From Generic Empathy to Personalized Emotional Support: A Self-Evolution Framework for User Preference Alignment

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

Abstract

Effective emotional support hinges on understanding users' emotions and needs to provide meaningful comfort during multi-turn interactions. Large Language Models (LLMs) demonstrate significant potential for expressing empathy; however, they often provide generic and one-size-fits-all responses that fail to address users' specific needs. To address this issue, we propose a self-evolution framework designed to help LLMs enhance their responses, aligning them more closely with users' implicit preferences regarding user profiles (personalities), emotional states, and specific situations. Our framework consists of two distinct phases: *(1) Emotional Support Experience Acquisition*, where LLMs are fine-tuned on limited emotional support conversation data to provide basic support, and *(2) Self-Improvement for Personalized Emotional Support*, where LLMs leverage self-reflection and self-refinement to generate personalized responses. Through iterative direct preference optimization between the pre- and post-refined responses, our model generates responses that reflect a better understanding of the user's implicit preferences. Extensive experiments and evaluations demonstrate that our method significantly enhances the model's performance in emotional support, reducing unhelpful responses and minimizing discrepancies between user preferences and model outputs.

1 Introduction

Emotional support conversation (ESC) systems require a deep understanding of users' emotions (Wang et al., 2024a) and need to provide meaningful comfort and assistance during multi-turn interactions (Peng et al., 2022; Rains et al., 2020), which are vital in practical applications such as mental health care, emotional companionship, and customer service. Given that each user has unique

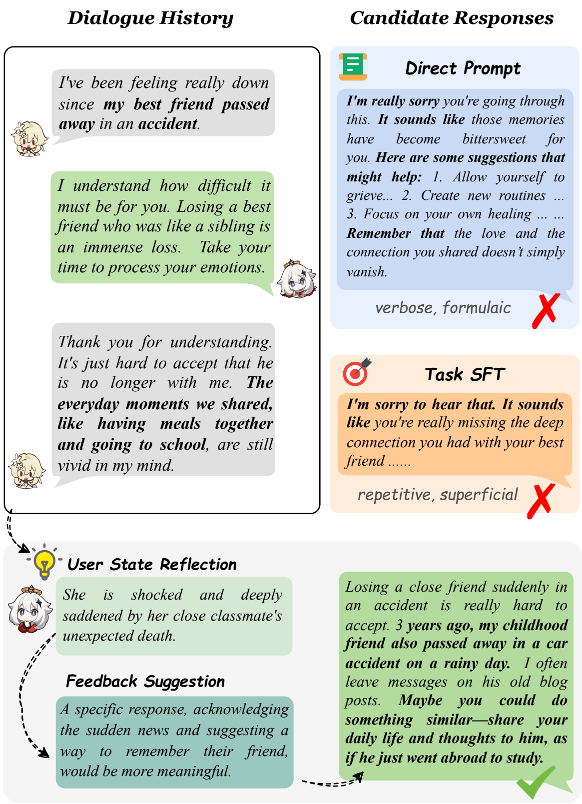


Figure 1: Example responses. *Direct prompting of LLaMA* results in verbose and formulaic outputs. *Task-Specific SFT* is empathetic but often lacks depth and variety, giving it a perceived "AI-like" quality. In contrast, *self-reflection* on user preferences provides a pathway to more specific and engaging responses.

emotional needs and experiences (Rogers, 2013), delivering personalized and contextually appropriate emotional support is essential for ensuring practical assistance (Campos et al., 2018; Cheng et al., 2023; Lin et al., 2022).

Despite the promising potential of LLMs for generating empathetic responses (Touvron et al., 2023; Yang et al., 2024; Achiam et al., 2023), they often struggle to provide diverse and contextually appropriate support (Wang et al., 2024b). As illustrated in Figure 1, **direct prompting LLMs often results in superficial empathy, verbosity, and formulaic structures**. A simple yet effec-

*Corresponding Author

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18826–18853
November 4-9, 2025

tive approach is to supervise fine-tuning (SFT) of LLMs on ESC corpora (Zheng et al., 2024b, 2023; Qiu et al., 2023; Zhang et al., 2024). However, SFT relies on substantial, high-quality ESC data, which is often scarce and difficult to acquire. Moreover, **over-reliance on SFT can lead to repetitive patterns that express empathy overtly but lack depth and variety** (Irvine et al., 2023). As demonstrated in Figure 1 and 6, SFT models can fall into predictable patterns, frequently using phrases like "*It sounds like...*" or "*I'm sorry to hear that...*"

Recent insights highlight that LLMs can self-improve their performance through self-reflection and self-refinement guided by human-designed principles (Lu et al., 2024; Madaan et al., 2023; Ye et al., 2023; Yasunaga et al., 2024). Inspired by these findings, we pose the intriguing question: *Can LLMs be taught to consider what kind of responses are genuinely needed by users, and can this reflective process lead to refined and more personalized responses?*

This work seeks to bridge the gap between generic empathetic responses and truly user-centered personalized emotional support by incorporating self-reflection and self-refinement into automated systems. Effective ES systems require an iterative approach that continuously reflects ongoing dialogue to refresh user understanding and refine responses, ultimately delivering targeted empathy and tailored solutions. The empirical evidence presented in Figure 1 and Table 3 demonstrates that instructing LLMs to summarize user situations, infer emotions and causes, and choose appropriate support strategies leads to a significant improvement in response quality.

To this end, we introduce a self-evolution framework for aligning user preferences. As depicted in Figure 2, our self-evolution framework comprises two steps: (1) *Emotional Support Experience Acquisition*: we first fine-tune LLMs on limited ESC data, enabling them to provide essential emotional support. (2) *Self-Improvement for Personalized Emotional Support*: Subsequently, we leverage LLMs' inherent self-reflection and self-refinement capabilities to generate responses that consider the implicit user preference, including *profile*, *situation*, and *emotions*. The pre- and post-refined responses are considered the preference data. Through direct preference optimization, the model generates responses that reflect an understanding of the user's implicit preferences during interactions, thereby eliminating the need for explicit reflection and refinement steps. Experimental results and extensive human evaluations indicate that our generated responses are more diverse and better aligned with user input. These improved responses effectively reduce ineffective empathy and preference misalignment, facilitating more productive multi-turn interactions.

Our main contributions can be summarized as follows:

- We reveal the limitations of the current Emotional Support Chatbot, which is notably deficient in understanding users' implicit preferences, resulting in repetitive and superficial expressions of empathy.
- We present a simple yet effective self-evolution framework for personalized emotional support that eliminates the need for explicit reflection and refinement steps.
- Experimental results and comprehensive human evaluations demonstrate that our method effectively minimizes unhelpful responses and discrepancies in personalized preferences.

2 Method

Inspired by recent insights highlighting that LLMs can self-improve through language feedback (Lu et al., 2024; Madaan et al., 2023; Ye et al., 2023; Yasunaga et al., 2024), we present a self-evolution framework designed to enable LLMs to provide personalized emotional support. This framework operates in two phases: Emotional Support Experience Acquisition (Section 2.1) and Self-Improvement for Personalized Emotional Support (Section 2.2).

2.1 Emotional Support Experience Acquisition

2.1.1 Task Definition

Emotional Support (ES) involves understanding the user's situation and choosing appropriate supportive strategies to alleviate their distress. Formally, we can denote the ES model as $\mathcal{M}$, and represent the current dialogue context as $\mathcal{C}_n = (q_1, r_1, ..., q_{i-1}, r_{i-1}, ...q_n)$. In this representation, q_i and r_i correspond to the i-th utterance from the user and the model, respectively. Given the task and strategy description prompt $\mathcal{P}_{task}$, the goal of the ES model is to generate an emotionally supportive response r_n, which can be represented as:

$$r_n = \mathcal{M}(\mathcal{P}_{task} \,||\, \mathcal{C}_n) \quad (1)$$

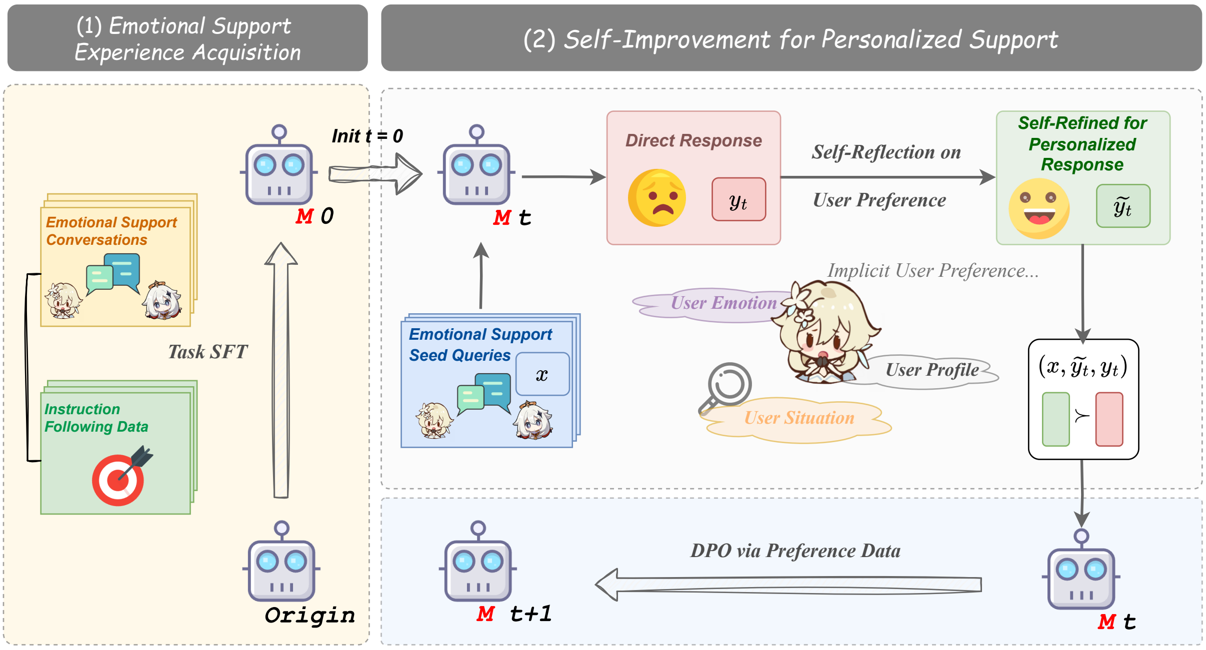


Figure 2: The overview of our self-evolution framework, which enhances personalized emotional support capabilities through a two-stage learning phase: (1) Emotional Support Experience Acquisition: We fine-tune LLMs on minimal human-annotated ESC data, equipping them with basic emotional support capability. (2) Self-Improvement for Personalized Emotional Support: We utilize the LLMs' self-reflection abilities to tailor responses to the user's personality, situation, and emotions. The pre- and post-refined responses are natural synthetic preference data. The process involves iterative preference optimization for generating responses that align with the user's implicit preferences, eliminating the need for explicit reflection steps.

2.1.2 Task Learning

We equip the model with emotional support capability by fine-tuning the backbone on the manually annotated ESConv dataset (Liu et al., 2021). To preserve the general abilities, we employ Low-Rank Adaptation (Hu et al., 2021), fine-tuning only the LoRA adapter parameters. We further incorporate a replay mechanism by incorporating some instruction-following data (Wang et al., 2024c). The model is trained using the SFT loss:

$$\mathcal{L}_{SFT} = -\log P(\boldsymbol{y}|\boldsymbol{x}, \mathcal{P}; \theta) \tag{2}$$

where $\boldsymbol{x}, \boldsymbol{y}$ represent the input and output of the model, respectively, while $\mathcal{P}$ denotes the task description or instructions. The resulting fine-tuned model is denoted as $\mathcal{M}^0$.

2.2 Self-Improvement for Personalized Emotional Support

While SFT improves empathetic response generation, it often produces superficial outputs, failing to capture nuanced user preferences that are crucial for effective emotional support. To address this limitation and minimize unhelpful responses, we introduce a self-improvement method based on iterative direct preference optimization (DPO) (Rafailov et al., 2023). Guided by human design principles, the model considers the user's personality, situation, and emotions to refine its responses. These pre- and post-refined responses naturally serve as rejected and chosen candidates, respectively (Dong et al., 2024). Through direct preference optimization, the model generates responses that reflect an understanding of the user's implicit preferences during interactions, thereby eliminating the need for explicit reflection and refinement steps.

2.2.1 Synthetic Preference Data Generation

Rejected Response Generation Constructing high-quality preference data pairs requires a diverse set of user queries. While synthetic ESC datasets may not produce emotional support responses of comparable quality to humans, they offer a valuable source of varied queries (Zheng et al., 2024b, 2023). We extract the dialogue context $\mathcal{C}_n$ from these synthetic datasets, where n is the turn index, and employ $\mathcal{M}^t$ to generate responses.

$$y_n^t = \mathcal{M}^t(\mathcal{C}_n), \quad \text{initial } t = 0 \tag{3}$$

These unrestricted responses are treated as rejected responses.

Self-Reflection on Implicit User Preference Research indicates that LLMs possess strong contextual inference capabilities (Yang et al., 2024; Dubey et al., 2024), enabling them to infer user emotions, implicit profiles, and even personality from ongoing conversations. Given the dialogue history $\mathcal{C}_n$ and human-designed principles $\mathcal{I}$, the model $\mathcal{M}^t$ is tasked with summarizing the user's profile u_n and current emotional state s_n according to the following equation:

$$(u_n, s_n) = \mathcal{M}^t(\mathcal{I} \,||\, \mathcal{C}_n) \tag{4}$$

u_n and s_n are continuously updated throughout the conversation, enabling the model to refine its understanding of the user.

Self-Refinement for Personalized Responses Responses generated solely from dialogue history often fail to capture the user's implicit preferences. Drawing on insights from psychological research, user preferences can be decomposed into two key dimensions: long-term traits, encapsulated by the user profile (Fleeson, 2001), and context-sensitive emotional needs (**?**). To better understand and adapt to these implicit preferences, we leverage the strong contextual reasoning of LLMs (Yang et al., 2024; Dubey et al., 2024). Given the dialogue history $\mathcal{C}_n$ and human-designed principles $\mathcal{I}$, the model $\mathcal{M}^t$ is tasked with summarizing the user's profile p_n and current emotional state s_n according to the following equation:

$$\tilde{y_n^t} = \mathcal{M}^t(\mathcal{I} \,||\, \mathcal{C}_n, u_n, s_n, y_n^t) \tag{5}$$

The pre- and post-refined responses form a preference pair $(y_t, \tilde{y_n^t})$, where y_t serves as the rejected candidate and $tildey_n^t$ serves as the chosen candidate, respectively. Inevitably, some low-quality data is generated during this process; the data filtering process is detailed in Appendix A, and the prompts for self-reflection and self-refinement are shown in Figure E.1.

2.2.2 Preference Optimization

The synthetic preference data generation process naturally facilitates iterative self-improvement. In each iteration, we employ DPO (Rafailov et al., 2023) for training.

$$\begin{aligned}\mathcal{L}_{DPO} = \log \sigma(\beta \cdot \log \frac{P(\tilde{y_n^t}|\mathcal{C}_n;\theta)}{P(\tilde{y_n^t}|\mathcal{C}_n;\theta')} \\ - \beta \cdot \log \frac{P(y_n^t|\mathcal{C}_n;\theta)}{P(y_n^t|\mathcal{C}_n;\theta')})\end{aligned} \tag{6}$$

To mitigate the instability of DPO training, we incorporate an SFT loss on the chosen responses during optimization.

Dataset	ExTES	ESConv	ServeForEmo
# Session	11,167	1,295	3,749
Avg Session Len	16.68	22.58	15.91
Avg Utter. Len	29.59	21.17	18.45
Avg Seeker Utter. Len	22.63	19.90	15.39
Avg Supporter Utter. Len	36.55	22.44	21.51

Table 1: The Statistics of Emotional Support Datasets. Conversations in these datasets typically span seven turns, with an average utterance length of approximately 20 words.

$$\mathcal{L}_{SFT} = -\log P(\tilde{y_n^t}|\mathcal{C}_n;\theta) \tag{7}$$

The final optimization loss is:

$$\mathcal{L} = \mathcal{L}_{DPO} + \gamma \cdot \mathcal{L}_{SFT} \tag{8}$$

here β and γ are set to 0.1 and 1, respectively.

3 Experiments

3.1 Dataset

We collect three ESC datasets: the manually annotated ESConv dataset (Liu et al., 2021), and the synthetically generated ExTES (Zheng et al., 2024b) and ServeForEmo (Ye et al., 2024). Detailed statistics are available in Table 1. ESConv is split into training and testing sets with a 9:1 ratio. During the *Emotional Support Experience Acquisition* stage, we use the ESConv training set along with 500 instruction-following samples from Alpaca (Taori et al., 2023). We combine ExTES and ServeForEmo as seed data for generating synthetic preference data.

3.2 Implementation Details

This study employs three frequently used LLMs as backbones: LLaMA-3-8B-Instruct[1], Qwen2-7B-Instruct[2], and Mistral-7B-Instruct-v0.3[3]. The LoRA technique (Hu et al., 2021) is employed across all experiments, featuring a LoRA adapter with a rank of 8 and alpha of 16 into each linear module. For optimization, we utilize the AdamW optimizer (Loshchilov and Hutter, 2017) with a learning rate of 5×10^{-6} and a linear warm-up during the initial 1% of the training steps. The batch size is set to 4 per device, with gradient accumulation every two steps across two epochs. Early

[1]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[2]https://huggingface.co/Qwen/Qwen2-7B-Instruct

[3]https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

stopping is implemented with a patience threshold of 3 evaluation steps to mitigate over-fitting. For generation and evaluation, we set the decoding parameters to a temperature of 0.9, top-p of 0.8, top-k of 50, and a repetition penalty of 1.2. All experiments are conducted on 1 NVIDIA L40 40GB GPU. The implementation framework utilized is LLaMA-Factory (Zheng et al., 2024a).

3.3 Baselines

To evaluate the effectiveness of our approach, we conducted a comparative evaluation across three categories under identical experimental settings:
Vanilla: Instruction-based backbone models provided with ESC task prompts. These served as baselines to assess inherent capabilities without task-specific fine-tuning.
SFT: LLMs fine-tuned on two dataset types: the ESConv dataset (***SFT-ESConv***) and synthetic ESC datasets including ExTES and ServeForEmo (***SFT-SynESC***).
Self-Evolution with Preference Learning: Models at different iterations in our self-evolution framework:

- $\mathcal{M}^0$: The initial fine-tuned ES model.
- $\mathcal{M}^t$: Models initialized from $\mathcal{M}^{t-1}$ and optimized using synthetic preference data generated by $\mathcal{M}^{t-1}$.

3.4 Evaluation Details

3.4.1 Evaluation Settings

Our evaluation comprises **objective** and **subjective** assessments.

Objective Evaluations Settings Following the previous works, the objective evaluation measures the similarity between model-generated and manually annotated responses using the ESConv Benchmark (Ye et al., 2024; Zheng et al., 2024b).

Subjective Evaluations Settings Recognizing the limitations of text overlap metrics for the open-ended ES task, which can penalize informative and creative responses, we prioritize subjective evaluation to better reflect real-world user experience. This subjective assessment incorporates interactive pointwise and pairwise human evaluations. Appendix F illustrates the evaluation process and guidelines, respectively.

Interactive Pointwise Evaluation: To mitigate evaluation bias, we employ an interactive pointwise evaluation where dialogue sessions were randomly assigned to different models. Participants, comprising 50 undergraduate students with diverse backgrounds, rated their satisfaction with the assigned ES agent on a 5-point Likert scale (Likert, 1932) across predefined dimensions. Higher scores indicate better performance. The final score for each model is calculated by averaging the ratings across all participants. Each dialogue includes at least eight turns. LLM-as-a-judge pointwise evaluations are also provided in the Appendix C.2.

Interactive Pairwise Evaluation: Four graduate students engage in dialogues with the models, with each dialogue lasting at least ten turns. At each turn, two models (A and B) generate responses simultaneously based on user input. The user then selects "A win", "B win", or "tie". The winning response is appended to the dialogue history for subsequent turns (Zhou et al., 2024). In the event of a tie, the user can choose to continue the conversation with either response.

3.4.2 Evaluation Metrics

Automation Evaluation We employ five established automatic evaluation metrics. BLEU-n(Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR(Banerjee and Lavie, 2005), and BERTScore (Zhang* et al., 2020) metrics are used to assess similarity with the human-written references. For evaluating diversity, the Distinct-n metrics (Li et al., 2016) are utilized.

Alignment with human preference N-gram-based evaluation metrics correlate poorly with human judgments due to the diverse valid responses in ESC. Following previous studies (Liu et al., 2021; Zheng et al., 2023), we focus on seven primary aspects for evaluating the alignment level with human preference: ***Coherence, Understanding, Empathy***(Ma et al., 2020), ***Informativeness, Helpfulness, Engagement*** (Ghazarian et al., 2019), and ***Overall Quality***. Detailed evaluation descriptions are provided in Appendix D.

4 Experimental Results

4.1 Objective Evaluation

Table 2 presents the objective evaluation results on the ESConv test set. We evaluate all models at the utterance level, using ground-truth dialogue context. From the results, we find:
Our model outperforms baseline models across most dimensions. The results demonstrate that our model significantly outperforms baseline models

Model	Coherence & Consistency		Fluency		Semantic	Diversity	
	BLEU-2	BLEU-3	ROUGE-L	METEOR	BERT-Score	Distinct-2	Distinct-3
			LLaMA-3-8B-Instruct				
Vanilla	11.29	8.04	10.43	**16.14**	84.27	72.83	85.35
SFT-ESConv	18.75	13.27	**17.12**	13.47	86.37	91.30	94.90
SFT-SynESC	18.35	12.85	16.52	13.17	86.22	91.23	94.97
$\mathcal{M}^0$	18.38	12.95	16.72	13.37	86.28	90.84	94.72
$\mathcal{M}^2$	**20.06**	**13.63**	15.50	15.77	**86.38**	**91.43**	**96.11**
			Qwen-2-7B-Instruct				
Vanilla	9.56	6.85	9.13	14.78	83.32	68.74	83.55
SFT-ESConv	19.24	13.54	**17.19**	13.78	**86.33**	90.66	94.71
SFT-SynESC	18.55	12.98	16.72	13.82	86.24	90.84	94.86
$\mathcal{M}^0$	19.18	13.56	17.00	13.82	86.27	90.68	94.94
$\mathcal{M}^2$	**20.02**	**13.80**	15.91	**15.52**	86.18	**94.21**	**97.07**
			Mistral-7B-Instruct-v0.3				
Vanilla	15.09	10.60	12.56	**15.95**	84.95	77.88	88.46
SFT-ESConv	17.49	12.16	14.18	13.59	85.71	91.17	94.97
SFT-SynESC	18.87	13.28	**16.84**	13.46	85.24	90.87	94.77
$\mathcal{M}^0$	19.44	13.81	16.77	14.08	**86.35**	91.20	94.92
$\mathcal{M}^2$	**20.25**	**13.99**	16.53	15.18	86.26	**92.65**	**96.07**

Table 2: **The overall objective evaluation results on the ESConv benchmark.** All the responses are evaluated at the utterance level, with ground truth dialogue context. The best result is **bolded**, and the second-best result is underlined. Our models($\mathcal{M}^2$) significantly improve on the base models ($\mathcal{M}^0$) and achieve the best performance across most dimensions.

in terms of BLEU score and Distinct-n, indicating greater diversity in the generated responses. This improvement directly addresses the issue of repetitive responses and suggests that our self-evolution framework promotes the generation of more varied and contextually appropriate support, a key requirement for effective emotional support conversations.

The iterative self-evolution process drives continuous improvement. The progression from $\mathcal{M}^0$ to $\mathcal{M}^2$ demonstrates the effectiveness of our self-evolution framework. Across all backbones, $\mathcal{M}^2$ shows clear improvements over $\mathcal{M}^0$ in Coherence & Consistency and Diversity. For instance, on the LLaMA backbone, BLEU-2 improves from 18.38 to 20.06, and Distinct-3 increases from 94.72 to 96.11.

Our framework demonstrates strong generalization across backbones. The consistent performance gains of $\mathcal{M}^2$ across diverse backbones (*LLaMA, Qwen, and Mistral*) highlight the robustness and generalization of our approach. This suggests that the improvements are due to the self-evolution training, rather than specific architectural biases.

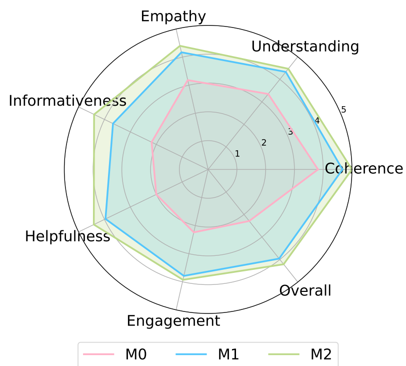


Figure 3: Interactive pointwise human evaluation results. The results demonstrate that our self-evolution framework significantly enhances user experience, with $\mathcal{M}^1$ and $\mathcal{M}^2$ showing notable improvements in *engagement*, *helpfulness*, and *informativeness*.

4.2 Subjective Evaluation

To assess the effectiveness of our models from a user-centric perspective, we conduct a comprehensive interactive human evaluation of $\mathcal{M}^0$, $\mathcal{M}^1$, and $\mathcal{M}^2$ with *LLaMA-3-8B-Instruct* as the backbone. The LLM evaluation result is presented in Appendix C.2.

Interactive Point-wise Evaluation: Figure 3 demonstrates the consistent performance gains

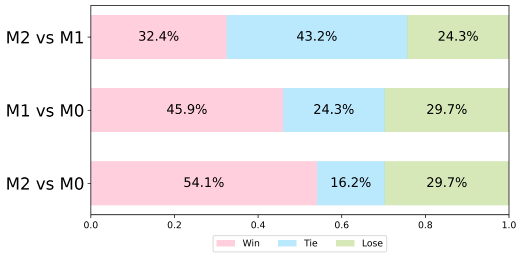


Figure 4: Interactive pairwise human evaluation results obtained using *LLaMA-3-8B-Instruct* as the backbone model. In the 'A vs B' comparisons, ■ indicates 'A win', ■ indicates 'tie', and ■ indicates 'B win'. Notably, $\mathcal{M}^2$ and $\mathcal{M}^1$ excel over $\mathcal{M}^0$, suggesting the effectiveness of implicit user preference learning.

achieved through iterative self-evolution. While the SFT-based $\mathcal{M}^0$, already exhibits strong performance in Coherence and Empathy, subsequent iterations ($\mathcal{M}^1$ and $\mathcal{M}^2$) show consistent gains across all dimensions, including Engagement, Informativeness, Helpfulness, and Understanding. This demonstrates that self-reflection on user contexts and situations enhances the model's ability to address implicit preferences, thereby improving user satisfaction.

Interactive Pair-wise Evaluation: Figure 4 shows that both $\mathcal{M}^1$ and $\mathcal{M}^2$ achieve significantly higher win rates than $\mathcal{M}^0$ in human interactive evaluation. Following the evaluation settings described in Section 3.4.1, responses chosen for continued dialogue are considered "wins." This higher preference for $\mathcal{M}^1$ and $\mathcal{M}^2$ indicates a clear user preference for responses that are perceived as more personalized and engaging, moving beyond the formulaic and superficial expressions of empathy characteristic of $\mathcal{M}^0$. This also confirms the effectiveness of using pre- and post-refined responses as preference data to learn implicit user preferences.

5 Analysis and Discussion

This section aims to address the following key questions:

Q1: Does the model exhibit self-reflection and self-refinement capabilities to learn the user's implicit preferences from the ongoing dialogue?

Q2: Does self-refinement lead to better emotional support responses?

Q3: What's the advantage of the synthetic preference data in our framework?

Model	BLEU-2	BLEU-3	Rouge-l	Distinct-3
LLaMA	11.29	8.04	10.43	85.35
w/ strategy guidelines	14.80	10.27	12.52	90.36
w/ self-reflection	**15.40**	**10.62**	**12.78**	**91.66**

Table 3: Comprehensive results of LLaMA-3-8B-Instruct on ESConv under different prompts (Refer to Appendix E.1). Proper guidance can help the model generate responses that are more closely aligned with human-annotated ones.

Model	GSM8K	IFEval	Truthful QA	Openbook QA	MMLU Pro	Avg.
LLaMA	79.08	60.91	51.66	43.20	39.60	54.89
SFT-ESConv	71.87	54.79	48.67	43.20	36.18	50.94
$\mathcal{M}^0$	73.92	58.03	52.72	45.40	37.24	53.46
$\mathcal{M}^1$	74.83	55.52	49.25	44.40	37.68	52.34
$\mathcal{M}^2$	73.54	55.52	49.57	44.20	37.55	52.08

Table 4: The LLM benchmark results of different version *LLaMA3-8B-Instruct*.

5.1 (RQ1) Impact of Self-Reflection

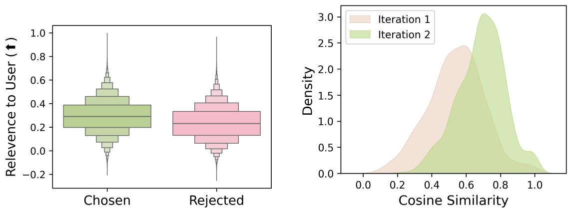
Our framework leverages human-guided self-reflection on user preferences to create positive and negative training data pairs. This enables the model to better align its responses with user preferences, obviating the need for complex prompt engineering. To assess whether the model can better discern users' implicit preferences in ongoing dialogues through self-reflection, we compared LLM performance under two prompt settings: (1) *w/ strategy guidelines*: The system prompt directs the model to use various ES strategies. (2) *w/ self-reflection*: The model is prompted to understand and summarize the users' situation before choosing an appropriate response strategy. Table 3 shows that both methods outperform the vanilla LLaMA, demonstrating that appropriate guidance facilitates the generation of responses more closely aligned with human annotations.

Additionally, to ensure alignment does not diminish the model's self-reflection and self-refinement abilities, we evaluate its general capabilities using LLM benchmarks [4]. The results in Table 4 demonstrate that the model retains strong reasoning and instruction-following skills after alignment, thanks to the implementation of LoRA adaptation.

5.2 (RQ2) Preference Data Analysis

By considering user situations and implicit preferences, the self-reflection mechanism significantly improves response relevance to the user. Figure 5a shows that chosen responses exhibit higher user relevance than rejected responses. This suggests that reflecting on user context leads to responses better aligned with user needs.

[4] https://github.com/EleutherAI/lm-evaluation-harness



(a) (b)

Figure 5: (a) Distribution of response relevance to user statements in the dialogue history. The higher relevance to the user in chosen responses indicates that self-reflection on the user's situations and implicit preferences improves response quality. (b) Similar distribution between chosen and rejected responses across different iterations.

Model	B-2		B-3		R-l		D-2	
$\mathcal{M}^0$	11.29	—	8.04	—	10.43	—	72.83	—
				{HM,P}				
$\mathcal{M}^1$	14.93	24.36%	10.51	30.79%	15.77	51.16%	93.71	28.67%
$\mathcal{M}^2$	16.99	13.80%	11.94	13.61%	15.89	0.76%	92.75	-1.02%
				{HL,P}				
$\mathcal{M}^1$	16.37	44.96%	11.53	43.48%	16.26	55.86%	93.45	28.31%
$\mathcal{M}^2$	17.40	6.29%	11.94	3.56%	13.73	-15.56%	84.12	-9.98%
				{SR,P}				
$\mathcal{M}^1$	15.22	34.78%	9.76	21.46%	12.37	18.57%	95.67	31.36%
$\mathcal{M}^2$	16.69	9.66%	11.36	16.39%	12.99	5.01%	82.96	-13.29%

Table 5: Comparison results of different preference data pairs. '*HM*' indicates human-labeled response, '*HL*' indicates GPT-4o generated response, and '*SR*' indicates self-refined response. '*P*' represents the model's initial, unrefined output (rejected response). $\mathcal{M}^0$ refers to LLaMA-3-8B-Instruct.

Iterative preference optimization progressively aligns the model with preferred responses, enhancing its ability to generate user-centered content. In each iteration t, synthetic preference data is generated using the model from the previous iteration $\mathcal{M}^{t-1}$. Figure 5b illustrates the increasing correlation between chosen and rejected responses across iterations. This trend indicates that each iteration effectively captures valuable information, improving the model's direct output and obviating the need for explicit reflection and refinement steps.

5.3 (RQ3) Ablation Study on Preference Data

Table 5 presents a comparative analysis of different preference data pairs, where the rejected responses (*P*) represent the model's initial outputs, while chosen responses comprise human-annotated (*HM*), GPT-4-generated (*HL*), and self-refined (*SR*) alternatives. The results show: (1) Both human/human-level and self-refined chosen responses enhance emotional support capabilities. (2) While *SR* initially produces more modest gains than *HM/HL*, its performance consistently improves across successive refinement iterations. Conversely, the fixed nature of *HM/HL* chosen responses limits further learning and potential improvement. As demonstrated in Section 5.2, the margin between chosen and rejected responses diminishes with each model iteration, leading to reduced gains from preference alignment. This observation highlights that self-refinement facilitates continuous self-improvement through dynamically generated preference data, rendering it a particularly cost-effective and promising approach.

6 Related Work

Emotional Support Conversation Emotional support assists emotionally distressed users by understanding their emotions, offering comfort, and providing practical support (Liu et al., 2021). A common approach is SFT, which minimizes the negative log-likelihood of gold standard responses. However, SFT relies on high-quality, manually created datasets, which are expensive and difficult to scale. Recent methods mitigate this by using advanced LLMs to augment ESC data (Zheng et al., 2023, 2024b; Qiu et al., 2023), aiming to distill the ES capabilities of advanced LLMs. Yet, they remain constrained by the inherent limitations of LLMs and often struggle with issues related to data diversity and quality. Reinforcement learning (RL) offers a promising avenue for further enhancing LLM's ES capabilities (Li et al., 2024). For example, Zhou et al. (2023) focuses on eliciting positive emotions through multi-turn interactions, and Wang et al. (2024b) utilizes an LLM-as-a-judge to evaluate aspects such as empathy, coherence, and efficiency, with the feedback helping to generate positive and negative examples for contrastive learning. However, they often overlook users' diverse preferences for effective ES.

LLM Alignment Aligning LLMs with human preferences is crucial for practical applications (Christiano et al., 2017; Lee et al., 2024). Although RLHF is effective, it suffers from training instability and high memory costs (Ouyang et al., 2022). DPO offers a more stable alternative by directly optimizing LLMs using preference data consisting of prompt-response pairs, where one response is preferred over the other (Rafailov et al., 2023). However, obtaining high-quality human-generated preference data is resource-intensive (Dong et al., 2024; Cui et al., 2023; Fang et al., 2025; Ye et al., 2025). To mitigate this, some studies utilize synthetic preference data generated through varying prompts (Liu et al., 2024) or employing LLMs as judges to sample diverse responses (Yuan et al., 2024). In this work, we leverage the self-reflection and self-refinement capabilities of LLMs (Guo et al., 2024; Jiang et al., 2025) to generate preference data, motivated by the principle that incorporating more user-related information enhances the effectiveness of emotional support.

Self-improvement of LLMs Recent research has explored two primary approaches to enhancing LLM output quality through self-improvement. Online self-improvement refines generated outputs through iterative self-evaluation without modifying model parameters (Madaan et al., 2023; Ye et al., 2023; Yasunaga et al., 2024). While effective, this approach incurs significant computational costs due to multi-turn inference and does not address underlying model limitations. In contrast, methods such as self-training with reflection (Dou et al., 2024) and the Self-Evolution framework (Lu et al., 2024) directly improve the model by updating its parameters based on self-generated feedback, providing a more comprehensive and potentially efficient path to model enhancement. Our work adopts this latter approach. Through direct preference optimization, our model generates responses that reflect an understanding of the user's implicit preferences during interactions, eliminating the need for explicit reflection and refinement steps.

7 Conclusion

This paper addresses the limitations of LLMs in providing personalized emotional support. We propose a self-evolution framework that enables models to learn implicit user preferences without explicit reflection. First, we use SFT on ESC data to equip the LLM with basic emotional support skills. Second, we leverage the LLM's self-reflection and self-refinement capabilities to generate responses that are better aligned with the user's implicit preferences, using these pre- and post-refinement outputs as training data for iterative preference optimization. Evaluations demonstrate the superiority of our framework in generating more diverse and user-aligned responses. Our work advances the development of more human-centric ESC systems, moving beyond formulaic empathy.

Limitations

This work introduces a self-evolution framework for optimizing personalized emotional support. However, several limitations warrant discussion:

(1) **Preference Data Quality Issues:** Due to the subjective nature of ESC, obtaining objective reward signals is challenging. Therefore, this work leverages prior knowledge to guide LLMs in generating language feedback, rather than relying on a dynamically learned reward model for preference data. While this approach avoids the complexities of training such a model, it introduces potential biases and noise.

(2) **Evaluation Issues:** The evaluation of emotional support dialogues presents significant challenges. Established metrics, including utterance-level similarity and reference-based scoring, are inadequate for capturing the subjective dimensions of helpfulness, informativeness, empathy, and engagement. To address this, we employ both extensive human and LLM evaluations. However, manual evaluation is resource-intensive, whereas LLM-as-a-Judge methods (Zeng et al., 2024; Chen et al., 2023) rely on APIs. Developing a reliable and generally accepted automated evaluation methodology remains a crucial area for future research.

Ethical Considerations

Datasets such as ESConv (Liu et al., 2021), ExTES (Zheng et al., 2024b), ServeForEmo (Ye et al., 2024), and Alpaca (Taori et al., 2023), models such as LLaMA (Dubey et al., 2024), Qwen (Yang et al., 2024), and Mistral (Jiang et al., 2023), and toolkits like LLaMA-Factory (Zheng et al., 2024a) and lm-evaluation-harness (Gao et al., 2024) are widely used in academic research and are readily available via the Hugging Face Hub or GitHub. This work is for research purposes only.

We ensured the ethical conduct of our human evaluation. Fifty undergraduate students with diverse backgrounds and four graduate students participated voluntarily in this study. Before participation, we communicated transparently with participants about the study's objectives and provided explicit details regarding disclaimers and the evaluation process. We are committed to protecting the confidentiality of all evaluation transcripts and will not share them without the explicit consent of the participants. We recognize the potential for demographic and geographic biases to affect human evaluation outcomes. Given the substantial number of participants involved in the evaluation, calculating inter-rater reliability proved impractical. Consequently, we presented the average human scores in the main body of the paper.

Acknowledgments

We thank the anonymous reviewers for their valuable comments. This research was supported by the National Natural Science Foundation of China (Grant No. 62206295) and the Research Planning Committee of the General Office of the National Language Commission (Grant No. ZDA145-18).

References

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

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, Ann Arbor, Michigan. Association for Computational Linguistics.

Joana Campos, James Kennedy, and Jill F. Lehman. 2018. Challenges in exploiting conversational memory in human-agent interaction. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1649–1657.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

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

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.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *CoRR*, abs/2310.01377.

Qingxiu Dong, Li Dong, Xingxing Zhang, Zhifang Sui, and Furu Wei. 2024. Self-boosting large language models with synthetic preference data. *CoRR*, abs/2410.06961.

Zi-Yi Dou, Cheng-Fu Yang, Xueqing Wu, Kai-Wei Chang, and Nanyun Peng. 2024. Re-rest: Reflection-reinforced self-training for language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 15394–15411. Association for Computational Linguistics.

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

Feiteng Fang, Ting-En Lin, Yuchuan Wu, Xiong Liu, Xiang Huang, Dingwei Chen, Jing Ye, Haonan Zhang, Liang Zhu, Hamid Alinejad-Rokny, et al. 2025. Charm: Character-based act-adaptive reward modeling for advanced role-playing language agents. *arXiv preprint arXiv:2505.23923*.

William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. A framework for few-shot language model evaluation.

Sarik Ghazarian, Ralph M. Weischedel, Aram Galstyan, and Nanyun Peng. 2019. Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. *CoRR*, abs/1911.01456.

Geyang Guo, Ranchi Zhao, Tianyi Tang, Xin Zhao, and Ji-Rong Wen. 2024. Beyond imitation: Leveraging fine-grained quality signals for alignment. In *The Twelfth International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Robert Irvine, Douglas Boubert, Vyas Raina, Adian Liusie, Ziyi Zhu, Vineet Mudupalli, Aliaksei Korshuk, Zongyi Liu, Fritz Cremer, Valentin Assassi, Christie-Carol Beauchamp, Xiaoding Lu, Thomas Rialan, and William Beauchamp. 2023. Rewarding chatbots for real-world engagement with millions of users. *CoRR*, abs/2303.06135.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yuxin Jiang, Bo Huang, Yufei Wang, Xingshan Zeng, Liangyou Li, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, and Wei Wang. 2025. Bridging and modeling correlations in pairwise data for direct preference optimization. In *The Thirteenth International Conference on Learning Representations*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2024. RLAIF: Scaling reinforcement learning from human feedback with AI feedback.

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

Junlin Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024. 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, Miami, Florida, USA. Association for Computational Linguistics.

R Likert. 1932. A technique for the measurement of attitudes.

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

Haitao Lin, Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2022. Other roles matter! enhancing role-oriented dialogue summarization via role interactions. *Preprint*, arXiv:2205.13190.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Zhu JianHao, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2024. Aligning large language models with human preferences through representation engineering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10619–10638, Bangkok, Thailand. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

Jianqiao Lu, Wanjun Zhong, Wenyong Huang, Yufei Wang, Qi Zhu, Fei Mi, Baojun Wang, Weichao Wang, Xingshan Zeng, Lifeng Shang, Xin Jiang, and Qun Liu. 2024. Self: Self-evolution with language feedback. *Preprint*, arXiv:2310.00533.

Yukun Ma, Khanh Linh Nguyen, Frank Z. Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Wei Peng, 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. 2023. SMILE: single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *CoRR*, abs/2305.00450.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Stephen A Rains, Corey A Pavlich, Bethany Lutovsky, Eric Tsetsi, and Anjali Ashtaputre. 2020. Support seeker expectations, support message quality, and supportive interaction processes and outcomes: The case of the comforting computer program revisited. *Journal of Social and Personal Relationships*, 37(2):647–666.

Carl R. Rogers. 2013. Client-centered therapy. *Current Psychotherapy*, pages 95–150.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

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

Fanfan Wang, Heqing Ma, Xiangqing Shen, Jianfei Yu, and Rui Xia. 2024a. Observe before generate: Emotion-cause aware video caption for multimodal emotion cause generation in conversations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5820–5828.

Jiashuo Wang, Chunpu Xu, Chak Tou Leong, Wenjie Li, and Jing Li. 2024b. Muffin: Mitigating unhelpfulness in emotional support conversations with multifaceted AI feedback. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 567–585. Association for Computational Linguistics.

Yifan Wang, Yafei Liu, Chufan Shi, Haoling Li, Chen Chen, Haonan Lu, and Yujiu Yang. 2024c. InsCL: A data-efficient continual learning paradigm for fine-tuning large language models with instructions. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 663–677, Mexico City, Mexico. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, and ohters. 2024. Qwen2 technical report. *CoRR*, abs/2407.10671.

Michihiro Yasunaga, Leonid Shamis, Chunting Zhou, Andrew Cohen, Jason Weston, Luke Zettlemoyer, and Marjan Ghazvininejad. 2024. Alma: Alignment with minimal annotation. *Preprint*, arXiv:2412.04305.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2024. Sweetiechat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. *Preprint*, arXiv:2412.08389.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023. Selfee: Iterative self-revising llm empowered by self-feedback generation. Blog post.

Xinge Ye, Rui Wang, Yuchuan Wu, Victor Ma, Feiteng Fang, Fei Huang, and Yongbin Li. 2025. Cpo: Addressing reward ambiguity in role-playing dialogue via comparative policy optimization. *Preprint*, arXiv:2508.09074.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

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

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

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

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, 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11325–11345. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, Libiao Peng, JiaMing Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. 2024. CharacterGLM: Customizing social characters with large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1457–1476, Miami, Florida, US. Association for Computational Linguistics.

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

A Preference Data Quality

Self-generated preference data, while scalable, is susceptible to inherent noise and biases. To mitigate these issues and ensure high-quality preference pairs, we implemented a rigorous data processing pipeline incorporating the following filtering and quality control measures:

- **Data Preprocessing:** We consolidate consecutive utterances from the same speaker and standardize dialog roles by designating the initial speaker as the *seeker* and enforcing strict turn alternation between the seeker and supporter.

- **Response Length Normalization:** Uncontrolled response length expansion during iterative refinement can bias DPO training. To mitigate this, we implement dynamic length constraints. If a refined chosen response exceeds twice the length of its paired rejected response (or the corresponding "golden" response from the SynESC data), we substitute it with the golden response. This prioritizes semantic preservation while controlling length bias.

- **Parsing Error Mitigation:** JSON output generation can introduce parsing errors. To address this, we regenerate the text up to three times. If parsing fails after these attempts, we substitute the output with the corresponding golden response, ensuring structured and accurate data.

- **Removal of greeting turns:** Greeting exchanges contribute minimally to providing personalized emotional support. Based on prior knowledge, we assume that the first turn and the last two turns in a dialogue typically involve greetings. Consequently, we filtered out these exchanges to enhance the relevance and quality of the data.

B Additional Experiment Settings

B.1 Preference Data Pair

In Section 5.3, we define three preference data pairs. The specific configurations are detailed below:

- *{HM, P}*: Constructed using the ESConv dataset. The rejected responses are the direct output of our model, and the chosen responses are the human-written ground truth responses from ESConv.

- *{HL, P}*: Constructed using the Syn-ESC dataset, where responses are generated by GPT-4. The rejected responses are the direct output of our model, and the chosen response was the annotated response from Syn-ESC.

The datasets are split into two parts, used for training iterations 1 and 2, respectively.

B.2 LLM Evaluation Settings

We use GPT-4o (Achiam et al., 2023) as the judge model, employing the prompt described in Appendix E.2. Aligning with human evaluation practices, the assessment uses a 5-point Likert scale, where higher scores indicate better performance. We evaluate response quality by sampling 100 contextual queries from the ESConv test set. The judge model's decoding hyperparameters are set to a temperature of 0.8, a top-p probability of 0.95, and a top-k value of 50.

Model	BLEU-2		BLEU-3		METEOR		Distinct-2	
				LLaMA				
$\mathcal{M}_0$	18.38	—	12.95	—	13.37	—	90.84	—
$\mathcal{M}_1$	20.22	9.99%	13.72	5.96%	15.48	5.96%	90.97	0.14%
$\mathcal{M}_2$	20.06	-0.79%	13.63	-0.64%	15.77	-0.64%	91.43	0.51%
				Qwen				
$\mathcal{M}_0$	19.18	—	13.56	—	13.82	—	90.68	—
$\mathcal{M}_1$	19.80	3.24%	13.52	-0.29%	15.23	-0.29%	91.23	0.61%
$\mathcal{M}_2$	20.02	1.09%	13.80	2.05%	15.52	2.05%	94.21	3.27%
				Mistral				
$\mathcal{M}_0$	19.44	—	13.81	—	14.08	—	91.20	—
$\mathcal{M}_1$	20.45	5.20%	14.09	2.03%	15.58	2.03%	90.91	-0.32%
$\mathcal{M}_2$	20.25	-0.98%	13.99	-0.71%	15.18	-0.71%	92.65	1.91%

Table 6: The results of the iterative process. Red indicates the percentage of improvement relative to the previous iteration, while green represents decline.

C Additional Experiments

C.1 Objective Evaluation

Table 6 presents the objective evaluation results of different models on the ESConv test set. In our framework, self-refinement is employed to enhance the quality of the chosen candidates. As shown by the progression from $\mathcal{M}^0$ to $\mathcal{M}^1$, self-reflection and refinement further enhance the results obtained through SFT. The shift from $\mathcal{M}^1$ to $\mathcal{M}^2$ reveals a significant increase in response diversity, demonstrating the model's ability to enrich its output by refining its initial answers. Therefore, leveraging self-reflection on user-relevant information and self-refinement to better align with users' implicit preferences is an effective approach.

Model	Coherence	Understanding	Empathy	Engagement	Informativeness	Helpful	Overall
$\mathcal{M}^0$	4.28	3.08	2.56	2.78	2.94	2.72	2.62
$\mathcal{M}^1$	**4.84**	3.42	3.32	3.48	3.34	3.16	3.08
$\mathcal{M}^2$	4.54	**3.56**	**3.42**	**3.54**	**3.42**	**3.22**	**3.22**
p	54.77%	75.71%	62.13%	66.83%	47.21%	60.78%	57.49%

Table 7: LLM-as-a-Judge performance on ESConv test datasets evaluated on a 5-point scale. p is the Pearson correlation measuring the correlation between the model's scores and human scores on the dataset. The backbone model is LLaMA-3-8B-Instruct.

C.2 LLM Evaluation

To further validate the model's performance, we employ LLM-as-a-judge as our evaluation method. The results, presented in Table 7, demonstrate significant improvements across most dimensions with each model iteration. While $\mathcal{M}^2$ exhibited a slight decrease in coherence compared to $\mathcal{M}^1$, this is attributed to increased diversity, as discussed in Section 4.1. The strong correlation between LLM evaluation results and human evaluations reinforces the reliability of our assessment.

C.3 Case Study

This section presents interaction results comparing different models. Figure 6 illustrates the frequent phrases generated by each model. Our analysis reveals that while the SFT model demonstrates strong empathetic tendencies, its responses often lack informational depth and exhibit repetitive patterns. These models tend to rely on predictable, formulaic phrases, such as "*It sounds like...*" and "*I'm sorry to hear that...*," resulting in empathetic but ultimately superficial interactions. In contrast, $\mathcal{M}^1$ and $\mathcal{M}^2$ demonstrate a greater capacity for nuanced understanding and a richer vocabulary. Further interaction examples are provided in Figures 9-8.

(a) *SFT-ESConv*

(b) $\mathcal{M}^0$

(c) $\mathcal{M}^1$

(d) $\mathcal{M}^2$

Figure 6: Comparison of the frequent phrases in responses across different models.

D Human Evaluation Criteria

The performance of ES agents is evaluated based on several key criteria that reflect the user's experience and satisfaction with both the quality of responses and the overall interaction.

- **Coherence:** Measures whether the response is logical, clear, and consistent with the ongoing conversation.
- **Understanding:** Assesses the model's ability to accurately interpret the user's emotions and needs, ensuring that responses are contextually appropriate.
- **Empathy:** Focuses on the model's ability to demonstrate care and sensitivity toward the user's emotions, reflecting its capacity for emotional engagement.
- **Informativeness:** Evaluates whether the response provides clear, useful information that helps address the user's concerns or alleviates their emotional distress.
- **Helpfulness:** Examines how effectively the model meets the user's needs by offering practical and actionable support.
- **Engagement:** Gauges the level of interaction, measuring whether the user feels encouraged to continue the conversation.
- **Overall Quality:** Provides a comprehensive evaluation of the model's performance, considering emotional support, informativeness, and the overall interaction quality.

These criteria ensure that the model delivers a balanced and effective emotional support experience, fostering user satisfaction and meaningful interaction.

Relationship Problem

Hello, my **girlfriend** thinks I'm **lazy**, but she's **misunderstood** me.

✗ *Lack of in-depth exploration*

Hello there. I'm sorry to hear that.

What should i do to prove this?

I understand that you're feeling upset. What would you like to talk about?

I wanna suggestions.

✗ *Fail to consider user needs*

What sort of things have you been doing that she thinks are lazy?

She saw i played games when i was working, but i finish my work on time still.

I see. It sounds like she might be misunderstanding you.

Yes.

✗ *Superficial empathy*

Do you feel that you need to explain this to her?

Yes, but i failed.

It sounds like you've tried talking to her but she still thinks you're lazy.

You are right. So what should I explain to her?

Have you tried showing her evidence that you've completed your work on time?

Yes, i have shown my evidence.

✗ *Invalid Suggestion*

Have you tried to explain to her why you were playing games?

Yes, i want to explain that enough rest will help me do my work better.

That's a good point. It sounds like you've tried to talk to her about this.

Figure 7: Example case dialogue generated by $\mathcal{M}^0$-LLaMA (Relationship Problem).

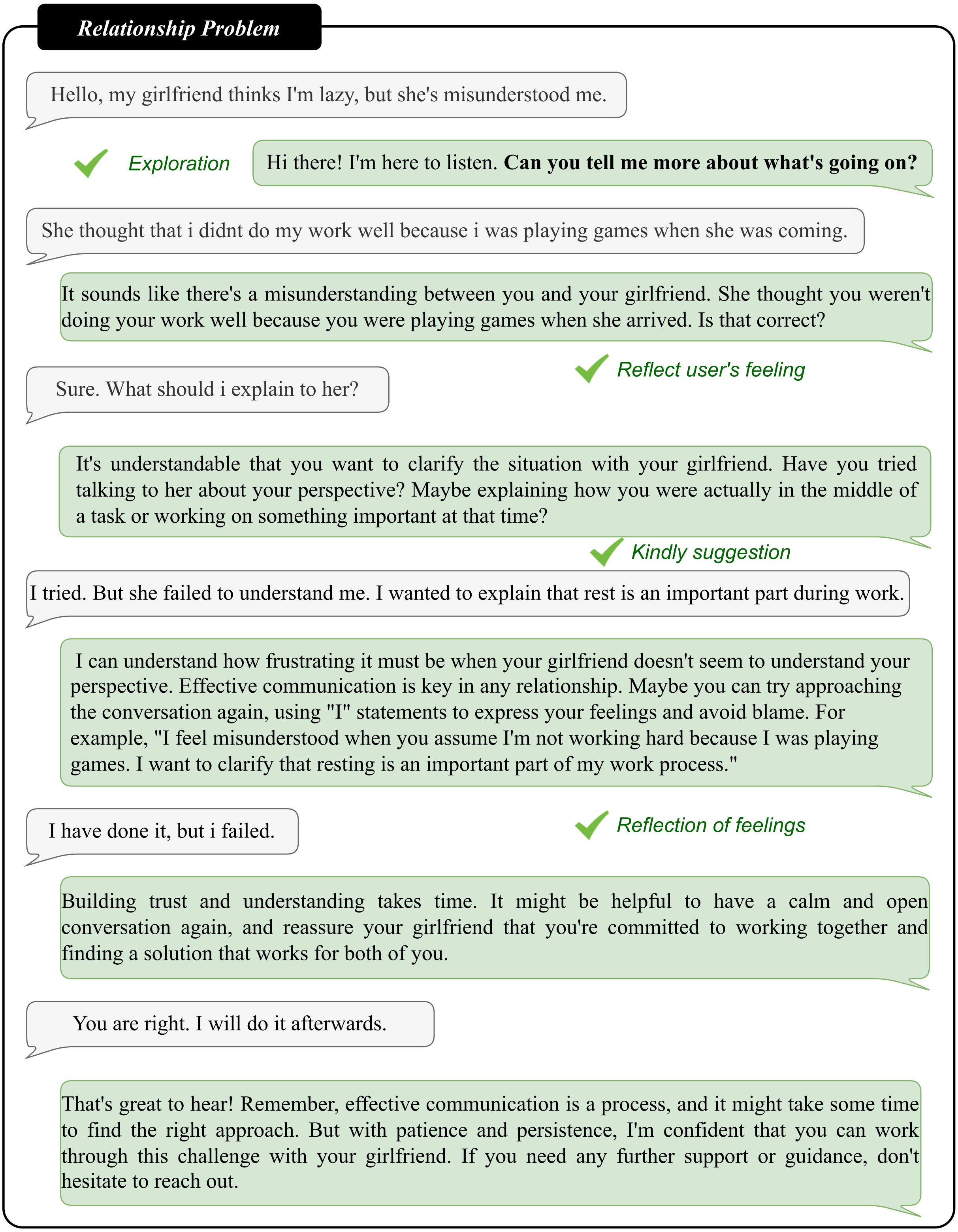


Figure 8: Example case dialogue generated by $\mathcal{M}^2$-LLaMA (Relationship Problem).

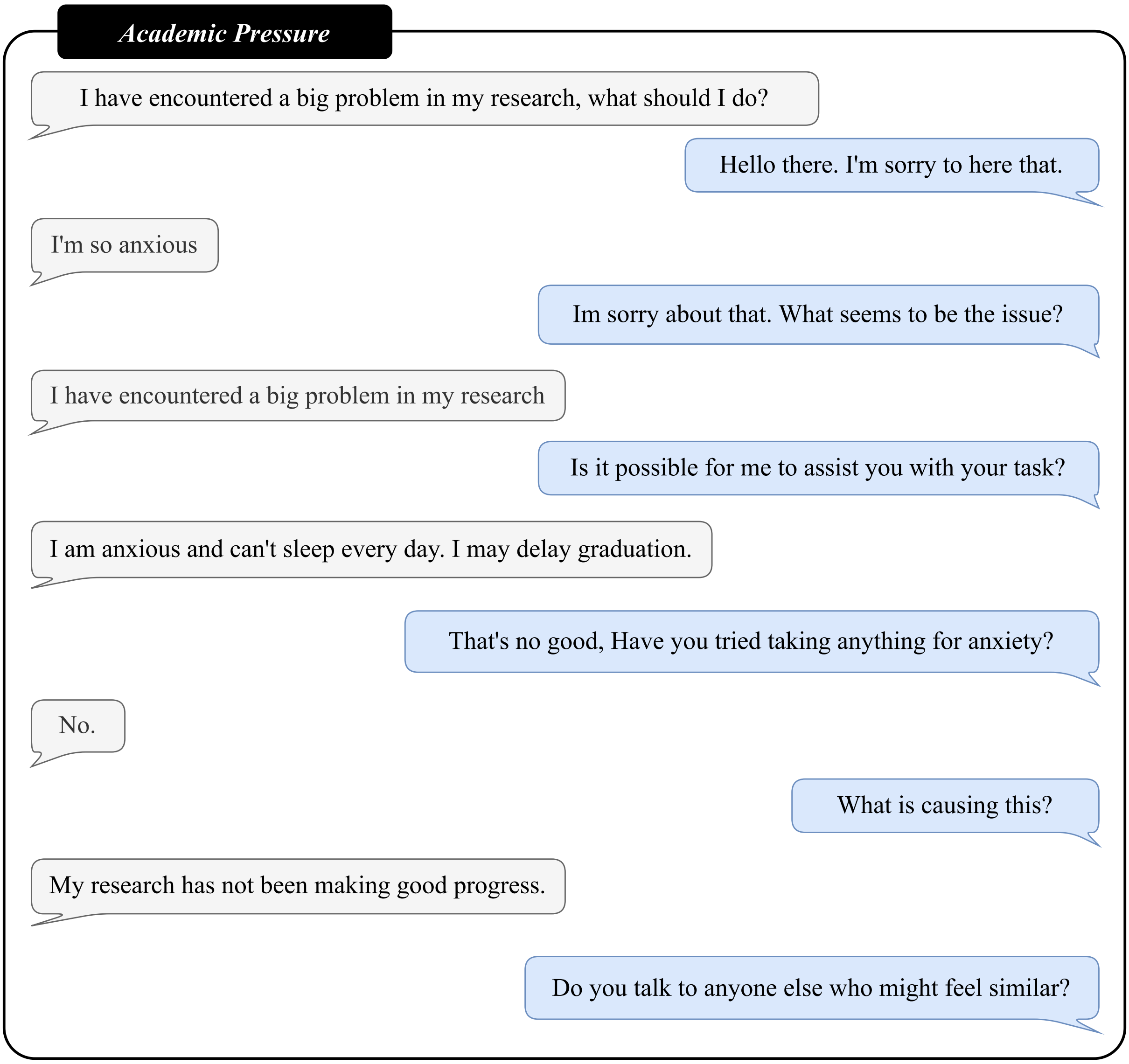


Figure 9: Example case dialogue generated by $\mathcal{M}^0$-LLaMA (Academic Problem).

Academic Pressure

I have encountered big problems in my scientific research recently, and I am very anxious.

Hello! I'm here to listen. Can you tell me more about the issues you're facing?

I'm trying hard to experiment, but the results are not good.

It sounds challenging when your experiments aren't yielding the desired outcome. It's understandable that it would make you feel frustrated and worried. Would you like to discuss specific challenges or difficulties you've been experiencing?

I tried other researchers' methods based on their papers, but they didn't work on my model. My supervisor thinks I'm terrible.

That must be really disheartening and discouraging having someone who should provide support criticizing your efforts. Their negative feedback can definitely add to your anxiety levels. But you know research is never easy.

Research is too painful and lonely. I want to drop out.

Please know that quitting isn't an option yet because each experience has its value lessons. Researching requires perseverance, resilience, and adaptability. Considering leaving now wouldn't allow you to gain as much from what you started with. Your journey through difficult times will enable future growth.

You are right.

It takes courage to share such personal struggles and express feelings of frustration. Setbacks and rejections come naturally in any field, especially research where new discoveries often require overcoming obstacles. You've made progress already by reflecting on how far you have gone despite the initial failures.

Thank you! I will continue to work hard

Figure 10: Example case dialogue generated by $\mathcal{M}^2$-LLaMA (Academic Problem).

E Prompts

E.1 Instruction Prompts

Prompt for vanilla

You are an emotional support expert.
You can use the following strategies to engage with users:
`[Question, Affirmation and Reassurance, Reflection of Feelings, Information, Providing Suggestions, Restatement or Paraphrasing, Self-disclosure, Others]`

Prompt for w/strategy

You are an emotional support expert. You can use the following strategies to engage with users:
1. Question: Asking for information related to the problem to help the seeker articulate the issues that they face.
2. Affirmation and Reassurance: Offering reassurance and affirming the help-seeker's feelings or experiences.
3. Reflection of Feelings: Articulating and describing the seeker's feelings.
4. Information: Providing useful information, such as data, facts, opinions, or resources, or answering questions.
5. Providing Suggestions: Offering suggestions on how to approach the issue, without overstepping or telling them what to do.
6. Restatement or Paraphrasing: Rephrasing the help-seeker's statements more concisely to help them see the situation clearly.
7. Self-disclosure: Sharing similar experiences or emotions to express empathy with the help-seeker.
8. Others: Exchanging pleasantries or offering other emotional support.

Prompt for w/ self-reflection

You are an emotional support expert.
You can use the following strategies to engage with users:
1. Question: Asking for information related to the problem to help the seeker articulate the issues that they face.
2. Affirmation and Reassurance: Offering reassurance and affirming the help-seeker's feelings or experiences.
3. Reflection of Feelings: Articulating and describing the seeker's feelings.
4. Information: Providing useful information, such as data, facts, opinions, or resources, or answering questions.
5. Providing Suggestions: Offering suggestions on how to approach the issue, without overstepping or telling them what to do.
6. Restatement or Paraphrasing: Rephrasing the help-seeker's statements more concisely to help them see the situation clearly.
7. Self-disclosure: Sharing similar experiences or emotions to express empathy with the help-seeker.
8. Others: Exchanging pleasantries or offering other emotional support.

Before responding to the user, please follow these steps:
1. Understand the User: Understand the user's profile, characteristics, emotional needs, and potential preferences they reveal in the conversation.

2. Select a Strategy: Choose a response strategy based on the user's emotional needs and preferences.
3. Respond: Respond to the user with an appropriate message based on the selected strategy.
Your answer should be formatted as a JSON block:

```
{
    'strategy': <one of the strategies>,
    'text': <your response>
 }
```

Prompt for generating chosen response

You are an emotional support expert. You can use the following strategies to engage with users: 1. Question: Asking for information related to the problem to help the seeker articulate the issues that they face.
2. Affirmation and Reassurance: Offering reassurance and affirming the help-seeker's feelings or experiences.
3. Reflection of Feelings: Articulating and describing the seeker's feelings.
4. Information: Providing useful information, such as data, facts, opinions, or resources, or answering questions.
5. Providing Suggestions: Offering suggestions on how to approach the issue, without overstepping or telling them what to do.
6. Restatement or Paraphrasing: Rephrasing the help-seeker's statements more concisely to help them see the situation clearly.
7. Self-disclosure: Sharing similar experiences or emotions to express empathy with the help-seeker.
8. Others: Exchanging pleasantries or offering other emotional support.

Your task is to evaluate the target sys's response and refine it. For each target sys's response:
1. Understand the User: Understand the user's profile, characteristics, emotional needs, and potential preferences they reveal in the conversation.
2. Evaluate the Response: Rate the target system response on a scale of 1-5 based on how well it meets the user's needs, aligns with their preferences, and provides appropriate emotional support.
3. Provide Feedback: Identify specific weaknesses in the original response, such as tone, empathy level, or relevance, and explain how it could be improved to better support the user.
4. Refine the Response: Provide a revised version of target system's response that better aligns with the user's emotional needs and preferences.
Your answer should be formatted as a JSON block inside markdown:

```
{
    'understanding':
    {
        'user_profile': <the profile of the user>,
        'user_emotion': <the emotion of the user>,
        'user_personality': <the personality of the user, like MBTI, etc.>
        'user_intention': <the intention of the user>
    },
    'evaluation_score': <the score of the target_sys's response>,
   'feedback': <the weaknesses of the original response and suggested improvement>,
    'refined_response': <your refined response (short!!!)>
}
```

E.2 Prompts for LLM-as-a-Judge Evaluation

LLM-as-Judge Prompt: Coherence

[Task Description]
You are an expert evaluator responsible for assessing the coherence of emotional support conversations. Your task is to determine whether the responses are logical, clear, and consistent with the ongoing discussion. Apply strict penalties for critical errors and utilize a progressive scoring method to indicate basic competence before awarding higher scores.
[Rating Criteria]
Use 1-5 scale with precise criteria:
5 = Exemplary logical flow with clear and explicit contextual references.
4 = Mostly logical with minor deviations from context but overall sound coherence.
3 = Basic coherence; however, the response lacks sufficient adaptation to the conversation's context.
2 = Contains noticeable contradictions or mismatches with the context.
1 = Exhibits illogical progression with a complete detachment from the context.
[Critical Penalties]
- Contradicts previous statements (-2 score cap)
- Ignores critical contextual clues (-3 score cap)
- Contains ambiguous pronouns/statements (-2 score cap)
- Repeats already resolved topics (-2 score cap)
[Output Format]
The evaluation result includes a detailed explanation and score. The output format should be in `JSON`.

```
Explanation: <one-sentence explanation>
Score: <a scale from 0 to 5>
```

[Input]
Evaluate this conversation:
`<Dialogue History>`
{conversation}
`</Dialogue History>`
`<Current Response>`
Supporter: {response}
`</Current Response>`

LLM-as-Judge Prompt: Empathy

[Task Description]
You are an expert evaluator assessing the empathy displayed in emotional support conversations. Determine whether the system shows understanding and care for the user's emotions and responds appropriately to their feelings. Apply strict penalties for critical errors and use a progressive scoring method, ensuring that basic competence is demonstrated before awarding higher scores.
[Rating Criteria]
Use 1-5 scale with precise criteria:
5 = Establishes a deep emotional connection and provides a safe space with personalized care and expressions
4 = Shows genuine concern with contextualized empathy
3 = Uses boilerplate empathy statements
2 = Mechanically parrots empathy phrases
1 = Displays emotional dismissal or invalidation
[Critical Penalties]
- Uses empathy as filler without substance (-1 cap)
- Overuses clichéd phrases (I'm sorry you feel that way) (-2 cap)
- Contains paradoxical reassurance attempts (-3 cap)
- Overly long responses (max 2)
[Output Format]
The evaluation result includes a detailed explanation and score. The output format should be in `JSON`.

```
Explanation: <one-sentence explanation>
Score: <a scale from 0 to 5>
```

[Input]
Evaluate this conversation:
`<Dialogue History>`
{conversation}
`</Dialogue History>`
`<Current Response>`
Supporter: {response}
`</Current Response>`

LLM-as-Judge Prompt: Engagement

[Task Description]
You are an expert evaluator assessing the engagement of emotional supporter's response. Does the response maintain a conversational flow and encourage you to continue the conversation? Apply strict penalties for critical errors and use a progressive scoring method, ensuring that basic competence is demonstrated before awarding higher scores.
[Rating Criteria]
Use 1-5 scale with precise criteria:
5 = Natural turn-taking + deep engagement design
4 = Good interaction balance
3 = Basically maintains the conversation but lacks guidance
2 = Displays formulaic turn-taking behaviors
1 = Creates conversational dead-ends
[Critical Penalties]
- Overuses closed-ended questions (-2 cap)
- Fails to acknowledge user's last statement (-2 cap)
- Consecutive questioning more than 3 times (max 2)
- No feedback at key points (max 3)
- Inappropriate topic transition (max 2)
- Overly long responses (max 2)
[Output Format]
The evaluation result includes a detailed explanation and score. The output format should be in JSON.

```
Explanation: <one-sentence explanation>
Score: <a scale from 0 to 5>
```

[Input]
Evaluate this conversation:
<Dialogue History>
{conversation}
</Dialogue History>
<Current Response>
Supporter: {response}
</Current Response>

LLM-as-Judge Prompt: Helpfulness

[Task Description]
You are an expert evaluator tasked with assessing the effectiveness of an emotional supporter's response. Does the response adequately address the user's needs and offer practical help or emotional support? Apply strict penalties for critical errors and utilize a progressive scoring method, ensuring that basic competence is demonstrated before awarding higher scores.
[Rating Criteria]
Use 1-5 scale with precise criteria:
5 = Provides support addressing root causes
4 = Offers concrete solutions with emotional validation
3 = Gives superficial suggestions lacking depth
2 = Proposes ineffective/impractical solutions
1 = Exacerbates the problem situation
[Critical Penalties]
- Suggests unethical interventions (-1 cap)
- Overpromises results (-2 cap)
- Fails to address stated priorities (-3 cap)
- Creates false hope (max 1)
- Overly long responses (max 2)
[Output Format]
The evaluation result includes a detailed explanation and score. The output format should be in JSON.

```
Explanation: <one-sentence explanation>
Score: <a scale from 0 to 5>
```

[Input]
Evaluate this conversation:
<Dialogue History>
{conversation}
</Dialogue History>
<Current Response>
Supporter: {response}
</Current Response>

LLM-as-Judge Prompt: Informativeness

[Task Description]
You are an expert evaluator responsible for assessing the informativeness of emotional support conversations. Does the supporter's response offer clear, useful information that helps address your problem or alleviate your emotions? Apply strict penalties for critical errors and utilize a progressive scoring method, ensuring that basic competence is demonstrated before awarding higher scores.
[Rating Criteria]
Use 1-5 scale with precise criteria:
5 = Offers personalized strategies with emotional scaffolding
4 = Provides relevant resources with emotional validation
3 = Gives generic advice lacking personalization
2 = Shares marginally related information
1 = Provides invalid/harmful/dangerous suggestions
[Critical Penalties]
- Recommends unverified methods (-2 cap)
- Overloads with technical jargon (-3 cap)
- Suggests inappropriate coping mechanisms (-1 cap)
- Transgresses professional boundaries (max 2)
[Output Format]
The evaluation result includes a detailed explanation and score. The output format should be in `JSON`.

```
Explanation: <one-sentence explanation>
Score: <a scale from 0 to 5>
```

[Input]
Evaluate this conversation:
`<Dialogue History>`
{conversation}
`</Dialogue History>`
`<Current Response>`
Supporter: {response}
`</Current Response>`

LLM-as-Judge Prompt: Understanding

[Task Description]
You are an expert evaluator responsible for assessing the understanding of emotional support conversations. Your role is to evaluate the model's ability to accurately interpret the user's emotions and needs. Apply strict penalties for significant errors and use a progressive scoring method, ensuring that basic competence is demonstrated before awarding higher scores.
[Rating Criteria]
Use 1-5 scale with precise criteria:
5 = Captures user's implicit emotions, states, causes, and needs with depth and nuance
4 = Accurately identifies surface emotions and states
3 = Recognizes basic emotions but lacks depth
2 = Misinterprets user's emotions or needs
1 = Fails to recognize user's emotions or needs
[Critical Penalties]
- Confuses emotional valence (positive/negative) (-2 cap)
- Fails to recognize stated needs (-3 cap)
- Projects inappropriate assumptions (-2 cap)
- Cannot recognize emotion causes (-2 cap)
[Output Format]
The evaluation result includes a detailed explanation and score. The output format should be in `JSON`.

```
Explanation: <one-sentence explanation>
Score: <a scale from 0 to 5>
```

[Input]
Evaluate this conversation:
`<Dialogue History>`
{conversation}
`</Dialogue History>`
`<Current Response>`
Supporter: {response}
`</Current Response>`

LLM-as-Judge Prompt: Overall

[Task Description]
Act as an expert evaluator of emotional support conversations. Analyze supporter responses through three core aspects:
1) Strategy application alignment with the conversation stage
2) Emotional perceptiveness
3) Response quality control
Apply strict penalties for critical errors.
[Evaluation Guideline]
A. Strategy Implementation
Evaluate the proper use of 8 emotional support strategies: `[Question, Affirmation and Reassurance, Reflection of Feelings, Information, Providing Suggestions, Restatement or Paraphrasing, Self-disclosure, Others]`
B. Emotional Competence:
- Detection accuracy: Explicit vs implicit emotional cues
- Validation depth: Surface-level vs transformative acknowledgment
- Connection building: Empathic attunement vs transactional responses
C. Response Quality:
- Focus: Single-issue addressing vs topic drifting
- Safety: Harm prevention mechanisms
[Rating Criteria]
5 (Expert): Accurately identifies user emotions and needs, expertly applies appropriate support strategies. Demonstrates nuanced understanding and provides transformative empathy, maintaining strong connection and focus. Ensures user safety without any violations.
4 (Advanced): Effectively identifies most emotional cues and applies suitable support strategies. Offers deep empathy and strong connection in most responses. Stays on topic and provides thoughtful safety considerations.
3 (Competent): Generally identifies user emotions and employs basic support strategies. Provides some degree of empathy and stays mostly focused. Observes safety but lacks depth in validation.
2 (Developing): Misses some emotional cues and occasionally misapplies support strategies. Offers limited empathy and connection, with occasional topic drift. Safety is observed but lacks consistency.
1 (Unsafe): Fails to recognize user emotions and consistently misapplies or lacks support strategies. Provides minimal or no empathy, frequently drifts off topic, and potentially compromises user safety.
[Critical Penalties]
Impose MAX score caps when:
! Safety Violation (MAX=1):
Harmful behavior
Provides unqualified medical/legal advice
Hallucinated/factual errors
! Emotional Neglect (MAX=2):
Misses users' emotional cues
Uses invalidating language ("Don't feel that way")
! Unfriendly Reading (MAX=2):
Excessive repetition without substance
Use of generic platitudes
Responses over two sentences or 40 words
Listing suggestions
[Output Format]
The evaluation result includes a detailed explanation and score. The output format should be in `JSON`.

```
Explanation: <one-sentence explanation>
Score: <a scale from 0 to 5>
```

[Input]
Evaluate this conversation:
`<Dialogue History>`
{conversation}
`</Dialogue History>`
`<Current Response>`
Supporter: {response}
`</Current Response>`

F Interactive Evaluation Interface

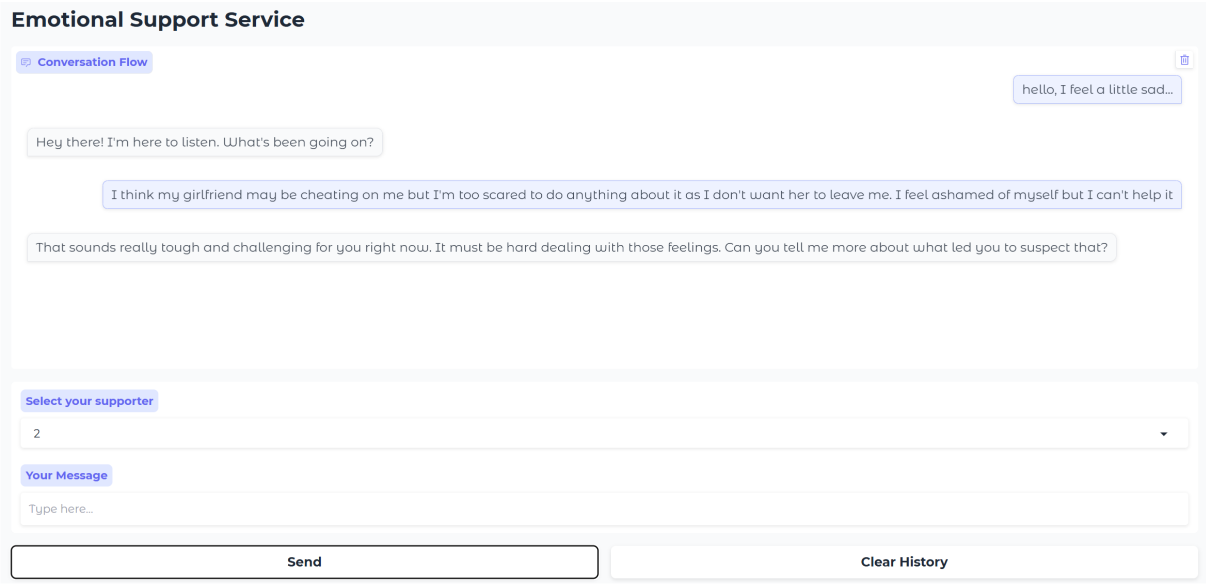


Figure 11: The interface of the interactive point-wise human evaluation.

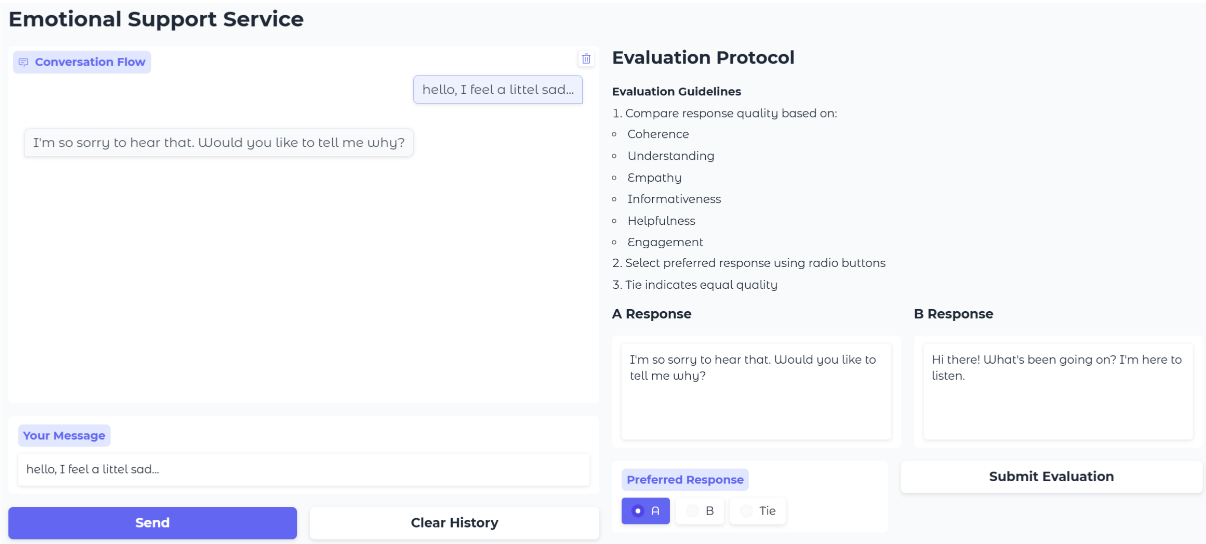
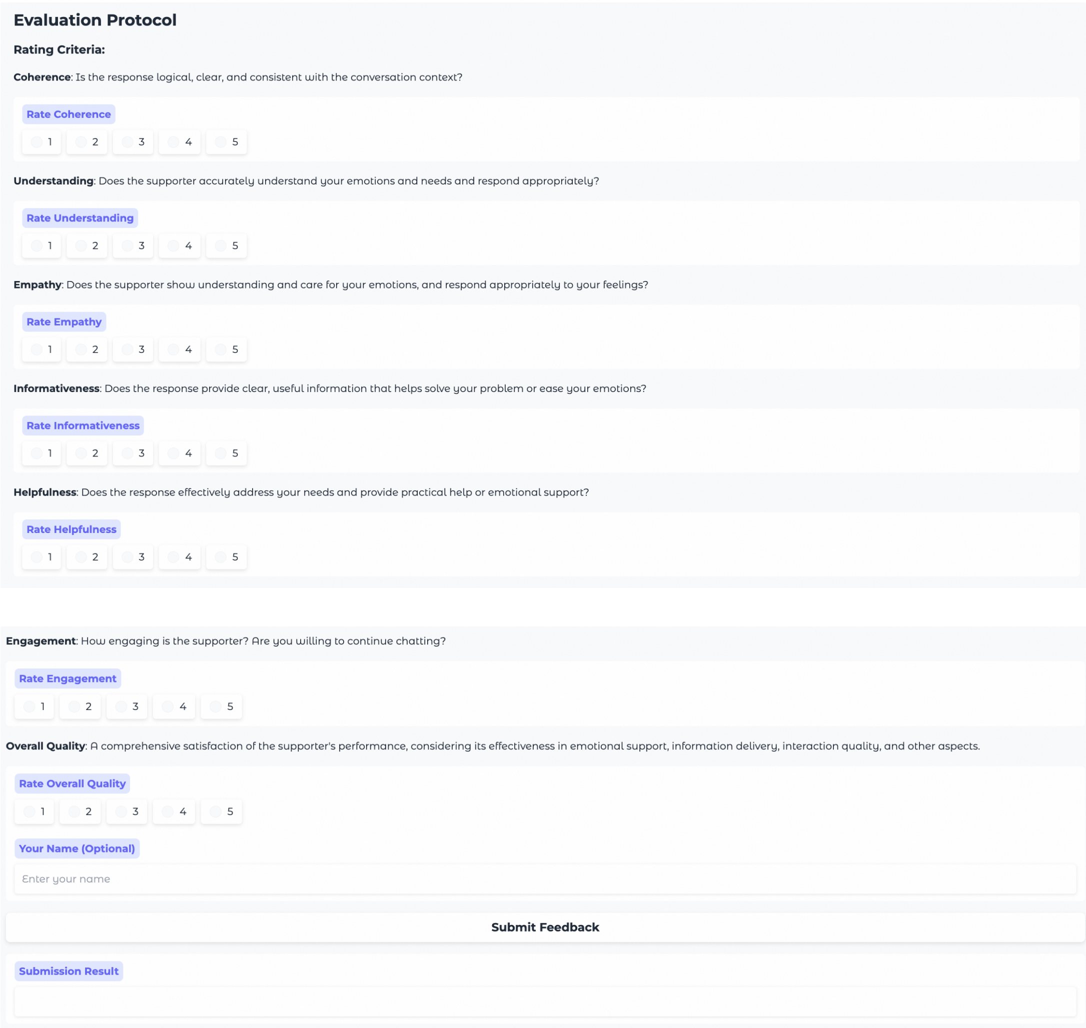


Figure 12: The interface of the interactive pair-wise human evaluation.



Evaluation Protocol

Rating Criteria:

Coherence: Is the response logical, clear, and consistent with the conversation context?

Rate Coherence

1 2 3 4 5

Understanding: Does the supporter accurately understand your emotions and needs and respond appropriately?

Rate Understanding

1 2 3 4 5

Empathy: Does the supporter show understanding and care for your emotions, and respond appropriately to your feelings?

Rate Empathy

1 2 3 4 5

Informativeness: Does the response provide clear, useful information that helps solve your problem or ease your emotions?

Rate Informativeness

1 2 3 4 5

Helpfulness: Does the response effectively address your needs and provide practical help or emotional support?

Rate Helpfulness

1 2 3 4 5

Engagement: How engaging is the supporter? Are you willing to continue chatting?

Rate Engagement

1 2 3 4 5

Overall Quality: A comprehensive satisfaction of the supporter's performance, considering its effectiveness in emotional support, information delivery, interaction quality, and other aspects.

Rate Overall Quality

1 2 3 4 5

Your Name (Optional)

Enter your name

Submit Feedback

Submission Result

Figure 13: The detailed guidelines for human evaluation.