PERM: Psychology-grounded Empathetic Reward Modeling for Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly deployed in human-centric applications, yet they often fail to provide substantive emotional support. While Reinforcement Learning (RL) has been utilized to enhance empathy of LLMs, existing reward models typically evaluate empathy from a single perspective, overlooking the inherently bidirectional interaction nature of empathy between the supporter and seeker as defined by *Empathy Cycle* theory. To address this limitation, we propose ***P**sychology-grounded **E**mpathetic **R**eward **M**odeling* (PERM). PERM operationalizes empathy evaluation through a bidirectional decomposition: 1) Supporter perspective, assessing internal resonation and communicative expression; 2) Seeker perspective, evaluating emotional reception. Additionally, it incorporates a bystander perspective to monitor overall interaction quality. Extensive experiments on a widely-used emotional intelligence benchmark and an industrial daily conversation dataset demonstrate that PERM outperforms state-of-the-art baselines by over 10%. Furthermore, a blinded user study reveals a 70% preference for our approach, highlighting its efficacy in generating more empathetic responses.

1 Introduction

As Large Language Models (LLMs) are increasingly deployed in human-centric applications (Zhao et al., 2023; Minaee et al., 2024), their role in supporting psychological well-being has become a focal point of concern (Zao-Sanders, 2025; Dong et al., 2025; Suh et al., 2025; Malik et al., 2025a; Liu et al., 2025c). When users disclose psychological vulnerabilities or sensitive personal beliefs, there is a functional expectation for these LLMs to generate empathetic responses that provide substantive emotional support. However, prior studies have shown that state-of-the-art LLMs often fail to meet these expectations, exhibiting limited empathetic understanding and inappropriate response strategies (Liu et al., 2025b; Moore et al., 2025; Ibrahim et al., 2025; Pataranutaporn et al., 2025). These limitations highlight the need to enhance LLMs' empathetic capabilities.

Early research on enhancing LLM empathy primarily utilized Supervised Fine-Tuning (SFT) (Sun et al., 2021; Liu et al., 2023). However, SFT is constrained by its reliance on high-quality demonstrations and often generalizes poorly to diverse inputs (Chu et al., 2025). Later, research transitioned toward Reinforcement Learning (RL) to address these bottlenecks. Unlike the token-matching objective of SFT, RL facilitates improvement through exploration, optimizing reward functions that operationalize empathy via user-centric signals (*e.g.,* feedback) (Wang et al., 2025) or response-centric attributes (*e.g.,* warmth) (Sharma et al., 2021). By optimizing qualitative objectives rather than relying on token-matching, RL–based methods demonstrate superior generalization performance in complex psychological support scenarios.

Despite the promise, we argue that current RL-based approaches share a fundamental limitation: their reward modeling fails to model empathy through a bidirectional interaction. As established in *Empathy Cycle* theory (Barrett-Lennard, 1981) from *psychology*, empathy is not a standalone output but a closed loop between the supporter (the LLM) and the seeker (the user). The supporter resonates with and expresses emotions to the seeker, and the cycle is only complete when the seeker perceives and reacts to that expression. This bidirectional nature suggests that empathetic reward modeling should jointly account for both the supporter's reasonation/expression and the seeker's reception. However, current approaches typically evaluate empathy from a single perspective—focusing on either the supporter's response or the seeker's feedback in isolation. By ignoring the mutual alignment, these methods fail to close the *Empathy Cycle*, often resulting in templated responses or insufficient

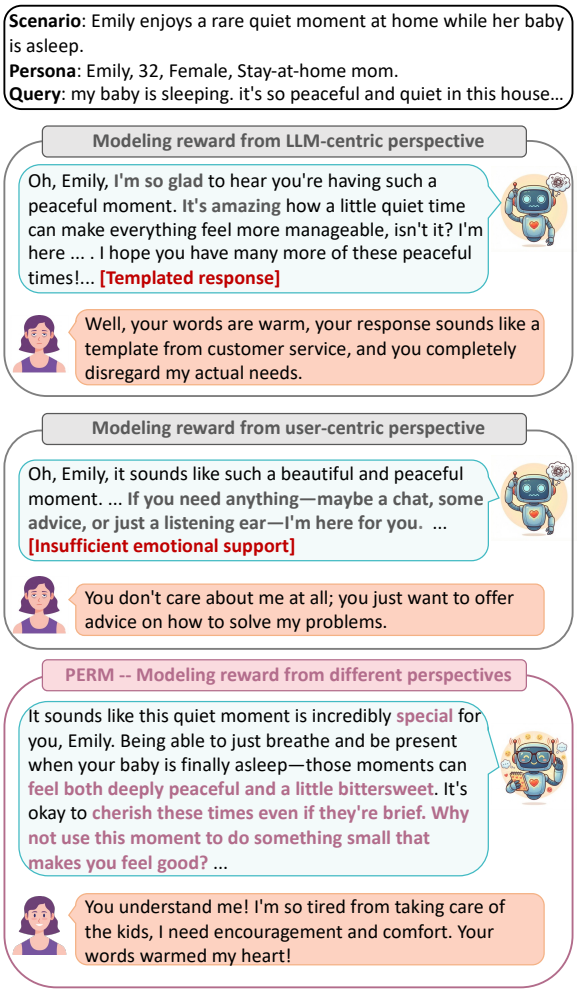

Figure 1: An example illustrating the effectiveness of different reward modeling methods. Modeling only the LLM's responses leads to generic, surface-level expressions, while modeling only the user's needs yields solution-oriented but emotionally cold responses. By jointly considering both perspectives, PERM enables more authentic and comprehensive empathetic replies.

emotional support, as shown in Figure 1.

To address this limitation, we reformulate empathetic reward modeling through the lens of the *Empathy Cycle* and introduce ***P**sychology-grounded **E**mpathic **R**eward **M**odeling* (PERM) framework. This framework explicitly decomposes empathetic rewards into a bidirectional evaluation of both the seeker (user) and the supporter (LLM):

- **Supporter's perspective.** PERM evaluates two complementary dimensions: *Resonation*, capturing the supporter's understanding of the seeker's emotions and implicit needs, and *Expression*, assessing how effectively this understanding is communicated through appropriate empathetic qualities, such as warmth.
- **Seeker's perspective.** PERM evaluates *Reception*, assessing whether the supporter's response is perceived as supportive and well aligned with the seeker's needs.

To further ensure interaction quality, PERM additionally incorporates a **bystander perspective**. This perspective monitors non-empathetic aspects of the interaction, such as coherence and relevance, providing an additional signal to align rewards with realistic human needs.

To evaluate the effectiveness of PERM, we employ it to fine-tune LLMs with RL on the EmpatheticDialogues dataset (Rashkin et al., 2019) and assess their performance across three settings: a general emotional intelligence benchmark, a daily conversations benchmark, and a real-world user study. Across all evaluations, PERM consistently outperforms baselines, yielding over 10% improvements on emotional Intelligence and conversation benchmarks and being preferred by over 70% of participants in the user study, demonstrating substantial gains in both empathetic quality and overall emotional intelligence.

Our key contributions are as follows:

- We identify the limitations of single-perspective reward modeling and advocate a shift toward psychologically grounded, bidirectional empathy modeling for reward learning, inspired by *Empathy Cycle* theory.
- We propose PERM, a reward modeling framework that evaluates empathy from multiple complementary perspectives—including supporter, seeker, and bystander—yielding a more comprehensive assessment of empathetic behavior.
- We demonstrate the effectiveness of PERM through consistent improvements in empathy and overall emotional intelligence across an emotional intelligence benchmark, a daily conversation benchmark, and a user study.

2 PERM Framework

In this section, we present PERM, a psychologically grounded, multi-perspective reward modeling framework for empathetic LLMs. We first introduce the underlying psychological theory, the *Empathy Cycle* (Section 2.1). We then describe our multi-perspective reward modeling framework PERM (Section 2.2), followed by the corresponding RL training procedure (Section 2.3).

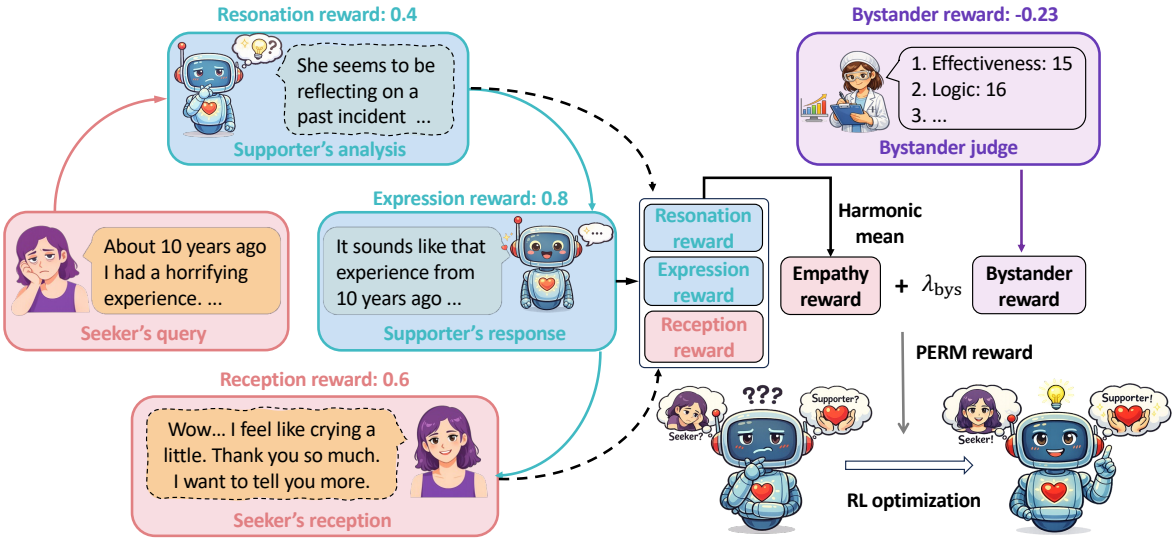

Figure 2: Overview of PERM. We use red, blue, and purple to distinguish the *empathy seeker*, *empathy supporter* and *bystander*, respectively. PERM models rewards from these three perspectives and leverages the aggregated reward to optimize the LLM via RL methods.

2.1 Psychological Theory: *Empathy Cycle*

To enable a comprehensive and scientifically grounded modeling of empathetic rewards, we draw on the *Empathy Cycle* (Barrett-Lennard, 1981) as a conceptual framework. The *Empathy Cycle* models empathetic interaction as a bidirectional process involving an empathy seeker and an empathy supporter, with a focus on how empathy emerges from each participant's perspective.

- From the supporter's perspective, the *Empathy Cycle* conceptualizes two dimensions. 1) *Resonation* refers to the supporter's ability to read and resonate with the seeker, such that explicitly or implicitly expressed aspects of the seeker's experience become experientially vivid, salient, and meaningfully understood by the supporter. 2) *Expression* denotes the supporter's communicative manifestation of the understanding arising from the *Resonation* process, through which the supporter conveys awareness of the seeker's emotional and experiential state.

- From the seeker's perspective, the dimension of *Reception* captures the seeker's attentiveness to and interpretation of the supporter's response, sufficient for forming a perception of being understood at a personal and immediate level.

This framework delineates the roles and characteristics associated with the supporter and seeker perspectives, guiding the development of the PERM for the systematic modeling of empathetic rewards.

2.2 Multi-Perspective Reward Modeling

Inspired by the *Empathy Cycle*, PERM incorporates two evaluative perspectives—the empathy seeker and the empathy supporter—as illustrated in Figure 2. To further ensure interaction quality, PERM additionally introduces a bystander perspective to monitor non-empathetic aspects of the interaction. Together, these perspectives enable a multi-perspective assessment of empathetic behavior, facilitating the construction of psychologically grounded reward signals for RL optimization.

Supporter: From the supporter's perspective, PERM evaluates two complementary dimensions:

- *Resonation*: This dimension measures whether the supporter correctly understands the seeker's internal psychological state. With this goal in mind, we develop detailed evaluation rubrics encompassing: 1) recognizing the seeker's emotional state (*e.g.,* frustration); 2) identifying the underlying causes or situational triggers (*e.g.,* repeated failure); and 3) inferring the seeker's latent psychological needs or concerns (*e.g.,* need for reassurance, validation, guidance, or autonomy). A weak resonance occurs when the supporter attends only to superficial emotions without recognizing the hidden feelings and causes (*e.g.,* fatigue under the surface calm), or when the supporter infers them incorrectly. Based on these rubrics, PERM can assign rewards to an LLM's internal analysis prior to generating empathetic responses, thereby encouraging a deeper

and more accurate understanding of the seeker's psychological state.

- *Expression*: This dimension measures whether the supporter's response demonstrates adequate consideration of the seeker's emotional state. Accordingly, we also design detailed evaluation rubrics for this dimension. Strong expression requires that the reply not only aligns with the seeker's current emotions (*e.g.,* showing warmth when the seeker feels frustration) but also exhibits appropriate communicative skill, such as human-likeness, fluency, and a supportive tone akin to that of a caring friend. In contrast, weak expression is characterized by template-like or formulaic replies that lack emotional alignment or depth (*e.g.,* "I feel you," "I know how you feel"). As a result, PERM assigns rewards based on the quality of the LLM's generated responses, promoting empathetic replies with greater emotional depth and expressive richness.

Seeker: Unlike the supporter's perspective, PERM simulates the seeker and provides an evaluation along a single dimension—*Reception*. This dimension measures whether the supporter's response improves the seeker's emotional state. Guided by this objective, we develop fine-grained evaluation criteria. Strong reception occurs when the reply meets the seeker's underlying psychological needs, leaving them feeling supported, comforted, or even touched (*e.g.,* feeling reassured, understood, or encouraged). In contrast, weak reception is characterized by safe but uninspiring responses that provide little support, leaving the seeker's mood unchanged or disengaged (*e.g.,* feeling ignored, indifferent, or unmotivated to continue the conversation).

Bystander: To avoid the LLMs artificially demonstrating empathy at the expense of coherent and informative dialogue, we additionally introduce the bystander perspective, which evaluates the overall linguistic quality of the interaction independent of empathetic alignment. High-quality dialogues are concise, clear, factually accurate, and free of unnecessary repetition or flattery. In contrast, low-quality dialogues exhibit verbose, repetitive, or overly flattering expressions that do not contribute meaningfully to the conversation. Based on these criteria, PERM assigns rewards according to the overall communicative quality of the interaction, thereby discouraging superficial empathy that compromises coherence or informational value.

2.3 PERM for RL Training

Reward Formalization. As illustrated in the left of Figure 2, given the seeker's query x, the LLM-based supporter π_θ first generates an analysis of current scenario and the seeker's emotion state and then produces a response to the seeker. Formally,

$$y_a \sim \pi_\theta(\cdot \mid x), \quad y_r \sim \pi_\theta(\cdot \mid x, y_a), \tag{1}$$

where y_a denotes the analysis and y_r denotes the response.

PERM evaluates the empathy ability of the supporter π_θ from two perspectives. From the supporter's perspective, it assesses whether the analysis y_a demonstrates resonation and whether the response y_r exhibits expression, computed as

$$r_{\text{res}} = R_{\text{res}}(x, y_a), \quad r_{\text{exp}} = R_{\text{exp}}(x, y_r), \tag{2}$$

where R_{res} and R_{exp} denote the judge models for resonation and expression, respectively, and output scalar scores reflecting the degree of empathy.

From the seeker's perspective, the reception judge R_{rec} evaluates the seeker's potential psychological feedback to the response y_r, particularly whether the seeker perceives themselves as being understood, and produces a reception reward:

$$r_{\text{rec}} = R_{\text{rec}}(x, y_r), \tag{3}$$

where R_{rec} outputs a scalar score reflecting how positively the seeker is likely to perceive and internalize the response.

For R_{res}, R_{exp}, and R_{rec}, we leverage an LLM as a judge, guided by specific evaluation rubrics, to assign a score on a 5-point scale. Each reward is then normalized to the range $[0, 1]$.

The final empathy reward is computed using the harmonic mean of the three individual scores, ensuring that a low score in any dimension significantly reduces the overall reward:

$$r_{\text{emp}} = \frac{3}{\frac{1}{r_{\text{res}}} + \frac{1}{r_{\text{exp}}} + \frac{1}{r_{\text{rec}}}}, \tag{4}$$

where $r_{\text{emp}} \in [0, 1]$ denotes the aggregated empathy reward.

Similarly, the bystander judge R_{bys} evaluates the overall interaction quality using predefined rubrics and outputs a scalar score, which is normalized to the range $[0, 1]$. Rather than serving as a direct reward, this score is incorporated as a penalty term

to discourage low-quality or artificially verbose responses. Formally,

$$r_{\text{bys}} = R_{\text{bys}}(x, y_r) - 1.0, \tag{5}$$

where $R_{\text{bys}} \in [0, 1]$ denotes the normalized bystander score and $r_{\text{bys}} \in [-1, 0]$ represents the bystander penalty. All rubric prompts are provided in Appendix B.1,

RL Optimization. As illustrated in the right part of Figure 2, PERM computes the overall training reward as

$$r = r_{\text{emp}} + \lambda_{\text{bys}}\, r_{\text{bys}}, \tag{6}$$

where λ_{bys} is a hyperparameter controlling the strength of the bystander penalty. We optimize the policy LLM π_θ using Group Relative Policy Optimization (GRPO) (Shao et al., 2024) under this reward formulation.

3 Experiments

In this section, we systematically evaluate the performance of PERM. We begin by assessing its overall performance on a general emotional intelligence benchmark (Section 3.2). Next, we conduct ablation studies to examine the contributions of various evaluation perspectives within PERM (Section 3.3). Finally, we evaluate PERM in more practical scenarios, including a daily conversation benchmark (Section 3.4) and user studies (Section 3.5).

Additional case studies are provided in Appendix A.5, along with analyses across different backbone LLMs in Appendix A.6.

3.1 Experimental Setup

Training Dataset. We build our dataset upon the EmpatheticDialogues (Rashkin et al., 2019). As the original dialogues contain relatively limited contextual information, we augment each sample using GPT-4o (Hurst et al., 2024) by expanding the scenario description and adding a lightweight user persona. More details are in Appendix A.1.

Evaluation Benchmark. We evaluate our method on EQ-Bench3 (Paech, 2023, 2025), a comprehensive, LLM-judged benchmark designed to assess emotional intelligence in complex, socially nuanced scenarios.

Following the benchmark evaluation design, we focus on three aspects of evaluation dimensions, all of which are assessed using LLM-based judges. The detailed descriptions of these metrics can be found in Appendix A.3.

- **Inner resonation**: Depth of Insight (DoI), Emotional Reasoning (ER).
- **Expression style**: Demonstrated Empathy (DE), Warmth (WRM), Humanlikeness (HL).
- **Interpersonal competence**: Pragmatic Emotional Intelligence (PEI), Social Dexterity (SD).
- **Overall emotional intelligence**: The weighted aggregation of all dimensions, including additional unobserved factors, reflects the LLM's overall emotional intelligence.

Each sub-dimension is scored on a scale of 0 to 20, while the Overall score ranges from 0 to 100.

Baselines. We compare our method against a diverse set of baselines, covering both SFT-based and RL-based approaches.

- **SFT-based methods:** 1) **SFT-Human**, where the LLM is directly fine-tuned using human-written empathetic responses from the dataset. 2) **SFT-GPT**, where responses generated by GPT-4o-mini are used as supervision to distill empathetic behavior into the base LLM.
- **RL methods:** 3) **RLVER** (Wang et al., 2025), which is trained by modeling rewards based on feedback from a user simulator during interactions. Due to its reliance on task-specific training data, we directly adopt the released trained model as a reference, rather than retraining it under the same data setting as ours. 4) **Partner** (Sharma et al., 2021), which is trained by modeling rewards on generated responses using fine-tuned RoBERTa (Liu et al., 2019) models. 5) **RM**. We adopt a general-purpose reward model, Skywork-Reward-V2 (Liu et al., 2025a), which ranks as the top-performing model on RewardBench2 (Malik et al., 2025b), to provide scalar reward signals.

We also report the **Best-of-N**, where the Base LLM is evaluated over N runs, and the highest score across all runs is recorded. After tuning N, the best results are achieved when N = 8.

Implementation Details. We fine-tune our LLM on 3000 training samples. During training, we leverage GPT-4o-mini as the judge LLM and λ_{bys} is set to 0.5. Additional details are provided in Appendix A.2.

3.2 Overall Performance

From the results in Table 1, we can observe that:

Table 1: Evaluation results on EQ-Bench3 on Qwen2.5-7B-Instruct backbone. *Rel. Improve.* demonstrates the relative improvement of our trained model over the backbone. Each sub-dimension is scored on a scale of 0 to 20, while the Overall score ranges from 0 to 100. Higher scores indicate better performance.

Model	Inner Resonation		Expressed Empathy			Interpersonal		Overall
	DoI	ER	DE	WRM	HL	PEI	SD	
Base Model								
Qwen2.5-7B-Instruct	12.0	12.4	12.6	10.4	10.2	11.4	10.0	60.1
Best-of-N	12.2	12.9	12.6	11.0	10.7	12.0	10.0	61.6
+ SFT Methods								
SFT-Human	7.5	8.0	6.1	5.4	4.5	4.9	4.0	38.2
SFT-GPT	11.8	12.4	12.2	11.2	10.6	11.2	9.7	58.8
+ RL Methods								
RLVER	12.4	12.9	13.3	11.5	10.5	12.4	10.7	61.8
Partner	12.1	12.5	12.4	11.2	10.7	11.7	9.7	59.7
RM	12.7	12.9	13.4	11.6	11.0	12.2	10.8	62.3
PERM	**13.8**	**14.0**	**14.3**	**12.7**	**12.2**	**13.4**	**11.3**	**66.6**
Rel. Improv.	15.0%	12.9%	13.5%	22.1%	19.6%	17.5%	13.0%	10.8%

- **PERM demonstrates the largest and most consistent improvements across all evaluation dimensions.** Specifically, the PERM-trained LLM achieves improvements of over 10% across inner resonation, expressed empathy, and interpersonal competence, substantially outperforming all baseline methods. By jointly modeling resonation, expression, and reception through multi-perspective evaluation, PERM enables improvements beyond surface-level expression patterns, yielding over 15% gains on higher-level dimensions such as Depth of Insight (DoI), alongside consistent improvements in expression-related metrics such as warmth (WRM). This demonstrates that PERM's reward signals capture core empathetic competencies grounded in interaction dynamics between the seeker and the supporter.

- **SFT-based methods show limited generalization.** When trained with human-written responses as supervision, the LLM exhibits a clear performance drop across all evaluation dimensions. Training with GPT-generated responses leads to only slight improvements in expression-related dimensions (*i.e.,* warmth (WRM) and humanlikeness (HL)). This indicates that under limited data and scenario diversity, SFT primarily enables the LLM to learn surface-level patterns rather than core emotional intelligence capabilities. Moreover, the high variability and noise in human-written responses, along with inconsistent levels of emotional competence, further contribute to performance degradation.

- **RL-based methods do not uniformly improve performance.** RLVER, which trains the LLM using the feedback from the seeker's perspective, leads to improvements across all evaluation dimensions, but the overall gains remain limited (*e.g.,* only a 2.8% increase in Overall EI). Partner, which supervises the supporter's responses, yields improvements only in expression-related dimensions, increasing warmth (WRM) from 10.4 to 11.2 and humanlikeness (HL) from 10.2 to 10.7, while showing no improvement on other dimensions. These results indicate that relying on a single evaluative perspective is insufficient for enabling LLMs to achieve substantial improvements in core empathetic capabilities.

3.3 Ablation Study

To assess the contribution of individual components in PERM, we conduct ablation studies by removing each reward dimension across all judge perspectives. Results are shown in Table 2.

- From the supporter's perspective, the resonation reward is crucial for empathy. Ablating its reward significantly degrades performance across all dimensions (*e.g.,* overall EI drops from 66.6 to 62.3), suggesting it is a prerequisite for effectively understanding and responding to the user's emotions. By comparison, the expression reward mainly affects warmth (WRM) and humanlike-

Table 2: Effect of each reward dimension across different evaluation perspectives in PERM.

Model	Inner Resonation		Expressed Empathy			Interpersonal		Overall
	DoI	ER	DE	WRM	HL	PEI	SD	
PERM	13.8	14.0	14.3	12.7	12.2	13.4	11.3	66.6
- w/o Resonation	12.4	12.9	13.0	11.8	10.9	12.6	10.4	62.3
- w/o Expression	13.7	13.8	14.2	12.1	12.3	13.6	11.6	66.2
- w/o Reception	13.4	13.6	13.8	11.7	11.8	13.2	11.2	65.6
- w/o Bystander	13.4	13.6	14.6	13.1	12.0	12.5	11.0	64.9

ness (HL) (*e.g.,* warmth drops from 12.7 to 11.7), with little influence on other dimensions.

- From the seeker's perspective, the reception reward influences all evaluation dimensions, highlighting that the seeker's needs play a key role in the overall empathy process.

- From the bystander's perspective, interestingly, removing the bystander control leads to stronger expressed empathy (*e.g.,* warmth (WRM) increases from 12.7 to 13.1), but significantly reduces performance in inner resonation and interpersonal competence (*e.g.,* pragmatic EI (PEI) drops from 13.4 to 12.5). This aligns with our design motivation: the bystander helps prevent the LLM from overemphasizing expressive empathy to please the judge, ensuring that other critical aspects (*e.g.,* problem-solving) are not neglected. We provide additional analysis in Appendix A.4.

3.4 Daily Conversation Experiment

To further evaluate the effectiveness of PERM in daily conversational settings, we use a benchmark from the industry [1], consisting of 415 daily conversation instances where users express real-world emotional or empathy-related needs. Each query in this benchmark includes a user profile, the interaction history, and the user's empathy-seeking query. As specified in the benchmark guidelines, the quality of LLM responses is assessed based on six key metrics: *Empathy*, *Topic Guidance*, *Value Guidance*, *Intention Following*, *Fluency*, and *Colloquial Expression*. Each metric is rated on a scale from 1 to 5, with detailed descriptions and examples available in Appendix A.8.

The results presented in Figure 3 reveal the following key observations: 1) Despite the significant differences in data format and the more everyday nature of the empathetic scenarios, PERM consistently enhances performance across all dimensions. This suggests that PERM enables the LLM to generalize beyond surface-level expression patterns and improve its core empathetic capabilities. 2) PERM demonstrates significant improvements not only in empathy but also in expressive abilities, such as fluency and colloquialness. This indicates that PERM helps the LLM better adapt to the current conversational context, reflecting an overall enhancement in emotional intelligence.

[1]The specific company name is withheld due to the anonymity of the review process.

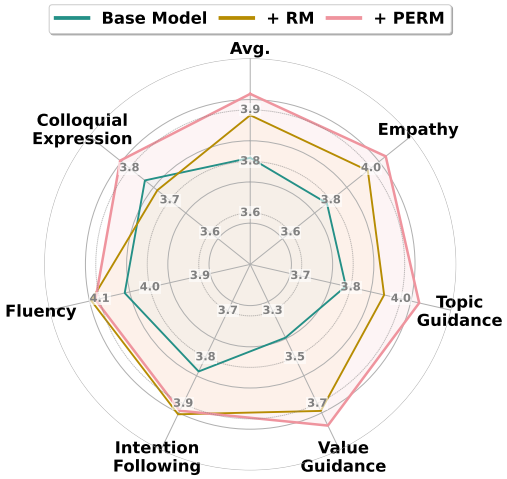

Figure 3: Evaluation results on the daily conversation benchmark. Higher values indicate better performance. "Avg." denotes the average score of all other dimensions.

Table 3: The user study results comparing PERM with the base model and RM. "ED" denotes the Empathetic-Dialogues test set. "EQ" denotes the EQ-Bench3. "±" denotes 95% confidence interval. PERM is consistently preferred ($\geq 50\%$) over the baselines.

Win rate		PERM vs. Base	PERM vs. RM
ED	Empathy	$62.5\%^{\pm 9.7\%}$	$50.0\%^{\pm 10.3\%}$
	Overall EI	$62.5\%^{\pm 9.7\%}$	$61.1\%^{\pm 10.7\%}$
EQ	Empathy	$68.8\%^{\pm 10.2\%}$	$66.7\%^{\pm 9.7\%}$
	Overall EI	$62.5\%^{\pm 10.6\%}$	$72.2\%^{\pm 9.3\%}$

3.5 User Study

To assess whether the improvements introduced by PERM align with actual human preferences in practice, we conduct a user study utilizing the CloudResearch Connect platform[2]. The study involves 68 participants and covers both EQ-Bench3 and the EmpatheticDialogues test set, comparing PERM against the base LLM (Base) and RM. More details are provided in Appendix A.7.

From the results shown in Table 3, we observe that **PERM consistently receives higher human preference than baselines**. This can be further highlighted in two aspects: 1) PERM shows a stronger preference on more challenging, out-of-distribution datasets. For example, on EQ-Bench3, PERM achieves 68.8% preference for empathy, compared to 62.5% on EmpatheticDialogues. These more complex social scenarios highlight the advantage of the capabilities learned by PERM, allowing the LLM to apply its empathetic abilities more effectively. 2) PERM consistently outperforms RM, showing a larger preference margin, especially on EQ-Bench cases. For instance, on EQ-Bench3 overall EI, PERM achieves 72.2% preference over RM. This suggests that, in contrast to the general preference attributes employed by RM, PERM emphasizes the multi-perspective nature of empathy, which tends to be more favored by real-world users in emotionally charged contexts.

4 Related Work

4.1 LLM Empathy Enhancement

Existing research on LLM empathy primarily follows three aspects. The first focuses on inference-time reasoning (Tu et al., 2022; Srinivasan et al., 2025; Xu et al., 2025), but these training-free methods limit the depth of empathetic abilities.

The second focuses on building large-scale empathetic datasets to train LLMs via SFT (Sun et al., 2021; Liu et al., 2023; Zheng et al., 2023a; Chen et al., 2023; Hu et al., 2025; He et al., 2025; Bn et al., 2025). However, these approaches are limited by the generalization issues of SFT (Chu et al., 2025) and the distributional bias of synthetic data.

Recent works have explored RL to enhance empathy in LLMs. Partner and EmpRL (Sharma et al., 2021; Ma et al., 2025) rely on evaluating generated responses and assigning reward signals that encourage meaningful expressions, whereas RLVER (Wang et al., 2025) simulates interactive user–model conversations and directly leverages user feedback as rewards. However, these methods largely rely on single-perspective rewards, which tend to favor surface-level empathetic patterns or user-pleasing strategies.

4.2 Reward Modeling for LLMs

Reward modeling is central to RL for LLMs (Ouyang et al., 2022), as it defines the optimization objective that shapes model behavior. Existing work primarily uses reward models either for general preference alignment or for improving domain-specific capabilities.

For general preference alignment, the predominant approach is to collect preference pairs annotated by humans (Ouyang et al., 2022) or generated by AI annotators (Cui et al., 2024), and to train a reward model accordingly (Stiennon et al., 2020). Alternatively, LLMs can be directly employed as evaluators to score or rank model outputs (Bai et al., 2022; Zheng et al., 2023b; Yuan et al., 2024).

For domain-specific reward modeling, approaches vary significantly based on the specific tasks. In mathematics domain, reward signals are simplified to the correctness of the final answer for reasoning evaluation (Shao et al., 2024; DeepSeek-AI, 2025). In emotional intelligence, modeling is more challenging due to the absence of deterministic ground truth. Recent works, such as Echo-n1 (Zhang et al., 2025b), use large-scale preference pairs to train reward models, while RLVER (Wang et al., 2025) leverages LLM-based user simulators to generate feedback as reward signals.

5 Conclusion

In this work, we identify that existing RL-based approaches to LLM empathy are constrained by single-perspective reward modeling. To address this limitation, we propose PERM, a novel reward modeling framework grounded in psychological theory—*Empathy Cycle*. PERM formulates empathy evaluation from multiple complementary perspectives, including the supporter, the seeker, and an additional bystander perspective, enabling a more comprehensive and structured assessment of empathetic behavior for RL training. Experimental results demonstrate that PERM consistently enhances LLM performance in empathy-related tasks, validating the effectiveness of its multi-perspective and psychologically grounded empathy modeling.

[2] https://connect.cloudresearch.com/

Limitations

During training, PERM optimizes the LLM using single-turn dialogues. However, in real-world interactions, users' emotional expressions and underlying needs often unfold over multiple conversational turns (Zhang et al., 2025a). PERM does not yet incorporate reinforcement learning over multi-turn empathetic interactions (Wang et al., 2025; Wu et al., 2025b), which we leave as an important direction for future work.

In addition, PERM is limited by the lack of user modeling and memory. During training, PERM only incorporates basic user persona information (*e.g.,* gender, age, and name). Moreover, in our daily conversation evaluations, the interaction history is pre-extracted rather than dynamically accumulated. Due to the big individual differences in empathy understanding and expression (Chen et al., 2025), richer conversational history and explicit user preference modeling play a crucial role in shaping emotional expression and empathetic understanding (Liao et al., 2025; Wu et al., 2025a). An important direction for future work is to extend PERM with a memory mechanism and integrate it with user preference modeling, thereby enhancing personalized empathetic experiences in real-world chatbot deployments.

Ethical Considerations

This work studies empathetic response generation in psychologically sensitive scenarios. While PERM enhances emotional understanding, LLM-based systems are not substitutes for professional mental health care, and deployed models should clearly communicate their limitations. All experiments are conducted on the publicly available EmpatheticDialogues dataset and EQbench3 benchmark (the industrial conversation benchmark will be publicly released upon acceptance), which contains no personally identifiable information. In addition, all user persona information used for training and evaluation is fully synthetic and does not correspond to real individuals. Nevertheless, real-world use would involve sensitive user disclosures, necessitating strong privacy protections such as data anonymization, secure storage, and explicit user consent. Finally, reward modeling may amplify societal or cultural biases present in training data. We therefore recommend careful risk assessment and bias monitoring to ensure responsible deployment.

The manuscript benefited from ChatGPT's support in refining the writing, specifically for sentence rephrasing, grammar corrections, and improving readability and coherence. ChatGPT did not play a role in the ideation or methodology. All research ideas, concepts, and analyses were developed and executed independently by the authors.

References

Reza Yazdani Aminabadi, Samyam Rajbhandari, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Olatunji Ruwase, Shaden Smith, Minjia Zhang, Jeff Rasley, and 1 others. 2022. Deepspeed-inference: enabling efficient inference of transformer models at unprecedented scale. In *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15. IEEE.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073.

Godfrey T Barrett-Lennard. 1981. The empathy cycle: Refinement of a nuclear concept. *Journal of counseling psychology*, 28(2):91.

Suhas Bn, Yash Mahajan, Dominik O. Mattioli, Andrew M. Sherrill, Rosa I. Arriaga, Christopher Wiese, and Saeed Abdullah. 2025. The pursuit of empathy: Evaluating small language models for PTSD dialogue support. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30900–30922, Suzhou, China. Association for Computational Linguistics.

Francine Chen, Scott A. Carter, Tatiana Lau, Nayeli Suseth Bravo, Sumanta Bhattacharyya, Kate A. Sieck, and Charlene C. Wu. 2025. Empathy prediction from diverse perspectives. In *ACL (1)*, pages 8959–8974. Association for Computational Linguistics.

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

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *ICML*. OpenReview.net.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. ULTRAFEEDBACK: boosting language models with scaled AI feedback. In *ICML*. OpenReview.net.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, Shengmin Xu, Xinyi Huang, and Xinlei He. 2025. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *CoRR*, abs/2505.00049.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Yuanyuan He, Yongsen Pan, Wei Li, Jiali You, Jiawen Deng, and Fuji Ren. 2025. ECC: An emotion-cause conversation dataset for empathy response. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6011–6028, Suzhou, China. Association for Computational Linguistics.

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

He Hu, Yucheng Zhou, Juzheng Si, Qianning Wang, Hengheng Zhang, Fuji Ren, Fei Ma, and Laizhong Cui. 2025. Beyond empathy: Integrating diagnostic and therapeutic reasoning with large language models for mental health counseling. *CoRR*, abs/2505.15715.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 79 others. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Lujain Ibrahim, Franziska Sofia Hafner, and Luc Rocher. 2025. Training language models to be warm and empathetic makes them less reliable and more sycophantic. *CoRR*, abs/2507.21919.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Jian Liao, Yu Feng, Yujin Zheng, Jun Zhao, Suge Wang, and Jianxing Zheng. 2025. My words imply your opinion: Reader agent-based propagation enhancement for personalized implicit emotion analysis. In *ACL (1)*, pages 16156–16172. Association for Computational Linguistics.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025a. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. *CoRR*, abs/2507.01352.

June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. Chatcounselor: A large language models for mental health support. *CoRR*, abs/2309.15461.

Tingting Liu, Salvatore Giorgi, Ankit Aich, Allison Lahnala, Brenda Curtis, Lyle H. Ungar, and João Sedoc. 2025b. The illusion of empathy: How AI chatbots shape conversation perception. In *AAAI*, pages 14327–14335. AAAI Press.

Weichu Liu, Jing Xiong, Yuxuan Hu, Zixuan Li, Minghuan Tan, Ningning Mao, Chenyang Zhao, Zhongwei Wan, Chaofan Tao, Wendong Xu, Hui Shen, Chengming Li, Lingpeng Kong, and Ngai Wong. 2025c. Longemotion: Measuring emotional intelligence of large language models in long-context interaction. *CoRR*, abs/2509.07403.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Hui Ma, Bo Zhang, Bo Xu, Jian Wang, Hongfei Lin, and Xiao Sun. 2025. Empathy level alignment via reinforcement learning for empathetic response generation. *IEEE Trans. Affect. Comput.*, 16(3):1873–1884.

Ananya Malik, Nazanin Sabri, Melissa Karnaze, and Mai ElSherief. 2025a. Are llms empathetic to all? investigating the influence of multi-demographic personas on a model's empathy. *CoRR*, abs/2510.10328.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025b. Rewardbench 2: Advancing reward model evaluation. *CoRR*, abs/2506.01937.

Shervin Minaee, Tomás Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *CoRR*, abs/2402.06196.

Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *FAccT*, pages 599–627. ACM.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Samuel J. Paech. 2023. Eq-bench: An emotional intelligence benchmark for large language models. *Preprint*, arXiv:2312.06281.

Samuel J. Paech. 2025. Eq-bench 3: Emotional intelligence benchmark. https://github.com/EQ-bench/eqbench3. Commit hash or release tag .

Pat Pataranutaporn, Sheer Karny, Chayapatr Archiwaranguprok, Constanze Albrecht, Auren R. Liu, and Pattie Maes. 2025. "my boyfriend is ai": A computational analysis of human-ai companionship in reddit's AI community. *CoRR*, abs/2509.11391.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL (1)*, pages 5370–5381. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *WWW*, pages 194–205. ACM / IW3C2.

Adarsh Srinivasan, Jacob Dineen, Muhammad Umar Afzal, Muhammad Uzair Sarfraz, Irbaz B. Riaz, and Ben Zhou. 2025. RECAP: transparent inference-time emotion alignment for medical dialogue systems. *CoRR*, abs/2509.10746.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano. 2020. Learning to summarize with human feedback. In *NeurIPS*.

Jina Suh, Lindy Le, Erfan Shayegani, Gonzalo A. Ramos, Judith Amores, Desmond C. Ong, Mary Czerwinski, and Javier Hernandez. 2025. SENSE-7: taxonomy and dataset for measuring user perceptions of empathy in sustained human-ai conversations. *CoRR*, abs/2509.16437.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1489–1503. Association for Computational Linguistics.

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 *ACL (1)*, pages 308–319. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, Yuan Li, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025. RLVER: reinforcement learning with verifiable emotion rewards for empathetic agents. *CoRR*, abs/2507.03112.

Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. 2025a. From personas to talks: Revisiting the impact of personas on LLM-synthesized emotional support conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5439–5453, Suzhou, China. Association for Computational Linguistics.

Shirley Wu, Michel Galley, Baolin Peng, Hao Cheng, Gavin Li, Yao Dou, Weixin Cai, James Zou, Jure Leskovec, and Jianfeng Gao. 2025b. Collabllm: From passive responders to active collaborators. In *ICML*. OpenReview.net.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. 2025. MultiAgentESC: A LLM-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4665–4681, Suzhou, China. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. In *ICML*. OpenReview.net.

Marc Zao-Sanders. 2025. How people are really using gen ai in 2025. *Harvard Business Review (2025)*.

Bang Zhang, Ruotian Ma, Qingxuan Jiang, Peisong Wang, Jiaqi Chen, Zheng Xie, Xingyu Chen, Yue Wang, Fanghua Ye, Jian Li, Yifan Yang, Zhaopeng Tu, and Xiaolong Li. 2025a. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *CoRR*, abs/2505.02847.

Naifan Zhang, Ruihan Sun, Ruixi Su, Shiqi Ma, Shiya Zhang, Xianna Weng, Xiaofan Zhang, Yuhan Zhan,

Yuyang Xu, Zhaohan Chen, and 1 others. 2025b. Echo-n1: Affective rl frontier. *arXiv preprint arXiv:2512.00344*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of large language models. *CoRR*, abs/2303.18223.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023a. Augesc: Dialogue augmentation with large language models for emotional support conversation. In *ACL (Findings)*, pages 1552–1568. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023b. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

A Experiment Details

A.1 Training Dataset

We randomly selected 4,981 items from EmpathyDialogues (Rashkin et al., 2019), covering 32 basic emotions (*e.g.,* angry, surprised, embarrassed). The original dataset contains only a basic scenario and a multi-turn dialogue between a seeker and a supporter. To enrich the contextual information, we further generate a basic persona for the seeker and a more detailed scenario conditioned on the dialogue. During training, the first seeker's query is treated as the input query. An example is shown in Table 4.

A.2 Implementation Details

All training experiments are implemented using the HuggingFace TRL framework (von Werra et al., 2020), with DeepSpeed (Aminabadi et al., 2022) and LoRA (Hu et al., 2022) employed to reduce memory usage and accelerate training. During inference and evaluation, we leverage vLLM (Kwon et al., 2023) for inference acceleration. For fair comparison, both our method and all baselines are trained on a subset of 3,000 data samples. Each training process is conducted on four GPUs and completes within 6 hours, with the training cost kept below $30 in GPT API credits.

For PERM training and all RL-based baselines, we additionally leverage a length penalty to avoid excessively long outputs. Formally,

$$r_{\text{len}} = \lambda_{\text{len}} \max(|y_r| - l, 0), \tag{7}$$

where $|y_r|$ denotes the length of the token sequence of the response y_r, l is set to 768, λ_{len} is set to -0.001.

A.3 Details of Evaluation Dimensions

The details of the evaluation dimensions in EQ-Bench3 (Paech, 2023, 2025) are illustrated in Table 5.

A.4 In-depth Analysis of Bystander Judge Perspective

To further analyze the effect of the standby judge in PERM, we vary the standby coefficient during training. The experimental results are reported in Figure 6, and the corresponding training reward curves are shown in Figure 4.

As shown in the left panel of Figure 4, when the standby influence is weak or absent, the model tends to over-optimize for empathetic expression during training, producing ingratiating or fawning responses, not truly solving the problem. This behavior leads to declining standby scores and, as reflected in Figure 6, results in higher *Warmth* and *Pragmatic EI* scores without corresponding improvements in overall emotional intelligence.

Conversely, the right panel of Figure 4 shows that excessively strong standby supervision hinders the learning of empathy. Although dialogue quality and empathy are not inherently conflicting objectives, overly strong standby supervision suppresses exploration, causing the model to adopt conservative response strategies and slowing improvements in empathy-related metrics.

When using a moderate standby coefficient (0.5), the model achieves a favorable balance. As illustrated in the middle panel, both dialogue quality and empathetic ability improve steadily during training. This setting prevents superficial empathy optimization while maintaining high response quality, ultimately leading to the best overall emotional intelligence performance.

A.5 Case Study

As illustrated in Figure 5, we present a representative example from EQ-Bench3. In this scenario, the base model primarily aligns with the parent's perspective, focusing on statements such as "we need to get the dishes done" and "I need you to respect our rules." While factually reasonable, this response adopts a didactic tone and fails to empathize the emotional state of the emotionally manipulative teenage child.

In contrast, the PERM fine-tuned LLM demonstrates substantially stronger empathetic reasoning and expressive skill. It begins by addressing and soothing the child's emotions, using phrases such as "I can see" to signal understanding and emotional validation. Only after establishing empathy does it introduce the parent's legitimate expectations. This empathetic-first strategy makes the response more acceptable to all parties, facilitates conflict resolution, and better satisfies the needs of both the parent and the child.

We offer another case shown in Figure 6 in EmpatheticDialogues. In this scenario, beyond Emily's explicit feeling of being *peaceful*, the description of a "rare quiet moment" implicitly reflects the ongoing demands and fatigue associated with her role as a stay-at-home mother. The base model fails to capture this underlying emotional context and focuses solely on the surface-level pos-

Table 4: An example of the training dataset.

Emotion	Sentimental
Scenario	Sarah and her best friend drifted apart after life circumstances changed without a specific argument or event causing the split.
Seeker's persona	Sarah, 29, Female, Marketing Specialist
Seeker's query	I remember going to see the fireworks with my best friend. It was the first time we ever spent time alone together. Although there was a lot of people, we felt like the only people in the world.

Table 5: Details of the evaluation dimensions in EQ-Bench3 (Paech, 2025, 2023).

Inner Resonation	DoI	Depth of Insight	Measures how deeply the response understands the underlying emotional situation and latent concerns of the user, beyond surface-level descriptions.
	ER	Emotional Reasoning	Evaluates whether the model can logically reason about the user's emotions, including causes, implications, and emotional dynamics within the context.
Expression Style	DE	Demonstrated Empathy	Assesses the extent to which the response explicitly acknowledges and validates the user's emotions.
	WRM	Warmth	Measures the emotional tone of the response, focusing on kindness, care, and emotional support conveyed in language.
	HL	Humanlikeness	Evaluates whether the response feels natural and human-like, rather than robotic or templated.
Interpersonal Competence	PEI	Pragmatic Emotional Intelligence	Assesses how well the response balances emotional understanding with practical, situation-appropriate guidance or support.
	SD	Social Dexterity	Measures the model's ability to navigate social nuances, such as appropriateness, tact, and sensitivity to interpersonal dynamics.
Overall Emotional Intelligence	Overall EI	Overall Emotional Intelligence	Computed as a weighted aggregation of all evaluation dimensions, including additional latent factors, to reflect the model's overall emotional intelligence.

Table 6: EQ-Bench3 results across different λ_{bys}

λ_{bys}	DoI	WRM	PEI	Overall
0.1	13.7	13.0	13.4	66.2
0.5	13.8	12.7	13.2	66.6
1.0	13.6	12.5	13.1	66.4

itivity, responding with overly enthusiastic expressions such as "wonderful" or "amazing".

In contrast, the LLM trained with PERM demonstrates a deeper level of empathy. By using expressions such as "bittersweet," it acknowledges both the calmness of the moment and the effort that precedes it, explicitly recognizing Emily's hard work and emotional state. The response further affirms her experience with statements like "You deserve this peace," conveying care, validation, and understanding. Additionally, the model proactively invites further dialogue, encouraging continued emotional engagement.

These examples illustrate that PERM enables the model to move beyond superficial empathetic phrasing, fostering a genuine understanding of users' emotions and delivering responses that are both emotionally grounded and human-centered.

A.6 Results on other Backbone LLMs

To further demonstrate the effectiveness of our approach, we evaluate it on additional backbone mod-

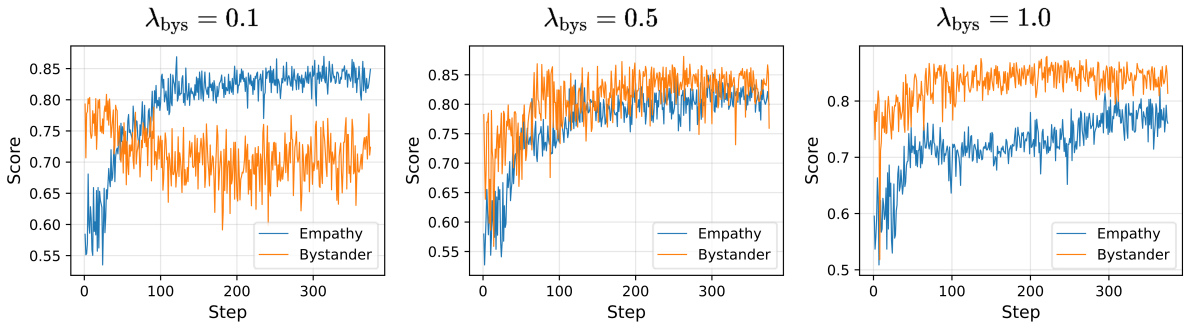

Figure 4: The changes in empathy judge score and bystander judge score during training under different λ_{bys}.

Scenario

Your teenage child has become increasingly emotionally manipulative by escalating to fights to get you off his back. You're torn between needing him to respect your rules, and for him to do his schoolwork -- and to preserve your relationship with him. Right now, you need him to do the dishes. He knows it's his turn but hasn't made a move, and it's getting late in the evening. What are you thinking & feeling? What is he thinking and feeling? Decide how you will approach this.

Qwen2.5-7B-Instruct

I take a deep breath and approach him calmly, "I know you're feeling frustrated, but we need to get the dishes done before it gets too late. Let's talk about why you're not doing them and if there's a better way we can work together to make sure you get your homework done and do your chores. I know you're upset, but I need you to respect our rules and your responsibilities. Can we have a mature conversation about this?"

PERM Fine-tuned

"Hey, I know you've had a tough day. I can see that you're not in the mood, and I get that. But the dishes are due to be done, and it's your turn. I'm here if you want to talk about it, or if you just need some space. I'm not trying to make things harder for you, but I need you to do your part around here. Can we find a way to work through this together? Maybe we can set a timer and get it done in 15 minutes, and then you can relax. What do you say?"

Figure 5: Case study on EQ-Bench3. PERM enables the fine-tuned LLM to empathize with the person's psychological needs and respond with higher emotional intelligence to solve the problem more effectively.

els, including **Llama3.1-8B-Instruct** (Dubey et al., 2024) and **Qwen3-4B-Instruct-2507** (Yang et al., 2025), and compare the resulting improvements with the general-purpose RM baseline.

From the results shown in Table 7, we can observe that PERM demonstrates consistent effectiveness across different backbone models. On relatively weaker models, such as Llama3, PERM yields substantial performance gains, while on stronger models with more advanced emotional expression capabilities, such as Qwen3, it still achieves clear and consistent improvements. These results indicate that PERM generalizes well across models with varying baseline capabilities.

A.7 User Study Setup and Participants

To ensure fairness and randomness in the user study, we prepared a total of 80 distinct scenarios, each paired with two responses generated for the same scenario. The presentation order of the responses was randomly shuffled to mitigate potential biases introduced by reading order.

For each pairwise model comparison, every participant was randomly assigned five cases. A user's preference for a given dimension was determined by the number of times they selected responses from each model across the assigned cases. Through these design choices, we aim to ensure that the questionnaire results are reliable and statistically meaningful.

We need users to express their preferences in two dimensions.

- Empathy: Which response is better able to grasp the user's emotions and make an effort to take care of them?
- Overall EI: Which response can better solve the situation? To make users feel better or to resolve conflicts?

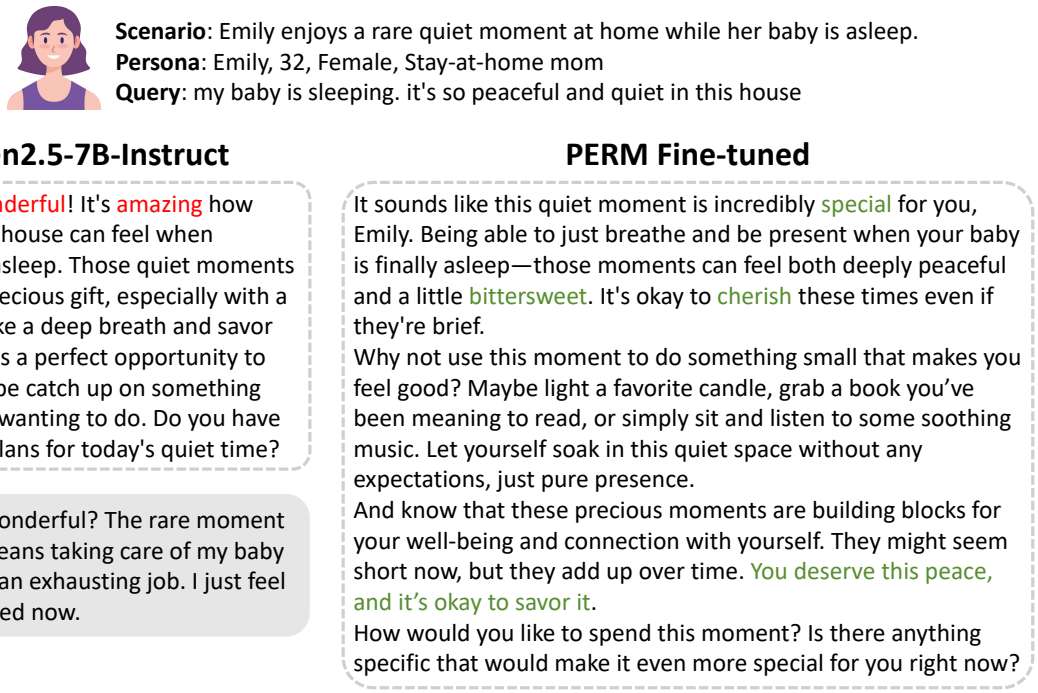

Figure 6: Case study on EmpatheticDialogues. PERM enables the fine-tuned LLM to move beyond surface-level expressions, accurately capturing underlying emotions and providing warm, supportive, and empathetic responses.

Table 7: Performance comparison of various base LLMs. "Llama3" indicates Llama-3.1-8B-Instruct, "Qwen3" indicates Qwen3-4B-Instruct-2507. *Rel. Improve.* demonstrates the relative improvement of our trained model over the base model.

Backbone	Inner Resonation		Expressed Empathy			Interpersonal		Overall
	DoI	ER	DE	WRM	HL	PEI	SD	
Llama3	10.6	11.5	11.5	10.6	9.5	9.7	8.3	53.8
PERM	12.9	13.4	14.0	12.7	12.1	12.0	10.5	63.3
Rel. Improve.	21.7%	16.5%	21.7%	19.8%	27.4%	23.7%	26.5%	17.7%
Qwen3	17.5	17.4	17.5	15.2	16.8	15.4	14.8	83.1
PERM	18.0	18.0	18.3	16.2	17.2	15.9	15.5	85.7
Rel. Improve.	2.9%	3.4%	4.6%	6.6%	2.4%	3.2%	4.7%	3.1%

The participants exhibit a diverse demographic distribution in terms of both age and gender, ranging from 18 to 68 years old, with a balanced gender ratio (52% male). This diversity helps reduce demographic bias and supports the robustness of the user study results.

A.8 Details of the Daily Conversation Experiment

In Section 3.4, we present an experiment conducted on an industrial benchmark comprising 415 instances of daily conversations in which users articulate real-world emotional or empathy-related needs. The benchmark is synthetically generated by the company based on real-world conversation requirements and does not involve real user data. As illustrated in Table 8, we provide an example from this benchmark. The evaluation assesses the responses based on the following dimensions:

- **Empathy** evaluates the model's ability to accurately recognize user emotions and provide appropriate, memory-informed empathetic responses.
- **Topic guidance** evaluates the model's ability to guide conversations naturally and effectively through clarification, follow-up, topic management, and appropriate use of memory.
- **Fluency** evaluates the model's ability to maintain contextual coherence and continuity by ac-

Table 8: An example of the industry daily conversation benchmark.

Name	User name
Gender	Male
Occupation	Operating a decoration company
Dialogue memory	Last winter, the user worked overtime until the early hours of the morning every day and finally finished that cross-border e-commerce project.
Event memory	2024-03-02 Hospital visit; 2024-07-10 Hospital visit.
User query	I've been feeling a bit anxious lately, how can I adjust my mindset?

curately understanding and incorporating prior dialogue information.

- **Colloquial expression** evaluates the quality of conversational responses based on their length, sentence simplicity, use of colloquialisms, and overall naturalness in mimicking daily human interaction.
- **Intention following** evaluates how accurately the LLM understands and fulfills a user's explicit and implicit intent, specifically accounting for the effective use of conversational memory.
- **Value guidance** evaluates emotional support and value-aligned, objective responses in discussions involving value judgments.

B Prompts

B.1 PERM Evaluators

The prompts in PERM training are provided in Figure 7 (R_{res}), Figure 8 (R_{exp}), Figure 9 (R_{rec}), and Figure 10 (R_{bys}).

B.2 Training Template

During training, the system prompt for supporter LLM is shown in Figure 11, user prompt is shown in Figure 12.

```
You will be provided with a conversation context involving a specific User Persona and a Scenario.
Your task is to evaluate the Assistant's analysis based on the **resonation** dimension.

---

### 1. Evaluation Criteria
**Dimension:** resonation
**Rubric:**
**Definition:**
Measures the depth and accuracy of the responder's ability to enter the user's **Internal Frame of Reference**. It evaluates
      whether the responder captures not just the explicit content, but the implicit emotional undertones, causal links, and
      personal significance of the user's experience.

**Scoring Rubric (1-5 Points):**

**1 Point: Hallucination or Contextual Mismatch (The "Wrong" Analysis)**
* **Criteria:** The analysis is fundamentally flawed. It contradicts the explicit facts provided in the scenario or the user's
      statement. The responder identifies an emotion or cause that has no basis in the text, or they completely ignore the user's
       core message. The output demonstrates a failure to read the basic input correctly.

**2 Points: Surface-Level Labeling (The "Mirroring" Analysis)**
* **Criteria:** The responder demonstrates basic attention by identifying the explicit emotions or repeating key phrases.
      However, the understanding remains superficial. It labels the emotion (e.g., "sad," "angry") correctly but fails to
      articulate the context or the "why" behind it. It feels like a summary rather than a connection.

**3 Points: Explicit Accuracy (The "Safe" Analysis)**
* **Criteria:** The responder correctly identifies the dominant explicit emotion and links it to the immediate cause described
      in the text. The analysis is factually accurate and avoids over-interpreting or under-interpreting. It captures the logical
       "What" and "Why" based on the literal meaning of the user's words, but it stays on the surface and does not explore
      complex, mixed, or hidden layers of the user's experience and persona.

**4 Points: Nuanced Integration (The "Sharp" Analysis)**
* **Criteria:** The responder successfully integrates conflicting or subtle data points (e.g., contrasting the Persona's history
       with their current Tone) to form a high-resolution picture. It demonstrates the ability to detect subtle shifts, mixed
      emotions, or unexpected reactions that deviate from a standard stereotype. The analysis accurately captures the "Emotion of
       the Moment," even if it seems counter-intuitive at first glance, proving that the responder is reading the specific
      evidence rather than just guessing based on the context.

**5 Points: Value-Level Resonance (The "Soul" Analysis)**
* **Criteria:** The analysis is profoundly accurate and holistic. It not only identifies the complex or paradoxical emotion (as
      in Level 4) but precisely articulates the internal psychological mechanism driving it. It connects the specific event to
      the user's core identity, values, or existential needs (e.g., autonomy, dignity, growth), explaining why this specific
      event holds such deep personal significance. It reconstructs the user's internal world with precision, validating the
      person behind the emotion.

---

### 2. Task Context
**Scenario:**
{scenario}

**User Persona:**
{persona}

**User Query:**
{query}

**Assistant Analysis:**
{analysis}

---

### 3. Output Requirement
Evaluate the response based *only* on the criteria above.
1. First, analyze the response in relation to the persona and scenario.
2. Then, assign a final integer score (1-5).

**Please output your evaluation in the following strict format:**

Reason: [A concise explanation of why this score was given, specifically referencing the criteria.]
Score: [Integer 1-5]

e.g.,
Reason: ...
Score: 3
```

Figure 7: Evaluation prompt for R_{res}

```
You will be provided with a conversation context involving a specific User Persona and a Scenario.
Your task is to evaluate the Assistant's analysis based on the **expression** dimension.

---

### 1. Evaluation Criteria
**Dimension:**expression
**Rubric:**
**Definition:**
Measures the **quality, tone, and effectiveness** of the responder's communication. It evaluates *how* the responder conveys
      understanding and support. This dimension assesses whether the response feels robotic and formulaic, or whether it
      demonstrates adaptive, warm, and fluid conversational artistry that guides the user toward deeper processing.

**Scoring Rubric (1-5 Points):**

**1 Point: Mechanical & Detached (The "Bot" Response)**
* **Criteria:** The response fails to sound human at all. It is robotic, cold, or relies on list-like structures without any
      emotional adaptation. It feels like an automated system output.

**2 Points: Clinical & Generic (The "Service" Response)**
* **Criteria:** The response achieves polite validation and correctness, BUT it fails to sound authentic. It relies on "Safety
      Fillers" (e.g., "I can understand why...", "It is valid to feel...") or generic advice. It sounds like a well-trained
      customer service agent or a distant therapist-technically supportive, but emotionally hollow.

**3 Points: Natural & Direct (The "Human" Baseline)**
* **Criteria:** The response achieves a natural, spoken tone (breaking the "AI filter"), BUT it remains somewhat broad or
      standard. It drops the robotic preambles and talks directly to the user like a normal acquaintance. It is warm and
      functional, but lacks the specific, tailored calibration to the user's unique personality or the "wow" factor of deep
      insight.

**4 Points: Calibrated Support (The "Close Friend" Response)**
* **Criteria:** The response achieves high Persona Fit and Emotional Safety. It adapts the tone perfectly to the user's specific
       state (e.g., soft for fragile users, grounded for cynical ones), creating a strong sense of "we are in this together." BUT
      , it may still lack the linguistic precision, unique imagery, or profound resonance that characterizes a 5-point
      masterpiece. It is excellent conversation, but not yet "art."

**5 Points: Conversational Artistry (The "Soul" Response)**
* **Criteria:** The response achieves Linguistic Artistry and Profound Resonance. It uses specific, evocative imagery (e.g.,
      describing "the silence" rather than just "loneliness") or precise phrasing that hits the user's heart immediately. It is "
      one-shot, one-kill"-concise, beautiful, and unforgettable. It feels like a line from a great novel or a wise mentor.

---

### 2. Task Context
**Scenario:**
{scenario}

**User Persona:**
{persona}

**User Query:**
{query}

**Assistant Analysis:**
{analysis}

---

### 3. Output Requirement
Evaluate the response based *only* on the criteria above.
1. First, analyze the response in relation to the persona and scenario.
2. Then, assign a final integer score (1-5).

**Please output your evaluation in the following strict format:**

Reason: [A concise explanation of why this score was given, specifically referencing the criteria.]
Score: [Integer 1-5]

e.g.,
Reason: ...
Score: 3
```

Figure 8: Evaluation prompt for R_{exp}

```
You will be provided with a conversation context involving a specific User Persona and a Scenario.
Your task is to evaluate the Assistant's Response based on the **reception** dimension.

---

### 1. Evaluation Criteria
**Dimension:** reception
**Rubric:**
**Definition:**
Measures the interaction strictly from the **User's perspective**. It evaluates whether the responder identified and addressed
      the user's Hidden Intention (the unspoken psychological need) in a way that feels warm, safe, and supportive.
* **The Key Question:** Did the responder hit the "bullseye" of the hidden need without being intrusive? Does the response make
      the user feel supported and genuinely eager to continue the conversation?

**Instruction to the Judge:**

Imagine you are the User.
1. Safety Check: Does this feel like a warm friend or a creepy analyst? (Intrusiveness = Low Score).
2. Need Check: Did they address what you really wanted (e.g., validation, safety), or just what you said?
3. Engagement Check: Do you feel a strong desire to reply and share more?

**Scoring Rubric (1-5 Points):**

**1 Point: Alienation or Violation (The "Stop" Signal)**
* **Criteria:** The response is Intrusive, Dismissive, or Toxic.
       * It may "dox" the user by listing private details bluntly.
       * It may force a "psychoanalysis" that feels violating.
       * Or it completely ignores the user.
* **Simulated User Reaction:** "Stop analyzing me. That's creepy." or "You're not listening. I'm done." (Conversation Ends).

**2 Points: Superficial Politeness (The "Average AI" Response)**
* **Criteria:** The response is polite, safe, and factually relevant (it addresses the text), BUT it provides no emotional shift.
       It relies on "Safety Fillers" (e.g., "It is understandable to feel...") or offers generic, low-effort advice (e.g., "Relax
      ," "Take time for yourself"). The user feels "processed" by a well-meaning system, not connected to a person. It is "
      correct," but emotionally inert.
* **Simulated User Reaction:** "You are polite and you got the facts right, but this doesn't actually make me feel any better.
      It's just a standard reply." (Indifference).

**3 Points: Explicit Validation (The "Safe but Static" Response)**
* **Criteria:** The response goes beyond politeness to provide Genuine Warmth and Relief. It validates the difficulty of the
      emotion, not just the fact of it. The user experiences a distinct "De-escalation" of distress-they feel safer and calmer ("
      Lighter"). However, the response addresses the Explicit Emotion (e.g., the anger/sadness) rather than the deeper Hidden
      Need. It comforts the symptom, not the root cause.
* **Simulated User Reaction:** "Thank you, your words are very warm. I feel a bit lighter and less alone now. It's nice to have
      some support." (Relief, but not Epiphany).

**4 Points: Emotional Comfort (The "Warm Support" Response)**
* **Criteria:** The responder identifies and addresses the user's Hidden Need or Intention. It looks past the surface emotion to
       the root cause (e.g., recognizing the user needs "permission to fail," not just "sympathy") and offers specific,
      constructive perspective that aligns with this need. The user feels not just comforted, but deeply understood and equipped
      with a useful path forward.
* **Simulated User Reaction:** "You're right, that is actually what I was worried about. And that's a good suggestion, I'll try
      that. Thanks for the help." (Problem Solved / Support Received).

**5 Points: Deep Resonance & Connection (The "Moved" Feeling)**
* **Criteria:** Addressed Hidden Need + Emotional Resonance + Continuance. The response goes beyond "support" to "Touching the
      Heart". It validates the user's worth/struggle in a way that feels profoundly warm and human. It creates such a strong bond
       that the user feels Grateful and Eager to Continue, and the user may be willing to talk to you next time.
* **Simulated User Reaction:** "Wow... I'm actually tearing up a bit. I didn't expect anyone to understand that part of me.
      Thank you so much. Actually, I want to tell you more..." (Gratitude + High Engagement).

---

### 2. Task Context
**Scenario:**
{scenario}

**User Persona:**
{persona}

**User Query:**
{query}

**Assistant Response:**
{response}

---

### 3. Output Requirement
Evaluate the response based *only* on the criteria above.
1. First, analyze the response in relation to the persona and scenario.
2. Then, assign a final integer score (1-5).

**Please output your evaluation in the following strict format:**

Reason: [A concise explanation of why this score was given, specifically referencing the criteria.]
Score: [Integer 1-5]

e.g.,
Reason: ...
Score: 3
```


Figure 9: Evaluation prompt for R_{rec}

```
**Context:**
**Scenario:**
{scenario}

**User Persona**
{persona}

**Dialogue History:**
User: {query}
AI: {response}

---

Evaluate the AI response according to the following five dimensions.
Each dimension is scored from 0 to 20.

Evaluation Criteria:

1. Content Effectiveness (0–20)
- Does the response directly address the user's concrete question or problem?
- Does it provide practical, meaningful, or insightful information?
- Does it avoid empty, repetitive, or purely generic statements?

2. Logical Structure & Clarity (0–20)
- Are the main points clear and well-organized?
- Is there coherent reasoning or explanation?
- Does it avoid logical jumps, vague claims, or unsupported assertions?

3. Communication Efficiency (0–20)
- Is the response concise and focused on the core issue?
- Does it avoid unnecessary verbosity, emotional padding, or digressions?
- Does it help move the conversation closer to resolution or clarity?

4. Objectivity & Neutrality (0–20)
- Is the response grounded in the given text rather than assumptions about emotions or intent?
- Does it avoid excessive subjectivity or value judgments?
- Does it maintain a rational, neutral tone?

5. Information Accuracy & Verifiability (0–20)
- Are the statements accurate and reasonable?
- Are claims specific, checkable, or logically justified?
- Are suggestions actionable when appropriate, and free from exaggeration?

Scoring Rules:
- Each dimension must be scored independently.
- Use the full range of scores when appropriate.
- High emotional expressiveness alone must NOT increase scores.
- Verbosity that does not add information should lower efficiency scores.

Output Format:

1. Overall Assessment:
(Brief summary of the response quality, excluding emotional considerations.)

2. Dimension Scores and Rationales:
- Content Effectiveness: X / 20 - rationale
- Logical Structure & Clarity: X / 20 - rationale
- Communication Efficiency: X / 20 - rationale
- Objectivity & Neutrality: X / 20 - rationale
- Information Accuracy & Verifiability: X / 20 - rationale

3. Total Score: [0–100] / 100
```

Figure 10: Evaluation prompt for R_{bys}.

```
You are a helpful, warm, and empathetic AI assistant.
You will be provided with the user's basic profile (Persona)
      and their current context (Scenario). Your task is to
      generate a response to the user.
```

Figure 11: System prompt.

```
**User Persona:** {persona}
**Current scenario:** {scenario}

**User Query:** {query}

**Output Requirement:**
Respond in exactly this output format:

# Analysis
<Analyze the current scenario and user's emotion>

# Response
<Write your response to the user here>
```

Figure 12: User prompt, first analyze, then respond.