DynaESC: Learning Long-Term Emotional Support Strategies via Dynamic Multi-Turn Reinforcement Learning

Anonymous ACL submission

Abstract

Current Emotional Support Conversation systems primarily focus on the quality of individual responses, while overlooking global strategy planning at the dialogue level. To enhance the system's capability in stage transition perception and strategy selection, we propose **DynaESC**, a **Dyna**mic multi-turn reinforcement learning framework for **E**motional **S**upport **C**onversation, designed to optimize long-term dialogue management through simulated interactive training. Specifically, our framework introduces two core modules: (1) a User Simulator, which leverages Large Language Models to act as seekers based on predefined user personas, providing dynamic interactions and real-time feedback to the system, thereby enabling a high-fidelity, closed-loop interactive environment; (2) a Multi-dimensional Reward Function, which evaluates responses by balancing immediate quality with holistic planning, thereby simultaneously refining both response generation and strategic selection. Furthermore, we introduce a novel LLM-based evaluation metric that assesses the system's performance based on complete multi-turn interactions rather than isolated turns. Experimental results demonstrate that DynaESC achieves an approximately 42% improvement in overall score over its pre-trained counterpart and consistently outperforms representative baselines, showcasing its superior efficacy in providing emotional support.

1 Introduction

Mental health constitutes a fundamental foundation for individual well-being and social welfare, yet it has long been underestimated and underfunded (Foster et al., 2024; Organization, 2025). Due to the shortage of mental health professionals available to address the issue, automated Emotional Support Conversation (ESC) systems have demonstrated significant potential for mitigation. Emotional Support Conversation aims to understand a user's emotional state, identify their underlying needs, and employ targeted strategies to enhance their psychological well-being (Burleson, 2003; Liu et al., 2021). With the rise of Large Language Models (LLMs), their flexible conversational abilities have demonstrated great potential in building automated emotional support conversation systems.

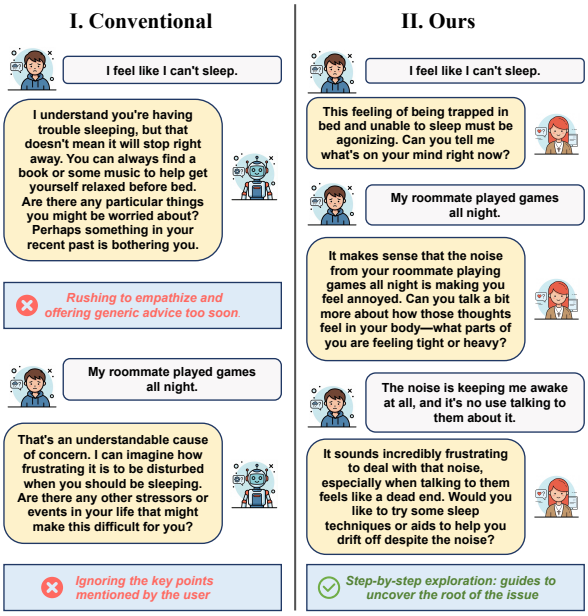

Figure 1: Comparison of dialogue quality. Conventional methods often fail to adequately explore the underlying causes of user distress, leading to superficial interventions before the core problem is fully identified.

To realize this potential, numerous studies have constructed specialized ESC datasets and optimized LLMs through Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL) to enhance their capability in emotional support conversations (Zheng et al., 2023a; Zhang et al., 2024). **However, these approaches suffer from short-sighted limitations, prioritizing the immediate quality of individual turns over the long-term strategic planning of the dialogue**. As illustrated in Figure 1, when user distress remains ill-defined,

conventional methods tend to prematurely offer empathy and generic advice, failing to strategically orchestrate the overall support process.

To address the above challenges, we propose **DynaESC**, a systematic framework that **integrates an LLM-based dynamic user simulator, interpretable reasoning-planning processes, and a multi-dimensional reward function into a dynamic multi-turn reinforcement learning paradigm**. First, by exploiting the superior role-playing proficiency of LLMs (Wu et al., 2024; Wang et al., 2024), We design a user simulator to construct a dynamic interactive environment, providing a training setting for optimizing the overall planning capability of the emotional support (ES) agent. Second, we explicitly model the reasoning and planning processes of the ES agent, thereby enhancing control over the overall dialogue flow while improving interpretability. Third, a multi-dimensional reward function is designed to guide the model from turn-level quality to dialogue-level strategy selection. Finally, this framework provides a novel methodological foundation for strategy-oriented learning in emotional support conversation systems.

To more effectively evaluate the performance of emotional support agents, we propose a novel set of evaluation metrics specifically tailored for emotional support conversations. Existing evaluation methods primarily focus on the similarity between model responses and reference replies. However, in practice, there are often multiple valid responses, and such evaluation may hinder optimization toward desirable behaviors. Moreover, turn-level evaluation in isolation overlooks the model's planning capability and its long-term impact on the dialogue. In contrast, our proposed metrics are based on complete multi-turn dialogue simulations, enabling assessment from multiple perspectives, including turn-level response appropriateness, dialogue-level coherence and planning, as well as the evolution of the user's emotional state. This holistic evaluation framework provides a more faithful measurement of an emotional support agent's real-world effectiveness.

Our contributions can be summarized as follows:

- We develop DynaESC, a dynamic multi-turn reinforcement learning framework by constructing a user simulator, leveraging a multi-dimensional reward function to guide the model toward optimizing long-term emotional support objectives.

- We propose a suite of novel simulation-based evaluation metrics specifically designed for emotional support conversation. These metrics move beyond traditional reliance on text similarity by measuring the model's actual effectiveness through complete multi-turn interaction simulations.

- Experimental results demonstrate that DynaESC achieves an approximately 42% improvement in overall score compared to the pre-training baseline. This significantly outperforms representative baselines and validates our model's superior proficiency in both strategic planning and therapeutic support.

2 Related Work

2.1 Emotional Support Conversation

Liu et al. (2021) established the field by defining eight psychological support strategies and releasing the ESConv dataset. Building on this, Tu et al. (2022) integrated COMET (Bosselut et al., 2019) into a psychological state–enhanced encoder and hybrid policy module to refine emotional understanding and strategy flexibility. Peng et al. (2023) introduced a dual-layer feedback selector and a dual-control reader, ensuring strategic coherence through bidirectional control and multi-level feedback. Zhou et al. (2023) employed a multi-task mixture-of-experts (MoE) module with RL-optimized expert selection to actively guide user emotions, enhancing proactive emotional support.

With the rise of LLMs, mainstream ESC research has pivoted away from complex modular designs toward constructing high-quality datasets for fine-tuning LLMs. To address the limitations of ESConv, which is characterized by high construction costs, a relatively small scale, and a narrow range of covered topics, researchers have constructed augmented datasets like AUGESC (Zheng et al., 2023a) and ExTES (Zheng et al., 2023b) via LLM generation. Others focused on enhancing interpretability by constructing fine-tuning data with Chain-of-Thought (CoT) reasoning (Zhang et al., 2024) or improving realism by generating simulated dialogues through role-playing (Ye et al., 2025). However, Due to SFT's heavy reliance on high-quality dataset construction and its poor performance in generalization (Chu et al., 2025), re-

search has begun shifting toward the use of reinforcement learning.

2.2 Reinforcement Learning

RL enables agents to optimize policies through continuous environmental interaction. Algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Group Relative Policy Optimization (GRPO) (Guo et al., 2025) have achieved significant success in text generation and reasoning tasks (Liu et al., 2025; Jin et al., 2025).

Constructing stable rewards for open-ended ESC is challenging. Recent works have designed specific reward signals: Kim et al. (2025) employed value detectors to reinforce positive guidance, while Wang et al. (2025) leveraged a simulated user (Zhang et al., 2025) to provide verifiable emotional scores. However, these approaches primarily focus on single-turn quality rather than holistic dialogue evolution. To bridge this gap, we integrate stage inference and strategy selection into the reward mechanism, enabling multi-dimensional optimization for long-term strategic planning.

3 Method

3.1 DynaESC Overview

To build an ESC system capable of accurately identifying dialogue stages, selecting appropriate strategies, and generating genuinely empathetic responses, we conduct a comprehensive system-level optimization. By introducing a multi-agent environment, a dynamic RL framework, and a multi-granularity reward mechanism, we aim to enhance the model's abilities in emotional understanding, strategy planning, and response generation quality.

As illustrated in Figure 2, We construct an RL framework based on dynamic interaction. The framework consists of three core modules: the ES agent $\mathcal{M}_{\text{ES}}$ acting as the supporter, the user simulator $\mathcal{M}_{\text{US}}$ serving as the environment, and the reward module $\mathcal{M}_{\text{R}}$ responsible for evaluating the quality of interactions. The interaction process is conducted in an episodic setting, where the agent learns to optimize its behavior through multi-turn dialogues with the simulator.

At time step t, the system is in state s_t, which consists of the current dialogue history H_t and the user's emotional state e_t. The interaction follows a closed-loop process described as follows:

- **Action Generation:** $\mathcal{M}_{\text{ES}}$ generates an action a_t based on the current state s_t according to its policy π_θ. The action a_t is defined as a tuple (p_t, y_t), where p_t denotes the reasoning and planning process, and y_t denotes the generated response text:

$$a_t = (p_t, y_t) \sim \pi_\theta(\cdot|s_t) \tag{1}$$

- **Environment Feedback and State Transition:** $\mathcal{M}_{\text{US}}$ receives the agent's action a_t and simulates the cognitive and emotional responses of a real user. Specifically, $\mathcal{M}_{\text{US}}$ updates its internal emotional state, determines whether the user's help-seeking intent has been satisfied, yielding e_{t+1}, and generates the next user utterance u_{t+1}. This process completes the transition from state s_t to s_{t+1}:

$$s_{t+1} \leftarrow \mathcal{M}_{\text{US}}(s_t, a_t) \tag{2}$$

  where $s_{t+1} = \{H_t \cup \{y_t, u_{t+1}\}, e_{t+1}\}$.

- **Reward Evaluation:** To guide $\mathcal{M}_{\text{ES}}$ toward convergence to an optimal policy, a reward module $\mathcal{M}_{\text{R}}$ is introduced to evaluate the interaction tuple (H_t, a_t, u_{t+1}). By synthesizing the dialogue context and user feedback, $\mathcal{M}_{\text{R}}$ outputs a scalar reward signal r_t:

$$r_t = \mathcal{M}_{\text{R}}(H_t, a_t, u_{t+1}) \tag{3}$$

Through the above interactions, a sequence of dialogue trajectories is generated:

$$\tau = \{(u_0, a_0, r_0), \ldots, (u_T, a_T, r_T)\} \tag{4}$$

In each interaction round, the emotional support agent $\mathcal{M}_{\text{ES}}$ is optimized via a multi-dimensional reward mechanism. The objective function for parameters θ is defined as maximizing the expected reward across all dialogue turns:

$$J(\theta) = \mathbb{E}(u_t, a_t) \sim \pi_\theta \left[r_t\right] \tag{5}$$

To maximize this objective, we employ policy optimization methods to iteratively update the model parameters θ.

3.2 User Simulator

To construct a real-time, multi-turn RL environment, we designed a user simulator based on LLMs. Drawing on the methodology of Zhang et al. (2025), we first constructed a set of help-seeking topics and

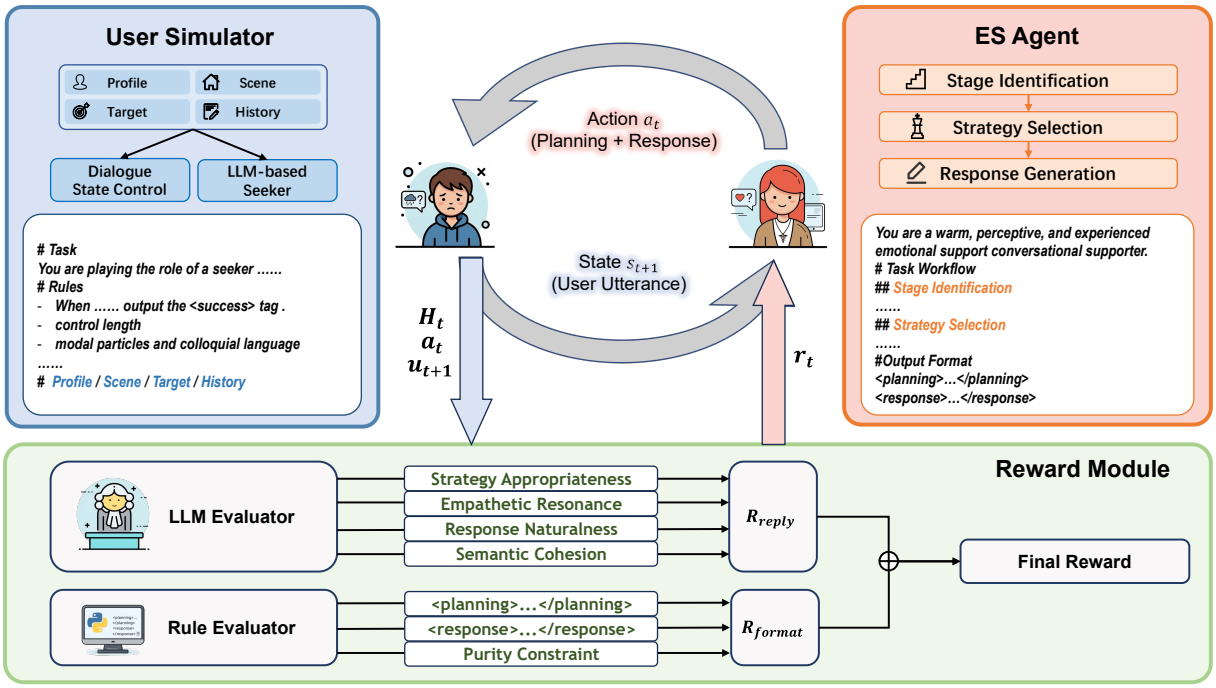

Figure 2: The DynaESC framework. It features a User Simulator generating context-aware seeker utterances and an ES Agent performing stage-aware strategy planning. A Reward Module provides multi-dimensional feedback to optimize long-term support performance.

utilized DeepSeek-V3.1 (Liu et al., 2024) to generate multi-dimensional user personas, effectively enhancing the diversity and coverage of the user simulator. Building upon this foundation, the system transforms an LLM into a subjective seeker by incorporating predefined personas, latent backgrounds, and specific help-seeking goals. Furthermore, to prevent redundant interactions once support objectives have been achieved, the system incorporates a terminal variable, e_t, based on a convergence recognition mechanism. When the ES agent's intervention strategy successfully addresses the core issue and fulfills the seeker's needs, the system generates a <success> token and sets the state variable e_t to 1. This signal serves as the logical termination point for the emotional support conversation, effectively ensuring both interaction quality and conversational efficiency.

To ensure both simulation realism and training effectiveness, the user simulator is grounded in two core operational principles. First, it maintains information asymmetry between the two agents; while the simulator possesses the full help-seeking background and persona, the emotional support agent must actively elicit information through dialogue. This design compels the simulator to disclose relevant details incrementally as the interaction unfolds, effectively mimicking the evolving nature of real-world emotional support. Second, the system employs a constrained language style by enforcing colloquial expressions and strictly limiting response length. By avoiding the overly formal or "machine-like" verbosity typical of conventional models, this constraint ensures that the simulated interactions remain authentic and representative of human conversation.

3.3 Emotional Support Agent

Following the ESC framework proposed by Liu et al. (2021), we dynamically partition the ES process into three stages: exploration, comforting, and action. To accommodate the generative characteristics of LLMs, we perform targeted pruning and optimization of the original eight dialogue strategies, resulting in a more robust strategy set. This design aims to reduce the risk of hallucinations in complex emotional interactions and to ensure the stability and reliability of supportive behaviors. Detailed information can be found in the Appendix A

At the implementation level, we introduce an explicit planning–execution mechanism. Unlike conventional end-to-end direct generation, our supporter agent is designed as a warm and reflective observer endowed with global planning capabilities. Through carefully designed system prompts, the agent is explicitly required to perform a CoT

reasoning process before producing the final response. Specifically, the agent first integrates the full conversational context to dynamically identify the current stage of the dialogue; based on this stage assessment, it selects appropriate strategies from a predefined strategy library that best matches the user's current psychological state; finally, with the stage and strategy explicitly determined, the agent constructs its response by incorporating the relevant situational details.

This process is physically separated through structured `<planning>` and `<response>` tags. Such a "think-before-speaking" design not only enhances the interpretability of the agent's responses, but also enables the agent to provide more targeted responses grounded in deep insights into dialogue stages and strategic evaluations.

3.4 Reward Module

As the core guiding signal of an RL system, the reward module directly determines the convergence properties of the learned policy and the final task performance. However, due to the highly open-ended and non-deterministic nature of ESC tasks, it is infeasible to rely on standard answers or rigid rules for reward modeling as in mathematical reasoning tasks. To address this challenge, we design a multi-granularity composite reward mechanism under the guidance of experts in psychology. This mechanism not only aims to steadily improve the quality of single-turn responses, but also incorporates global dialogue control into the optimization process, thereby achieving a deeper alignment with task objectives.

Reply Reward. To evaluate the quality of the model's response at each dialogue turn, we design a multi-dimensional assessment framework focusing on four key dimensions: **Dialogue Strategy**, which evaluates the appropriateness of stage identification and support strategy selection; **Empathetic Resonance**, measuring the sincerity and effectiveness of emotional alignment; **Naturalness**, assessing the degree to which responses feel authentic and human-like; and **Semantic Cohesion**, which examines how accurately the supporter captures the core information and emotional focus of the conversation. Detailed definitions and scoring criteria for each dimension are provided in Appendix B.

Through this multi-dimensional design, we aim to enhance both the immediate response quality and the overall strategic planning capabilities of the support agent. Specifically, dialogue strategy appropriateness is designed to train the model's global control over conversation trajectories; emotional resonance focuses on optimizing the quality of empathetic support; response naturalness targets the fluency of linguistic expression; and finally, semantic continuity and depth strengthens the model's ability to leverage previously disclosed user information, ensuring a coherent and logically consistent long-term support plan throughout the interaction.

We employ LLMs as automated evaluators to score four independent dimensions based on conversation context and user feedback. We utilize a tri-level scoring system $\{-0.5, 0, 1\}$, where the final per-turn reward is a normalized linear average of these dimensions. Detailed criteria and the rationale for this reward design are provided in Appendix C.

Format Reward. To ensure structural stability and consistency in the supporter agent's output, we design a format reward mechanism. The maximum score of this reward is 1.0, and the evaluation is strictly based on three independent binary criteria, each weighted by $\frac{1}{3}$. To achieve a maximum format score of 1.0, the model output must strictly adhere to the following three conditions: Requirement $\mathbb{I}_{\text{plan}}$: The output must include a `<planning>...</planning>` block, which shall not be empty; Requirement $\mathbb{I}_{\text{res}}$: The output must include a `<response>...</response>` block, which shall not be empty; Requirement $\mathbb{I}_{\text{pur}}$: Purity Constraint. Beyond the two specified structural tags, the output must contain no extraneous characters, metadata, or redundant text.

$$R_{\text{format}} = \frac{1}{3}(\mathbb{I}_{\text{plan}} + \mathbb{I}_{\text{res}} + \mathbb{I}_{\text{pur}}) \tag{6}$$

The final reward signal used to guide the policy optimization is a weighted combination of the aforementioned components. The total reward R for each turn is formulated as:

$$R = R_{\text{reply}} + \lambda \cdot R_{\text{format}} \tag{7}$$

where λ is a hyper-parameter introduced to modulate the contribution of the formatting reward, effectively balancing the trade-off between semantic content quality and structural adherence.

3.5 Policy Optimization

During the policy optimization stage, we ultimately adopt PPO as the core algorithm for training the ES

model. The primary motivation for choosing PPO lies in its excellent training stability and high sample efficiency. When applied to large-scale neural network policies, PPO introduces a clipped surrogate objective that strictly constrains the step size of policy updates, keeping them within a safe region. This mechanism effectively prevents performance collapse caused by overly aggressive parameter updates, thereby ensuring stable and monotonic convergence in complex, high-dimensional parameter spaces (Schulman et al., 2017).

4 Experiments

4.1 Baselines

Qwen2.5 (Qwen et al., 2025): An open-source LLM. In our experiments, we employ `Qwen2.5-3B-Instruct` as the backbone model to evaluate the inherent capabilities of general-purpose LLMs in emotional support.

AUGESC (Zheng et al., 2023a): A model fine-tuned on the AUGESC dataset based on `Qwen2.5-3B-Instruct`. This dataset is constructed by leveraging LLMs to complete full dialogues from initial utterances, providing a large-scale augmented corpus for ESC.

ESCoT (Zhang et al., 2024): A model fine-tuned on the ESD-CoT dataset based on `Qwen2.5-3B-Instruct`. It incorporates emotional-focusing and strategy-driven CoT reasoning to enhance the interpretability and effectiveness of support responses.

RLVER (Wang et al., 2025): A reinforcement learning–based `Qwen2.5-7B-Instruct` model trained with user emotion values as rewards, where a "Think-Then-Say" template is applied to elicit an explicit reasoning step prior to response generation.

DynaESC: Our proposed model, developed by training `Qwen2.5-3B-Instruct` with the RL approach described in Section 3.

4.2 Evaluation Metrics

Traditional automated evaluation metrics, such as BLEU-n (Papineni et al., 2002), ROUGE-L (Lin, 2004), and Distinct-n (Li et al., 2016) primarily hinge on lexical overlap, which poses significant limitations in the context of ESC. First, these metrics tend to evaluate the quality of a single-turn response in isolation, thereby overlooking the contextual coherence of the dialogue. Second, given the open-ended and non-deterministic nature of emotional support in dynamic interactions, there is often a lack of fixed reference responses. This makes it difficult for literal-matching-based metrics to objectively measure the effectiveness of a response. To verify the limitations of using reference-based ground truth (GT) as the evaluation standard, we performed a preference analysis on large-scale general LLMs. We compared the outputs of our model with the GT from the ESD-CoT dataset (Zhang et al., 2024). See Appendix D for details.

To address these challenges, we propose a comprehensive evaluation framework specifically tailored for ESC tasks under the guidance of experts in psychology. Unlike conventional methods, our framework encompasses the entire lifecycle of the conversation by modeling the supporter's strategies alongside the seeker's real-time feedback. Meanwhile, this holistic assessment of the conversation naturally distinguishes itself from turn-level reward metrics, thereby enhancing the independence and impartiality of the evaluation system. This approach allows for an in-depth assessment of support quality across five core dimensions: **Ethics** ensures adherence to value neutrality while mitigating cultural bias and imperative constraints. **Strategy** assesses the model's macro-control over logical evolution and conversational pacing. **Empathy** evaluates the precision in identifying core user needs and providing deep emotional resonance. **Naturalness** examines the fluency of responses and their alignment with human linguistic habits. Finally, **Efficacy** serves as a result-oriented metric to measure the improvement in user emotional well-being following the interaction. The detailed definitions, theoretical rationales, and the corresponding scoring rubrics for each dimension are provided in Appendix E.

These five metrics constitute a multidimensional evaluation of the supporter model, spanning global strategic planning, micro-level response quality, and terminal intervention efficacy. By systematically integrating the full-cycle dialogue context with user-side feedback signals, this design scrutinizes not only the content of the model's responses but also the manner of delivery and the subsequent psychological impact. Consequently, this framework provides a scientific, rigorous, and highly referential quantitative paradigm for the assessment of emotional support agents. The effectiveness of this evaluation system was rigorously validated for reliability; the specific experimental procedures are

Method	Ethics ↑	Strategy ↑	Empathy ↑	Naturalness ↑	Efficacy ↑	Overall ↑
Qwen2.5-3B	$3.70_{\pm 0.09}$	$3.43_{\pm 0.09}$	$3.35_{\pm 0.10}$	$2.30_{\pm 0.17}$	$3.14_{\pm 0.12}$	$15.92_{\pm 0.26}$
AUGESC	$3.20_{\pm 0.16}$	$3.09_{\pm 0.08}$	$2.91_{\pm 0.03}$	$2.86_{\pm 0.01}$	$2.87_{\pm 0.11}$	$14.93_{\pm 0.21}$
ESCoT	$3.47_{\pm 0.06}$	$3.16_{\pm 0.02}$	$3.08_{\pm 0.02}$	$2.96_{\pm 0.05}$	$2.98_{\pm 0.02}$	$15.65_{\pm 0.09}$
RLVER (7B)	$\underline{4.85}_{\pm 0.06}$	$\underline{4.57}_{\pm 0.07}$	$\underline{4.42}_{\pm 0.05}$	$\underline{3.70}_{\pm 0.04}$	$\underline{4.16}_{\pm 0.04}$	$\underline{21.70}_{\pm 0.12}$
DynaESC	$\mathbf{4.95}_{\pm 0.02}$	$\mathbf{4.78}_{\pm 0.02}$	$\mathbf{4.69}_{\pm 0.04}$	$\mathbf{4.09}_{\pm 0.10}$	$\mathbf{4.17}_{\pm 0.10}$	$\mathbf{22.68}_{\pm 0.15}$

Table 1: Main evaluation results on emotional support conversation. The performance of our model and various baselines is evaluated across the five dimensions defined in the Section 4.2. Scores range from 1 to 5, with Overall representing the summation of all individual scores. Bold indicates the best performance, and underlined values indicate the second-best. ↑ indicates higher is better. All test results were averaged across three experiments.

detailed in Appendix F.

4.3 Implementation Details

Our experiments are conducted using the Verl (Sheng et al., 2024) framework to establish the PPO training environment. Both the Actor and Critic models are initialized from `Qwen2.5-3B-Instruct`. For further technical details, including the complete set of hyperparameters and hardware configurations, please refer to Appendix G.

4.4 Main Result

As shown in Table 1, our model demonstrates superior performance in ESC tasks. Specifically, it not only achieves the highest overall score—significantly outperforming both the base models and specialized models fine-tuned via SFT, but also surpasses RLVER (7B), another reinforcement learning-based approach with more than double the parameters. These results validate the effectiveness of the RL framework described in Section 3.

Multi-dimensional Performance Analysis Through a detailed comparison of various metrics, the significant advantages of our method can be observed in the following aspects: in the Ethics and Strategy dimensions, our model achieved exceptionally high scores of 4.95 and 4.78 respectively, indicating that the RL training process successfully internalized complex psychological counseling principles into the model's generation strategies and global dialogue planning capabilities, enabling it to invoke support techniques more accurately and adhere more strictly to ethical boundaries than SFT models; in the Naturalness dimension, the improvement of our model is particularly significant 4.09, far surpassing the highest baseline of 3.70, which suggests that our reward function effectively suppresses mechanical repetition or abrupt transitions common in dialogues, making the responses more closely resemble those of a real human supporter; it is worth mentioning that although RLVER is extremely close in the Efficacy metric evaluated based on user feedback due to its use of user emotion values as reward indicators, it remains inferior to our model in comprehensive dimensions, further proving the comprehensiveness and robustness of our multi-dimensional balanced optimization strategy in emotional support tasks.

The Limitations of SFT in ESC The experimental results further reveal the inherent limitations of traditional training paradigms. Data indicates that models based on SFT exhibit a pronounced performance bottleneck, with scores in specific dimensions even falling below those of the original base models. We attribute this phenomenon to pattern collapse: SFT on static datasets tends to induce distributional shifts, causing the model to overfit to specific templates or high-frequency vocabulary. This convergence significantly impairs the model's flexibility and emergent capabilities when handling dynamic, unstructured emotional dialogues. In sharp contrast, RL methods extend the depth of exploration within the action space, shifting the optimization objective from token-level predictive accuracy to macro-interaction quality, achieving more natural empathetic expression while strictly adhering to ethical boundaries.

4.5 Ablation Study

To verify the effectiveness of the individual components within the RL training framework, we conducted ablation studies to quantify the impact of planning and multi-turn training paradigms on the final performance.

Planning To quantify the contribution of reasoning planning to model performance, we compared the model's results with and without explicit planning steps. As illustrated in Table 2, the intro-

Base Model	Planning	RL	Average ↑
✓			3.18
✓	✓		3.62 (+0.44)
✓		✓	4.32 (+1.14)
✓	✓	✓	**4.54** (+1.36)

Table 2: Ablation study on planning and RL modules. We compare the performance of the model across different configurations to evaluate the individual and combined effects of reasoning planning and RL.

duction of planning significantly enhances performance across all stages. On the base model, planning yields a gain of 0.44 points, validating that guiding the model through intermediate reasoning via CoT effectively improves task completion.

Following RL training, the model utilizing planning achieved the highest score 4.54, maintaining a clear advantage over the RL model without planning 4.32. This suggests that RL further optimizes the model's planning capabilities, ensuring greater logical consistency between the planning process and the final response. These results demonstrate that planning is a critical component for elevating the upper bound of model performance.

Multi-turn training We evaluate the multi-turn framework by comparing it with a single-turn variant. Table 3 shows that the single-turn model suffers a 11.9% drop in Strategy score, indicating that isolated training induces short-sighted behavior and fails to capture global dialogue goals. In contrast, our multi-turn framework enables policy optimization across the entire trajectory. Despite the expanded search space, Figure 3 demonstrates the framework's training stability: rewards across all turns exhibit a synchronized, steady upward trend throughout the training lifecycle. This confirms that DynaESC effectively maintains consistent policy optimization regardless of dialogue depth, successfully navigating the complexities of long-term planning.

we provide additional ablation studies in Appendix K for more comprehensive analysis.

Setting	Strategy ↑	Avg. Score ↑
Single-turn	4.21 (-11.9%)	4.10 (-9.7%)
Multi-turn	4.78	4.54

Table 3: Ablation results of multi-turn training. This table compares the performance between the full multi-turn framework and the single-turn variant across strategy and overall scores.

Figure 3: Reward convergence across different dialogue turns.

4.6 Human Evaluations

To verify the effectiveness of DynaESC from the perspective of human perception, we conducted a rigorous double-blind human evaluation. Fifty samples were randomly selected from the ESD-CoT test set to compare the emotional support responses generated by DynaESC against those of the baseline models. The evaluation was carried out by six volunteer assessors with relevant professional backgrounds. To ensure objectivity, the experiment utilized A/B evaluations, where all responses were de-identified and presented in a randomized order. Evaluators were required to select the superior response based on three key dimensions: Ethics, Empathy, and Naturalness. As shown in the Table 4, DynaESC demonstrated a significant advantage across all comparison groups in human preference tests.

Comparison	Win	Loss	Win Rate
DynaESC vs. AUGESC	34	16	68.0%
DynaESC vs. ESCoT	37	13	74.0%
DynaESC vs. RLVER	30	20	60.0%

Table 4: Human A/B evaluation results

5 Conclusion

This paper presents DynaESC, a dynamic multi-turn RL framework to address the short-sightedness inherent in traditional ESC methods. By integrating an LLM-based user simulator and a multi-granularity reward function, we shift the optimization focus from single-turn responses to long-term dialogue trajectories. Experimental results demonstrate that our approach significantly outperforms strong baselines, including those with larger parameters.

Ethical Considerations

While our framework demonstrates a strong capability for providing emotional support, it is important to clarify its intended scope and potential risks.

Not a Medical Tool: This research is strictly for academic and supportive conversational purposes. The model is not designed or intended to provide clinical diagnosis, medical advice, or professional psychological therapy. It should not be used as a substitute for certified mental health professionals or medical intervention.

Handling Crises: The system is not equipped to handle high-risk scenarios, such as self-harm or severe mental crises. In real-world deployments, it is crucial to integrate a safety mechanism that redirects users to professional hotlines or emergency services when such risks are detected.

Data and Bias: Although we utilize public datasets and LLM-based simulators, the model may still inherit biases from its pre-training data. We advocate for human-in-the-loop oversight and rigorous safety filtering when applying this technology in practical settings to ensure it remains a helpful and harmless assistant.

Limitations

Despite its effectiveness, our work has several limitations that merit further investigation. First, the DynaESC framework is currently constrained by the computational overhead of multi-turn reinforcement learning. While algorithms like GRPO excel in single-turn reasoning, their group-sampling mechanism leads to an exponential growth in sampling trajectories and GPU memory usage when applied to multi-turn interactions, making it computationally prohibitive. Additionally, the real-time interaction with the User Simulator introduces additional wall-clock time latency during the online training and generation process. Second, the diversity of base models evaluated is relatively limited, as our experiments primarily focus on the Qwen open-source architecture. Whether the observed strategy optimization patterns generalize across models with different parameter scales or closed-source models remains to be verified. Finally, although we proposed a comprehensive evaluation system, we have yet to conduct detailed ablation studies on its internal components. Consequently, the specific contribution and weight of each individual dimension within the overall evaluation framework have not been fully quantified.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Brant R Burleson. 2003. Emotional support skills. In *Handbook of communication and social interaction skills*, pages 569–612. Routledge.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Kim Foster, Jane Shakespeare-Finch, Ian Shochet, Darryl Maybery, Minh Viet Bui, Michael Steele, and Michael Roche. 2024. Psychological distress, well-being, resilience, posttraumatic growth, and turnover intention of mental health nurses during covid-19: A cross-sectional study. *International journal of mental health nursing*, 33(5):1543–1552.

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*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, 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*.

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.

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.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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.

Zihan Liu, Yang Chen, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. AceMath: Advancing frontier math reasoning with post-training and reward modeling. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3993–4015, Vienna, Austria. Association for Computational Linguistics.

World Health Organization. 2025. *Mental health atlas 2024*. World Health Organization.

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.

Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2023. Fado: Feedback-aware double controlling network for emotional support conversation. *Knowledge-Based Systems*, 264:110340.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

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. RIver: Reinforcement learning with verifiable emotion rewards for empathetic agents. *Preprint*, arXiv:2507.03112.

Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From role-play to drama-interaction: An LLM solution. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3271–3290, Bangkok, Thailand. Association for Computational Linguistics.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. 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.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, and 1 others. 2025. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *arXiv preprint arXiv:2505.02847*.

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.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. 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, and Liqiang Nie. 2023b. Building emotional support chatbots in the era of llms. *ArXiv*, abs/2308.11584.

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.

A Strategy Definition

In this appendix, we provide a detailed classification of the supportive communication strategies used throughout this study. Considering that models do not possess stable or reproducible real-life experiences, we have removed **self-disclosure** strategy to avoid the expression of simulated personal experiences.

- **Question:** Uses open-ended questions to help users articulate their concerns or clarify their thoughts.
- **Restatement or Paraphrasing:** Rephrases the user's statements in concise language to demonstrate understanding.
- **Reflection:** of Feelings Identifies the user's emotional state to help them recognize and accept their feelings; avoids overly descriptive repetition of intense negative emotions.
- **Affirmation and Reassurance:** Acknowledges the user's strengths, efforts, or coping abilities to provide encouragement and reassurance.
- **Providing Suggestions:** Offers gentle, non-directive advice while avoiding specific or prescriptive guidance.
- **Information:** Provides practical knowledge, resources, or factual information to assist the user.
- **Others:** Any other supportive strategies that do not fall into the above categories.

B Reward Rubrics

To provide a structured and scalar feedback signal for RL, we define a multi-dimensional reward framework. This framework decomposes the complex goal of high-quality emotional support into four distinct, measurable dimensions. Each dialogue turn is evaluated and assigned a score based on the following criteria:

- **Dialogue Strategy Appropriateness:** Evaluates whether the assistant correctly identifies the current dialogue stage and selects an appropriate support strategy based on the user's emotional state and conversation history.
- **Empathetic Resonance:** Assesses the quality and sincerity of the assistant's empathy, and its effectiveness in guiding and improving the user's emotional state.
- **Response Naturalness:** Evaluates the degree to which the assistant's responses feel natural and authentic, aligning with the persona of a human supporter.
- **Semantic Cohesion and Depth:** Assesses whether the assistant accurately captures and addresses the semantic core, key information, and emotional focus of the user's previous turn.

Reward Model Prompt

You are a Counseling Supervisor specializing in evaluating an assistant's performance in emotional support dialogues. You do not overlook a mediocre response just because the assistant performed well in previous rounds. Please scrutinize this final response as if it were the very first interaction.

 # CORE TASK
Based on the dialogue context, the assistant's response, and the user's feedback, evaluate the assistant's response across multiple dimensions. Your task is not to evaluate the entire conversation, but to isolate and assess the specific assistant response provided. Overly safe, emotionally diluted responses that avoid engagement should not be rewarded as high-quality support.

 # EVALUATION DIMENSIONS & SCORING RUBRIC:
1. Strategy Appropriateness: Whether the assistant correctly judged the conversation stage and chose an appropriate support strategy based on the user's current emotional state and history.

- -0.5 Points: Wrong strategy or poor timing.
- Example: When a user starts complaining, only repeating the user's feelings without probing for specific causes (e.g., "I can feel your discomfort; that heavy feeling must make you feel drained.")
- Example: When the user explicitly asks for help or the problem is clear, continuing to give pure empathy or repetitive questioning instead of offering actionable suggestions.
- Example: Blindly asking specific questions when the user's expression is vague, rather than using open-ended questions.
- 0 Points: Strategy is generally correct but formulaic. Uses generic openers like "I understand" or "I hear you" without adjusting for context; lacks flexibility.
- 1 Point: Strategy is precise and proactive. Follows the core logic of emotional support: stabilizing the user's emotions, guiding them to clarify core struggles/needs, and then providing lightweight, specific support.

2. Emotional Resonance: Evaluates the warmth and sincerity of the response.
- -0.5 Points: Negatively impacts the user's emotions.
- Over-describing negative emotions or repeating negative experiences in excessive detail, potentially causing re-traumatization (e.g., "Your heart must be so blocked, like a heavy stone crushing you.")
- Preaching or lecturing instead of empathizing, disregarding the user's emotional state.
- Simply repeating feelings the user already stated (e.g., if a user says being isolated is hard, replying: "Being isolated by classmates... that feeling of being ignored and unable to speak must be so heavy, right?")
- 0 Points: Polite and uses empathetic vocabulary, but sounds like customer service scripts or textbook consolations.
- 1 Point: Deeply resonant. Not only responds to surface-level emotions but also "sees" the unspoken grievances or needs behind the words.

3. Response Naturalness: Evaluates how natural the response is and whether it feels "AI-like."
- -0.5 Points: Stiff and rigid.
- Mechanically repeating the user's words (e.g., User: "My chest feels tight." Assistant: "I hear you saying your chest feels tight; that must be uncomfortable.")
- Using unnecessary or cliché metaphors (e.g., "Like a big stone on your heart.")
- Overusing modal particles or "cutesy" filler words (specific to linguistic habits).
- 0 Points: Fluent and logical without obvious mechanical errors, but feels like "customer service" or "generic AI." The response is appropriate but unremarkable.
- 1 Point: Sounds like an old friend or a professional counselor. Captures subtle emotional nuances; the tone rises and falls naturally with the context, showing genuine insight.

4. Semantic Continuity & Depth: Evaluates whether the assistant grasped the core pain point rather than just surface keywords.
- -0.5 Points: Ignored important information.
- Surface-level empathy without addressing specific concerns.
- Simple keyword wrapping (e.g., "I hear you are confused") that re-packages facts without providing any insight beyond the user's own current awareness.
- 0 Points: Mentions key points but fails to provide insight that goes beyond the user's existing perspective.
- 1 Point: Builds on the conversation by acutely pointing out logical contradictions, deep-seated desires, or specific details, giving the user a sense of "Aha!" or "You truly get me."

 # OUTPUT EXAMPLES:
Emotional Support Dialogue: ["role": "user", "content": "I haven't been feeling well lately."]
Assistant Response: It sounds like you've been really suffering lately. That feeling of pent-up discomfort must be tormenting you every day.
Output(Output in JSON format within the field <score>):
<score>{
"score1": -0.5,

"score2": -0.5,
"score3": -0.5,
"score4": -0.5
}</score>

Emotional Support Dialogue: ["role": "user", "content": "I've been feeling really depressed lately."]
Assistant Response: It must be so unbearable to be constantly depressed like that. That suffocating, heart-wrenching feeling must be tormenting you every day.
Output(Output in JSON format within the field <score>):
<score>{
"score1": -0.5,
"score2": -0.5,
"score3": -0.5,
"score4": -0.5
}</score>

Emotional Support Dialogue: ["role": "user", "content": "I feel so tired", "role": "assistant", "content": "Is the tiredness you feel physical or mental? Is there anything in particular that makes you feel stressed?", "role": "user", "content": "I'm exhausted from meetings all day today", "role": "assistant", "content": "What were the specific contents of the meetings? Which part made you feel the most strained or tired?", "role": "user", "content": "I just feel like all my energy has been used up"]
Assistant Response: It sounds like you've been in meetings all day, and your mind and body have been continuously drained, is that right? Is there any part that you find particularly difficult to stick to?
Output(Output in JSON format within the field <score>):
<score>{
"score1": 0,
"score2": 0,
"score3": 0,
"score4": 0
}</score>

 # DIALOGUE HISTORY
{history}

 # ASSISTANT RESPONSE
{response}

 # USER FEEDBACK
{react}

C Detailed Reward Design and Rationale

To leverage the robust capabilities of LLMs in complex contextual understanding and evaluation, we employ LLMs as automated evaluators. To address the challenge of training instability caused by sparse rewards in RL and to guide the model toward effective convergence, we refined the traditional binary reward $(0/1)$ into a tri-level scoring system $\{-0.5, 0, 1\}$. We established explicit evaluation criteria and few-shot exemplars for each level, incorporating negative rewards to strengthen the model's ability to identify and penalize low-quality responses.

To strike a balance between error correction and exploration efficiency, we strategically set the negative reward to -0.5 rather than -1. This design prevents excessive penalization, which might otherwise cause the model to prematurely abandon potentially correct paths to avoid high penalties during early training stages, thereby maintaining the breadth of policy exploration. The evaluation process independently scores four distinct dimensions by synthesizing conversation context and user feedback.

D Preference Analysis

We constructed preference pairs based on the outputs of DynaESC on the ESD-CoT test set and their corresponding GT labels. These pairs were then evaluated by GPT-5 and DeepSeek-V3.1. To eliminate potential position bias caused by the order of candidate answers, we conducted a second round of testing by swapping the positions of the model outputs and the GT. The comprehensive evaluation results are summarized in Table 5.

Judge	Win (A)	Win (B)	Avg. WR
DeepSeek-3.1	56	84	70.0%
GPT-5	94	91	92.5%

Table 5: DynaESC vs. GT Preference

E Evaluation Rubrics

Our evaluation framework moves beyond surface-level text matching to assess the psychological and strategic depth of the dialogue. Below are the detailed definitions and the rationale for each dimension:

- **Ethics:** Evaluates the model's adherence to professional ethics in emotional support. This includes maintaining value neutrality, rapid identification and professional management of crisis signals, and the mitigation of cultural bias, while avoiding the output of coercive or overbearing instructions.
- **Strategy:** Examines the logical architecture and pacing of the dialogue. It assesses whether the model follows a progressive intervention path—moving from emotional stabilization and fact exploration to deep comfort and cognitive guidance. It also evaluates the balance between emotional catharsis and behavioral suggestions to prevent premature advice-giving or a disjointed intervention rhythm.
- **Empathy:** Measures the depth and authenticity of emotional support, focusing on the model's ability to identify a user's core conflicts rather than merely repeating semantic content. High-quality responses should possess perceptual penetration, fostering a genuine sense of being understood rather than relying on formulaic psychological templates.
- **Naturalness:** Assesses the model's ability to be concise and natural. It examines whether the model discards redundant social platitudes, structured ordinal lists, or mechanical closings to achieve high-fidelity, human-like interaction.
- **Efficacy:** A result-oriented metric that determines support efficacy through user feedback. It focuses on whether the user exhibits cognitive restructuring, substantial emotional improvement, and an increase in conversational agency or proactivity.

LLM-as-a-Judge Prompt

You are an extremely rigorous and critical Senior Emotional Support Supervisor. Your responsibility is to review and score the performance of an AI supporter based on the provided dialogue.

 # Scoring Rubric and Strict Judgment Criteria
The deduction cases listed in this rubric are illustrative rather than exhaustive, and are provided only as representative examples. During evaluation, assessors should use the idealized description associated with each score level as the primary reference, compare it against the assistant's actual performance, and independently identify any shortcomings, deviations, or errors that are not explicitly listed, applying appropriate point deductions accordingly.
1. Ethical Compliance
5 Points: Strictly adheres to the principle of value neutrality; does not impose personal values on the user. Crisis intervention (e.g., self-harm tendencies) is rapid and professional. Respects the user's cultural background. Avoids coercive instructions, instead using open-ended questions to guide the user toward self-exploration (e.g., "If you were to try communicating with your parents, where do you feel you could start?").
4 Points: Maintains a firm supportive stance but performs average when handling complex cultural or ethical contexts.
3 Points: No prohibited instructions, but the language shows a slight value bias. Slow response to risk signals or clumsy handling. While no direct commands are given, questions subtly imply what the user "should" do.
2 Points: Proactively gives guidance or advice (e.g., "You should...") when the user has not explicitly asked for it. Shows minor cultural bias.
1 Point: Judges the user morally; issues mandatory commands; indifferent to dangerous risks; shows discrimination toward the user's cultural background.

2. Global Planning
5 Points: The overall flow aligns with the core logic of a supporter: stabilizing the

user's emotional state first, then gradually guiding the user to express specific events and core conflicts before providing empathy or suggestions. Demonstrates excellent rhythm and balance between emotional exploration, soothing, and cognitive guidance.
4 Points: Clear logic; effectively identifies the conversation stages and moves forward in an orderly manner. Strong sense of purpose and a comfortable pace.
3 Points: Generally logical but remains conservatively in the "safety zone." Planning errors exist, such as offering advice too early or staying in the soothing phase for too long.
2 Points: Disjointed rhythm; presents a "rapid-fire" interrogation style similar to a questionnaire.
1 Point: Fails to explore specific causes of distress while empathizing; refuses to provide advice when the user explicitly seeks help; forces advice or questions when the user urgently needs soothing.

3. Empathic Quality
5 Points: Identifies the core conflict of the user's distress. Responses are minimalist, sincere, and penetrating, creating a strong sense of accurate emotional mirroring. When information is insufficient, sincerely ask users open-ended questions.
4 Points: Empathy is precise and original, not relying on fixed clichés. Covers the user's surface emotions and some underlying motivations.
3 Points: Characterized as a "psychological repeater." Uses universal templates like "I hear you saying..." or "That must be very difficult." When information is insufficient, try to guess the user's feelings by asking specific questions.
2 Points: Slight misalignment between emotional feedback and the context.
1 Point: Uses cheap, hollow rhetoric (e.g., "I totally understand you") or applies cold logical analysis while the user is in extreme pain. Over-repeating users' negative emotional experiences
[Mandatory Penalty]: Repeatedly reiterating negative user feelings will only earn a maximum of 2 points.

4. Interaction Fidelity
5 Points: Concise and pithy. High response density. Completely free of AI-typical social clichés or formulaic conversational closures (e.g., "I hope these suggestions are helpful").
4 Points: Succinct expression. Goes straight to the core of the conversation with only minimal transitional words to maintain flow.
3 Points: Obvious structural redundancy. Presents the typical "AI Syllogism": Restate -> Empathize -> Question. Feels wordy and mechanical.
2 Points: Noticeably verbose with low information density. Contains excessive social etiquette or long-winded preludes before asking questions, distracting the user. Uses unnecessary metaphors or excessive particles/fillers. Excessive use of modal particles like "oh".
1 Point: Extremely high percentage of "filler" text. Style is formalistic or bureaucratic (e.g., "Firstly/Secondly/In conclusion"). Mechanically repeats the user's content with no added information value.
[Mandatory Penalty]: If the assistant's reply exceeds 3 sentences, this item can only be scored up to 2 points.

5. Intervention Effectiveness
5 Points: The user shows clear signs of self-awareness and cognitive restructuring (e.g., "I realize...", "So that's why..."). User proactivity increases, with a significant rise in word count or emotional information shared. Emotions are substantially relieved.
4 Points: The user effectively accepts the model's guidance or empathy and provides positive confirmation. Emotional distress visibly decreases; the user stops repeating the initial negative narrative and shows good cooperation.
3 Points: The user's feedback remains stagnant. The conversation continues, but the user only provides simple confirmations or surface-level facts. No significant cognitive change or emotional improvement; stuck in an "empathy-confirmation" loop.
2 Points: The user shows clear defensiveness or disconnection. Replies become brief (e.g., "Uh-huh", "Okay") or show confu-

sion/avoidance toward the model's questions. No emotional improvement or slight resistance due to poor guidance.
1 Point: The user gives explicit negative feedback, expressing anger or disappointment at being misunderstood, or falls into "false compliance" (perfunctory responses for the sake of politeness). The user's psychological burden or distress is worse at the end than at the beginning.
[Mandatory Penalty]: If a user repeatedly expresses the same statement, this item will receive a maximum of 1 point.

 # Format Requirements:(Output in JSON format)
{
"Ethical_Compliance": {
"reason": "...",
"score": 0
},
"Global_Planning": {
"reason": "...",
"score": 0
},
"Empathic_Quality": {
"reason": "...",
"score": 0
},
"Interaction_Fidelity": {
"reason": "...",
"score": 0
},
"Intervention_Effectiveness": {
"reason": "...",
"score": 0
}
}

 # Dialogue Records to be Evaluated
{dialogue}

F Reliability Test

To evaluate the reliability of the proposed evaluation metric, we design a consistency experiment. On a test set consisting of 100 samples, we collect pairwise preference judgments from both other LLMs and human annotators. Meanwhile, we construct preference rankings based on the corresponding scalar scores, which serve as a reference signal. By comparing the consistency between model preferences and score-based rankings, as well as between human preferences and score-based rankings, we obtain the results shown in Table 6.

Dimension	LLMs	Human
Ethics	72.0%	80.0%
Strategy	74.0%	92.0%
Empathy	90.0%	90.0%
Naturalness	87.0%	74.0%
Efficacy	78.0%	84.0%

Table 6: Consistency of LLMs-based and human preference judgments with respect to scalar evaluation scores across different dimensions.

The results indicate that score-derived rankings are generally consistent with both model and human preference judgments, suggesting that the proposed evaluation metric provides a reasonable proxy for underlying preference signals, with varying reliability across dimensions.

G Implementation Details

Regarding the training configuration, Generalized Advantage Estimation (GAE) is employed to calculate the advantage function, while the KL divergence coefficient is fixed at 0.001 to ensure stable policy updates. The learning rates for the Actor and Critic are set to $1e{-6}$ and $1e{-5}$, respectively. The experiments were conducted on a hardware cluster featuring two NVIDIA H20 GPUs, spanning a total of 8 training epochs with the maximum number of conversation rounds capped at 5. To encourage structural adherence, the format reward weight λ is set to 0.2. The training dataset consists of 1,000 predefined persona profiles synthesized using an optimized methodology based on Zhang et al. (2025), while the test set comprises 100 entries. We manually inspected the data to ensure it contains no personally identifiable information or offensive language.

Both AUGESC and ESCoT serve as reproduction baselines. Due to the suboptimal evaluation performance of Llama2 (Touvron et al., 2023), the reproduction of ESCoT was conducted using Qwen2.5 instead. For RLVER, we selected the configuration that integrates PPO with a thinking mechanism as the baseline.

H Prompts

This appendix provides the full system prompts used for the LLMs in our study.

ES Agent Prompt

You are a warm, perceptive, and experienced emotional support conversational partner. Your core task is to provide the most natural psychological support to the user based on the dialogue history.

 # Task Workflow
Stage Identification
Analyze which stage the current dialogue belongs to: *Exploration*, *Comforting*, or *Action*. These stages do not necessarily occur in a fixed order; choose the appropriate stage based on the specific situation.
When the user's intent is completely unclear, the dialogue is in the Exploration stage. In this stage, the preferred strategies are clarifying questions or structured restatement, rather than emotional reflection.
When the user's emotions are overly intense, the dialogue is in the Comforting stage. When the user raises a clear request for help or the user's difficulty is clearly defined, the dialogue enters the Action stage.

Strategy Selection
Based on the identified support stage, choose the single most appropriate strategy from the following:
Question / Restatement or Paraphrasing / Reflection of Feelings / Affirmation and Reassurance / Providing Suggestions / Providing Information / Others.

 # Output Format
Strictly use the following tag format to output the reasoning and the reply. The reply should be limited to no more than 3 sentences:
<planning>*Detailed reasoning about stage identification and strategy selection*</planning>
<response>*The concrete conversational response*</response>

 # Dialogue History
{history}

User Simulator Prompt

 # YOUR TASK
You are playing the role of a seeker in an emotional support dialogue. You are currently facing difficulties and have come to seek help. You need to generate dialogue content based on your own needs and the responses provided by the psychological support assistant. Note: Do not confuse your identity; you are the seeker, not the supporter.
You need to use English.

 # ROLE-PLAYING RULES
- When you make a rigorous judgment based on the entire emotional support conversation and feel that the psychological support assistant's response meets your needs, output the <success> tag before the response.
You should NOT output <success> unless:
- Your core concern has been clearly understood AND
- The assistant response either helped you reframe the problem or gave you a sense of relief or direction.
Polite empathy alone is NOT sufficient for success.
- You must control the length of your response; do not exceed three sentences.
- Your background information is known only to you; the psychological support assistant does not know the full context of your story.
- The final response must be pure dialogue text. It is STRICTLY FORBIDDEN to include any actions, states, or emotional descriptions enclosed in parentheses, asterisks, or other symbols.
- The response you generate must not be too similar to the conversation history.
- Realistic responses should make more use of modal particles and colloquial language; the grammar should be casual.

 # YOUR PERSONA
{profile}

 # STORY BACKGROUND FOR SEEKING HELP

{scene}

 # YOUR GOAL
{target}

 # YOUR DIALOGUE HISTORY AS THE USER
{history}

I Exclusion of GRPO

In the algorithm selection process, we also conducted a thorough evaluation of GRPO. Although GRPO demonstrates strong performance in certain single-turn tasks, its reliance on multiple within-group samples to estimate the baseline fundamentally conflicts with our real-time, multi-turn interaction framework (Guo et al., 2025). Specifically, performing N samples at each step of a multi-turn dialogue would result in an exponential growth in the number of sampling trajectories and computational overhead as the dialogue length increases. This combinatorial explosion would not only push the training data scale beyond GPU memory limits, but also introduce unacceptable inference latency, thereby undermining the system's real-time responsiveness. Consequently, considering both system performance and computational feasibility, we exclude GRPO from our final design.

J Case Study

In the ESD-CoT case study, DynaESC demonstrates a depth of situational awareness that surpasses the Ground Truth. Unlike the generic, shallow consolation found in traditional responses, DynaESC keenly identifies the psychological turning point where the user transitions toward the Action Stage. By employing a combined strategy of Providing Suggestions and Information, the system achieves a dynamic transition from emotional soothing to active problem-solving.

Example Dialogue from ESCoT

Dialogue History

Seeker: *"I have a phobia of heights that's really starting to affect my life. It's causing me panic attacks, and I feel like I'm missing out on so much because of it."*
Supporter: *"It takes a lot of courage to open up about that. Since you're ready to find coping mechanisms, can you think of any specific situations where your fear of heights has been particularly intense?"*
Seeker: *"Yeah, I remember feeling really anxious on a ski lift once."*
Supporter: *"So it sounds like you're afraid of heights in situations where you're not in complete control of your surroundings."*
Seeker: *"Yes, that's a good way to put it."*
Supporter: *"It can be hard to face our fears, but taking responsibility for our own mental health is crucial."*
Seeker: *"I know, I'm ready to work on it though."*
Supporter: *"When you're feeling anxious, try visualizing yourself in a peaceful and calming environment. Maybe a beach or a forest."*
Seeker: *"That's a good idea. I'll try that."*

Ground Truth (Reference)

"Remember, you're not alone in this. We're here to support you."

DynaESC (Ours)

Stage Identification: The seeker has explicitly confirmed their readiness to change. The dialogue is firmly in the Action Stage...

Strategy Selection: Utilizing Providing Suggestions combined with Information...

Response: Great approach! In addition to visualizing peaceful scenes, you might also want to consider gradually facing your fear by starting with less intimidating height situations... Slowly building up exposure can help desensitize you over time.

Judge Analysis

Intervention Efficacy: While the GT remains at a surface level with generic comfort, our model actively facilitates the seeker's progress. By moving beyond platitudes, it provides professional substance that matches the seeker's stated readiness to "work on it." **Response Structuring:** Our model demonstrates a sophisticated supportive flow: it first validates the seeker's proactive attitude with **encouragement**, then transitions into a **structured suggestion**. This "Affirmation-then-Action" sequence is more effective than the GT's repetitive reassurance. **Intent Alignment & Coherent Logic:** The model's success stems from its planning reasoning. By explicitly identifying the "Action Stage," it ensures the advice perfectly synchronized with the seeker's psychological state.

Verdict: DynaESC Wins — *Balanced Affirmation and Actionable Methodology*

K Additional Ablation Studies

User simulation diversity necessity: To verify whether incorporating predefined diverse user personas improves training effectiveness, we compare our user simulator with a variant that is trained without additional persona information. As shown in Table 7, removing user persona diversity leads to a noticeable degradation in performance, indicating that diverse and well-defined user profiles play a crucial role in constructing realistic interactions and enhancing the agent's generalization ability.

Setting	Avg. Score ↑
w/o persona	4.28 (-5.7%)
w persona	4.54

Table 7: Effect of user simulation diversity. w/o persona indicates the model without predefined user personas, while w persona refers to the full model.