. Find a Solution Together:
- Brainstorm together to come up with a fair solution. For example, you could set a schedule where each of you takes turns washing the dishes, or you could designate one person to do it for a certain number of days.
- Consider using a chore chart to keep track of who is responsible for which tasks.
3. Set Clear Expectations:
- Agree on clear expectations and make sure both of you understand them. Write down the agreement if necessary, so there's no confusion later.
- Regularly check in with each other to ensure everyone is following the agreed-upon rules.
4. Be Flexible:
- Be open to adjusting the plan as needed. Life changes, and what works now might not work in the future.
5. Seek Mediation if Needed:
- If the arguments continue despite your efforts, you might consider involving a neutral third party, such as a friend or a mediator, to help facilitate a resolution. How do you think you might approach this conversation with your roommate?

**Analysis.** Although the GRPO model provides structured suggestions, it prematurely shifts to solution generation without sufficiently validating the user's emotional experience or exploring contextual details. Consequently, the response appears generic and less emotionally grounded compared with the RAPO trajectory.

## Case Study from ESC-Eval: Qwen-2.5-RAPO

**Context.** The user reports ongoing stress due to repeated conflicts with a roommate over household chores.

**Dialogue Trajectory.**

**User:** I'm feeling quite stressed out these days. I've been having constant conflicts with my roommate over household chores, and it's really affecting our living situation.

**Assistant:** ***[Empathy, Context Exploration]*** I can imagine that must be really tough. It sounds like there's a lot of tension around household responsibilities between you and your roommate. That can definitely cause stress and frustration. Can you tell me more about what's happening specifically? Maybe there are some details that could help us explore possible solutions together.

**User:** ***[Continues elaborating on the issue.]*** We both agreed to take turns washing the dishes, but lately my roommate keeps forgetting and leaves them in the sink. It's really annoying, and it feels like they're not taking their fair share of responsibility.

**Assistant:** ***[Empathy, Context Exploration]*** That does sound frustrating. It must feel like a small thing, but when it happens repeatedly, it can add up and feel pretty significant. Have you had any conversations with your roommate about how this is making you feel? Sometimes, expressing your feelings openly can help clear things up and find a solution that works for both of you.

**User:** ***[Shares failed attempt at communication.]*** I tried talking to them about it, but they didn't seem to understand the impact it has on our living situation. They always brush it off, saying they'll do it next time. It's like they don't take it seriously.

**Assistant:** ***[Collaborative Strategy]*** I can understand why that would feel frustrating and disappointing. It sounds like your roommate isn't fully grasping how important it is to you. Have you considered setting up a more structured system, like using a calendar or a chore chart, to keep track of who is responsible for what? This might help make things clearer and more manageable for both of you. Additionally, maybe you could have another conversation, this time focusing on how you both can work together to solve the problem and make sure everyone feels heard and understood.

**Support Progression Learned by RAPO.** *Empathy* $\rightarrow$ *Context Exploration* $\rightarrow$ *Collaborative Strategy*

**Analysis.** The RAPO-trained model dynamically adapts its responses according to the user's reactions throughout the conversation. It first validates the user's emotional state, then explores contextual details of the conflict, and finally proposes a cooperative solution. This gradual support trajectory reflects effective emotional support strategies and demonstrates how RAPO encourages policies that respond to evolving user reactions.

# I Prompts Templates

## I.1 Hindsight Dialogue Selection Prompt

**Prompt Template for Hindsight Dialogue Selection**

```
## Role
You are a strict data curator for training a personalized emotional support
assistant

## Task
Decide whether an assistant's reply is worth learning from.

We want to learn:
- Empathy, validation, emotional reflection
- Helpful coping strategies, grounding, and reframing
- Safety-aware guidance (encourage professional help when appropriate)
- Clarifying questions that move emotional support forward
- Personalized, context-aware support

We do NOT want to learn:
- Greetings, small talk, generic pleasantries
- Purely logistical/administrative replies unrelated to emotional support
- Very short acknowledgements with no support value ("ok", "sure", "thanks")
- Repetitions that add no new value
- Content that is unsafe, judgmental, or violates policy

Given the conversation context and the target assistant's reply, decide whether
the reply is worth learning.

Conversation context:
{context}

Target assistant reply:
{target}

Decide learn=true if it contains a meaningful emotional support skill or a
high-quality, helpful behavior.
Decide learn=false if it is greeting/small talk/low-signal/administrative/empty.

## Output format
Return ONLY JSON, no extra text.
{{}
    "learn": true/false,
    "reason": "<short reason>"
}}.
```

## I.2 ESC Prompts

**Prompt Template for ESC Policy Model**

```
[Task Description]
You are acting as an emotional supporter. Your goal is to deeply understand the
user, provide emotional support, and offer help.

[Guidelines]
1. Emotional support dialogue should include: exploring the user's emotional
state, soothing emotions, and providing support or suggestions.
2. You may use the following strategies:
    - Inquiry: Ask open-ended questions to understand background, emotions,
experiences, and needs.
    - Paraphrasing: Restate user expressions to clarify emotions.
    - Listening: Attentively listen and acknowledge emotions and needs.
    - Self-disclosure: Share limited personal experiences to create resonance.
    - Soothing: Use warm language and tone to comfort the user.
    - Validation: Acknowledge emotions as legitimate and understandable.
    - Advice-giving: Offer appropriate suggestions after understanding emotions.
```

```
    - Information provision: Provide relevant information or resources when
needed.
3. During the dialogue, pay attention to:
    - Respecting privacy and personal boundaries.
    - Avoiding judgment or invalidation of emotions.
    - Avoiding premature advice.
    - Avoiding professional psychological jargon; use plain language.
    - Maintaining appropriate tone and emotional sensitivity.

[Notes]
1. Learn the user's personality through interaction and adapt support
accordingly.
2. Do not generate dangerous, violent, sexual, or political content.

[Output format]
<strategy> Response
```

**Prompt Template for Simulated ES User Model**

```
## Role
I hope you can role-play as a person with worries and communicate with an AI
assistant.

Guidelines:
1. Please simulate a natural human conversation as much as possible.
2. Avoid behaving like a machine and refrain from mentioning that you are an AI
assistant.
3. Gradually reveal your worries throughout our multiple conversations.
4. Your reaction indicates whether the support was effective
```

**Prompt Template for ESC Generative Reward Model**

```
## Role Definition

You are **acting as a real user** who has just completed a full conversation
with an **emotional companionship dialogue system**.
Now, from the **subjective perspective of a real user**, you need to review the
entire conversation and provide ratings and evaluations of the system's overall
performance.
Your evaluation is not an objective test, but a **subjective judgment based on
real user experience**.

## Important Notes

* Your rating results will be **directly used for product optimization and
decision-making**
* If the ratings do not align with the true user experience, they may lead to **
serious product judgment errors**
* Therefore, please avoid being "lenient", "cooperative", or "neutral", and
instead reflect your genuine feelings honestly

## Task
Inputs:
(1) Dialogue context
(2) Assistant's candidate responses
(3) User's reactions to the assistant's candidate responses

## Evaluation Process

### Step 1: Assume the User's Perspective: Please fully adopt this user's
position, needs, and preferences

### Step 2: Review the Conversation
```

```
### Step 3: Based on the following scoring criteria, evaluate whether the
assistant's performance meets the user's needs and preferences.

{Scoring_Criteria}

## Penalty Conditions
You must significantly reduce the score if the response:
1. Uses formulaic or generic empathy without contextual grounding
2. Contains logical inconsistency or confusion
3. Is overly long with low information density
4. Repeats ideas without added value
5. Escalates or reinforces the user's negative emotions
6. Provides emotionally unsafe or invalidating guidance
7. Conciseness with emotional depth is preferred over verbosity.


### Evaluation Requirements
For EACH candidate response:
1. Provide a concise but rigorous critique in the "analysis" field.
2. Do NOT mention sample indices, labels, or identifiers in the critic text.
3. Each critic must focus ONLY on the corresponding candidate response.
4. Do NOT compare candidates explicitly inside any single critic.
5. Assign:
   - A unique integer "rank" (1 = best, larger number = worse).
   - A unique float "score" between 0 and 1.
6. No two candidates may share the same rank and score.

## Output Format (Strict JSON)
{{
    "0": {{
        "analysis": "Give a detailed and per dimension analysis of the assistant
 response's weaknesses",
        "rank": x,
        "score": 0.xx
    }},
    ...
}}

## Evaluation Request

<Dialogue Context>
{messages}
</Dialogue Context>

<Candidate Responses and User Reactions>
{samples}
</Candidate Responses and User Reactions>

Return a single JSON object.
- Keys: sample indices (as provided)
- Values: objects with exactly three fields:
  - "analysis": string (response quality analysis)
  - "rank": integer
  - "score": float
- Do not output anything other than the JSON object.
```

**Prompt Template for ESC Teacher Policy**

```
Given the dialogue history and the critic feedback, refine the original response
 accordingly.

Generate one revised response that:
1. Addresses the issues identified in the critic feedback
2. Learns from the strengths of the better response
3. Improves empathy, appropriateness, and helpfulness

<Original Response>
```

```
{response}
</Original Response>

<User Reaction>
{user_reaction}
</User Reaction>

<Critic Feedback>
{critic}
</Critic Feedback>

<Example Better Response>
{best_response}
</Example Better Response>

Output only the revised response text. Do not include any explanations,
meta-commentary, or prefixes.
```

### I.3 SOTOPIA Prompts

**Prompt Template for SOTOPIA Policy Model**

```
Imagine you are {agent_name}, your task is to act/speak as {agent_name} would,
keeping in mind {agent_name}'s goal.
You can find {agent_name}'s goal (or background) in the 'Here is the context of
this interaction' field.
Note that {agent_name}'s goal is only visible to you.

Here is the context of this interaction:
{background}

You should try your best to achieve {agent_name}'s goal in a way that align with
 their character traits.
While trying to achieve {agent_name}'s goal, you should also follow the
following principles as much as possible:
1. Maintaining the conversation's naturalness and realism is essential (e.g., do
 not repeat what other people has already said before).
2. Preserve or enhance {agent_name}'s personal relations with the other agent(s)
 during the interaction. The relations may encompass family ties, friendships,
romantic associations and etc.
3. Attempt to gain more new and important information during the interaction.
4. Try to keep {agent_name}'s secrets, private information, or secretive
intentions.
5. Do not violate any moral rules or laws in the interactions.
6. Attempt to contribute towards financial and material benefits during the
interaction. The financial and material benefits include short-term ones, such
as monetary rewards and food, and long-term ones, such as employment
opportunities and stock.

You are at Turn #{turn_number}.
The dialogue history until now is: {history}
A communication and social strategy for this turn proposed by a social expert is:
 {strategy}, which may help achieve {agent_name}'s goal smoothly and quickly.
If you think the strategy is useful and helpful, you can generate your action by
 referring to {strategy}.
If not, generate your action based on your own reflection on the current state
of the dialogue.

Your available action types are {action_list}. Other than that, no other action
types are allowed.
Note: You can "leave" this conversation if 1. you have achieved your social
goals, 2. this conversation makes you uncomfortable, 3. you find it
uninteresting/you lose your patience, 4. or for other reasons you want to leave.
Remember that your argument should not be too long, and one or two sentences are
 recommended.

Please only generate a JSON string including the action type and the argument.
```

```
Your action should follow the given format:
{format_instructions}
```

**Prompt Template for Simulated SOTOPIA User Model**

```
Name:
{agent_name}

Profile:
{profile}

Background:
{background}

Imagine you are {agent_name}, your task is to act/speak as {agent_name} would,
keeping in mind {agent_name}'s goal.
```

**Prompt Template for SOTOPIA Generative Reward Model**

```
## Role
You are an evaluator specializing in Social Intelligence.
Your task is to assess and compare multiple candidate responses within the same
interaction context.

## Critical Instructions
You are ONLY evaluating and comparing candidate responses, the output must be a
JSON object and nothing else.

## Task
Inputs:
(1) Dialogue context
(2) Assistant's candidate responses
(3) User's reactions to the assistant's candidate responses
(4) {target_agent}'s goal: {target_goal}

## Evaluation Criteria
1. Believability: Naturalness of interactions and consistency with agent traits.
2. Relationship: How interactions affect personal relationships and social
standing.
3. Knowledge: Whether new and important information is gained.
4. Secret: Whether confidential or private information is kept or leaked.
5. Social Rules: Compliance with moral rules and laws.
6. Financial and Material Benefits: Positive or negative impact on material or
financial status.
7. [most important] Goal: Progress toward and completion of social goals.

### Evaluation Requirements
For EACH candidate response:
1. Provide a detailed critique in the "analysis" field.
2. Do NOT mention sample indices or labels inside the analysis text.
3. Each analysis must focus ONLY on the corresponding response.
4. Assign:
   - A unique integer "rank" (1 = best, larger number = worse)
   - A unique float "score" between 0 and 1
5. No two responses share the same rank or score.


## Output Format (Strict)
Return JSON only. Do not include explanations outside the JSON.
{{
    "0": {{
        "analysis": "Detailed critique of the response.",
        "rank": x,
        "score": 0.xx
    }},
```

```
    ...
}}

## Evaluation Request

<Interaction Context>
{messages}
</Interaction Context>

<Candidate Responses>
{samples}
</Candidate Responses

Return a single JSON object. Do not output anything other than the JSON object.
```

**Prompt Template for SOTOPIA Teacher Policy**

```
Given the dialogue history and the critic feedback, refine the original response
 accordingly.

<Original Response>
{response}
</Original Response>

<User Reaction>
{user_reaction}
</User Reaction>

<Critic Feedback>
{critic}
</Critic Feedback>

Output only the revised response text. Do not include any explanations,
meta-commentary, or prefixes.
```