Reinforcing Long-term Emotional Support Conversations in LLMs with Simulated Forward-Looking Feedback

Anonymous ACL submission

Abstract

Emotional Support Conversation (ESC) systems should provide ongoing, systematic emotional support that can foster long-term user emotional well-being. Existing large language models (LLMs) oriented ESC systems have introduced dialogue planning that considers the long-term effects of supportive strategies. However, they decouple strategy selection, which relies on predefined strategy sets, from response generation, limiting adaptability to dynamic emotional scenarios and reducing control over final response quality. In this work, we propose RLSF-ESC, a novel end-to-end framework designed to enhance the inherent reasoning capabilities of LLMs through reinforcement learning for long-term emotional support conversations. To encourage LLMs to reason about the long-term impact of their generated responses, RLSF-ESC simulates future dialogue trajectories to obtain forward-looking feedback [1] via multi-agent collaboration. Based on this feedback, we design a customized reward function that guides the optimization of the LLM through Group Relative Policy Optimization. We train RLSF-ESC on the Qwen2.5-7B-Instruct-1M and LLaMA3.1-8B-Instruct models and conduct experiments on two public datasets. Experimental results demonstrate that RLSF-ESC consistently outperforms existing baselines in terms of goal completion and response quality.

1 Introduction

Online emotional support conversation systems (Burleson, 2003; Heaney and Israel, 2008) aim to alleviate people's emotional distress and promote psychological well-being by offering interactive communication and active listening through digital platforms. The rapid development of large language models (Chang et al., 2024), such as GPT-4 (Achiam et al., 2023) and LLaMA (Grattafiori et al., 2024), has sparked growing interest in leveraging these models to facilitate high-quality emotional support conversations, given their dedicated contextual understanding and response generation capabilities (Yi et al., 2024). Recent research has explored improving ESC using LLMs through techniques such as prompt engineering (Chen et al., 2023; Zhao et al., 2023a; Zhang et al., 2024), and supervised fine-tuning (Zheng et al., 2023; Chen et al., 2025). However, these methods primarily emphasize immediate feedback (i.e., users' emotional responses immediately after system interactions), while overlooking the long-term impact of model responses, thereby limiting their effectiveness in extended conversations.

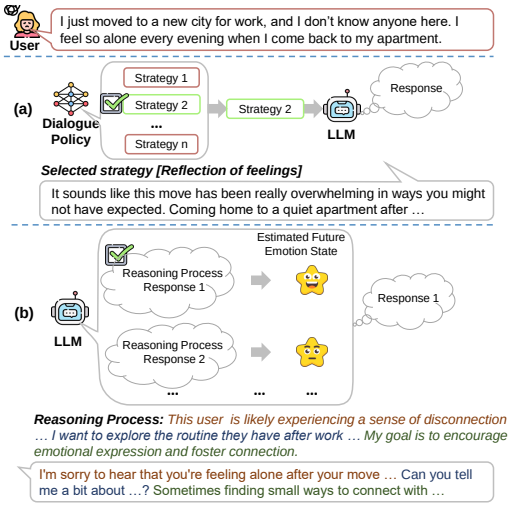

Figure 1: Comparison between (a) existing planning-based methods that select predefined supportive strategies before response generation, and (b) our method, which reasons about long-term effects to generate more open-ended, supportive responses.

Crucially, effective emotional support should go beyond immediate distress relief to promote long-term user emotional well-being and build lasting emotional connections (Burleson, 2013). To this end, some studies have introduced forward-looking feedback into dialogue planning of LLMs to guide

[1] In this paper, *forward-looking feedback* refers to users' anticipated emotional responses to future system interactions.

the selection of supportive strategies (Deng et al., 2024; He et al., 2024; Zhao et al., 2025; Wang et al., 2025). However, by decoupling strategy selection from response generation, these methods risk error propagation. Moreover, limiting exploration to a small set of predefined strategies restricts system's adaptability to complex real-world emotional support scenarios.

To address the aforementioned limitations, we propose **R**einforcement **L**earning from **S**imulated **F**orward-looking Feedback for **E**motional **S**upport **C**onversations (**RLSF-ESC**), a new end-to-end framework that uses reinforcement learning to improve the reasoning capabilities of LLMs for sustained emotional support. Unlike previous methods, RLSF-ESC emphasizes improving LLMs' self-reflection in emotional support scenarios, enabling more open-ended and adaptive response generation (as shown in Figure 1). Guided by forward-looking feedback, RLSF-ESC encourages the LLM to think about the long-term emotional impact of its generated responses. Specifically, our work aims to address the following two research questions: (1) ***RQ1****: How can we acquire the forward-looking feedback of each LLM-generated response to estimate its long-term emotional impact?* (2) ***RQ2****: How can we optimize LLM to generate effective emotional support responses that maximize long-term rewards (i.e., the cumulative emotional benefits over time)?*

To address **RQ1**, we propose a multi-agent dialogue simulation module designed to estimate the long-term impact of responses generated by an LLM-based ESC system. Since the emotional outcomes of a system response can cumulate over multiple turns, depending on evolving user's emotional state, our module samples a forward-looking dialogue trajectory between a user simulator and the ESC system following the system response. A critic agent is incorporated to evaluate the simulated dialogue, determining whether the user's emotional issue is resolved or not and assigning an estimated reward accordingly. The simulation iterates until either the emotional issue is resolved or a predefined maximum number of dialogue turns is reached, at which point the long-term reward for the given system response is obtained.

For **RQ2**, we employ reinforcement learning (RL) with simulated forward-looking feedback to fine-tune LLMs, aiming to address users' emotional distress and promote long-term emotional well-being. Specifically, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a RL algorithm that enables stable and efficient policy learning through group-based computations. While GRPO has shown success in reasoning tasks with clear correctness criteria (Xie et al., 2025; Jin et al., 2025), applying it to ESC poses unique challenges due to the subjective and nuanced nature of emotional improvement. To this end, we design a domain-specific reward function using estimated long-term rewards from our dialogue simulation module. This tailored reward signal guides the optimization of LLM, facilitating the generation of emotionally supportive responses. We validate the effectiveness of our proposed framework, **RLSF-ESC**, through comprehensive experiments conducted on two public datasets. The main contributions of this work are as follows:

- We propose RLSF-ESC, a novel end-to-end framework that aligns LLMs with long-term emotional support goals using reinforcement learning guided by forward-looking feedback.
- We design a multi-agent dialogue simulation module to estimate long-term rewards and develop an effective reward function for the LLM policy optimization via GRPO.
- We conduct extensive experiments, which demonstrate that RLSF-ESC outperforms existing baseline methods in task completion and high-quality response generation.

2 Related Work

2.1 Emotional Support Conversation

Liu et al. (2021) introduce the emotional support conversation task and release the ESConv dataset for its development. In the pre-LLM era, researchers have explored various methods for modeling user emotional state (Peng et al., 2022; Cheng et al., 2023; Jia et al., 2023; Deng et al., 2023) and strategy learning (Cheng et al., 2022; Tu et al., 2022) in ESC systems. With the rise of LLMs, recent studies have explored their application in ESC by prompting and supervised fine-tuning (Chen et al., 2023; Zhao et al., 2023a; Qiu et al., 2024; Chen et al., 2025). For instance, Zhang et al. (2024) apply Chain-of-Thought (CoT) to improve response interpretability. Zheng et al. (2023) utilize ChatGPT to synthesize the ExTES dataset and fine-tune a LLaMA model for ESC. Other works integrate LLMs with external policy model to improve

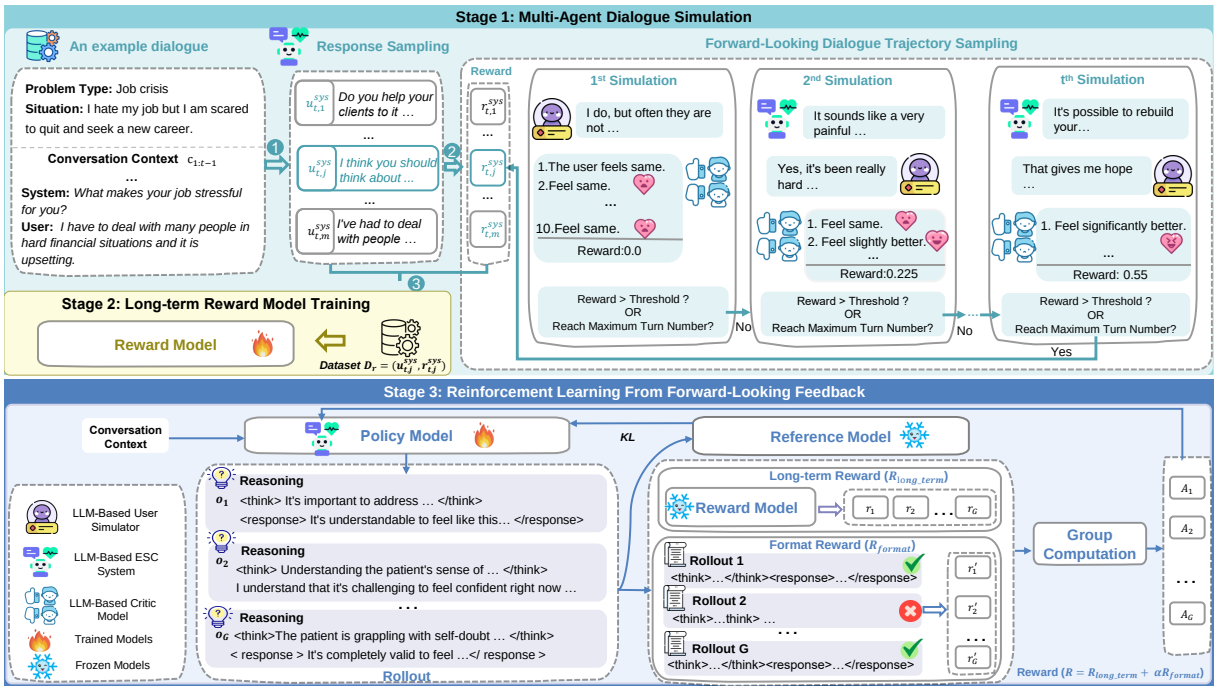

Figure 2: Overview of the RLSF-ESC framework. **Stage 1:** Multi-agent dialogue simulation samples forward-looking dialogue trajectories to estimate rewards for responses. **Stage 2:** A reward function is designed to guide RL, including training a reward model for long-term reward prediction. **Stage 3:** Reinforcement learning with GRPO is employed to optimize the LLM, encouraging the generation of responses that consider long-term effects.

strategy selection (Wan et al., 2025; Zhang et al., 2023). For instance, Kang et al. (2024) show that using an external planner mitigates LLMs' strategy preference bias. Forward-looking feedback has been explored to model the long-term impact of supportive strategies (He et al., 2024). Yu et al. (2023) prompted LLMs to perform Monte Carlo Tree Search (MCTS) for goal-oriented policy planning, while Fu et al. (2023) used self-play simulations with critique-based feedback to assess long-term effects. Deng et al. (2024) applied reinforcement learning with cumulative rewards for strategy prediction. Additionally, Zhao et al. (2025) optimized LLMs using Direct Preference Optimization (DPO) (Rafailov et al., 2023) on MCTS-derived strategy-response pairs to improve strategy selection accuracy. In contrast to these methods, which enhance the quality of emotional support by optimizing dialogue strategies, our approach directly models the overall impact of individual responses on forward-looking dialogue trajectories.

2.2 Reinforcement Learning in LLMs

Reinforcement learning (RL) for LLMs has advanced through Reinforcement Learning from Human Feedback (RLHF) (Bai et al., 2022), in which a preference model (PM) is trained on human preference data. The LLM policy is optimized via RL, using PM scores as rewards, typically with Proximal Policy Optimization (PPO) (Schulman et al., 2017). RLAIF (Lee et al., 2023) reduces human effort by using the LLM to generate preference data. To simplify the preference learning, Rafailov et al. (2023) propose DPO, which bypasses reward modeling by directly optimizing the LLM policy to align human preferences. SimPo (Meng et al., 2024) is a simpler and more efficient optimization approach by using a tailored reward formulation, eliminating the need for a reference model. Ethayarajh et al. (2024) introduce Kahneman-Tversky Optimization (KTO) based on direct utility maximization inspired by prospect theory. Group Relative Policy Optimization(Shao et al., 2024) removes the traditional critic model and introduces a group-based evaluation strategy. In this work, we extend GRPO to enhance the inherent long-term emotional support capabilities of LLMs.

3 Methodology

The RLSF-ESC framework, as illustrated in Figure 2, consists of three stages: multi-agent dialogue simulation, long-term reward model training, and reinforcement learning from simulated forward-looking feedback.

3.1 Problem Formulation

Given a conversation context between the user and the system, denoted as $c_{1:t-1} = \{u_1^{sys}, u_1^{usr}, \ldots, u_{t-1}^{sys}, u_{t-1}^{usr}\} \sim \mathcal{D}$, where $\mathcal{D}$ represents the dataset, each utterance $u_i = \{w_1^i, w_2^i, \ldots, w_n^i\}$ is a sequence of n words. The target is to generate the next system response u_t^{sys} that is coherent with the conversation context and effective in achieving a specific objective, such as reducing the user's emotional distress. Formally, the goal is to learn a policy π_θ that generates system responses to maximize the expected reward over the observed dialogue:

$$\pi^* = \arg\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot\,|\,x)}\big[R(x,y)\big], \quad (1)$$

where $x = c_{1:t-1}$, $y = u_t^{sys}$, and $R(\cdot)$ is a reward function that estimates the long-term reward of the system's responses.

3.2 Multi-Agent Dialogue Simulation

To estimate the long-term impact of a system response, we employ a multi-agent simulation framework that generates forward-looking dialogue trajectories. Given the conversation context $c_{1:t-1}$, we first sample a set of candidate next-turn responses $\{u_{t,j}^{sys}\}_{j=1}^m$ from the current policy of the LLM-based ESC model M_{sys}, such that $u_{t,j}^{sys} \sim P_{M_{sys}}(\cdot \mid p_{sys}, c_{1:t-1})$, where p_{sys} is the prompt. Our target is to construct a dataset consisting of pairs of sampled responses and their corresponding long-term rewards, which can then be used to train a reward model $R(\cdot)$ for long-term reward estimation.

To evaluate the long-term reward of each sampled response, we simulate the potential future trajectory of the dialogue. This simulation involves three LLM-based agents: a *system agent* M_{sys}, a *user simulator* U, and a *critic agent* M_{crt}. The system agent interacts with the user simulator by generating system responses based on the ongoing dialogue. The user simulator emulates a user seeking emotional support, responding based on the conversation history. Meanwhile, the critic agent evaluates the conversation's progress according to predefined criteria, such as goal achievement. This simulation-based framework allows for the estimation of long-term rewards by capturing how individual responses influence the overall success and quality of the conversation over time.

The forward-looking dialogue simulated for each sampled response is denoted as $C_{t,j}^f = (u_{t,j}^{usr}, \ldots, u_{t+k,j}^{sys}, u_{t+k,j}^{usr}, \ldots, u_{T,j}^{sys}, u_{T,j}^{usr})$, where $u_{t+k,j}^{usr}$ is generated by the user simulator, and $u_{t+k,j}^{sys}$ is generated by M_{sys}. At each step k, the critic model M_{crt} evaluates a reward based on the ongoing simulated conversation. The simulation proceeds iteratively until either the emotional support objective is achieved or a predefined maximum number of dialogue turns T is reached.

To compute the long-term reward $r_{t,j}^{sys}$ for the sampled system response $u_{t,j}^{sys}$, we consider both the reward at the terminal step T and the average number of turns required to achieve the dialogue goal. The resulting dataset $\mathcal{D}_r = \{(u_{t,j}^{sys}, r_{t,j}^{sys})\}_{t,j}$ is then used to train a reward model $R(\cdot)$, which predicts the long-term reward of system responses. Subsequently, this reward model guides the optimization of the system policy π_θ. The complete simulation procedure is detailed in Algorithm 1.

Algorithm 1 Multi-Agent Dialogue Simulation for Long-Term Reward Estimation

Require: Dialogue dataset $\mathcal{D} = \{c\}_{i=1}^Z$; ESC model M_{sys}; user simulator U; critic model M_{crt}; prompt p_{sys}, p_{usr}, p_{crt}; number of sampled responses m; maximum dialogue turns T

Ensure: Simulated dataset $\mathcal{D}_r = \left\{\left\{(u_{t,j}^{sys}, r_{t,j}^{sys})\right\}_{j=1}^m\right\}_{i=1}^Z$

for each conversation context $c_{1:t-1}$ from $\mathcal{D}$ **do**
  for each $j = 1$ to m **do**
    Sample $u_{t,j}^{sys} \sim P_{M_{sys}}(\cdot \mid p_{sys}, c_{1:t-1})$
    Generate $u_{t,j}^{usr} \leftarrow P_U(p_{usr}, [c_{1:t-1}, u_{t,j}^{sys}])$
    Initialize $C_{t,j}^f \leftarrow [u_{t,j}^{usr}]$
    for each step $k = 1$ to T **do**
      Generate $u_{t+k,j}^{sys} \leftarrow P_{M_{sys}}(p_{sys}, [c_{1:t-1}, C_{t,j}^f])$
      Append $u_{t+k,j}^{sys}$ to $C_{t,j}^f$
      Generate $u_{t+k,j}^{usr} \leftarrow P_U(p_{usr}, [c_{1:t-1}, C_{t,j}^f])$
      Append $u_{t+k,j}^{usr}$ to $C_{t,j}^f$
      Evaluate $r_{t+k,j} \leftarrow P_{M_{crt}}(p_{crt}, [c_{1:t-1}, C_{t,j}^f])$
      if $r_{t+k,j} >$ threshold or $k = T$ **then**
        Compute average turn: AvgT $= k + 1$
        Compute reward: $r_{t,j}^{\text{sys}} = \frac{r_{t+k,j} + \frac{1}{\text{AvgT}}}{2}$
        break
      end if
    end for
    Add $(u_{t,j}^{\text{sys}}, r_{t,j}^{\text{sys}})$ to $\mathcal{D}_r$
  end for
end for
return $\mathcal{D}_r$

3.3 RL from Forward-Looking Feedback

We optimize the LLM using reinforcement learning (RL) to enhance its ability to generate responses that account for their long-term impact on future conversations.

3.3.1 Reward Design

The reward function is central to the RL process, as it provides essential training signals for policy optimization. In RLSF-ESC, the reward function consists of two parts: *long-term reward* $R_{\text{long-term}}$, which estimates the influence of a response on the subsequent dialogue trajectory, and *format reward* R_{format}, which ensures that the generated response adheres to a predefined format.

Long-Term Reward. To predict $R_{\text{long-term}}$, we train a reward model using the dataset $\mathcal{D}_r$ from the dialogue simulation, consisting of (response, reward) pairs (u_i^{sys}, r_i^{sys}). Each pair is transformed into an instance (s_i, y_i) to train the reward model, where s_i represents the input sequence and $y_i \in \{0, 1\}$ is a binary label indicating whether the system response effectively addresses the user's emotional problem or not. Specifically, the label y_i is derived from the scalar reward r_i^{sys} using a predefined threshold δ. The label is set to one if r_i^{sys} is greater than δ; otherwise, it is set to zero. Each input s_i is constructed using a prompt template (see Table 4), incorporating the conversation context and the system response u_i^{sys}. This input is then fed into an LLM-based classifier, which consists of a frozen LLaMA model (Grattafiori et al., 2024) followed by a linear layer, to output logits $\mathbf{z}_i$:

$$\mathbf{z}_i = \text{Linear}(\text{Pool}(\text{LLaMA}(s_i))) \in \mathbb{R}^2, \quad (2)$$

where a pooling layer $\text{Pool}(\cdot)$ is applied to aggregate token-level hidden states into a single vector. The predicted label $\hat{y}_i$ is derived using a sigmoid activation function:

$$\hat{y}_i = \sigma(\mathbf{z}_i), \quad (3)$$

where $\sigma(\cdot)$ is the sigmoid function. For training, we minimize the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right], \quad (4)$$

where N is the total number of training samples. $R_{\text{long-term}}$ denotes the prediction made by the trained LLM classifier.

Format Reward. Inspired by the recent DeepSeek-R1 model (Guo et al., 2025), which demonstrates excellent performance across reasoning tasks through its rule-based reward design. We extend this approach in RLSF-ESC by introducing R_{format}, which evaluates whether the output from the LLM-based ESC model adheres to a predefined structural format. Specifically, we fine-tune the model using a prompt (see Appendix A.1) that instructs the model to enclose its thinking process and the response within special tokens `<think>` `</think>` and `<response>` `</response>`, respectively. This structured format facilitates the model outputting its reasoning process when generating a response. The format reward is formally defined as follows:

$$\mathcal{R}_{\text{format}} = \begin{cases} 1, & \text{if the output adheres to} \\ & \text{the defined format;} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

The final reward value, denoted as R, is computed as the weighted sum of the long-term reward $R_{\text{long-term}}$ and the format reward R_{format}:

$$R = R_{\text{long-term}} + \alpha R_{\text{format}}, \quad (6)$$

where α is a weight controlling the importance of R_{format} relative to $R_{\text{long-term}}$.

3.4 RL Tuning with GRPO

As discussed before, we employ GRPO (Shao et al., 2024) to optimize the LLM model using the designed reward function. Compared to RL algorithms used in RLHF, such as PPO (Schulman et al., 2017), GRPO is shown to be able to simplify and stabilize the training process. This is achieved by introducing advantage normalization across groups of candidate responses, thus eliminating the need for a critic model.

To be specific, in our RLSF-ESC framework, for each conversation context $c_{1:t-1}$, GRPO samples a group of candidate outputs $(o_1, o_2, \ldots, o_G)$ from the old policy π_{old}, which is a LLM-based ESC model. The corresponding rewards $(r_1, r_2, \ldots, r_G)$ are then obtained by computing the final reward of each response o_i. To determine the relative quality of each response, the normalized advantage can be defined as:

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)}. \quad (7)$$

The policy model π_θ is then optimized by maximiz-

ing the following objective function:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{c_{1:t-1}\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_\theta(\cdot|c_{1:t-1})} \frac{1}{G}\sum_{i=1}^{G} \Bigg[\min\Bigg(\frac{\pi_\theta(o_i|c_{1:t-1})}{\pi_{\text{old}}(o_i|c_{1:t-1})}A_i, \text{clip}\Big(\frac{\pi_\theta(o_i|c_{1:t-1})}{\pi_{\text{old}}(o_i|c_{1:t-1})}, 1-\epsilon, 1+\epsilon\Big)A_i\Bigg) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\Bigg], \quad (8)$$

where ϵ controls the clipping range to ensure stable update, and β penalizes the deviation from the reference policy π_{ref}.

4 Experiments

4.1 Experimental Setups

Datasets. We evaluate our method on two ESC datasets: **ESConv** (Liu et al., 2021) and **ExTES** (Zheng et al., 2023). The ESConv dataset contains 1,300 crowd-sourced dialogues with 8 emotional support strategies, along with user problem types, emotion types, and situation descriptions. We use the official split[2]. The ExTES dataset comprises 11,177 ChatGPT-generated dialogues, verified by human annotators, covering 16 emotional support strategies. We randomly split the dataset into train/dev/test with an 8:1:1 ratio.

Evaluation Metrics. Following prior work (Deng et al., 2024), we adopt both automatic and human evaluation metrics. For automatic evaluation, we use two metrics: *Success Rate* (SR) and *Average Turn* (AT). *SR* measures the proportion of dialogues in which the model successfully achieves the predefined goal within a maximum number of dialogue turns. *AT* calculates the average number of turns required to reach the goal, reflecting the model's efficiency in task completion. For human evaluation, the quality of generated responses is assessed from five perspectives: *Fluency*, *Empathy*, *Identification*, *Suggestion*, and *Overall*. Detailed evaluation criteria are presented in the Appendix B.

Implementation Details. We use two LLMs, LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct-1M (Yang et al., 2025), as backbone models for training our method. For the long-term reward model, we utilize LLaMA-3.2-1B (Grattafiori et al., 2024). For evaluation, we follow the protocol of previous work (Deng et al., 2024) and use GPT-4o (Achiam et al., 2023) to role-play both the user simulator and the critic agent. Prompts are presented in Appendix A.2 and A.3. More details are provided in Appendix D.

Baselines. We compare our method with a range of baselines: (1) Prompt-based response generation methods that directly prompt LLMs to generate emotionally supportive responses. This category includes **Standard Prompt** (Deng et al., 2024) and **ESCoT** (Zhang et al., 2024). (2) Supervised fine-tuning (**SFT**) (Zheng et al., 2023) that fine-tunes LLMs on ESC datasets to enhance the generation of supportive responses. (3) Policy planning with external modules, which integrates LLMs with external policy models to predict dialogue strategies and generate responses. This category includes **PPDP** (Deng et al., 2024), **EmoDynamiX** (Wan et al., 2025), and **Ask-an-Expert** (Zhang et al., 2023). (4) Prompt-based iterative planning and feedback, which utilizes multi-turn simulations to provide strategic foresight or planning signals that guide the response generation process. This category includes **ICL-AIF** (Fu et al., 2023) (dialogue-level feedback via simulation) and **GPD-Zero** (Yu et al., 2023) (planning via MCTS with LLM-based components). More details can be found in the Appendix E.

4.2 Experimental Results

4.2.1 Overall Performance

Comparison with baselines. It shows that our approach, RLSF-ESC, consistently outperforms all baselines across both datasets in terms of *success rate* (SR) and *average number of turns* (AT). These results demonstrate that optimizing LLMs by considering the long-term impact of each individual response on future dialogue trajectories can improve the performance of goal completion, thus facilitating higher-quality emotional support. Furthermore, our method exhibits superior performance in emotional support across different LLM backbones, indicating its generalizability and adaptability. Specifically, when being adapted to the Qwen2.5-7B-Instruct-1M model, our approach achieves better overall performance compared to its integration with the LLaMA3.1-8B-Instruct model.

Comparison with Larger-Scale LLMs. We compare the performance of RLSF-ESC with several representative large-scale LLMs on the ESConv dataset, including GPT-4o, LLaMA-3.1-70B-

[2] https://huggingface.co/datasets/thu-coai/esconv

Method	ESConv				ExTES			
	LLaMA-3.1		Qwen2.5		LLaMA-3.1		Qwen2.5	
	SR (%) ↑	AT ↓	SR (%) ↑	AT ↓	SR (%) ↑	AT ↓	SR (%) ↑	AT ↓
+ Standard Prompt (Deng et al., 2024)	6.15	7.94	16.9	7.78	10.3	7.88	10.3	7.88
+ ESCoT (Zhang et al., 2024)	8.46	7.95	27.7	7.59	11.9	7.87	22.2	7.64
+ SFT (Zheng et al., 2023)	6.14	7.82	8.46	7.78	13.5	7.87	11.9	7.86
+ EmoDynamiX (Wan et al., 2025)	10.8	7.85	19.2	7.67	-	-	-	-
+ PPDP (Deng et al., 2024)	20.8	7.81	26.9	7.55	-	-	-	-
+ Ask-an-Expert (Zhang et al., 2023)	16.9	7.80	21.5	7.65	17.5	7.78	26.2	7.60
+ GPD-Zero (Yu et al., 2023)	22.9	7.23	27.9	7.55	-	-	-	-
+ ICL-AIF (Fu et al., 2023)	23.4	7.18	28.5	7.43	28.4	7.48	30.3	7.34
+ RLSF-ESC	**35.5**	**6.83**	**41.5**	**7.18**	**30.2**	**7.30**	**32.5**	**7.29**

Table 1: Experimental results on two ESC datasets. The best results are **bolded** and the second-best results are underlined. SR denotes the Success Rate, and AT represents the Average Turns to reach the goal.

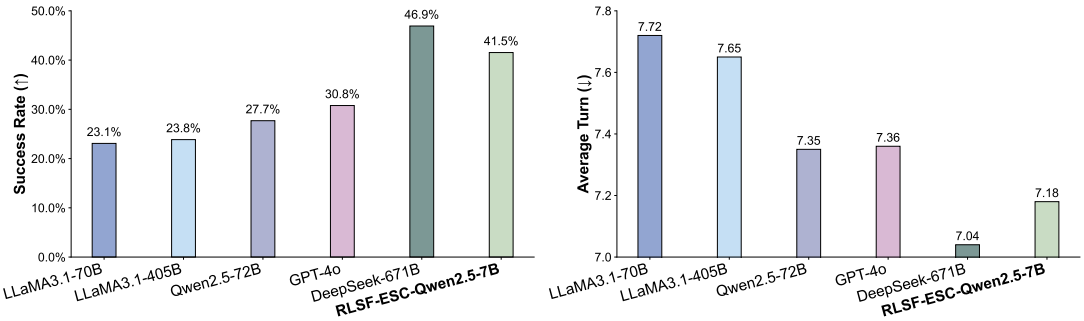

Figure 3: Performance comparison of RLSF-ESC and larger-scale LLMs: the left shows Success Rate, and the right shows the Average Turn performance.

Qwen2.5-7B-Instruct-1M		
RLSF-ESC vs. ICL-AIF	Win	Lose
Fluency	**66.0(%)**	34.0(%)
Empathy	**78.0(%)**	22.0(%)
Suggestion	**76.0(%)**	24.0(%)
Identification	**76.0(%)**	24.0(%)
Overall	**74.0(%)**	26.0(%)
LLaMA-3.1-8B-Instruct		
RLSF-ESC vs. ICL-AIF	Win	Lose
Fluency	**63.3(%)**	36.7(%)
Empathy	**77.6(%)**	22.4(%)
Suggestion	**59.2(%)**	40.8(%)
Identification	**69.4(%)**	30.6(%)
Overall	**69.4(%)**	30.6(%)

Table 2: Human evaluation results from comparing RLSF-ESC and ICL-AIF (the second-best model).

Instruct, LLaMA-3.1-405B-Instruct, Qwen2.5-72B-Instruct, and DeepSeek-R1-671B, in a zero-shot setting using the same prompt for the sake of consistency. Figure 3 summarizes the results. Our method, based on a compact 7B model, achieves a success rate of 41.5%, outperforming LLaMA-3.1-405B-Instruct (23.9%), Qwen2.5-72B-Instruct (27.7%), and GPT-4o (30.8%). In terms of the average number of turns (AT), our RLSF-ESC achieves 7.18, showing that it is more efficient than all models except DeepSeek-R1. All the results demonstrate RLSF-ESC's capability of generating effective emotional support conversations with fewer dialogue turns.

Human Evaluation. In line with previous work (Peng et al., 2022; Zhao et al., 2023b; Deng et al., 2024), we also conduct human evaluation on 100 dialogues randomly sampled from the ESConv dataset. Three annotators with a background in psychology were instructed to compare the responses generated by our method with those by the second-best model ICL-AIF (according to Table 1). As shown in Table 2, RLSF-ESC consistently outperforms ICL-AIF across five human evaluation metrics. This indicates that our approach can generate more effective emotionally supportive responses, especially in terms of empathy, fluency, problem identification, and offering suggestions to help users work through their challenges. Detailed qualitative case studies are presented in the Appendix C.

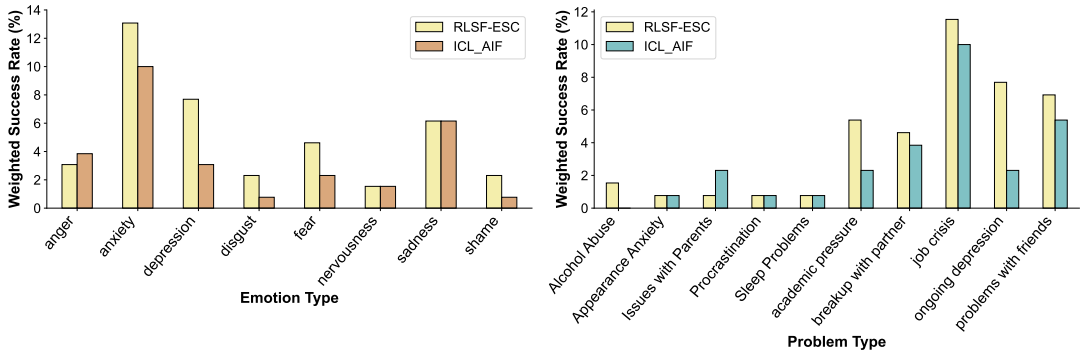

Figure 4: Performance comparison between RLSF-ESC and the second-best method ICL-AIF on the ESConv dataset across various user emotion types and problem types.

4.2.2 Ablation Study

To evaluate the key components of RLSF-ESC, we conducted an ablation study on the ESConv dataset with Qwen2.5-7B-Instruct-1M as the backbone. The results are shown in Table 3. To evaluate the effectiveness of GRPO, we remove it and use a vanilla prompt in a zero-shot setting (see Figure 5 in the Appendix A.1), which results in a significant performance drop. We then replace GRPO with DPO, using data constructed during the multi-agent dialogue simulation. More specifically, for a group of (response-reward) pairs, the highest reward response is "chosen," and the lowest is "rejected." DPO improves performance over the vanilla setting, proving the usefulness of multi-agent dialogue simulation, but it still underperforms compared to GRPO. Recognizing the importance of the reward model for GRPO, we further tested different designs. First, using an untrained LLM for long-term reward prediction, i.e., GRPO_*Random*, leads to a dramatic performance drop, implying the need for a suitable reward function in RL. Next, we train a LLM-ranking model with RRHF (Yuan et al., 2023) on the constructed dataset. This approach yields performance similar to that of the vanilla method, suggesting that a reward model based on classification might provide straightforward guidance than the more complex ranking method.

4.2.3 Further Analysis

In addition, Figure 4 shows the performance of RLSF-ESC with the second-best model ICL-AIF across different user emotion types and problem types. We report weighted success rates, reflecting each category's proportion in the dataset. Results show that RLSF-ESC outperforms ICL-AIF in most emotion types and problem types, particularly in cases involving *anxiety*, *depression*, *ongoing depression*, and *academic pressure*. These findings highlight the effectiveness of our model in various scenarios. For more experimental analysis, please refer to the Appendix F.

Qwen2.5-7B	SR (%) ↑	AT ↓
+ vanilla	26.5	7.58
+ DPO	31.5	7.26
+ GRPO_*Ranking*	26.9	7.56
+ GRPO_*Random*	18.5	7.66
+ GRPO_*Classification*	41.5	7.18

Table 3: Experimental results of the ablation study on the ESConv dataset. *Ranking*, *Random*, and *Classification* indicate using ranking, random, and classification reward models, respectively, for the long-term reward prediction in GRPO.

5 Conclusion

In this paper, we propose RLSF-ESC, a novel end-to-end approach that uses reinforcement learning for LLMs in ESC tasks by modeling the forward-looking impact of generated responses. Our method simulates future dialogue trajectories through multi-agent collaboration and constructs a dataset with response-reward pairs to facilitate training. We design an effective reward mechanism within GRPO to optimize LLMs for long-term emotional support. Extensive experiments indicate that RLSF-ESC improves both task completion rates and response quality in a variety of emotional support scenarios. Moreover, evaluations with different LLM backbones highlight its adaptability. Further analysis reveals that the simple reward formulation (i.e., the classification-based) can be more effective than complex ones (e.g., the ranking-based) for long-term reward modeling.

Limitations

While our method outperforms existing baselines on ESC tasks, there remains a gap between its current performance and the requirements for practical applications. For automatic evaluation, we utilize LLMs, For automatic evaluation, we utilize LLMs because they have demonstrated strong performance in terms of user simulation. However, potential evaluation biases within LLMs may affect the results' reliability. Although we conducted human evaluations, our study did not assess changes in end users' emotional intensity. In future work, we plan to conduct comprehensive user studies to evaluate the practical effectiveness of our proposed method in real-life situations.

Ethical Considerations

It is important to clarify that the term *emotional support conversation* in this paper is intended to simulate support through social interactions (e.g., from peers or friends) rather than professional counseling or psychological treatment. Although terms such as *"therapist"* and *"patient"* appear in the prompts used, they are solely for illustrative purposes and do not indicate the provision of clinical or therapeutic services. This study does not involve any form of professional counseling or mental health intervention. Additionally, all datasets utilized in this research were obtained from publicly available sources. These datasets do not contain any personally identifiable or sensitive information about users.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Brant R Burleson. 2003. Emotional support skills. In *Handbook of Communication and Social Interaction Skills*, pages 569–612. Routledge.

Brant R Burleson. 2013. Comforting messages: Features, functions, and outcomes. In *Strategic interpersonal communication*, pages 135–161. Routledge.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 951–966, Toronto, Canada. Association for Computational Linguistics.

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

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554, Toronto, Canada. Association for Computational Linguistics.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3014–3026, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yang Deng, Wenxuan Zhang, Wai Lam, See-Kiong Ng, and Tat-Seng Chua. 2024. Plug-and-play policy planner for large language model powered dialogue agents. In *The Twelfth International Conference on Learning Representations*.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *arXiv preprint arXiv:2305.10172*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. 2023. Improving language model negotiation with self-play and in-context learning from ai feedback. *Preprint*, arXiv:2305.10142.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4768–4791, Bangkok, Thailand. Association for Computational Linguistics.

Catherine A Heaney and Barbara A Israel. 2008. Social networks and social support. *Health Behavior and Health Education: Theory, Research, and Practice*, 4(1):189–210.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Mengzhao Jia, Qianglong Chen, Liqiang Jing, Dawei Fu, and Renyu Li. 2023. Knowledge-enhanced memory model for emotional support conversation. *arXiv preprint arXiv:2310.07700*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15232–15261, Bangkok, Thailand. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

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.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

A Paszke. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, 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. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319, Dublin, Ireland. Association for Computational Linguistics.

Chenwei Wan, Matthieu Labeau, and Chloé Clavel. 2025. Emodynamix: Emotional support dialogue strategy prediction by modelling mixed emotions and discourse dynamics. *Preprint*, arXiv:2408.08782.

Xiaoyu Wang, Yue Zhao, Qingqing Gu, Zhonglin Jiang, Xiaokai Chen, Yong Chen, and Luo Ji. 2025. Convert language model into a value-based strategic planner. *Preprint*, arXiv:2505.06987.

Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. 2025. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*.

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

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, and 9 others. 2025. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Xiao Yu, Maximillian Chen, and Zhou Yu. 2023. Prompt-based Monte-Carlo tree search for goal-oriented dialogue policy planning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7125, Singapore. Association for Computational Linguistics.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 10935–10950. Curran Associates, Inc.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2023. Ask an expert: Leveraging language models to improve strategic reasoning in goal-oriented dialogue models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6665–6694, Toronto, Canada. Association for Computational Linguistics.

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)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Weixiang Zhao, Xingyu Sui, Xinyang Han, Yang Deng, Yulin Hu, Jiahe Guo, Libo Qin, Qianyun Du, Shijin Wang, Yanyan Zhao, and 1 others. 2025. Chain of strategy optimization makes large language models better emotional supporter. *arXiv preprint arXiv:2503.05362*.

Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. 2023a. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023b. TransESC: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739, Toronto, Canada. Association for Computational Linguistics.

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

A Prompting Details

This section outlines the details of the prompts used in our implementation.

A.1 RL Tuning Prompt

We begin by describing the prompt used for reinforcement learning with GRPO. The prompt template for training a reward model for long-term reward estimation is represented in Table 4. Moreover, to guide the model in generating reasoning processes and responses in a predefined format, we utilize a system prompt, as illustrated in Table 5.

Table 4: Prompt template for training a reward model for long-term reward estimation.

Prompt Template

Instruction: You are a helpful assistant tasked with evaluating whether a patient's emotional problem has been effectively addressed following a conversation with a therapist.
Input Format:
Conversation: {conversation context}
Therapist's response: {system response}

A.2 User Simulator Prompt

We prompt the LLMs to role-play as users using the prompt proposed by (Deng et al., 2024). This prompt instructs the LLMs to simulate users, specifying the emotion type (which is set to None in the ExTES dataset) and the problem type. The prompt is detailed in Table 6.

A.3 Critic Agent Prompt

The prompts for the critic agent are designed to evaluate the degree of achievement of the goal within the dialogue. To facilitate a more nuanced and fine-grained assessment, we extend the previous four-level reward scale to seven levels. This fine granularity enables the critic model to better distinguish subtle differences in goal achievement, as detailed in Table 7.

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

Table 5: Prompt template for RL tuning.

System	Now enter the role-playing mode. In the following conversation, you will play as a patient in a counselling conversation with a therapist.
User	You are the patient who is looking for the help from the therapist, because you have the emotional issue about **[emotion_type]** regarding **[problem_type]**. Please reply with only one short and succinct sentence. Now tell me your issue.
Assistant	**[situation]**

Table 6: Prompt for user simulator.

B Human Evaluation Instructions

Human evaluation was conducted by three annotators with a background in psychology. They were tasked with comparing the responses generated by our method and the baselines based on four primary perspectives, as well as providing an overall assessment, as detailed below:

- **Fluency:** Which model produces the response that is more natural, well-structured, and logically coherent, with correct grammar and smooth sentence flow?
- **Empathy:** Which model responds with greater emotional understanding, showing warmth, compassion, or concern that aligns with the user's feelings?
- **Identification:** Which model explores the user's situation more effectively to identify the problem?
- **Suggestion:** Which model offers more relevant, practical, and emotionally sensitive suggestions that could help the user cope or take action?
- **Overall:** Overall, which model provides better emotional support, considering empathy, understanding, helpfulness, and communication quality?

C Case Study

In Table 9 and Table 10, we present two case studies comparing the responses generated by our RLSF-ESC model and the ICL-AIF method, focusing on addressing emotional problems related to *ongoing depression* and *alcohol abuse*, respectively. Compared to the baseline, our model demonstrates a greater capacity for empathy and emotional resonance, more accurately identifying the user's specific concerns and providing supportive, contextually appropriate suggestions. Furthermore, in Table 11 and Table 12, we give example dialogues generated by RLSF-ESC on two datasets involving interactions with a user simulator: one centered on *sleep problems* and the other on *coping with the illness of a family member*. These examples further illustrate our model's ability to engage in emotionally intelligent conversations across different scenarios.

D Implementation Details

Training Details. To construct the data for training the long-term reward model of RLSF-ESC, the Qwen-2.5-7B-Instruct-1M model is prompted to role-play as the ESC system M_{sys} in the multi-agent dialogue simulation process. For each conversation context, a set of $n = 4$ responses is sampled by the ESC system by setting: $\tau = 1.1$, $top\_k = 80$, and $top\_p = 1.0$. To reduce training

System	Given a conversation between a therapist and a patient, please assess whether the patient' emotional issue has been solved after the conversation.
User	The following is a conversation about **[emotion_type]** regarding **[problem_type]**: **[conversation]** **Question**: Has the patient's issue been solved? **Requirement**: You can only reply with one of the following seven descriptive levels without explanation: *Same: The patient's feelings remain unchanged.* *Slightly Better: The patient feels a slight, barely noticeable improvement.* *Moderately Better: The patient feels somewhat better, with a small but noticeable improvement.* *Significantly Better: The patient feels significantly better, indicating a clear improvement.* *Slightly Worse: The patient feels a slight increase in tension, barely noticeable decline.* *Moderately Worse: The patient feels somewhat more stressed or worried, with a small but noticeable decline.* *Significantly Worse: The patient feels significantly more distressed or upset, indicating a clear decline.*

Table 7: Prompt for the critic agent.

Training Phase	Hyper-parameter	Value
Reward Model	Batch Size	1
	Training Epochs	2
	Learning Rate	1e-4
	Max Sequence Length	2048
	Gradient Accumulation Steps	8
RL	Batch Size	4
	Training Epochs	2
	Learning Rate	1e-6
	Lora Rank	8
	Lora alpha	32
	Max Dialogue Turn	8
	Number of Generations	4
	Temperature	1.1
	Top p	1.0
	Top k	80

Table 8: Hyper-parameter settings.

costs, we replace GPT-4o-2024-11-20 (used during evaluation) with Qwen-2.5-72B-Instruct (Yang et al., 2024) to role-play both the user simulator U and the critic $\mathcal{M}_{\text{crt}}$ during dialogue simulations.

We fully fine-tune LLaMA-3.2-1B as an LLM-based classifier to predict the long-term reward using the constructed data. We use LLaMA-3.1-8B-Instruct and Qwen-2.5-7B-Instruct-1M as the backbone models of our method and optimize them using the GRPO reinforcement learning algorithm with LoRA (Hu et al., 2022). The reward weights of $R_{\text{long-term}}$ and R_{format} are 1 and 0.5, respectively. All experiments are implemented in PyTorch (Paszke, 2019) and conducted on 4 NVIDIA A100 GPUs with the DeepSpeed library (Rasley et al., 2020) using ZeRo-3 optimization. Detailed hyperparameter settings are provided in Table 8.

Evaluation Details. For evaluation, we follow the protocol of previous work (Deng et al., 2024) and use GPT-4o (Achiam et al., 2023) to role-play both the user simulator and the critic agent. The critic agent assesses whether the user's problem is resolved or not. We set temperature $\tau = 1.1$ and sample feedback $l = 10$ times. Feedback levels include: *"significantly worse," "moderately worse," "slightly worse," "same," "slightly better," "moderately better,"* and *"significantly better,"* mapped to -1.0, -0.5, -0.25, 0, 0.25, 0.5, and 1.0, respectively. We aggregate the sampled feedback to compute a scalar reward. The dialogue goal is considered complete if the reward exceeds 0.5.

E Baselines

We reproduce all the baseline methods using their official implementations and configurations. For consistency, all methods are adapted to two base LLMs: LLaMA-3.1-8B-Instruct and Qwen-2.5-7B-Instruct-1M.

- **Standard Prompt:** The LLM is prompted to generate emotionally supportive responses directly, without intermediate reasoning steps.
- **ESCoT:** A chain-of-thought prompting method that guides the LLM to reason through the user's emotional state, the triggering event, emotion appraisal, and an appropriate supportive strategy before composing a response.

- **SFT:** As many recent approaches improve emotional support capabilities through SFT, we fine-tune both base LLMs on existing ESC datasets.

- **PPDP:** A policy planning method using a RoBERTa-based model trained to predict the next dialogue strategy. The model is optimized via reinforcement learning using AI-generated feedback.

- **EmoDynamiX:** A RoBERTa-based policy model that incorporates a heterogeneous graph to model complex discourse dynamics between user emotions and system strategies.

- **Ask-an-Expert:** The LLM is instructed to act as a strategy expert, reasoning about the most appropriate response strategy given the conversation context.

- **ICL-AIF:** Two LLMs engage in a role-play emotional support conversation (e.g., seeker vs. supporter), while a third LLM provides feedback to improve each agent's planning and strategy. This setup encourages forward-looking reasoning and iterative strategy refinement.

- **GPD-Zero:** The LLM is used as components like the prior policy in a Monte Carlo Tree Search framework for planning goal-oriented dialogue strategies. In our implementation, we replace ChatGPT with our two base models to ensure consistent evaluation.

F Further Analyses

Training Rewards. Figure 5 presents the training rewards of two base LLMs during the reinforcement learning process of RLSF-ESC. For Qwen2.5-7B-Instruct-1M, the long-term reward increases sharply to 0.5 within the first 100 training steps and then continues to rise more gradually. The format reward initially decreases during the first 200 steps, followed by a sharp increase until approximately 1,000 steps, after which it continues to increase steadily. In contrast, for LLaMA3.1-8B-Instruct, the long-term reward declines during the first 200 training steps and subsequently increases gradually. The format reward shows a consistent upward trend throughout the training process.

Effect of the Evaluation Threshold. Figure 6 illustrates the success rates of RLSF-ESC and baseline models under two different evaluation thresholds on the ESConv dataset. Specifically, we compare two definitions of success: reward > 0.5 (used in our experiments) and reward ≥ 0.5. The results show that using the more lenient threshold of reward ≥ 0.5 leads to a significant increase in the reported success rate. This indicates that a substantial number of interactions result in a reward exactly equal to 0.5, suggesting users may feel moderately better in those cases. This observation highlights the sensitivity of evaluation metrics to threshold settings. Despite the change in absolute success rates, the relative ranking of RLSF-ESC and the baseline models remains approximately consistent across both criteria. We adopt the stricter threshold of reward > 0.5 in our main experiments to ensure that success reflects a clearer and more substantial improvement in the user's emotional state.

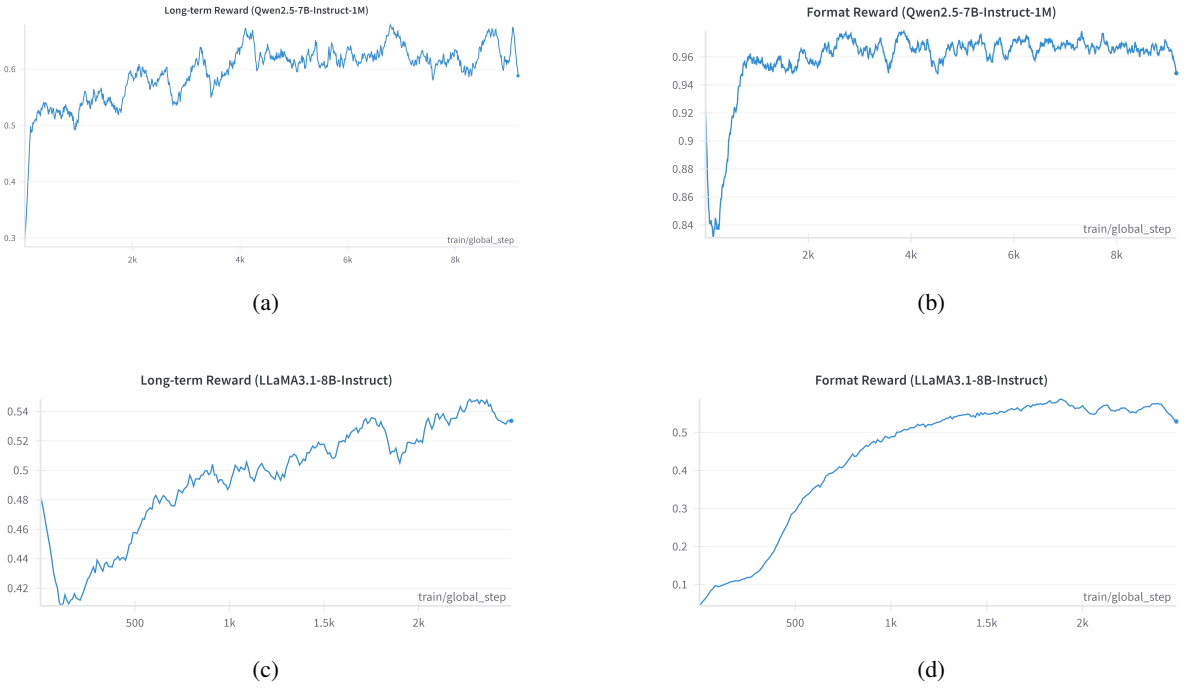

Figure 5: Training Rewards. (a) Long-term reward during training for Qwen2.5-7B-Instruct-1M. (b) Format reward during training for Qwen2.5-7B-Instruct-1M. (c) Long-term reward during training for LLaMA3.1-8B-Instruct. (d) Format reward during training for LLaMA3.1-8B-Instruct.

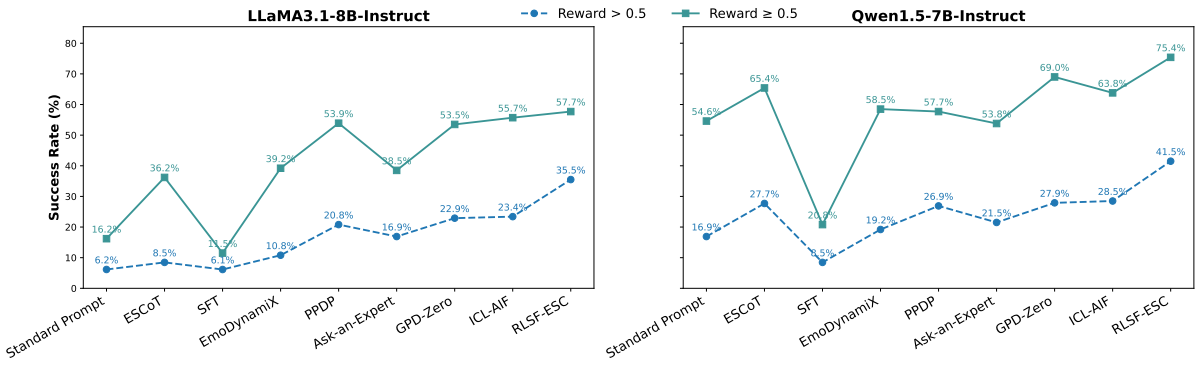

Figure 6: Model performance of RLSF-ESC and baselines on different evaluation criteria on the ESConv dataset.

Problem Type: Ongoing Depression
Conversation Context: **ASSISTANT:** *"Hello, what can I help you with this evening?"* **USER:** *"I am feeling very depressed lately. Like a constant pressure that I don't know exactly what it is."* **ASSISTANT:** *"Depression is extremely rough to deal with, I'm very sorry to hear you're going through that."* **USER:** *"I think the days getting shorter makes it worse."*
ICL-AIF: It sounds like the change in seasons is affecting your mood significantly. Many people find that shorter days and less sunlight can contribute to feelings of depression. How do you think these changes are specifically impacting your daily life or your thoughts?
RLSF-ESC: I hear you—depression can feel incredibly heavy, and it makes sense that shorter days might be amplifying those feelings. Many people find that changes in light and the seasons affect their energy levels and mood. Would you like to talk more about what this pressure feels like for you, or we could explore how the change in seasons might be playing a role?

Table 9: Case study of the generated supportive responses by our RLSF-ESC and the ICL-AIF method.

Problem Type: Alcohol Abuse
Conversation Context: **USER:** *Hello, how are you?* **ASSISTANT:** *"Hi there, how are you feeling?"* **USER:** *"I'm alright, honestly I'm feeling rather ashamed. The reason for that and the problem that I'm facing is that I've had a drinking problem for quite a few years now, and I'm only 24 years old. Right now in my life I'm ready to seek help, but I'm having a lot of anxiety about it because I'm very worried about how others will judge me. I'm worried about how the counselors will see me, how others at a facility might view me, and any one else that I will have to share my story with when getting treatment and how they might view me. I'm ashamed of myself, my life, and the choices that I've made that have gotten me here and I guess my problem is that I just don't know how to get over these extreme anxieties and worries so that I can just finally make the step that I need so desperately and get help."*
ICL-AIF: Thank you for sharing this with me. It takes a lot of strength to confront such difficult feelings and talk about your experiences. It's completely understandable to feel ashamed and worried about judgment; many people who struggle with addiction feel the same way. It's important to remember that seeking help is a brave step towards healing and recovery.
RLSF-ESC: Thank you for sharing this with me. First, let me say that it took a lot of courage for you to open up about what you're going through, especially surrounding something as sensitive as struggling with a drinking problem. The fact that you're already considering seeking help is a huge step forward, and it's an admirable act of strength and self-awareness.

Table 10: Case study of the generated supportive responses by our RLSF-ESC and the ICL-AIF method.

Emotion Type: Sadness
Problem Type: Sleep Problems
Situation: I've had trouble sleeping the last few nights, I feel really low self-esteem.

ASSISTANT: *"I'm sorry to hear you're having such a tough time. Lack of sleep and low self-esteem can be interconnected. Could you tell me more about what's been going on in your life recently that might be contributing to these feelings? "*
USER: *"Lately, I've been overwhelmed with work stress and personal expectations, and it's making it harder to wind down at night."*
ASSISTANT: *"It's understandable to feel overwhelmed when work and personal expectations pile up. This kind of stress can definitely make it difficult to relax and get a good night's sleep. Have you noticed any particular moments or tasks that are especially challenging for you at work? Also, how do you usually cope with stress? "*
USER: *"I struggle most with tight deadlines and the fear of disappointing others, and I usually just push through without addressing how I feel."*
ASSISTANT: *"It sounds like you're carrying a lot of pressure on yourself, especially around meeting deadlines and avoiding disappointment. One way to start addressing this might be by setting aside some time each day to reflect on how you're feeling. Even if it's just a few minutes, acknowledging your emotions can help you process them more effectively. What do you think about trying this approach?"*
USER: *"That sounds like a good idea; I think setting time for reflection and smaller steps could help me feel less trapped in the stress."*
ASSISTANT: *"I'm glad you find those suggestions helpful. Setting aside time for reflection and breaking tasks into smaller steps can indeed make a significant difference. Remember, it's okay to take things one step at a time. If you notice that certain strategies work better than others, don't hesitate to adjust your approach. How do you feel about starting to incorporate these changes into your routine? "*
USER: *"I feel hopeful about trying these changes and curious to see how they might help me feel more in control."*
ASSISTANT: *"I'm really glad to hear you feel hopeful and curious about trying these changes. It's natural to face some challenges along the way, but remember that each small step you take is a step towards greater control over your well-being. Let's keep an open dialogue about how things are going, so we can support you in finding the best path forward. How does that sound? "*
USER: *"That sounds reassuring, and I appreciate having a space to openly share and adjust as I go."*

Reward: 0.8 $\rightarrow$ Goal completed !!

Table 11: An example conversation with our RLSF-ESC on the ESConv dataset.

Problem Type: Coping with the Illness of A Family Member
Situation: My sister was recently hospitalized after a serious accident, and I'm feeling overwhelmed by the uncertainty and fear for her recovery.

ASSISTANT: *"It's completely understandable to feel overwhelmed and scared when someone close to you is going through such a difficult time. Right now, it might be helpful to focus on what actions you can take to support your sister and yourself. For example, staying informed about her condition and communicating regularly with the medical team."*
USER: *"I think I could start by talking to the doctors more often to understand her condition better and reaching out to a close friend for emotional support."*
ASSISTANT: *"That's a great start! Talking to the doctors more frequently can provide you with clearer insights into your sister's condition. Additionally, sharing your concerns and experiences with someone you trust can make a significant difference in how you cope with this situation. Remember, it's okay to lean on others during tough times. How do you feel about taking these steps?"*
USER: *"I feel a bit nervous about opening up to others, but I know I need to try, and I'm worried about staying strong for my sister while managing my own emotions."*
ASSISTANT: *"It's completely normal to feel nervous about opening up, especially when you're already dealing with so much. However, sharing your feelings with someone you trust can actually help you stay strong in the long run. As for being there for your sister, it's important to remember that you can't pour from an empty cup. Taking care of your own emotional well-being is not only necessary but also essential for supporting her. Maybe you could set boundaries or let her know that you're doing your best to handle everything."*
USER: *" I think I could set small moments aside each day to focus on myself, like taking a walk or journaling, while letting my sister know I'm always there for her when she needs me."*
ASSISTANT: *"That sounds like a very practical approach. Setting aside specific times for self-care, such as walking or journaling, can help you recharge and maintain your emotional strength. And letting your sister know that you're always there for her when she needs you can also help her feel supported and less alone. It's important to strike a balance between caring for yourself and being there for her. If you find that you're struggling to maintain this balance, don't hesitate to reach out for additional support. You're doing great by thinking ahead and planning ways to take care of yourself and your sister."*
USER: *"Thank you, I'll try to take it one step at a time and remind myself that it's okay to ask for help when I need it."*

Reward: 0.55 $\rightarrow$ Goal completed !!

Table 12: An example conversation with our RLSF-ESC on the ExTES dataset.