PEARL-CoT: Persona-Emotion Aware Reinforcement Learning via Chain-of-Thought for Emotional Support Conversation

Anonymous EMNLP submission

Abstract

Emotional Support Conversation (ESC) aims to ease seekers' emotional distress through empathic and personalized interactions. However, existing studies predominantly focus on fitting grounded responses, overlooking the cognitive reasoning process of human supporters and seekers' preferences. To address this, we propose PEARL-CoT, a reinforcement learning (RL) framework based on Group Relative Policy Optimization (GRPO), which incorporates emotion and persona reasoning via chain-of-thought (CoT). Specifically, instead of directly generating a response, our model first infers the seeker's emotion and persona, thereby constructing personalized empathic responses. This reasoning step is rewarded with an emotion accuracy reward and a persona consistency reward to ensure the correctness of the CoT process. Afterwards, we incorporate a helpfulness scoring reward, derived from a model trained on seeker feedback, to better align responses with seeker preferences. Additionally, a semantic relevance reward is applied to maintain consistency with human supporter responses. Experimental results demonstrate that PEARL-CoT excels at identifying seekers' concerns, delivering emotional support, and generating responses preferred by human annotators.[1]

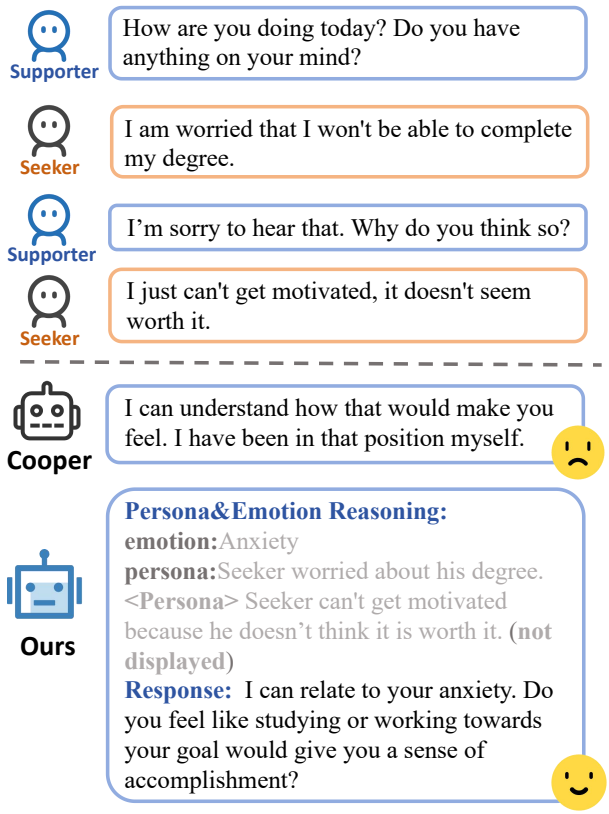

Figure 1: Example dialogue comparing responses from Cooper and ours. Our method first infers the seeker's emotion and persona, generating a response that is more empathic, personalized, and preferred by the seeker, while Cooper, by contrast, provides a generic response lacking emotional depth.

1 Introduction

The growing demand for accessible mental health care (Sharma et al., 2021) has brought increased attention to ESC, underscoring its significance in areas such as psychological counseling (Althoff et al., 2016; Shen et al., 2022) and motivational interviewing (Pérez-Rosas et al., 2016; Saha et al., 2022). In contrast to task-oriented or information-seeking dialogue systems, ESC agents must deliver responses that are emotionally appropriate, personalized, and helpful from the seeker's perspective. This makes ESC a particularly complex challenge in conversational AI, requiring not only a deep understanding of the seeker's emotional states (Spottswood et al., 2013) and personal traits (Rogers, 2012), but also careful adaptation to the seeker's preferences (Swift et al., 2018).

To this end, significant efforts have been made, e.g., MISC (Tu et al., 2022) integrates external commonsense knowledge and blends various strategies to generate supportive responses, while COOPER (Cheng et al., 2024) coordinates multiple

[1]Our code will be available at https://anonymous.4open.science/r/PEARL-CoT-88F4.

specialized agents to jointly promote distinct dialogue goals such as exploration, comforting, and action. Despite progress in this area, ESC systems still face two key challenges: **(a) Neglecting supervision of the cognitive reasoning process.** Recent works like ECOT (Li et al., 2024c) and CogChain (Cao et al., 2024) attempt to incorporate cognitive or emotional theories, e.g., emotional intelligence and structured reasoning chains, into the generation of supportive responses. However, these methods primarily focus on the final output and lack reliable mechanisms to supervise the intermediate reasoning steps. As a result, the generated responses often appear intuitive but ungrounded, failing to reflect a transparent reasoning trajectory. **(b) Overlooking individual seeker preferences.** While recent RL-based approaches such as Partner (Sharma et al., 2021) and SUPPORTER (Zhou et al., 2023) introduce various reward designs such as emotion elicitation, empathy shift, and mutual information to improve emotional coherence and response diversity, they treat seekers as a homogeneous group. Consequently, these methods fail to adapt to the unique emotional needs, persona, or preferences of individual seekers. As a result, the generated support remains generic and can even be misaligned with the specific context or expectations of the seeker.

To address these challenges, in this paper, we present **PEARL-CoT** (**P**ersona-**E**motion **A**ware **R**einforcement **L**earning via **C**hain-of-**T**hought), a novel RL framework that leverages CoT prompting for ESC. As illustrated in Fig. 1, PEARL-CoT introduces an intermediate reasoning step in which the model first infers the seeker's emotion and persona before generating a response. To train the model, we adopt RL based on GRPO. While GRPO was initially developed for outcome-driven and mathematical reasoning tasks (Shao et al., 2024), we extend its application to ESC by designing four domain-specific rewards.

Specifically, to go beyond mere outcome-oriented generation and supervise the cognitive reasoning process, we introduce the following rewards: **(a) Emotion Accuracy:** aligning inferred emotions with ground-truth labels; **(b) Persona Consistency:** assessing semantic alignment with annotated persona traits. Moreover, to generate responses that conform to the unique preferences of individual seekers, we have further designed the following rewards: **(c) Helpfulness Scoring:** derived from a reward model trained on real seeker feedback to reflect human preferences; and **(d) Semantic Relevance:** ensuring contextual alignment with human supporter responses. By integrating these rewards, PEARL-CoT generates responses that are emotionally appropriate, personally tailored, and aligned with individual seeker preferences, thereby enhancing the overall effectiveness of emotional support. The novelty and contributions of our work are highlighted below.

1) We introduce PEARL-CoT, a CoT-based ESC framework that first reasons over emotion and persona before generating responses, enabling interpretable reasoning and enhancing both empathy and personalization.

2) We introduce a multi-aspect reward scheme that supervises both the intermediate reasoning steps and the final response, addressing the limitations of outcome-only optimization and integrating seeker-specific preference alignment.

3) Experiments show that PEARL-CoT achieves state-of-the-art performance in ESC, producing empathic and personalized responses that better align with seeker preferences.

2 Related Work

2.1 Emotional Support Conversation

Initial research in ESC focused on single-turn empathic response generation and support strategy modeling (Rashkin et al., 2018; Sharma et al., 2020). The release of the ESConv dataset by Liu et al. (2021), with multi-turn dialogues annotated with support strategies, advanced ESC by enabling more realistic conversation modeling. Building on this, Tu et al. (2022) incorporated commonsense reasoning and mixed strategies to produce contextually grounded responses. Peng et al. (2022) proposed hierarchical graph networks to capture global emotional causes and local user intentions, while Zhao et al. (2023) improved coherence and fluency by modeling semantic, emotional, and strategic transitions across turns. Cheng et al. (2024) furthered this line by coordinating multiple specialized agents to jointly promote distinct dialogue goals. However, existing studies focus on grounded generation without modeling the cognitive reasoning process of human supporters. Our work fills this gap by introducing explicit reasoning over emotion and persona before generation, enabling more empathic and personalized support.

2.2 RL for ESC

RL (Kaelbling et al., 1996) has become central to aligning language models with human intent, especially via Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017). RLHF uses algorithms like Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Direct Preference Optimization (DPO) (Rafailov et al., 2023) to refine outputs based on human preferences. In the ESC domain, Sharma et al. (2021) first applied RL, employing transformer-based policies to enhance empathy and fluency. Cheng et al. (2022b) later introduced look-ahead planning to anticipate seeker needs, while Zhou et al. (2023) proposed dynamic expert selection for more adaptive strategies. Li et al. (2024b) further advances this direction by integrating the cognitive relevance principle into emotional support agents. However, these approaches fall short in modeling individual preferences and lack diverse, fine-grained reward signals needed for truly empathic and personalized responses. To address this, we propose a multi-aspect reward design to better align with seeker-specific needs.

2.3 CoT for ESC

CoT prompting, introduced by Wei et al. (2022), enables Large Language Models (LLMs) to perform complex reasoning via intermediate steps. Several approaches have adapted CoT prompting to the ESC setting. CogChain (Cao et al., 2024) presents a cognitively motivated framework that decomposes the supporter's reasoning into phases like issue analysis, internal inference, and support strategy selection to mirror human cognitive processes. To improve interpretability, ESCoT (Zhang et al., 2024) builds a dataset with manually verified reasoning chains covering emotional stimuli, cognitive appraisal, and strategy justification. ECoT (Li et al., 2024c) instead aligns the reasoning process with human emotional intelligence guidelines using a plug-and-play prompting strategy. Though not tailored for ESC, Cue-CoT (Wang et al., 2023) introduces a two-stage reasoning mechanism to infer the seeker's mental states before generating responses. However, existing CoT-based methods remain outcome-centric, with limited supervision over intermediate steps, which restricts interpretability and control. Our work addresses this by supervising both reasoning steps and final responses with a multi-aspect reward scheme, enhancing interpretability and alignment.

3 Methods

In this section, we elaborate on our framework, **PEARL-CoT**, which integrates structured CoT reasoning over emotion and persona with a multi-aspect reward scheme. In Sec. 3.1, we define the task and key notations. Sec. 3.2 describes the high-quality emotion and persona annotation process for the ESConv dataset (Liu et al., 2021). Finally, Sec. 3.3 details our fine-grained RL approach, including a policy warm-up phase and a reasoning-to-response GRPO phase.

3.1 Definition

Given a CoT prompt template $\mathcal{T}_{\text{cot}}$ (refer to C.3) and a multi-turn dialog context $C_t = \{u_1^A, u_1^B, \ldots, u_t^A\}$, where each u_i^A and u_i^B represent the seeker's and supporter's utterances at turn i, respectively, the model first infers the seeker's emotion E_{pred} and persona P_{pred}, then generates a supportive response r_{gen} based on C_t, E_{pred}, and P_{pred}. The final output a is thus a composite textual sequence consisting of the inferred emotion, persona, and response, i.e., $a = (E_{\text{pred}}, P_{\text{pred}}, r_{\text{gen}})$.

3.2 High-Quality Emotion and Persona Annotation for ESConv

We begin by detailing the emotion and persona annotation process for the ESConv dataset (Liu et al., 2021), which provides high-quality supervision signals to support structured reasoning and reward-based optimization in supportive dialogue generation. Given the strong inferential capabilities of LLMs in emotion and persona recognition, we adopt GPT-4o to automatically label the ESConv dataset with the seeker's emotion and persona, followed by manual correction to ensure the quality and reliability of the annotations.

Emotion Annotation. Given the dialogue context $\mathcal{C}_t$, we assign an emotion label $E_{\text{gt}} \in \mathcal{E}$[2] to the current seeker utterance u_t^A, conditioned on the full preceding context.

To generate context-aware emotion annotations, we employ GPT-4o (denoted as $\mathcal{M}_{\texttt{GPT-4o}}$) as an initial annotator:

$$E_{\text{gt}} = \mathcal{M}_{\texttt{GPT-4o}}(\mathcal{T}_{\text{emo}}(\mathcal{C}_t)) \quad (1)$$

where $\mathcal{T}_{\text{emo}}$ is a prompt template tailored for emotion annotation (see Appendix C.1 for details).

[2]We add *Neutral* and *Positive* to the original 7-class schema in ESConv: *Anxiety*, *Depression*, *Sadness*, *Anger*, *Fear*, *Disgust*, *Shame*.

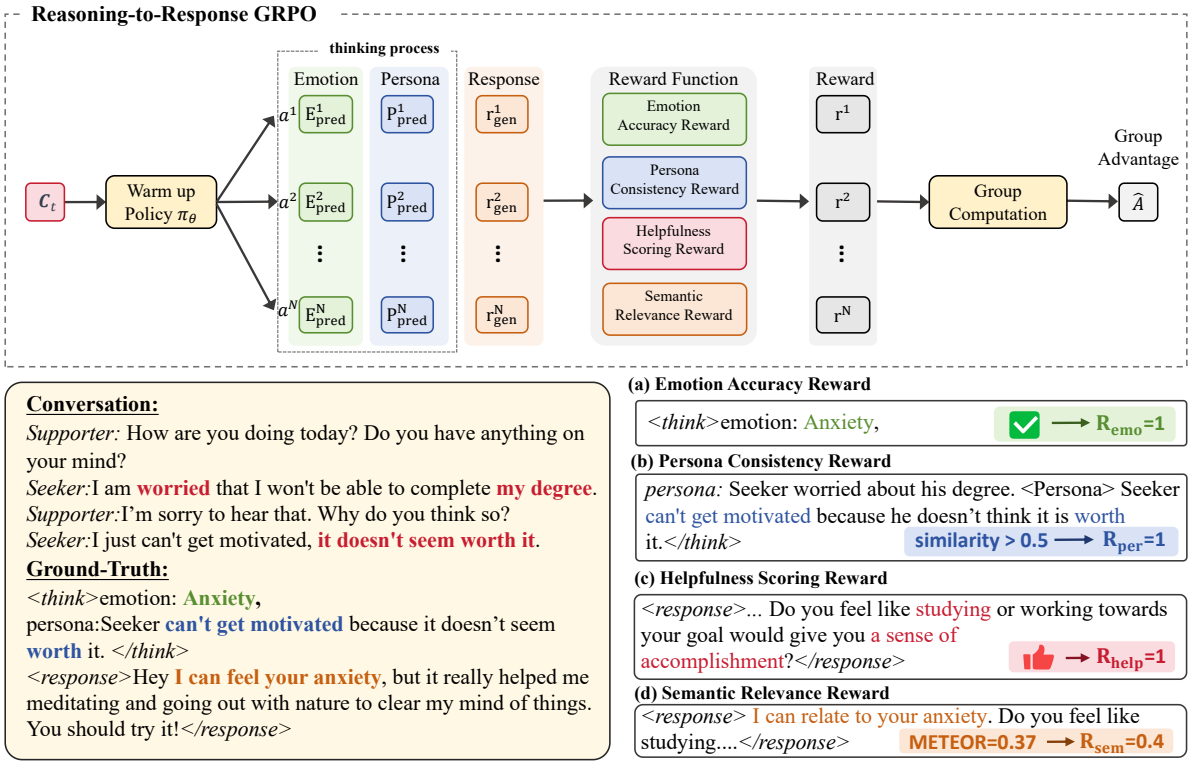

Figure 2: An overview of our RL framework, which operates in two main stages. First, the policy model π_θ undergoes a warm-up phase to establish a stable initialization. In the subsequent optimization phase, π_θ produces a batch of candidate outputs, each conditioned on the dialogue context. Both the reasoning process and the final response are evaluated with four distinct reward signals: emotion accuracy, persona consistency, helpfulness scoring, and semantic relevance, exemplified in panels (a), (b), (c), and (d). To guide learning, each reward is standardized against the batch distribution to compute the normalized advantage $\hat{A}$ used in policy updates.

Persona Annotation. Unlike prior works that infer seeker persona directly from seeker utterances alone (Cheng et al., 2022a), we adopt a *response-centric* annotation strategy. Given the dialogue context $\mathcal{C}_t$ and the supporter's response u_t^B, we identify a minimal set of persona traits P_{gt} that are either *reflected in* or *relevant to* u_t^B:

$$P_{\text{gt}} = \mathcal{M}_{\texttt{GPT-4o}}(\mathcal{T}_{\text{per}}(\mathcal{C}_t, u_t^B)), \quad P_{\text{gt}} \subseteq \mathcal{P} \tag{2}$$

Here, $\mathcal{T}_{\text{per}}$ is the prompt used for persona extraction (refer to Appendix C.2), and $\mathcal{P}$ is the full set of seeker persona traits possibly mentioned throughout the context. Only those elements from $\mathcal{P}$ that are actually leveraged by the supporter in u_t^B are included in P_{gt}.

Each trait is rewritten in concise third-person form (e.g., *the seeker cares deeply about his family*) and separated using the delimiter `<Persona>`. If no relevant persona is used, the result is explicitly annotated as `None`.

Manual Correction. To ensure high-quality annotation, three trained annotators manually corrected GPT-4o outputs. Corrections enforce consistency with the predefined emotion list and verify that persona traits are grounded in the dialogue context and actually reflected in the supporter's response. See Appendix B for detailed correction guidelines and common error patterns.

Final Dataset. Each annotated instance in the resulting dataset $\mathcal{D}_s$ consists of a dialogue context $\mathcal{C}_t$ and a response sequence τ, defined as:

$$\tau = (E_{\text{gt}}, P_{\text{gt}}, r_{\text{gt}}) \tag{3}$$

The final dataset is represented as:

$$\mathcal{D}_s = \{\mathcal{C}_t^n, \tau^n\}_{n=1}^N \tag{4}$$

3.3 Fine-Grained RL with Structured Reasoning and Multi-Aspect Reward

Based on the annotated emotion and persona labels, we propose a fine-grained RL framework to enhance the emotion and persona alignment of supportive responses through structured reasoning and multi-aspect rewards. As illustrated in Fig. 2, our

framework consists of two phases: (a) a policy warm-up phase and (b) a reasoning-to-response GRPO phase. This design enables the model to first acquire basic reasoning capabilities and then improve through fine-grained rewards.

3.3.1 Policy Warm Up

To prepare the policy model π_θ with essential reasoning capabilities prior to RL, we introduce a warm-up phase. In this stage, the model learns to infer both the seeker's emotion and persona, followed by the generation of a supportive response. The policy model is fine-tuned on a small subset of the training data, denoted as $\mathcal{D}'_s = \{C_t, \tau\}_{n=1}^{N'}$, which is held out from the subsequent RL stage. Here, we linearize τ into a token sequence $\{y_1, \ldots, y_T\}$. The warm-up objective is defined as:

$$\mathcal{L}_{\text{warm}} = -\mathbb{E}_{\tau \sim \mathcal{D}'_s}\Big[\sum_{t=1}^{T} \log \pi_\theta(y_t \mid C_t, y_{<t})\Big] \quad (5)$$

3.3.2 Reasoning-to-Response GRPO

After policy warm-up, π_θ generates a collection of outputs $\{a^i\}_{i=1}^{N}$. To better align generation with ESC objectives, we design multi-aspect reward functions that assess both the reasoning and response of each output from four perspectives: emotion accuracy, persona consistency, helpfulness scoring, and semantic relevance. Below, we detail the design and implementation of each component.

Emotion Accuracy Reward (R_{emo}). This reward captures the model's ability to identify and respond appropriately to the seeker's emotional state. To compute this reward, we first extract the ground-truth emotion E_{gt} using GPT-4o. Separately, the predicted emotion E_{pred} is parsed directly from the CoT reasoning process. We then define the emotion accuracy reward as:

$$R_{\text{emo}} = \mathbb{I}\left(E_{\text{pred}} = E_{\text{gt}}\right) \quad (6)$$

The reward is 1 if the predicted and ground-truth emotions match exactly, and 0 otherwise.

Persona Consistency Reward (R_{per}). Supportive responses should demonstrate consistency with the seeker's persona traits. To assess this, we define a persona consistency reward based on the semantic similarity between the inferred persona P_{pred} and the reference persona P_{gt} annotated by GPT-4o. Both are encoded into dense vector representations using a sentence embedding model[3] as follows:

$$\begin{aligned} \mathbf{e}_{\text{pred}} &= \text{Embed}(P_{\text{pred}}) \\ \mathbf{e}_{\text{gt}} &= \text{Embed}(P_{\text{gt}}) \end{aligned} \quad (7)$$

where $\text{Embed}(\cdot)$ represents the embedding function. The persona reward is then defined by computing the cosine similarity $\cos(\cdot, \cdot)$ between the two embeddings as:

$$R_{\text{per}} = \mathbb{I}\left(\cos(\mathbf{e}_{\text{pred}}, \mathbf{e}_{\text{gt}}) > 0.5\right) \quad (8)$$

A reward of 1 is assigned if the similarity exceeds a threshold of 0.5, and 0 otherwise.

Helpfulness Scoring Reward (R_{help}). This reward reflects the perceived helpfulness of the generated response. Unlike prior work such as Li et al. (2024b), which quantifies the positive effect of an utterance by computing the change in helpfulness score over a dialogue sequence, we adopt a turn-level reward that directly assigns a helpfulness score. Specifically, we define the reward as:

$$R_{\text{help}} = \text{Helpful}(\text{context}_{\leq 7}, r_{\text{gen}}) \quad (9)$$

where $\text{context}_{\leq 7}$ represents the most recent sequence of up to seven utterances preceding the response; and $\text{Helpful}(\cdot)$ is a pre-trained helpfulness model (see Sec. 4.3.1) that predicts a helpfulness score in $\{-1, 0, 1\}$, corresponding to unhelpful, neutral, and helpful, respectively.

Semantic Relevance Reward (R_{sem}). This reward assesses whether the generated response preserves the semantic intent and informativeness of a human-written reference response. To measure this, we employ the METEOR metric (Banerjee and Lavie, 2005), which has been shown to correlate well with human judgment in dialogue tasks, making it suitable for our evaluation. The reward is defined as:

$$R_{\text{sem}} = \phi\left(\text{METEOR}(r_{\text{gen}}, r_{\text{gt}})\right) \quad (10)$$

To normalize and discretize the reward signal, we introduce a rounding function $\phi(\cdot)$ that rounds the METEOR score to a single decimal place, yielding a reward value in the range $\{0.0, 0.1, \ldots, 1.0\}$.

[3] sentence-transformers/multi-qa-distilbert-cos-v1

Optimization with rewards. After obtaining the four individual rewards, we aggregate them to compute the total reward for each generated output as $r^i = R^i_{\text{emo}} + R^i_{\text{per}} + R^i_{\text{help}} + R^i_{\text{sem}}$. Given a set of N sampled responses, we calculate the standardized advantage $\hat{A}_i$ for each response to assess its relative quality within the batch:

$$\hat{A}^i = \frac{r^i - \text{mean}(\{r^1, r^2, \ldots, r^N\})}{\text{std}(\{r^1, r^2, \ldots, r^N\})} \quad (11)$$

This advantage score guides the policy update by highlighting how each response compares to its peers. To regularize policy updates, we include a Kullback–Leibler (KL) divergence term that penalizes deviation from a fixed reference policy π_{ref}, which is initialized as a copy of the policy and held constant during training. Following Shao et al. (2024), we approximate the KL divergence using:

$$D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(a^i \,|\, C_t)}{\pi_\theta(a^i \,|\, C_t)} - \log \frac{\pi_{\text{ref}}(a^i \,|\, C_t)}{\pi_\theta(a^i \,|\, C_t)} - 1 \quad (12)$$

Combining the advantage-weighted importance scores and a KL regularization term, the final RL objective is formulated as:

$$\mathcal{L}_{\text{RL}} = -\mathop{\mathbb{E}}_{C_t \in \mathcal{D}_s} \Big[\frac{1}{N} \sum_{i=1}^{N} \frac{\pi_\theta(a^i \,|\, C_t)}{[\pi_\theta(a^i \,|\, C_t)]_{\text{no grad}}} \hat{A}^i \quad (13)$$

$$- \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Big]$$

4 Experiment Setup

4.1 Dataset

Following previous works, we conduct our experiments on ESConv (Liu et al., 2021), a widely adopted multi-turn emotional support dialogue corpus. In each conversation, a seeker discloses a personal emotional difficulty, while the supporter responds with comforting and empathic messages aimed at alleviating the seeker's distress. To evaluate the perceived helpfulness of the support, seekers provide a 5-point feedback rating after every two supporter utterances.

In addition to the original data, we incorporate emotion and persona annotations to enhance the modeling of emotional understanding and personalization. A summary of dataset statistics, including those related to the annotations, is provided in Appendix A. We follow the official train, validation, and test split defined in the original ESConv repository of Liu et al. (2021).

	Categories	Num	Proportion
Feedback	1(Very Bad)	245	2.8%
	2(Bad)	353	4.1%
	3(Average)	1385	16.0%
	4(Good)	2524	29.1%
	5(Excellent)	4161	48.0%
	Overall	**8668**	**100.0%**
Score	-1(Unhelpful)	1312	33.3%
	0 (Neutral)	1312	33.3%
	1 (Helpful)	1312	33.3%
	Overall	**3936**	**100.0%**

Table 1: Statistics of origin ESConv corpus combined with failed ESConv examples, including the origin seeker's feedback and our processed supporter's score.

4.2 Baselines

Our baseline comparisons encompass both prompt-based LLMs and previous state-of-the-art approaches on the ESConv dataset (Liu et al., 2021). Specifically, we prompt Qwen2.5-1.5B-Instruct[4] and Llama3.2-1B-Instruct[5] using a concise task description along with the dialogue history (see Fig. 8) to generate responses. Additionally, we reproduce several domain-specific models, including MISC (Tu et al., 2022), TransESC (Zhao et al., 2023), Cooper (Cheng et al., 2024), and Emstremo (Li et al., 2024a). Detailed descriptions of these baselines can be found in Appendix D.

4.3 Implementation Details

4.3.1 Helpfulness Score

As shown in Tab. 1, we construct a labeled dataset by combining the original ESConv corpus with all failed ESConv examples[6], using the seeker's feedback, provided after every two supporter responses, as the score of supporter responses. Due to the highly imbalanced distribution of raw feedback scores, we map the original 5-point scale into three categories: scores of 1–2 are relabeled as -1 (unhelpful), 3 as 0 (neutral), and 4–5 as 1 (helpful). To mitigate class imbalance, we perform label-aware downsampling to obtain a more balanced dataset.

A BERT-base-uncased model[7] is fine-tuned to predict the discretized feedback score based on

[4] Qwen/Qwen2.5-1.5B-Instruct

[5] meta-llama/Llama-3.2-1B-Instruct

[6] Details about the failed examples are available in https://github.com/thu-coai/Emotional-Support-Conversation

[7] google-bert/bert-base-uncased

Model	D-1	D-2	D-3	METEOR	Fluency	Diversity	Empathy	Suggestion	Humanoid	Skillful	Overall	Average
Cooper	3.80	18.81	35.86	8.20	28.72	21.68	28.95	18.36	**22.26**	21.76	13.61	20.18
Emstremo	3.56	16.37	30.27	7.25	28.64	21.17	28.53	17.98	21.09	21.21	13.64	19.06
MISC	3.33	15.34	29.18	6.85	28.55	21.16	28.49	17.86	21.25	21.11	13.69	18.80
TransESC	2.92	13.08	25.10	6.62	28.52	21.10	28.51	17.82	21.32	21.04	13.66	18.15
Qwen2.5-1.5B-Instruct†	3.23	23.81	47.70	8.04	21.26	16.43	21.60	15.26	16.35	17.92	10.39	18.36
Llama3.2-1B-Instruct‡	3.13	22.01	45.61	9.68	25.51	19.57	25.62	18.29	19.83	21.17	13.17	20.33
PEARL-CoT (feat. Qwen)	**4.29**	**29.00**	**55.87**	8.89	**29.00**	21.60	**28.97**	18.76	21.80	22.24	13.91	**23.12**
PEARL-CoT (feat. Llama)	3.13	20.60	42.13	**9.77**	28.59	**21.71**	28.82	**19.13**	21.90	**22.96**	**14.05**	21.16

Table 2: Automatic evaluation results. † and ‡ indicates that only 1,597 and 1,901 out of 2,178 responses were valid due to formatting issues in the model outputs, respectively. Invalid responses were treated as empty strings during metric computation to ensure consistency in the evaluation.

the preceding eight utterances in the conversation. The model outputs one of $\{-1, 0, 1\}$ for each input instance. The checkpoint with the highest validation accuracy is selected for use in the RL phase. We employ the AdamW optimizer with an initial learning rate of 2e-5 and apply a cosine annealing scheduler (Loshchilov and Hutter, 2016) to reduce the risk of overfitting.

4.3.2 Implementation of Policy Warm Up

We perform supervised fine-tuning (SFT) on a small subset of the training data, which is excluded from the RL stage. This phase uses Qwen2.5-1.5B-Instruct and Llama3.2-1B-Instruct as the base models, which are trained for three epochs. We adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient fine-tuning, setting the rank to 8, the LoRA alpha to 32, enabling dropout, and specifying the task type as `CAUSAL_LM`. The fine-tuning process uses a batch size of 4 and a gradient accumulation step of 4. We initialize the learning rate at 1e-4 and employ a cosine learning rate scheduler. All experiments are conducted on two NVIDIA RTX A6000 GPUs with a maximum input sequence length of 3072 tokens.

4.3.3 Implementation of GRPO

Following SFT, we conduct RL using the GRPO algorithm. The training setup maintains the same LoRA configuration, number of epochs, learning rate, and cosine learning rate scheduler as used during the supervised stage. We set the batch size to 8 and the gradient accumulation step to 1. We set `num_generations` to 4 to sample 4 candidate responses for each input prompt. To reduce GPU memory consumption and accelerate training, we integrate FlashAttention-2 and DeepSpeed Stage 3 into our training pipeline.

4.4 Evaluation Metrics

For context-free evaluation, we utilize **Distinct** scores (Li et al., 2015) to quantify the lexical diversity of generated responses, and the **METEOR** score (Banerjee and Lavie, 2005) to assess their similarity to corresponding ground-truth responses.

To evaluate responses in context, we adopt a set of metrics tailored for ESC assessment, including evaluations of **Fluency**, **Diversity**, **Empathy**, **Suggestion**, **Humanoid**, **Skillful**, and **Overall**[8].

5 Results and Analysis

5.1 Automatic Evaluation

To validate the effectiveness of our PEARL-CoT framework, we compare it against several baseline models, with detailed results presented in Tab. 2.

First, our model achieves superior performance in both lexical diversity and reference similarity. These improvements are largely driven by the incorporation of a semantic relevance reward that encourages responses to remain coherent and closely aligned with the reference content. Notably, Qwen2.5-1.5B-Instruct and LLaMA3.2-1B-Instruct exhibit high lexical diversity scores due to their longer average response lengths (72.65 and 74.47 tokens, respectively), as compared to the shorter responses (16.11–40.42 tokens) produced by ESC-specific models.

Second, our approach consistently outperforms baselines on most context-aware metrics. This can be attributed to two key design choices. One contributing factor is our CoT-based prompting strategy, which guides the model to explicitly reason about the seeker's emotion and persona prior to response generation, thereby fostering interpretable and contextually grounded outputs. In addition, the RL phase incorporates fine-grained reward signals that supervise both the intermediate reasoning steps

[8] haidequanbu/ESC-RANK

and the final response, further enhancing empathy, personalization, and alignment with seeker preferences. Although Cooper slightly surpasses our method on the Humanoid metric due to its modular design with explicit dialogue tracking mechanisms, our model achieves competitive results and establishes new state-of-the-art performance across most evaluation dimensions.

5.2 Human Evaluation

To assess the quality of generated responses beyond automatic metrics, we conduct human evaluation following the protocol of prior work (Gao et al., 2021; Peng et al., 2022). Specifically, we compare outputs from two models across five dimensions: 1)*Fluency*: which response is more natural and grammatically correct? 2)*Identification*: which better understands or identifies the seeker's underlying problem? 3)*Comforting*: which response provides more emotional comfort? 4)*Suggestion*: which offers more useful and informative advice? 5)*Overall*: which response is generally more favorable?

We randomly select 100 dialogue samples from the test dataset. For each comparison, three annotators independently judge which model performs better, using a Win/Tie/Lose format. The human evaluation results are summarized in Tab. 3.

PEARL-CoT	**MISC**			**TransESC**		
(feat. Qwen)	Win	Lose	Tie	Win	Lose	Tie
Flu.	**33.3**‡	21.7	45.0	**35.3**‡	20.3	44.3
Ide.	**50.7**‡	14.7	34.7	**51.7**‡	11.0	37.3
Com.	**45.0**‡	20.7	34.3	**51.3**‡	16.7	32.0
Sug.	**39.3**‡	15.7	45.0	**42.7**‡	14.0	43.3
Ove.	**45.3**‡	26.7	28.0	**52.3**‡	21.7	26.0

Table 3: Human evaluation results(%). ‡ denotes p-value < 0.05 (statistical significance test).

We conduct a comparative analysis between PEARL-CoT (feat. Qwen) and two baseline models, TransESC and MISC. The results indicate that our solution consistently outperforms both baselines across all evaluation metrics, which verifies that its notable strength in identifying the seeker's underlying issues can be attributed to the integration of persona reasoning before response generation. Furthermore, its strong performance in emotional support stems from accurately detecting the seeker's emotion. The overall effectiveness is further reinforced by our helpfulness scoring reward and semantic relevance reward.

5.3 Ablation Studies

To investigate the individual contributions of each reward signal in our RL framework, we conduct a series of ablation experiments by removing one reward at a time from the full model.

Model	**D-1**	**METEOR**	**Fluency**	**Diversity**	**Empathy**	**Overall**
PEARL-CoT	4.29	**8.89**	**29.00**	**21.60**	**28.97**	**13.91**
w/o EReward	3.96	8.53	28.94	21.55	28.74	13.86
w/o PReward	3.84	8.49	28.76	21.49	28.82	13.88
w/o HReward	4.06	8.28	28.90	21.43	28.90	13.82
w/o SReward	**4.96**	6.86	28.78	21.31	28.53	13.74

Table 4: Results of ablation study over Qwen2.5-1.5B-Instruct. EReward/PReward/HReward/SReward refer to the Emotion Accuracy/Persona Consistency/Helpfulness Scoring/Semantic Relevance Reward, respectively.

As presented in Tab. 4, removing any single reward leads to a performance decline, underscoring the necessity of each component in our framework. Specifically, excluding the Emotion Accuracy Reward (**w/o EReward**) results in a notable drop in the Empathy score, highlighting its critical role in producing emotionally attuned responses. The removal of the Persona Consistency Reward (**w/o PReward**) leads to a decrease in D-1, indicating that persona grounding is essential for generating diverse and personalized outputs. Removing the Helpfulness Scoring Reward (**w/o HReward**) causes a noticeable decline in the Overall score, demonstrating its significance in aligning responses with seeker preferences and enhancing overall response quality. Lastly, omitting the Semantic Relevance Reward (**w/o SReward**) yields the most substantial reduction in METEOR, reinforcing its importance in preserving content coherence. Interestingly, this variant also achieves the highest D-1 score, suggesting a trade-off between semantic alignment and lexical diversity.

6 Conclusion

In this paper, we propose PEARL-CoT, an RL framework that incorporates CoT reasoning to generate empathic and personalized emotional support responses. By first inferring seeker emotion and persona, the model enables interpretable and controllable generation. A multi-aspect reward scheme supervises both reasoning and response to better align with seeker preferences. Experiments demonstrate that PEARL-CoT achieves state-of-the-art performance, validating the effectiveness of integrating structured reasoning and RL for ESC.

Limitations

Despite its effectiveness, PEARL-CoT exhibits two notable limitations: **(a) Dependency on Accurate Emotion and Persona Annotations.** PEARL-CoT's effectiveness relies heavily on precise emotion and persona labels to guide its reasoning process. However, in practical applications, obtaining such annotations can be costly and error-prone, limiting the scalability and robustness of the framework. Future work may focus on developing unsupervised or weakly supervised methods to efficiently extract seeker-specific attributes. **(b) Computational Overhead and Reasoning Noise from CoT Prompting.** While the CoT prompting enhances reasoning capability, it introduces extra computational cost and may produce incorrect or noisy intermediate reasoning steps. This not only increases inference time but may also complicate interpretability. Future research could explore optimizing or constraining the reasoning process to improve efficiency and reliability.

Ethical Considerations

Our experiments utilize the ESConv dataset, a publicly released resource curated for studying emotional support dialogues. This dataset was carefully constructed to exclude sensitive content, personal identifiers, and any language deemed unethical or harmful, ensuring participant privacy is thoroughly safeguarded. Our work aims to develop a conversational system that offers supportive responses in typical, non-critical emotional contexts, consistent with the scope of ESConv. It is important to emphasize that this system is not designed to handle urgent or high-risk scenarios, such as those involving suicidal ideation or self-harm. We do not position our model as a substitute for professional psychological care. We ensure all user feedback used during training was anonymized, and no identifiable information was incorporated at any stage. Furthermore, access to our feedback model and supporter system will be limited exclusively to academic research and will not be released for commercial or non-scholarly applications.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Yaru Cao, Zhuang Chen, Guanqun Bi, Yulin Feng, Min Chen, Fucheng Wan, Minlie Huang, and Hongzhi Yu. 2024. Enhancing emotional support conversation with cognitive chain-of-thought reasoning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 175–187. Springer.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2022a. Pal: Persona-augmented emotional support conversation generation. *arXiv preprint arXiv:2212.09235*.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022b. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. *arXiv preprint arXiv:2210.04242*.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17853–17861.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.

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

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Junlin Li, Bo Peng, and Yu-Yin Hsu. 2024a. Emstremo: Adapting emotional support response with enhanced emotion-strategy integrated selection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5794–5805.

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

Zaijing Li, Gongwei Chen, Rui Shao, Yuquan Xie, Dongmei Jiang, and Liqiang Nie. 2024c. Enhancing emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

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*.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2016. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51.

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Carl Rogers. 2012. *Client centered therapy (new ed)*. Hachette UK.

Tulika Saha, Saichethan Miriyala Reddy, Sriparna Saha, and Pushpak Bhattacharyya. 2022. Mental health disorder identification from motivational conversations. *IEEE Transactions on Computational Social Systems*, 10(3):1130–1139.

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, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the web conference 2021*, pages 194–205.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Siqi Shen, Verónica Pérez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107.

Erin L Spottswood, Joseph B Walther, Amanda J Holmstrom, and Nicole B Ellison. 2013. Person-centered emotional support and gender attributions in computer-mediated communication. *Human Communication Research*, 39(3):295–316.

Joshua K Swift, Jennifer L Callahan, Mick Cooper, and Susannah R Parkin. 2018. The impact of accommodating client preference in psychotherapy: A meta-analysis. *Journal of clinical psychology*, 74(11):1924–1937.

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. *arXiv preprint arXiv:2203.13560*.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. *arXiv preprint arXiv:2305.11792*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. Escot: Towards interpretable emotional support dialogue systems. *arXiv preprint arXiv:2406.10960*.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. Transesc: smoothing emotional support conversation via turn-level state transition. *arXiv preprint arXiv:2305.03296*.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. *arXiv preprint arXiv:2307.07994*.

Appendices

A Dataset

Tab. 5 presents the statistical overview of the ESConv dataset. Fig. 3 illustrates the distribution of average persona annotations. Fig. 4 depicts the distribution of emotion categories. Both sets of annotations were generated using GPT-4o.

Category	Official Division		
	Train	Dev	Test
Number of Supporter Utterances	10679	2257	2389
Number of Seeker Utternaces	10497	2210	2363
Avg. words per supporter utterance	25.42	25.66	24.10
Avg. words per seeker utterance	23.66	24.39	22.80
Avg. turns per dialogue	23.27	22.91	24.37

Table 5: Statistics of ESConv Dataset.

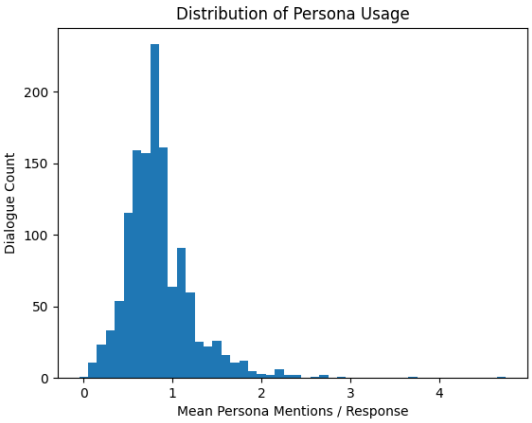

Figure 3: Distribution of average persona traits annotated via GPT-4o.

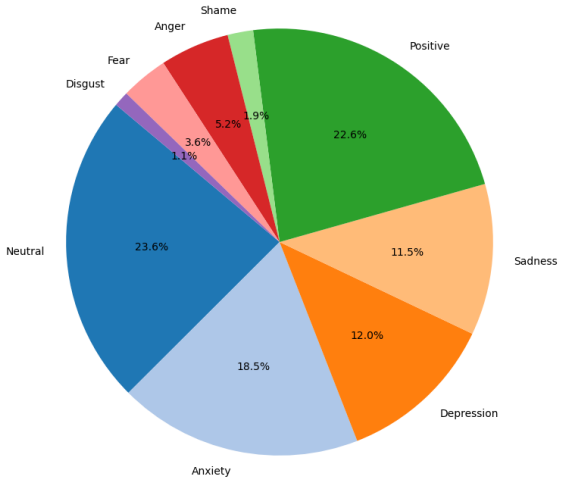

Figure 4: Distribution of emotion categories annotated via GPT-4o.

Annotation Criteria
Each emotion label must correspond to exactly one valid entry from a predefined emotion list and accurately reflect the seeker's emotional state within the context of the current utterance and overall dialogue.
Common Correction Cases
1. Use of invalid emotion labels not included in the predefined list (e.g., *mad*).
2. Incorrect grammatical form, such as adjectives used instead of nouns (e.g., *anxious* instead of *anxiety*).
3. Assignment of multiple emotions to a single utterance (e.g., *anxiety and sadness*).
4. Inclusion of unnecessary modifiers that reduce clarity (e.g., *much anger*).

Table 6: Guidelines and common correction cases for emotion annotation.

B Manual Correction Guidelines

We manually refined the GPT-4o-generated annotations with the help of three trained annotators, based on the following standardized criteria. Tab. 6 and Tab. 7 illustrate the guidelines for emotion annotation and persona annotation, respectively.

C Prompt Templates

C.1 Emotion Annotation

The prompt template used to elicit emotion annotations from GPT-4o is illustrated in Fig. 5. Given a multi-turn conversation between a seeker and a supporter, GPT-4o is prompted to infer the emotional tone of each seeker message, focusing on the message itself while also leveraging contextual cues from earlier turns when relevant.

C.2 Persona Annotation

Fig. 6 shows the prompt template designed to guide GPT-4o in identifying the minimal set of persona traits reflected in each supporter response within an emotional support dialogue.

C.3 CoT Prompt

The CoT prompt is provided in Fig. 7. It is structured to support both SFT and RL with Qwen2.5-1.5B-Instruct and Llama3.2-1B-Instruct.

Annotation Criteria
Persona labels should only include seeker background information that the supporter explicitly uses in their response and must be phrased concisely in the third person (e.g., *Seeker is a student*).
Common Correction Cases
1. Use of first-person phrasing rather than third-person (e.g., *I'm an engineer*). 2. Missing delimiters or tags to separate multiple persona items (e.g., absent `<Persona>` markers). 3. Inclusion of persona information irrelevant or unused in the response. 4. Factually incorrect or unsupported persona claims. 5. Overly complex statements that combine several facts into one line.

Table 7: Guidelines and common correction cases for persona annotation.

D Baselines

Fig. 8 presents the prompt templates utilized for implementing the prompt-based baseline(Qwen2.5-1.5B-Instruct and Llama3.2-1B-Instruct) on the ESConv dataset. Below, we provide a detailed overview of the remaining baselines in the fine-tuned category, along with their corresponding implementation strategies:

MISC is an ESC model that incorporates external commonsense knowledge and employs a mixture of response strategies to generate supportive responses. Specifically, MISC integrates knowledge from COMET and strategically selects a set of mixed strategies from multiple support strategies based on the seeker's needs and emotional state to generate contextually appropriate and emotionally supportive responses (Tu et al., 2022).

TransESC is an ESC model that improves response fluency and coherence by modeling turn-level transitions in semantics, strategy, and emotion. It uses a "transit-then-interact" mechanism and a transition-aware decoder to generate contextually aligned supportive responses (Zhao et al., 2023).

Emstremo is an ESC model that integrates emotion perception and support strategies to generate empathic, emotionally aligned responses. It emphasizes strategic control in aligning the supporter's tone with the seeker's emotional state, and has shown strong performance in both automatic and human evaluations (Li et al., 2024a).

COOPER is an ESC framework that coordinates multiple specialized agents to jointly promote distinct dialogue goals, including exploration, comforting, and action. By dynamically tracking and ranking dialogue aspects, COOPER generates strategically guided responses. It has shown superior performance in emotional support tasks, highlighting its effectiveness in managing complex conversational objectives (Cheng et al., 2024).

E Case Study

Tab. 8 present two examples to qualitatively evaluate the performance of different models.

Case 1: Short Dialogue with Clear Emotional Cues. In this low-turn conversation, the seeker expresses frustration that their friends are not taking COVID-19 precautions seriously. Models such as **Emstremo** and **TransESC** respond with vague suggestions like *"talking to them about it,"*, failing to specify what *"it"* refers to, thereby missing the core issue. **MISC** produces a response that is misaligned with the dialogue context, while **Cooper** merely echoes the seeker's words without offering meaningful support or demonstrating understanding. Although **Qwen2.5-1.5B-Instruct** recognizes the seeker's concern and offers targeted advice, its response is overly verbose, which may hinder effective engagement. In contrast, **PEARL-CoT (feat. Qwen)** accurately captures the seeker's emotion and context, offering empathic and concise guidance tailored to the seeker's situation and personal background.

Case 2: Long Dialogue with Complex Context. This example involves a high-turn conversation, introducing greater difficulty in tracking context. Under this more challenging setup, **Emstremo**, **TransESC**, and even **Qwen2.5-1.5B-Instruct** generate responses that contradict the dialogue history, indicating difficulty in maintaining coherence. **MISC** defaults to a generic response, while **Cooper** again parrots the seeker's statements and introduces logical inconsistencies. By contrast, **PEARL-CoT (feat. Qwen)** demonstrates a nuanced understanding of the conversation, responding in a way that aligns with both the seeker's emotional trajectory and their individual persona traits.

[Context]
Below is a conversation between a "Seeker" and a "Supporter".
<Here is context of conversation>
Each Seeker message is labeled with a number like "Seeker[1]:", "Seeker[2]:", etc.
[Task]
Your task is to infer the general emotional tone expressed by the Seeker in each message. Focus on the content and implied feeling of each Seeker[n] message. Use context from previous messages if it adds clarity. For each message, assign one emotional tone strictly from this list: [Neutral, Positive, Anxiety, Depression, Sadness, Anger, Fear, Disgust, Shame].
[Output]
Respond in this format, one line per message: Seeker[n]: [emotion]. Only output the labels. Do not provide explanations or comments.

Figure 5: Prompt template used for emotion annotation.

[Context]
Below is a dialogue between two roles: "Seeker" and "Supporter".
<Here is context of conversation>
Each Supporter message is labeled with a number, such as "Supporter[1]:", "Supporter[2]:", etc.
[Task]
For each Supporter[n], check whether it refers to any facts or details that were mentioned by the Seeker earlier. Process the dialogue in order. For each Supporter[n], only look at Seeker and Supporter messages before it. If Supporter[n] includes any reference to something the Seeker stated earlier, extract that reference. Rewrite the referenced content in a simple sentence starting with "Seeker", preserving original style as much as possible. If multiple items are used, join them with <Persona>. If nothing is referenced, output: "None".
[Output]
Output format(one line per message): Supporter[n]: Seeker... <Persona> Seeker... <Persona> ... (Or: Supporter[n]: None). Only output the result. No extra comments.

Figure 6: Prompt template used for persona annotation.

[Context]
Here is a conversation between a seeker and a supporter.
<Here is context of conversation>
[Task]
Based on the seeker's latrain message and conversation history, choose the most appropriate emotion from this list: [Neutral, Positive, Anxiety, Depression, Sadness, Anger, Fear, Disgust, Shame]]. Identify the seeker information from the conversation history that is relevant for crafting your response. Each piece of information should: Be written in third person (e.g., Seeker feels overwhelmed at work.). Be concise and relevant to the current message. Be separated by the tag '<Persona>' (e.g., Seeker feels stressed at work <Persona> Seeker is a nurse). If no relevant personal information can be extracted for the current response, output 'None' instead. Take the conversation, the seeker's emotion and the seeker's personal information into account. Role play as the supporter and generate a response from the first-person perspective.
[Output]
<think>emotion:(one emotion from the list) persona_info:(persona 1) <Persona> (persona 2) <Persona> ... OR None</think> <response>(keep your response concise and clear)</response>

Figure 7: Prompt template used for training and inference with Qwen2.5-1.5B-Instruct and Llama3.2-1B-Instruct. The structured prompt guides the model to perform emotion-persona-aware reasoning before generating responses.

[Context]
Here is a conversation between a seeker and a supporter.
<Here is context of conversation>
[Task]
Based on the conversation above, please role play as the supporter and generate a response from the first-person perspective.
[Output]
<response>(keep your response concise and clear)</response>

Figure 8: The prompt templates used for implement the prompt-based baseline(Qwen2.5-1.5B-Instruct, Llama3.2-1B-Instruct) on the ESConv dataset.

Seeker: Well, I'm really upset with my friends. **Supporter:** I am so sorry to hear that. Did something happen? **Seeker:** Yes, they aren't taking COVID seriously. **Supporter:** Oh that is so hard. I have family that isn't taking COVID seriously as well, and its so frustrating to see. Are you able to talk to them about your concerns? **Seeker:** Yes but they get angry with me. I don't know how to get through to them.
Cooper: I can tell you are really upset with your friends. I think it is important to get through to them and let them know that you are there for them. **Emstremo:** i would suggest talking to them about it, and see if they can understand. **MISC:** i can understand how you feel. i had a similar situation with my parents. they were very angry with each other and they were not understanding why. **TransESC:** i can understand how frustrating that would be. i would suggest talking to them in a more calm way. **Qwen2.5-1.5B-Instruct:** I understand it can be challenging to communicate with those who don't take the situation seriously. It might help if you could express your concerns calmly and respectfully, perhaps ...
PEARL-CoT (feat. Qwen): <think>emotion:Angry persona_info:Seeker doesn't know how to get through to their friends.</think><response>I understand it must be very difficult to try and convince people who disagree with you. Have you tried explaining why you think it is important?</response>
Ground-Truth: Some people express their frustrations in certain ways. Do you think there is another way to communicate with them? It's also okay to stand up for yourself and stand by what you feel is right.
Seeker: No I feel like she should come to me about it since she was caught. **Supporter:** What if she doesn't know you caught her? **Seeker:** I saw her do it. **Seeker:** She didn't really say a whole lot, she just left right afterwards.
Cooper: I would definitely suggest that she come to you about it and tell you exactly what she did. **Emstremo:** i think that would be a good idea. i would definitely ask her about it and see what she says. **MISC:** i can understand how that would be a difficult situation. i would definitely be upset too. **TransESC:** i think that would be a good idea. **Qwen2.5-1.5B-Instruct:** That's understandable. It sounds like she might not realize she's been caught yet. You could try sending her an anonymous message or email asking if everything is okay without giving away that you've seen her text ...
PEARL-CoT (feat. Qwen): <think>emotion:Sadness persona_info:Seeker saw their girlfriend texting another guy</think><response>I would encourage you to talk to her about it, but make sure not to push her away or try to scare her into telling you everything.</response>
Ground-Truth: I think it would definitely help to talk to her about it, it is never good to let situations like this build up without being talked about.

Table 8: Cases generated from baselines and PEARL-CoT (feat. Qwen). Blue highlights indicate key information from the dialogue history and corresponding mentions in the model responses. Red highlights mark vague, generic, or contradictory content.