Mitigating Strategy Preference Bias with Boundary-Aware Reward for Emotional Support Conversation

Anonymous ACL submission

Abstract

Emotional support conversation (ESC) aims to alleviate distress through empathetic dialogue, yet large language models (LLMs) face challenges in delivering effective ESC due to low accuracy in strategy planning. Moreover, there is a considerable preference bias towards specific strategies. Prior methods using fine-tuned strategy planners have shown potential in reducing such bias, while the underlying causes of the preference bias have not well been studied. In this work, we present an empirical analysis showing that strategy preference bias correlates with regions of low model confidence in strategy prediction. Based on this observation, we propose a boundary-aware reward to mitigate the bias by reinforcement learning, which optimizes strategy planning via both accuracy and entropy-based confidence for each region according to the estimated uncertainty. Experiments on the ESConv and ExTES datasets across multiple LLM backbones show that our approach consistently improves strategy selection accuracy while significantly reducing preference bias, without requiring external preference data or auxiliary modules.

1 Introduction

Emotional support conversation (ESC) aims to reduce distress through empathetic dialogue, supporting individuals facing personal challenges (Langford et al., 1997; Greene and Burleson, 2003; Heaney and Israel, 2008). Effective ESC not only requires nuanced strategy selection to generate contextually appropriate responses but also avoids low-quality responses that could lead to unintended psychological issues (Burleson, 2003). In recent years, large language models (LLMs), with their advanced conversational abilities, have been increasingly integrated into various dialogue systems (Jia et al., 2023; Lee et al., 2023). There is growing interest in harnessing LLMs for emotional support and professional counseling (Chen et al., 2023a; Zheng et al., 2023).

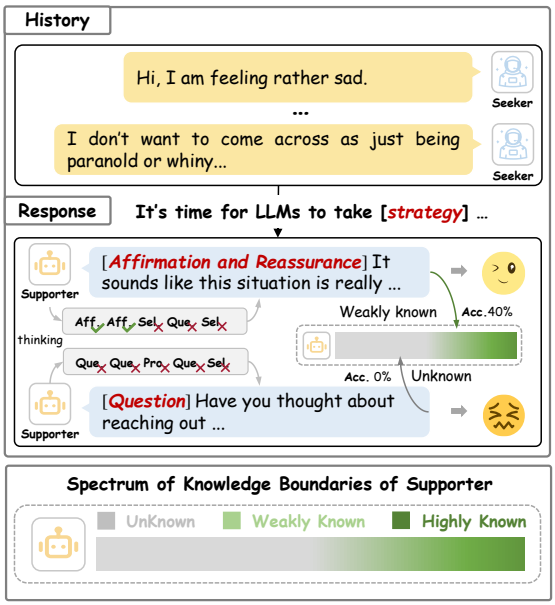


Figure 1: Examples of an LLM-based supporter with knowledge boundaries in emotional support. Weakly known areas reflect partial knowledge; unknown areas lie outside its correct knowledge boundaries.

ESC tasks involve both strategy selection and constrained response generation, where appropriate strategy choice is critical (Liu et al., 2021). Despite their strengths, LLMs often struggle to provide high-quality ESC due to limitations in strategy planning (Friedman et al., 2023; Chen et al., 2023b). First, LLMs exhibit low accuracy in strategy selection, frequently failing to identify contextually appropriate support strategies (Zhao et al., 2023; Farhat, 2024). Second, LLMs display a severe strategy preference bias, rigidly favoring certain general or familiar strategies over others (Kang et al., 2024; Zhao et al., 2025). Previous methods have introduced fine-tuned strategy planners to reduce bias in LLMs (Kang et al., 2024), or applied reinforcement learning with external preference datasets to improve strategy selection (Zhao et al.,

2025). Whereas, these methods rely on additional modules or external data to mitigate preference bias, and the underlying causes of strategy preference bias in LLMs have not well been studied.

Recent research has begun to investigate the limitations of LLMs by examining the internal knowledge distributions learned, which offers a new perspective to understand the origins of the preference bias. Previous studies show that the knowledge boundaries of LLMs are often ambiguous, encompassing substantial weakly known information (Gekhman et al., 2024) (depicted as the light green area in the spectrum, Figure 1). As a result, LLMs may struggle to confidently generate accurate response even when relevant information is present for the weakly known areas (Zhang et al., 2024b). In contrast, in the unknown areas out of the knowledge scope of LLMs (represented by the gray area in the spectrum, Figure 1), they usually exhibit overconfidence on the fabricated content (Zhang et al., 2024a). Furthermore, as LLMs incrementally learn new knowledge through unknown samples, their susceptibility to producing ungrounded content tends to increase (Gekhman et al., 2024). Inspired by these findings, we hypothesize that preference bias in ESC correlates with regions where LLMs exhibit low confidence in strategy prediction. As shown in Figure 1, the preference bias of strategies reflects a confidence calibration based on the strength of underlying knowledge. Therefore, we aim to approximate uncertainty regions aligned with strategy predictability of LLMs for ESC, and leverage this internal structure to mitigate strategy preference bias.

In this paper, we propose an approach that leverages uncertainty estimation to explicitly model model task-specific uncertainty regions aligned with strategy predictability in ESC, which improves strategy planning by mitigating the preference bias. We first identify and characterize the distinct regions within LLMs' knowledge distributions. Building on this, we employ reinforcement learning following supervised fine-tuning (SFT) with a dual reward function combining strategy accuracy and entropy confidence measures. This enables the model to optimize strategy planning across varying levels of knowledge certainty, prioritizing grounded responses while avoiding excessive dependence on overused strategies. Experiments on the ESConv and ExTES datasets demonstrate that our approach effectively improves the performance of LLMs in strategy planning for ESC. The main contributions of this paper are as follows:

- We provide an empirical analysis by linking task-specific uncertainty to strategy preference bias in ESC, and highlight the role of weakly known samples in bias mitigation.
- We propose an approach that models different knowledge regions with different rewards to balance strategy accuracy and diversity without requiring external preference data.
- We conduct extensive experiments and analyses on two benchmark datasets, demonstrating consistent improvements in both strategy proficiency and preference balance across multiple LLM backbones.

2 Related Work

2.1 Emotional Support Conversation

Emotional support conversation (ESC) involves interactions between a seeker experiencing emotional distress and a supporter aiming to alleviate it (Liu et al., 2021). Early approaches include global-to-local hierarchical graph networks to model dialogue context (Peng et al., 2022), incorporating commonsense knowledge for empathetic responses (Tu et al., 2022), and modeling emotions and semantics to enhance response relevance (Zhao et al., 2023). With the rise of LLMs, recent work leverages their conversational capabilities and apply SFT the LLM with ESC task specialized dialogues, which outperforms general-purpose LLMs (Chen et al., 2023b; Qiu et al., 2024). However, previous studies have suggested that they often struggle with selecting the correct strategy and a notable preference for a specific strategy (Kang et al., 2024). To make LLMs better emotional supporting, (Kang et al., 2024) use fine-tuned strategy planners reduce bias in both closed- and open-source LLMs and (Zhao et al., 2025) optimize strategy selection with external preference dataset. Unlike these methods, our approach do not require any additional plugs or data to improve LLMs' performance in strategy selection and response generation.

2.2 Knowledge Boundary

Knowledge boundaries of LLMs is critical for identifying their limitations in accurately expressing factual knowledge that they learned from the pretraining stage. While models are equipped with extensive parametric knowledge, some studies indicate their inability to discern the knowledge they

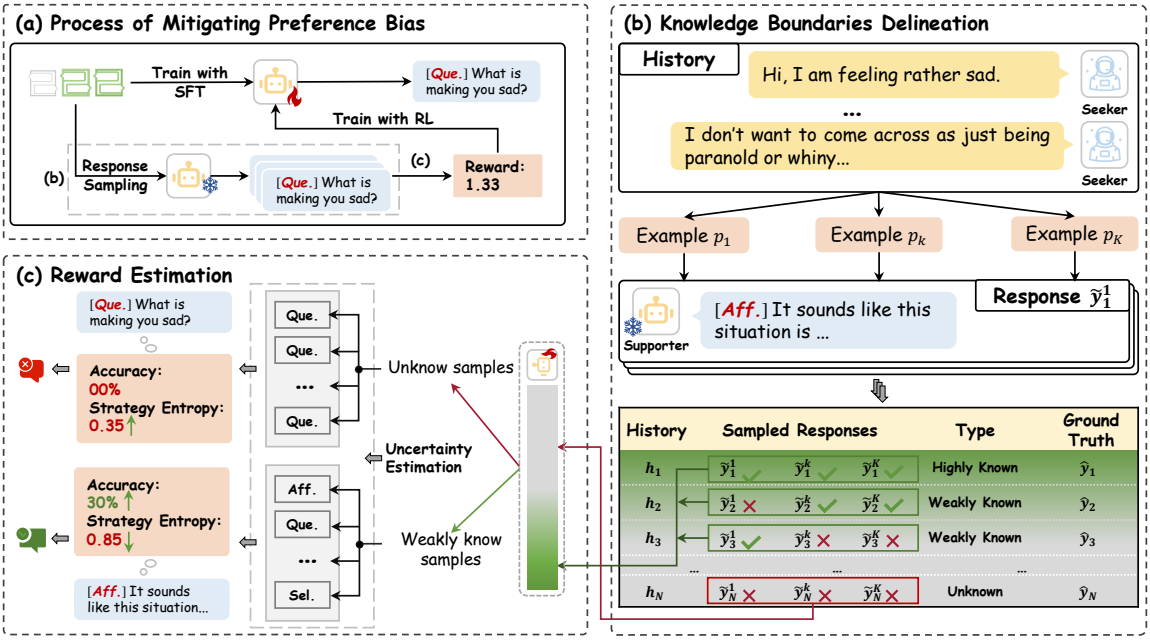


Figure 2: Overview of our approach. (a) illustrates the SFT and RL process for LLMs to enable strategy selection and empathetic response generation for ESC. (b) shows the delineation of knowledge boundaries into highly known, weakly known, and unknown knowledge areas. (c) presents the calculation of reward score, combining strategy selection accuracy and entropy confidence, tailored to distinct sample types relative to knowledge boundaries.

possess from what they lack, thus failing to articulate their knowledge boundary (Yin et al., 2023; Ren et al., 2025). To enhance a model's awareness of its knowledge boundary, prior work focus two areas: utilizing inherent knowledge to reduce the ratio of the model's "Unknown Knows" (Wei et al., 2022; Li et al., 2023; Tian et al., 2023) and acknowledging the knowledge it lacks to minimize the ratio of the model's "Unknown Unknows" (Zhang et al., 2024a; Yang et al., 2024). (Gekhman et al., 2024) find that LLMs struggle to acquire new factual knowledge through fine-tuning and that LLMs linearly increase the model's tendency to hallucinate by learning the examples with new knowledge. Focused on this aspect, our work investigates how to enable models to express the area of knows consistently, while also considering exploration uknows in knowledge boundaries.

3 Method

Figure 2 illustrates the overall procedure of our proposed approach. The framework for mitigating strategy preference bias is shown in Figure 2(a), which consists of two stages. In the first stage, we perform supervised fine-tuning (SFT) on a pre-trained LLM using the full dataset. In the second stage, reinforcement learning (RL) is applied to the finetuned model using different samples categorized according to the model's knowledge boundaries. Details about the ESC task and our method are presented as follows.

3.1 Problem Formulation

Following prior work (Liu et al., 2021), the effectiveness of ESC by LLMs largely depends on the selection of an appropriate support strategy. We formalize the ESC task as a generation problem for a pre-trained LLM $\mathcal{M}$, where the model directly produces a response that incorporates the intended support strategy. Given a flattened dialogue history h, the model generates a response $\tilde{y}$ consisting of a special strategy$s$ followed by the actual utterance y, formalized as $\tilde{y} = s \oplus y \sim P(\tilde{y} \mid h; \theta_{\mathcal{M}})$, where s corresponds to a predefined strategy in $S = \{s_1, s_2, \cdots, s_n\}$, and $\theta_{\mathcal{M}}$ denotes the parameters of the model. This formulation allows the model to implicitly select and express a strategy as part of the response generation process. As ESC is a strategy-centric task, (Kang et al., 2024) suggest that a good supporter LLM should satisfy two key properties: **Proficiency**, the ability to select appropriate strategies; and **Preference**, the ability to avoid overusing specific strategies.

3.2 Knowledge Boundaries Delineation

To categorize samples relative to the model's knowledge boundary, we define three regions: *highly known*, *weakly known*, and *unknown*. As illustrated in Figure 2(b), we estimate the location

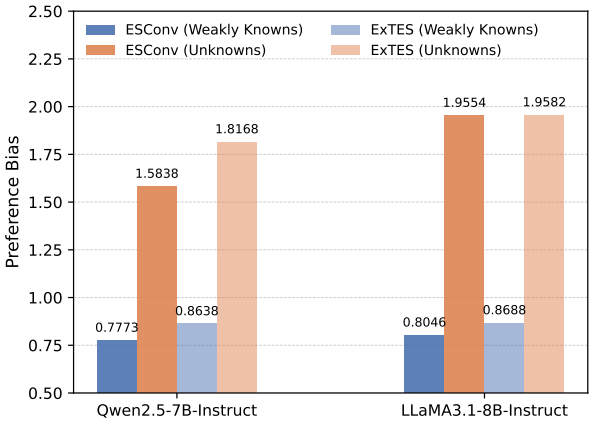


Figure 3: Preference bias separately within the weakly known and unknown regions.

of each dialogue history h_i within this boundary by analyzing the model's output probabilities. We emphasize that the proposed delineation does not aim to recover true knowledge boundaries of LLMs. Instead, it provides a pragmatic partition of dialogue histories into uncertainty regions based on observable strategy prediction consistency.

Response Sampling. To capture the variability in the model's behavior, we generate multiple responses for each dialogue history. Given a dataset $\mathcal{D} = \{(h_i, \hat{y}_i)\}_{i=1}^{N}$, where $\hat{y}_i = s_i \oplus y_i$ denotes the ground-truth strategy s_i and corresponding response content y_i, we employ few-shot prompting to balance response accuracy and diversity while mitigating prompt sensitivity. The prompt set P includes K distinct one-shot examples to facilitate in-context learning.

For each dialogue history h_i, the k-th sampling iteration combines a few-shot example $p_k \in P$ with h_i to prompt the model $\mathcal{M}$, producing the k-th response $\tilde{y}_i^k$. Repeating this process K times yields a response set $Y_i = \{\tilde{y}_i^k\}_{k=1}^{K}$. Corresponding labels $Z_i = \{z_i^k\}_{k=1}^{K}$ are assigned by comparing each generated strategy within response $\tilde{y}_i^k$ with the ground-truth $\hat{y}_i$, where $z_i^k \in 0, 1$ (1 indicates a correct response, and 0 indicates an incorrect one). This results in a data tuple $(h_i, Y_i, Z_i, \hat{y}_i)$ to calculate response accuracy subsequently.

Deriving Knowledge Categories. To assess the model's proficiency for each dialogue history h_i, we compute an accuracy-based confidence score c_i, defined as the proportion of responses with the correct strategy:

$$c_i = \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}(s_i^k = s_i), \tag{1}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Based on c_i, we categorize each example as follows:

- $c_i = 1$: Highly known samples, where all generated strategies are correct.
- $c_i = 0$: Unknown samples, where no generated strategies are correct.
- $0 < c_i < 1$: Weakly known samples, exhibiting partial correctness in strategy predictions.

Strategy Preference Bias. Having categorized dialogue histories into different knowledge boundary regions, we next investigate how strategy preference bias manifests within these regions. We adopt the Bradley–Terry model (Bradley and Terry, 1952), which has been widely used for preference estimation. To capture the overall imbalance of strategy usage, the *preference bias* $\mathcal{B}$ is defined as the standard deviation of preference scores (Kang et al., 2024), where a larger preference bias reflects a stronger skew toward specific strategies.

By computing preference bias separately within the weakly known and unknown regions, we empirically test our hypothesis that strategy preference bias is more severe in low-confidence regions. As shown in Figure 3, our results show that $\mathcal{B}$ is significantly higher in the unknown regions compared to the weakly known region, confirming that knowledge boundaries play a crucial role in the formation of strategy bias, which provides a solid foundation for the boundary-aware mitigation approach.

3.3 Mitigating Preference Bias

Supervised Fine-tuning. SFT is a widely used approach to adapt LLMs to specific tasks. To improve the proficiency of the LLM $\mathcal{M}$ in ESC, we first perform SFT using a curated dataset $\mathcal{D} = \{(h_i, \tilde{y}_i)\}_{i=1}^{N}$ consisting of dialogue history-response pairs. To ensure the model clearly understands the task, we prepend both a task description and a strategy description as a part of the prompt. Additionally, beyond the dialogue history, we include dialogue background, which encompasses the seeker's problem, emotion, and situation gathered from a pre-chat survey. The model is finetuned by minimizing the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}} = -\sum_{i=1}^{N} \log P_{\theta_{\mathcal{M}}}(\tilde{y}_i \mid h_i). \tag{2}$$

This objective encourages the model to generate responses containing appropriate emotional support strategies.

However, there are some examples that introduce new knowledge for LLMs, and LLMs struggle to acquire new knowledge through fine-tuning, as fine-tuning examples that introduce new knowledge are learned lower than those consistent with the model's knowledge. Moreover, as the examples with new knowledge are learned, it becomes prone to generating off-strategy responses. In the following subsections, we present our RL approach incorporating strategy accuracy-based and entropy-based rewards to reduce this limitation.

Reinforcement Learning. Inspired by recent insights on knowledge boundaries, we propose adapting dynamic rewards to different regions of the model's knowledge boundary. Unlike prior methods that apply a uniform reward function across all samples, our approach directs the model toward distinct learning objectives based on samples falling within different knowledge regions.

Our goal is to guild the LLM's generation toward desirable responses by maximizing an alignment objective $\mathcal{R}$, which can take various forms such as format measures. Formally, we aim to maximize the expected reward:

$$\mathbb{E}_{h\sim\mathcal{D},\tilde{y}\sim P_{\theta_{\mathcal{M}}}(\cdot|h)}[\mathcal{R}(h,\tilde{y})]. \tag{3}$$

We formulate this as an RL problem. We use the fine-tuned model to initialize the policy network as $\pi_0 = P_{\theta_{\mathcal{M}}}$ and update it using group relative policy optimization (GRPO). The token generation process for a response $\tilde{y}$ can be seen as a Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, r, \mathcal{P} \rangle$ with a state space $\mathcal{S}$, action space $\mathcal{A}$, reward function r, and state-transition probability $\mathcal{P}$. At each time step t, the agent (LLM) selects the next token from the vocabulary $\mathcal{V}$ based on the current policy network $\pi(\tilde{y}|h, \tilde{y}_{<t})$. The episode ends upon generation an end-of-sequence token, producing the complete response $\tilde{y}$. We can fine-tune the policy network π by maximizing the expected reward r:

$$\mathbb{E}_{\pi}[r] = \mathbb{E}_{h\sim\mathcal{D},\tilde{y}\sim\pi(\cdot|h)}[r(h,\tilde{y})]. \tag{4}$$

Recall that our goal is to maximize the proficiency of the model $\mathcal{M}$ and minimize its preference bias. For each input dialogue history h_i, we seek high confidence correct predictions, while also allowing for controlled exploration via alternative responses. To this end, we propose a reward estimation approach that dynamically adjusts the reward function r according to the model's estimated knowledge boundary for each input h_i.

3.4 Reward Estimation

The confidence score c_i serves as a reward signal to encourage the model toward higher strategy selection proficiency. To further promote consistency, we incorporate an entropy-based metric that quantifies the diversity of strategies expressed across the sampled responses. Specifically, we estimate the empirical distribution over strategies by computing the relative frequency of each strategy $s \in S$:

$$p(s \mid h_i) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{1}(s_i^k = s), \tag{5}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The corresponding strategy entropy e_i is calculated as:

$$e_i = -\sum_{s\in S} p(s \mid h_i) \log p(s \mid h_i), \tag{6}$$

which measures the degree of strategy variation in the model's outputs. To tailor reward optimization to different regions of the model's knowledge boundary, we define region-specific reward functions.

For *highly known* and *weakly known* samples, where the model exhibits full or partial proficiency, we prioritize consistency by a reward that emphasizes low diversity:

$$r_{known}(h_i, \tilde{y}_i) = 1 - \frac{e_i}{\log(|S|)}, \tag{7}$$

where $|S|$ is the number of predefined strategies. This encourages the model to converge toward the correct strategy with minimal variability.

Conversely, for *unknown* samples, where the model lacks proficiency, we encourage exploration to discover potentially correct responses. The reward function for these samples promotes high diversity:

$$r_{unknow}(h_i, \tilde{y}_i) = \frac{e_i}{\log(|S|)}, \tag{8}$$

which incentivizes the generation of a wider range of strategies that may lead to discovery of effective responses.

To keep the policy network π from moving too far from the initial model $P_{\mathcal{M}}$, we also add a KL-divergence penalty reward. Therefore, the final reward becomes:

$$r(h_i, \tilde{y}_i) = c_i + r_{region}(h_i, \tilde{y}_i) - \beta \log \frac{\pi(\tilde{y}_i|h_i)}{P_{\theta_{\mathcal{M}}}(\tilde{y}_i|h_i)}, \tag{9}$$

where r_{region} is either r_{known} or r_{unknow} based on the sample's classification, and the hyperparameter β is the coefficient of KL-penalty. This composite reward structure ensures that the model balances proficiency, consistency, and exploration while maintaining stability during optimization.

4 Experimental Setup

4.1 Datasets

We conduct experiments on two emotional support datasets: ESConv (Liu et al., 2021) and ExTES (Zheng et al., 2024b). **ESConv** collected through crowdworkers acting as help-seekers and supporters, categorizes emotional support strategies into 8 distinct types. **ExTES** generated iteratively using ChatGPT, leverages dialogue examples from authoritative emotional support datasets, which classify strategies into 16 types. For ESConv, we utilize the training set introduced by (Liu et al., 2021) for model training and knowledge boundary delineation. For evaluation, we employ the test set proposed by (Kang et al., 2024). For ExTES, we adopt an 8:2 split for training and test sets, respectively, and follow the test set construction methodology outlined in (Kang et al., 2024) for the test set to ensure consistency and comparability.

4.2 Evaluation Metrics

As described in prior studies (Kang et al., 2024), a good supporter LLM should satisfy two properties: Proficiency and Preference. For proficiency, we focus on strategy selection accuracy, measured by **macro F1** score $\mathcal{Q}$ and **weighted F1** score $\mathcal{Q}_{\mathcal{W}}$. For preference, we select **strategy preference bias** $\mathcal{B}$ as metric, which quantifies the deviation between the model's selected strategies and the ideal strategy preferences. Additionally, **ROUGE-L** (R-L) is also adopted to evaluate the semantic aspects of the generated response.

4.3 Baselines

To evaluate the effectiveness of our approach, we benchmark it against several baselines that designed to mitigate preference biases in LLMs using self-contact techniques. We implement three self-contact methods: (1) *Direct-Refine*, where the LLM iteratively refines its initial response; (2) *Self-Refine* (Madaan et al., 2023), which improves the initial response through self-generated feedback; and (3) *Few-Shots*, which samples some example from training sets to guide response generation. Given the prevalence of SFT in enhancing ESC capabilities, we also compare against full-parameters SFT models. Additionally, we evaluate a baseline using GRPO (Shao et al., 2024) with rewards based on response format and correctness to isolate the impact of general RL tuning from our region-aware reward design.

4.4 Implementation Details

We implement our method on LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Team, 2024). For SFT, we train the model via LLaMA-Factory (Zheng et al., 2024a) with full parameters. The input format follows the official chat templates. The models are trained for 2 epochs with a learning rate of $1 \times e^{-5}$. For RL, we train the fine-tuned model via VERL (Sheng et al., 2024) for 300 episodes, a batch size of 256, and a learning rate of $1 \times e^{-6}$. The KL-penalty coefficient is set to 0.001. All experiments are conducted on 2 NVIDIA Tesla A800 GPUs. We set maximum input length to 1024 tokens and maximum target length of 256 tokens across all backbones.

5 Results and Analyses

5.1 Main Results

Table 1 presents the performance comparison between our proposed method and several baselines. Our approach consistently improves strategy selection accuracy while effectively reducing strategy preference bias across all evaluated models. Specifically, while SFT and GRPO without explicit preference optimization enhance strategy selection accuracy, they simultaneously increase strategy preference bias, limiting their adaptability to seeker's diverse emotional needs. In contrast, our approach mitigates preference bias while enhancing accuracy, underscoring the importance of incorporating preference-aware optimization in ESC tasks. Moreover, Our method also enhances ESC performance across all backbone models, regardless of their initial capabilities. It enhances strategy accuracy and reduces bias in both Qwen2.5-7B-Instruct and LLaMA-3.1-8B-Instruct, demonstrating its versatility and robustness.

Our results also highlight the limitations of self-contact methods, which cannot achieve a steady increase in strategy accuracy. Although their strategy bias is low, they cannot accurately select the right strategy. This is slightly different from previous work (Kang et al., 2024), which have shown

Method	ESConv				ExTES			
	$\mathcal{Q}\uparrow$	$\mathcal{B}\downarrow$	$\mathcal{Q}_{\mathcal{W}}\uparrow$	**R-L** $\uparrow$	$\mathcal{Q}\uparrow$	$\mathcal{B}\downarrow$	$\mathcal{Q}_{\mathcal{W}}\uparrow$	**R-L** $\uparrow$
Qwen-2.5-7B-Instruct								
Standard Prompting	8.13	2.22	8.49	13.56	13.57	3.87	24.74	20.15
Few-Shot (2-shot)	12.41	1.19	13.84	13.47	11.78	3.83	20.75	20.41
Direct-Refine	12.51	1.22	14.55	13.78	13.31	3.86	24.38	19.75
Self-Refine	13.36	1.16	15.38	13.54	11.81	3.86	20.98	19.75
SFT	22.10	1.79	22.80	19.07	40.48	1.06	57.74	27.78
GRPO	**26.26**	1.73	26.46	19.12	45.96	0.65	61.27	28.62
Ours	26.16	**1.12**	**28.24**	**19.86**	**47.51**	**0.64**	**63.79**	**29.60**
LLaMA-3.1-8B-Instruct								
Standard Prompting	7.59	2.34	7.96	11.76	6.84	3.85	11.42	16.26
Few-Shot (2-shot)	12.61	1.81	12.00	12.98	11.98	3.81	19.29	19.52
Direct-Refine	9.45	1.69	10.04	12.41	6.75	3.85	11.51	16.82
Self-Refine	10.87	1.36	11.89	12.97	6.89	3.86	11.02	17.79
SFT	21.79	1.30	23.67	15.70	41.34	0.83	57.22	36.06
GRPO	22.69	1.27	25.74	15.79	45.83	0.75	60.04	37.69
Ours	**24.27**	**1.03**	**27.37**	**15.84**	**48.79**	**0.73**	**64.74**	**39.84**

Table 1: Automatic evaluation results on the ESConv and ExTES datasets using Qwen-2.5-7B-Instruct and LLaMA-3.1-8B-Instruct as backbones. $\uparrow$ indicates higher is better, $\downarrow$ indicates lower is better. The best results are highlighted in **bold**. The second-best results are underlined.

Ours vs. **GRPO**	Win	Lose	Tie	k
Acceptance	41.33	26.00	32.66	0.611
Effectiveness	38.67	27.33	34.00	0.650
Sensitivity	40.00	26.00	34.00	0.669
Satisfaction	41.33	24.67	34.00	0.729

Table 2: Human evaluation results. Cohen's kappa (k) indicates inter-annotator agreement.

that self-contact methods can cause increased policy bias and decreased policy accuracy.

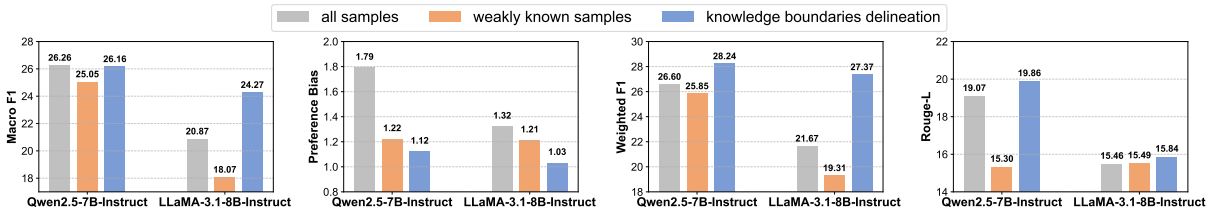
5.2 Human Evaluation Results

Table 2 presents the human evaluation results comparing our method with GRPO on 50 dialogue samples from the ESConv dataset. Three annotators rated each response pair across four criteria: *Acceptance*, *Effectiveness*, *Sensitivity*, and *Satisfaction*. Our method consistently outperforms GRPO, achieving higher win rates across all dimensions, ranging from 38.67% to 41.33%. These results suggest that our responses are generally perceived as more satisfying by human evaluators. The inter-annotator agreement ranges from 0.611 to 0.729, indicating substantial agreement among annotators. This reinforces the robustness of our method in improving emotional support quality.

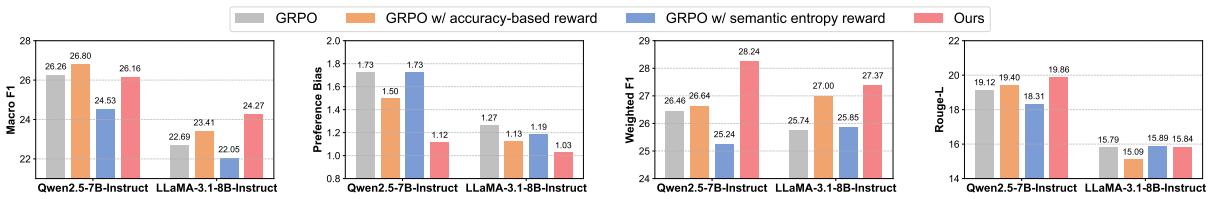
5.3 Ablation Study

Effect of Knowledge Boundary Delineation. We conduct an ablation study to assess the effectiveness of knowledge boundary delineation. Specifically, we compare three configurations: (1) our full method with dynamic rewards based on knowledge regions, (2) a baseline without boundary awareness, and (3) a variant that trains on weakly known samples using uniform rewards, simulating a coarse partition. All models are trained via SFT on the ESConv dataset and then optimized using RL with identical hyperparameters. As illustrated in Figure 4(a), our approach outperforms both variants across all backbones, achieving higher strategy accuracy and lower preference bias. Interestingly, training solely on weakly known samples reduces preference bias more effectively than the baseline, highlighting their value in reducing model overconfidence. However, this variant still underperforms our full method, likely due to its inability to leverage the exploratory potential of unknown-region examples. These results underscore the value of fine-grained knowledge boundary delineation in guiding reward design and improving alignment efficiency in ESC.

Effect of Uncertainty Estimation Reward We further conduct an ablation study to evaluate the impact of our two uncertainty-based reward components during the RL stage. Specifically, we compare: (1) a baseline using standard format and correctness-based rewards, (2) an accuracy-based reward that reflects the model's confidence alignment with ground-truth labels, and (3) a entropy-



(a) Performance comparison of different backbones using various samples of knowledge boundaries from the ESConv dataset.



(b) Performance of reward functions varying different uses of uncertainty measures on the ESConv dataset.

Figure 4: Effects of knowledge boundary delineation and uncertainty estimation rewards on ESConv.

Data Type	$\mathcal{Q}\uparrow$	$\mathcal{B}\downarrow$	$\mathcal{Q}_{\mathcal{W}}\uparrow$	**R-L** $\uparrow$
Qwen-2.5-7B-Intruct				
Full Dataset	22.10	1.79	22.80	19.07
Full Dataset(Ours)	26.16	1.12	28.24	19.86
Weakly Knows(Ours)	20.91	1.22	22.19	15.30
Weakly Knows(*Prompt*)	18.16	1.35	20.45	15.12
LLaMA-3.1-8B-Intruct				
Full Dataset	21.79	1.30	23.67	15.70
Full Dataset(Ours)	24.27	1.03	27.37	15.84
Weakly Knows(Ours)	18.07	1.11	19.31	15.89
Weakly Knows(*Prompt*)	17.19	1.27	19.30	15.50

Table 3: Performance metrics across different data types. *Prompt* refers a prompt-based uncertainty estimation.

based reward that quantifies strategy diversity via the entropy of the predicted strategy distribution. All models are fine-tuned using SFT on the ESConv dataset, followed by RL under identical settings. As shown in Figure 4(b), the accuracy-based reward consistently improves performance across all evaluation metrics for both LLMs. In contrast, using entropy alone leads to less stable performance, sometimes degrading specific metrics. Notably, combining both accuracy and entropy achieves the best overall performance, demonstrating the complementary nature of the two signals. These results validate the effectiveness of our dual uncertainty-guided reward design in enhancing strategy accuracy and reducing bias in ESC.

5.4 Analysis

Effect of Weakly Known Samples We analyze the effect of training on weakly known samples identified via uncertainty estimation. Specifically, we compare four configurations: (1) full dataset SFT, (2) our method applied to the full dataset, (3) SFT with only weakly known samples selected by our accuracy based estimator, and (4) a prompt based method where weakly known samples are self-reported by the model. As shown in Table 3, using weakly known samples consistently reduces strategy preference bias across all backbones. These findings indicate that weakly known samples can mitigate bias effectively. However, this reduction in bias comes at the cost of strategy proficiency. This trade-off likely stems from the reduced quantity of training examples when filtering for only weakly known samples. Notably, our accuracy-based method outperforms the prompt-based approach in both proficiency and bias reduction, demonstrating the reliability of our uncertainty estimation.

6 Conclusion

In this paper, we propose an approach to identify the knowledge boundaries of LLMs for the ESC task, and model different regions with a dual reward function combining strategy accuracy and entropy confidence measures for strategy planning. Extensive experiments on ESConv and ExTES datasets show that our approach outperforms the baselines in both proficiency and bias reduction. In particular, the results underscore the importance of weakly known samples to enhance preference bias mitigation. In the future, we will extend our approach to other dialogue tasks, and refine knowledge boundary definitions to adapt to downstream tasks.

Limitations

In real-world emotional support interactions, user needs and emotional states often evolve dynamically across conversations. While the datasets used in this work rely on predefined strategy annotations, which may not fully reflect the diversity and complexity of real-world emotional support interactions. Future work could incorporate more flexible or naturally occurring emotional support scenarios to better approximate real-world settings.

References

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

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

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

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

Faiza Farhat. 2024. Chatgpt as a complementary mental health resource: a boon or a bane. *Annals of Biomedical Engineering*, 52(5):1111–1114.

Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, et al. 2023. Leveraging large language models in conversational recommender systems. *arXiv preprint arXiv:2305.07961*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784.

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

John O Greene and Brant R Burleson. 2003. Handbook of communication and social interaction skills.

Catherine A Heaney and Barbara A Israel. 2008. Social networks and social support. *Health behavior and health education: Theory, research, and practice*, 4(1):189–210.

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

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

Catherine Penny Hinson Langford, Juanita Bowsher, Joseph P Maloney, and Patricia P Lillis. 1997. Social support: a conceptual analysis. *Journal of advanced nursing*, 25(1):95–100.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. 2023. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

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. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4324–4330. ijcai.org.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt

for mental health support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3697–3715, Abu Dhabi, UAE. Association for Computational Linguistics.

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

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*.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. In *The Twelfth International Conference on Learning Representations*.

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.

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.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. *Advances in Neural Information Processing Systems*, 37:63565–63598.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'i don't know'. In *2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, pages 7106–7132. Association for Computational Linguistics (ACL).

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024b. Self-alignment for factuality: Mitigating hallucinations in llms via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965.

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

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

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024a. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

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

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