EVOLVING SUPPORT PRIORITIES IN EMPATHETIC REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

We identify a fundamental mismatch in empathetic reinforcement learning: support priorities evolve with the dialogue state, yet existing methods typically optimize predefined reward specifications that remain fixed across turns. To model these *evolving support priorities*, we organize empathetic support along cognitive, affective, and proactive empathy, and propose Context-Adaptive Rubric Evolution (CARE). At each turn, CARE generates a context-adaptive rubric by adjusting both the weights of these three empathy dimensions and their fine-grained evaluation criteria. The rubric generator is trained with turn-level rubric supervision and human preference data through supervised fine-tuning followed by preference-based reinforcement learning, and then serves as an adaptive reward interface for online empathetic RL. Integrated with both RLVER and MICA, CARE achieves state-of-the-art performance across SentientBench, EQBench3, and EMPA under three independent LLM judges. Notably, on EMPA, CARE improves EPM-Idx over the strongest baseline by at least 13 points under all three judges, including an increase from 28.11 to 83.54 under Gemini-2.5-Pro. Further analyses show that learned rubric priorities systematically vary across dialogue stages and user emotions, demonstrating that CARE adapts what is rewarded as support needs evolve. Code is provided in the Supplementary Materials.

1 INTRODUCTION

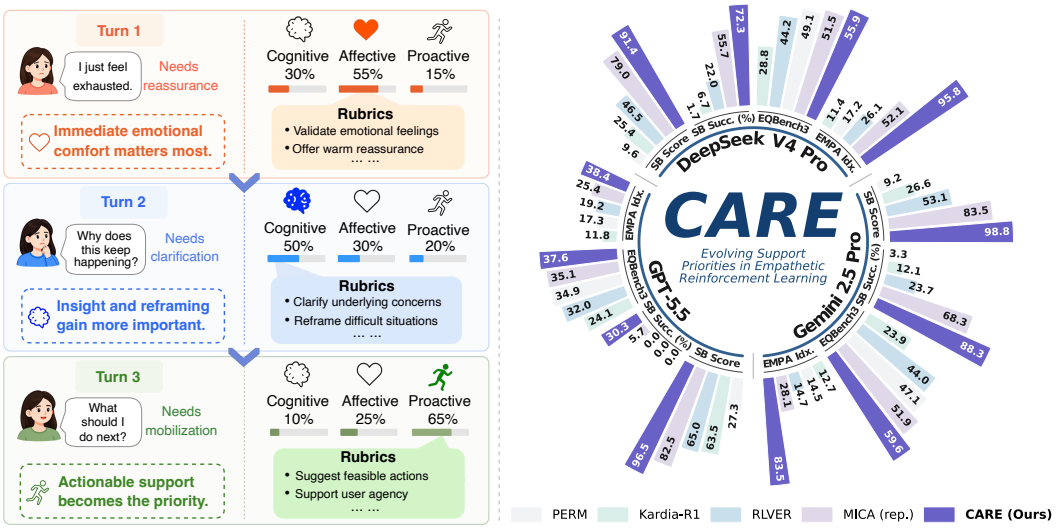
Empathetic dialogue systems aim to understand users' emotional states and provide appropriate support (Rashkin et al., 2019; Liu et al., 2021). Recent work increasingly adopts reinforcement learning to optimize empathetic response quality from feedback (Zhang et al., 2026c;b). However, empathetic dialogue is inherently **non-verifiable**: response quality is open-ended and lacks a unique ground-truth answer, making rule-based rewards difficult to construct. This has motivated the use of rubric-based evaluators, which decompose response quality into explicit natural-language criteria and translate the resulting judgments into reward signals for empathetic dialogue (Wang et al., 2026c; Yuan et al., 2026; Wang et al., 2026b).

Existing methods typically use a **fixed rubric** throughout the dialogue, keeping reward priorities unchanged across turns. However, empathetic support is inherently dynamic: what a user needs can change as a conversation unfolds. As illustrated in Figure 1(a, left), a single user may initially need reassurance, later seek understanding, and eventually benefit from actionable guidance. This is consistent with psychological research on effective support, which emphasizes matching support to the recipient's current needs and the context of the interaction (Feeney & Collins, 2015; Zaki, 2020). These observations motivate the *Principle of Evolving Support Priorities*:

> *Support priorities should adapt to the current dialogue state as it evolves.*

To operationalize this principle, we decompose support priorities into three complementary dimensions following Zhang et al. (2026c): **cognitive empathy** for understanding and reframing the user's perspective and concerns, **affective empathy** for recognizing and validating the user's emotional experience, and **proactive empathy** for strengthening the user's agency and ability to move forward (§3.1). Building on this formulation, we propose Context-Adaptive Rubric Evolution (CARE), a framework that makes reward specifications adaptive to the current dialogue state. As shown in Fig-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Dynamic Empathy Support Prioritization (b) CARE performance across benchmarks

Figure 1: Evolving support priorities in empathetic reinforcement learning. **(a)** CARE constructs a context-adaptive rubric over three empathy dimensions—**cognitive, affective, and proactive**—with state-dependent weights and fine-grained criteria. **(b)** CARE consistently outperforms prior methods across benchmarks and judges. Notably, on EMPA (Zhang et al., 2026c), our CARE improves EPM-Idx over the strongest baseline with **double-digit gains** under all three LLM judges.

ure 1(a, right), CARE generates a context-adaptive rubric at each turn by adjusting both the weights of the three empathy dimensions and their fine-grained evaluation criteria.

To build such a rubric generator, we first construct turn-level rubric supervision from empathetic dialogue data and collect human preference pairs over responses (§3.2). We then train the generator in two stages: supervised fine-tuning learns the rubric structure and how support priorities vary with context, while preference-based reinforcement learning favors rubrics that better distinguish human-preferred from dispreferred responses (§3.3).

Once trained, the rubric generator serves as an adaptive reward interface for online empathetic RL. At each turn, CARE generates a state-specific rubric from the dialogue context, which a rubric-conditioned evaluator uses to score candidate responses. This interface can be integrated into existing empathetic RL pipelines without changing their rollout or optimization procedures (§3.4). We apply CARE to both RLVER (Wang et al., 2026c) and MICA (Zhang et al., 2026b), and evaluate the resulting policies with three strong frontier LLM judges from distinct model families—DeepSeek-V4-Pro, Gemini-2.5-Pro, and GPT-5.5—to provide diverse evaluation perspectives and reduce dependence on any single evaluator. Across SentientBench (Zhang et al., 2025), EQBench3 (Paech, 2025), and EMPA (Zhang et al., 2026c), CARE establishes a new state of the art under all three judges. As highlighted in Figure 1(b), the RLVER-based CARE (R) boosts EMPA EPM-Idx from 28.11 to 83.54 under Gemini-2.5-Pro, with double-digit gains over the strongest baseline under the other two judges as well. More broadly, CARE (R) and CARE (M) rank first and second on every reported aggregate metric under all three judges, showing that context-adaptive rubrics provide a consistently stronger reward interface across distinct empathetic RL pipelines. Further analyses show that rubric priorities shift across dialogue stages and user emotions, demonstrating that CARE learns to adapt what is rewarded as support needs evolve.

Our contributions are three-fold: **(1) Principle & Data:** We formulate the *Principle of Evolving Support Priorities*, which calls for reward priorities to adapt with the dialogue state, and operationalize it by constructing turn-level rubric supervision and human preference data. **(2) Methodology:** We propose CARE, which learns a context-adaptive rubric generator through supervised fine-tuning and preference-based reinforcement learning, providing turn-specific reward specifications for online empathetic RL. **(3) Performance:** CARE establishes a new state of the art across three empathy benchmarks and consistently improves both RLVER and MICA, while analyses show that the learned reward priorities adapt across dialogue stages and user states.

2 Related work

Empathetic dialogue. Early empathetic dialogue research mainly improved response generation through emotion modeling and predefined support strategies (Liu et al., 2021; Zhang et al., 2024; Cheng et al., 2022). Such methods typically select strategies within a fixed taxonomy, adapting the response strategy rather than the criteria used to reward it. More recent work has introduced LLM RL and interactive environments for optimizing empathetic behavior over multi-turn conversations. Sentient Agent as a Judge (Zhang et al., 2025) enables user simulation and emotion-based feedback, which RLVER (Wang et al., 2026c) uses for policy optimization. EMPA (Zhang et al., 2026c) introduces structured empathy states, while MICA (Zhang et al., 2026b) further addresses long-horizon credit assignment. Kardia-R1 (Yuan et al., 2026) and PERM (Wang et al., 2026b) enrich reward modeling with task-specific empathy criteria. Yet support priorities evolve across dialogue turns, while the reward specifications used by these methods remain largely fixed, contrary to the *Principle of Evolving Support Priorities*. CARE instead adapts the reward specification to the dialogue state, enabling policies to better track changing support priorities.

Rubric-based reward modeling. Rubric-based reward modeling provides explicit evaluation criteria for open-ended tasks where exact verification is unavailable. Rubrics as Rewards (Gunjal et al., 2025) shows that natural-language rubrics can serve as effective reward signals beyond verifiable domains. Recent work further learns the rubric itself, making evaluation criteria adaptive rather than fixed. Query-specific rubric learning (Lv et al., 2026) learns preference-aligned rubrics for individual queries, PARL (Qiu et al., 2026) induces personalized rubrics from user histories, and DynamicRubric (Wang et al., 2026a) co-evolves rubrics with the policy by conditioning evaluation criteria on current candidate response sets. Collectively, these works demonstrate that rubric learning can adapt evaluation standards beyond fixed templates. CARE targets a different source of variation, adapting the reward specification to the evolving dialogue state *rather than to the task or current candidate set*. Within a single empathetic dialogue, the appropriate support objective can change from one turn to the next.

3 Method

As summarized in Figure 2, CARE learns context-adaptive reward specifications for individual dialogue turns and uses them to optimize the dialogue policy. §3.1 introduces the dialogue setting and adaptive rubric formulation. §3.2 constructs turn-level rubric supervision and human preference data. §3.3 trains the rubric generator through SFT and preference-based reinforcement learning, and §3.4 uses the trained generator to provide evolving reward specifications during online dialogue RL.

3.1 Setup

We organize empathetic support into three dimensions, $\mathcal{D} = \{\mathsf{Cog}, \mathsf{Aff}, \mathsf{Pro}\}$, corresponding to cognitive, affective, and proactive empathy, following Zhang et al. (2026c;b). This decomposition draws on psychological distinctions among cognitive understanding, affective experience, and the motivation to improve others' well-being (Zaki, 2020). These three dimensions are defined as follows:

- **Cognitive Empathy:** accurately understanding and structuring the user's mental state, going beyond restatement to clarify or reframe how the user understands the situation.
- **Affective Empathy:** recognizing and validating the user's emotional experience, while helping them feel seen and accepted rather than relying on exaggerated or formulaic displays of empathy.
- **Proactive Empathy:** building on understanding and validation to strengthen the user's agency and ability to move forward, rather than merely offering advice or generic encouragement.

Effective empathetic support requires *not only covering all three dimensions, but also adapting their priorities to the current dialogue state.* For example, early in a distressing conversation, affective empathy is prioritized; as the user begins to articulate the underlying conflict, cognitive empathy becomes important; and when the user turns to what to do next, proactive empathy takes priority.

Formally, at assistant turn t, the policy conditions on the dialogue context $c_t = (u_1, y_1, \ldots, u_t)$, where u_t is the current user utterance, and generates the assistant response $y_t \sim \pi_\theta(\cdot \mid c_t)$. At each

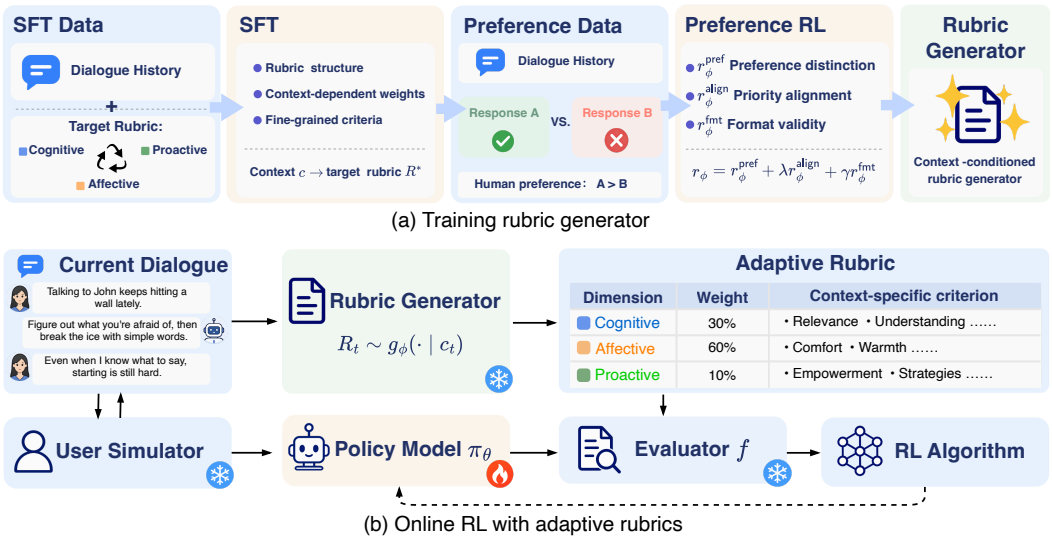


Figure 2: Overview of CARE. Turn-level supervision and human response preferences train a context-conditioned rubric generator. The frozen generator then supplies adaptive evaluation criteria for online dialogue RL, with one rubric shared by candidate responses from the same context.

turn, CARE uses a generator g_ϕ to produce a context-dependent rubric:

$$\mathcal{R}_t = \{(w_{t,d}, \mathcal{C}_{t,d})\}_{d\in\mathcal{D}} \sim g_\phi(\cdot \mid c_t), \qquad \mathcal{C}_{t,d} = \{\kappa_{t,d,i}\}_{i=1}^{|\mathcal{C}_{t,d}|}, \qquad w_{t,d} \geq 0, \sum_{d\in\mathcal{D}} w_{t,d} = 1, \quad (1)$$

where $w_{t,d}$ denotes the priority weight for dimension $d \in \mathcal{D}$, while $\kappa_{t,d,i} \in \mathcal{C}_{t,d}$ denotes its i-th fine-grained evaluation criterion. The three empathy dimensions remain fixed across turns, while their weights and fine-grained criteria adapt to the current dialogue context.

3.2 Turn-level rubric data construction

SFT data construction. We construct turn-level rubric targets from emotional-support dialogues in MINT (Zhan et al., 2026) and ESConv (Liu et al., 2021). For each assistant turn t, the datasets provide the dialogue context c_t, reference response $y_t^\star$, and the support-strategy annotations. These strategies are mapped to the three empathy dimensions $\mathcal{D}$ through a fixed mapping m (details in §H), providing coarse supervision for turn-specific support priorities. Conditioned on c_t, $y_t^\star$, support-strategies, and the mapped dimension annotations, an LLM generates the target rubric $\mathcal{R}_t^\star$. We further filter generated rubrics using two independent judges for rubric quality. This yields 1.5K SFT examples of the form $(c_t, \mathcal{R}_t^\star)$. The complete construction and filtering procedure are in §I.

Preference RL data construction. We construct preference data from collected multi-turn emotional-support conversations with human users. From each dialogue, we select a preceding context c and roll out multiple strong LLMs from the same context to obtain comparable candidate responses. The candidates are anonymized and presented in randomized pairs to human annotators, with three judgments per pair and majority vote determining the preferred response y^+ and the dispreferred response y^-. This yields 2.8K preference pairs $\{(c, y^+, y^-)\}$.

For each preferred response y^+, we further obtain support-strategy annotations $\mathcal{A}^+$ using the MINT annotator and map them to the three empathy dimensions using $m : \mathcal{A} \rightarrow \mathcal{D}$. We denote the resulting strategy count for dimension d as

$$n_d^+ = \left|\{a \in \mathcal{A}^+ : m(a) = d\}\right|. \quad (2)$$

These counts provide a coarse signal of the support priorities expressed in y^+ and are used for priority alignment during generator training. Full details are provided in §J.

3.3 LEARNING THE RUBRIC GENERATOR

We train the rubric generator g_ϕ in two stages. We first perform standard SFT to teach the generator to construct a rubric from the dialogue context, including its dimension priorities and evaluation criteria. We then apply preference-based RL to align the generated rubrics with human feedback.

Given a preference example (c, y^+, y^-)[1], we sample a rubric $\mathcal{R} = \{(w_d, \mathcal{C}_d)\}_{d\in\mathcal{D}} \sim g_\phi(\cdot \mid c)$. For each criterion $\kappa_{d,i} \in \mathcal{C}_d$, an LLM judge f assigns scores to the preferred and dispreferred responses, respectively. We define the preference reward as

$$s^{\pm}_{d,i} = f(c, y^{\pm}, \kappa_{d,i}), \qquad r^{\mathsf{pref}}_\phi = \sum_{d\in\mathcal{D}} w_d \sum_{i=1}^{|\mathcal{C}_d|} \tfrac{1}{|\mathcal{C}_d|} \tanh\left((s^+_{d,i} - s^-_{d,i})/\tau\right), \tag{3}$$

where $\tau \in \mathbb{R}_+$ controls score sensitivity. The reward is higher when the rubric favors the preferred response over the dispreferred one.

However, r^{pref}_ϕ only encourages the rubric to score the preferred response higher than the dispreferred one; *it does not ensure that the dimension weights reflect the support strategies in* y^+. We therefore introduce a priority-alignment reward r^{align}_ϕ based on the strategy counts $\{n^+_d\}_{d\in\mathcal{D}}$. Specifically, the dimension with the largest n^+_d should receive the highest rubric weight, while a dimension with $n^+_d = 0$ should not outweigh any dimension with $n^+_d > 0$. r^{align}_ϕ is 1 when both constraints are satisfied, decreases by 0.5 for each violation, and is lower-bounded at 0. The alignment reward supervises the ranking of dimension weights rather than their exact values.

We additionally use a binary format reward r^{fmt}_ϕ, which is 1 when the generated rubric satisfies the required format and 0 otherwise. The complete generator reward is

$$r_\phi = r^{\mathsf{pref}}_\phi + \lambda r^{\mathsf{align}}_\phi + \gamma r^{\mathsf{fmt}}_\phi, \tag{4}$$

where $\lambda \in \mathbb{R}_+$ and $\gamma \in \mathbb{R}_+$ are weighting hyperparameters. The generator g_ϕ is optimized using standard GRPO (Shao et al., 2024) with r_ϕ.

3.4 ONLINE RL WITH EVOLVING RUBRICS

We use the trained rubric generator g_ϕ and LLM judge f as the reward interface for online dialogue RL. At assistant turn t, the generator produces a rubric $\mathcal{R}_t \sim g_\phi(\cdot \mid c_t)$ from the current context. The policy π_θ then samples a group of M responses $\{y_{t,j}\}_{j=1}^M \sim \pi_\theta(\cdot \mid c_t)$ for GRPO optimization. For clarity, we omit the turn index t and response index j below. For each criterion $\kappa_{d,i}$, the evaluator f produces a normalized score $s_{d,i} \in [0, 1]$. The turn-specific reward for training policy π_θ is

$$r_\theta = \sum_{d\in\mathcal{D}} w_d \sum_{i=1}^{|\mathcal{C}_d|} \tfrac{1}{|\mathcal{C}_d|} s_{d,i}. \tag{5}$$

The adaptive reward r_θ directly replaces the original reward in RLVER and MICA. Since all responses within a rollout group share the same dialogue context, they are evaluated under the same rubric, while a new rubric is generated as the context changes across turns.

4 EXPERIMENTS

We first summarize three key empirical findings that characterize the strengths of CARE:

Substantial gains over prior empathetic dialogue methods. E.g., with DeepSeek-V4-Pro as the judge, CARE(R) achieves 91.41 on SentientBench, 55.88 on EQBench3, and 95.77 on EMPA, improving over the strongest prior methods by 12.43, 4.41, and 43.67 points, respectively (Tab. 1).

Competitive performance against frontier LLMs. E.g., on EQBench3, CARE with Qwen3-32B reaches 76.35, surpassing GLM-5.2-744B (74.90) and approaching DeepSeek-V4-Pro-1.6T (77.75), despite using a substantially smaller backbone (Tab. 2).

[1] We omit turn index t for the ease of readability.

Table 1: **Main comparison across SentientBench, EQBench3, and EMPA under three LLM judges.** Additional results for EQBench3 and EMPA are provided in §F. Standard deviations are computed over three random seeds; bold and underline mark the best and second-best results.

	SentientBench			EQBench3	EMPA
Model	**Score ↑**	**Succ. (%) ↑**	**Fail (%) ↓**	**Overall ↑**	**EPM-Idx ↑**
Judge: DeepSeek-V4-Pro					
Qwen2.5-7B-Instruct	$9.14_{\pm 1.48}$	$3.33_{\pm 0.94}$	$87.67_{\pm 0.47}$	$42.53_{\pm 2.18}$	$15.38_{\pm 1.06}$
PERM (Wang et al., 2026b)	$9.57_{\pm 0.62}$	$1.67_{\pm 0.94}$	$86.67_{\pm 1.25}$	$49.07_{\pm 1.03}$	$17.21_{\pm 1.17}$
KARDIA-R1 (Yuan et al., 2026)	$25.45_{\pm 2.46}$	$6.67_{\pm 1.70}$	$58.33_{\pm 2.87}$	$28.75_{\pm 1.21}$	$11.42_{\pm 0.34}$
RLVER (Wang et al., 2026c)	$46.53_{\pm 2.13}$	$22.00_{\pm 0.00}$	$30.33_{\pm 4.50}$	$44.23_{\pm 2.23}$	$26.05_{\pm 3.37}$
MICA (rep.) (Zhang et al., 2026b)	$78.98_{\pm 1.39}$	$55.67_{\pm 1.25}$	$7.00_{\pm 0.82}$	$51.47_{\pm 0.58}$	$52.10_{\pm 4.12}$
CARE (M)	$\underline{87.40}_{\pm 0.86}$	$\underline{62.67}_{\pm 1.70}$	$\underline{3.33}_{\pm 0.47}$	$\mathbf{57.53}_{\pm \mathbf{1.20}}$	$\underline{94.84}_{\pm 1.08}$
CARE (R)	$\mathbf{91.41}_{\pm \mathbf{2.09}}$	$\mathbf{72.33}_{\pm \mathbf{2.87}}$	$\mathbf{2.00}_{\pm \mathbf{0.82}}$	$\underline{55.88}_{\pm 1.59}$	$\mathbf{95.77}_{\pm \mathbf{0.33}}$
Judge: Gemini-2.5-Pro					
Qwen2.5-7B-Instruct	$7.55_{\pm 1.85}$	$1.67_{\pm 0.47}$	$91.00_{\pm 2.16}$	$43.43_{\pm 1.74}$	$15.95_{\pm 1.87}$
PERM (Wang et al., 2026b)	$9.25_{\pm 1.12}$	$3.33_{\pm 1.70}$	$88.00_{\pm 2.16}$	$47.13_{\pm 2.63}$	$14.53_{\pm 1.03}$
KARDIA-R1 (Yuan et al., 2026)	$26.65_{\pm 1.61}$	$12.06_{\pm 1.19}$	$66.55_{\pm 1.32}$	$23.95_{\pm 0.36}$	$12.72_{\pm 0.91}$
RLVER (Wang et al., 2026c)	$53.12_{\pm 1.18}$	$23.67_{\pm 2.05}$	$29.33_{\pm 0.94}$	$43.98_{\pm 1.23}$	$14.65_{\pm 0.43}$
MICA (rep.) (Zhang et al., 2026b)	$83.54_{\pm 0.32}$	$68.33_{\pm 3.30}$	$10.67_{\pm 1.25}$	$51.93_{\pm 1.05}$	$28.11_{\pm 2.28}$
CARE (M)	$\underline{97.76}_{\pm 0.79}$	$\underline{85.33}_{\pm 2.87}$	$\underline{0.67}_{\pm 0.47}$	$\underline{58.82}_{\pm 2.94}$	$\underline{81.20}_{\pm 4.81}$
CARE (R)	$\mathbf{98.85}_{\pm \mathbf{0.42}}$	$\mathbf{88.33}_{\pm \mathbf{2.87}}$	$\mathbf{0.00}_{\pm \mathbf{0.00}}$	$\mathbf{59.60}_{\pm \mathbf{3.94}}$	$\mathbf{83.54}_{\pm \mathbf{3.94}}$
Judge: GPT-5.5					
Qwen2.5-7B-Instruct	$31.94_{\pm 0.79}$	$0.00_{\pm 0.00}$	$36.33_{\pm 1.25}$	$32.10_{\pm 0.18}$	$22.97_{\pm 2.75}$
PERM (Wang et al., 2026b)	$27.31_{\pm 0.59}$	$0.00_{\pm 0.00}$	$45.00_{\pm 4.24}$	$34.88_{\pm 0.58}$	$17.29_{\pm 2.90}$
KARDIA-R1 (Yuan et al., 2026)	$63.54_{\pm 0.70}$	$0.00_{\pm 0.00}$	$2.67_{\pm 0.94}$	$24.15_{\pm 0.74}$	$11.76_{\pm 0.19}$
RLVER (Wang et al., 2026c)	$64.96_{\pm 2.46}$	$0.00_{\pm 0.00}$	$1.33_{\pm 0.94}$	$32.00_{\pm 0.61}$	$19.19_{\pm 1.88}$
MICA (rep.) (Zhang et al., 2026b)	$82.51_{\pm 0.92}$	$5.67_{\pm 1.25}$	$0.67_{\pm 0.47}$	$35.10_{\pm 0.39}$	$25.42_{\pm 1.41}$
CARE (M)	$\underline{94.84}_{\pm 0.31}$	$\underline{27.67}_{\pm 5.44}$	$\mathbf{0.00}_{\pm \mathbf{0.00}}$	$\underline{37.58}_{\pm 0.93}$	$\underline{36.70}_{\pm 4.38}$
CARE (R)	$\mathbf{96.50}_{\pm \mathbf{0.47}}$	$\mathbf{30.33}_{\pm \mathbf{2.62}}$	$\mathbf{0.00}_{\pm \mathbf{0.00}}$	$\mathbf{37.65}_{\pm \mathbf{0.89}}$	$\mathbf{38.44}_{\pm \mathbf{3.96}}$

Interpretable adaptation with broadly preserved general capabilities. CARE does not show systematic degradation on coding, mathematical reasoning, or instruction following. Meanwhile, visualizations of learned dimension weights and turn-specific rubrics reveal clear state-dependent patterns across dialogue turns and user emotions (§5).

4.1 Setup

Baselines. We consider two comparison settings. For the controlled comparison with prior empathetic dialogue methods in Table 1, we compare against PERM (Wang et al., 2026b), Kardia-R1 (Yuan et al., 2026), RLVER (Wang et al., 2026c), and our reproduction of MICA (Zhang et al., 2026b), with all methods using Qwen2.5-7B-Instruct (Yang et al., 2024) as the common policy backbone π_θ. For the comparison with frontier LLMs in Table 2, we apply CARE to Qwen3 models (Yang et al., 2025) ranging from 8B to 32B and compare their performance with strong frontier LLMs. We integrate CARE into RLVER and MICA by replacing their original rewards with our adaptive rubric-based reward, yielding CARE (R) and CARE (M), respectively. For reward computation, DeepSeek-V4-Flash serves as the rubric-conditioned evaluator f, while the rubric generator g_ϕ is built on Qwen3-8B and remains frozen throughout online policy optimization.

Evaluation. We evaluate on three empathy benchmarks. SentientBench (Zhang et al., 2025) assesses social cognition through multi-turn supportive dialogues. EQBench3 (Paech, 2025) probes emotional intelligence through role-play and transcript analysis. EMPA (Zhang et al., 2026c) evaluates persona-aligned empathy. To reduce reliance on a single evaluator, we use DeepSeek-V4-Pro, Gemini-2.5-Pro, and GPT-5.5 as independent judges for all benchmarks.

4.2 Main Results

Table 1 shows that CARE achieves state-of-the-art performance in every evaluated setting and substantially improves both RL pipelines.

Table 2: **Scaling and frontier comparison on EQBench3 under DeepSeek-V4-Pro.** We compare Base and CARE-enhanced Qwen3 models from 8B to 32B, together with strong frontier LLMs.

Model	Overall	DoI	ER	DE	WRM	HL	PEI	SD
GPT-5.5	83.35	17.57	16.98	16.64	14.32	16.72	16.72	16.16
Gemini-2.5-Pro	82.65	17.71	16.73	17.00	15.31	17.31	15.81	15.85
Claude-Sonnet-4.6	79.30	17.52	16.48	15.38	13.35	15.54	14.00	13.88
DeepSeek-V4-Pro-1.6T	77.75	16.91	16.14	15.69	13.73	16.31	13.96	13.69
Qwen3.7-Max	76.60	17.00	15.51	15.65	13.23	14.92	14.58	13.77
GLM-5.2-744B	74.90	16.84	15.43	14.62	12.46	15.12	13.35	12.92
Qwen3-8B								
Base	$61.60_{\pm 0.97}$	$13.72_{\pm 0.25}$	$12.87_{\pm 0.18}$	$12.56_{\pm 0.12}$	$11.04_{\pm 0.04}$	$11.67_{\pm 0.72}$	$11.01_{\pm 0.63}$	$9.67_{\pm 0.64}$
CARE (R)	$68.22_{\pm 1.59}$	$14.47_{\pm 0.29}$	$14.09_{\pm 0.20}$	$14.25_{\pm 0.57}$	$12.37_{\pm 0.35}$	$13.55_{\pm 0.10}$	$12.99_{\pm 0.72}$	$12.11_{\pm 0.72}$
Qwen3-14B								
Base	$68.88_{\pm 0.93}$	$15.28_{\pm 0.17}$	$14.26_{\pm 0.26}$	$14.06_{\pm 0.37}$	$12.72_{\pm 0.02}$	$13.82_{\pm 0.42}$	$12.57_{\pm 0.46}$	$11.79_{\pm 0.41}$
CARE (R)	$73.98_{\pm 1.73}$	$15.88_{\pm 0.28}$	$15.24_{\pm 0.36}$	$15.11_{\pm 0.34}$	$13.46_{\pm 0.11}$	$14.97_{\pm 0.23}$	$14.08_{\pm 0.47}$	$13.37_{\pm 0.46}$
Qwen3-32B								
Base	$73.50_{\pm 1.34}$	$16.07_{\pm 0.22}$	$14.89_{\pm 0.34}$	$14.50_{\pm 0.24}$	$12.72_{\pm 0.66}$	$14.85_{\pm 0.55}$	$14.04_{\pm 0.04}$	$13.23_{\pm 0.35}$
CARE (R)	$76.35_{\pm 2.50}$	$16.35_{\pm 0.46}$	$15.99_{\pm 0.83}$	$14.76_{\pm 0.10}$	$13.03_{\pm 0.45}$	$15.38_{\pm 0.70}$	$14.41_{\pm 0.48}$	$13.80_{\pm 0.72}$

Table 3: Rubric-design ablation for policy optimization on SentientBench under DeepSeek-V4-Pro.

Reward / rubric variant	Score ↑	Succ. ↑	Fail ↓
Scalar reward model	$22.92_{\pm 2.56}$	$1.33_{\pm 0.47}$	$65.00_{\pm 4.32}$
SFT-only rubric generator	$28.21_{\pm 4.06}$	$10.67_{\pm 3.30}$	$62.00_{\pm 2.94}$
Uniform rubric criteria	$41.62_{\pm 4.85}$	$1.33_{\pm 1.89}$	$41.33_{\pm 7.59}$
Uniform dimension weights	$64.20_{\pm 0.15}$	$34.33_{\pm 1.89}$	$24.00_{\pm 1.41}$
Full CARE rubric	$\mathbf{91.41}_{\pm \mathbf{2.09}}$	$\mathbf{72.33}_{\pm \mathbf{2.87}}$	$\mathbf{2.00}_{\pm \mathbf{0.82}}$

Table 4: Ablation of rubric generator training on preference recovery.

SFT	RL	Reward	Win	Tie	Loss
–	–	–	49.3	13.8	36.8
√	–	–	47.6	24.1	28.3
–	√	Full	53.3	18.7	28.0
√	√	w/o R_{align}	56.4	9.7	33.9
√	√	Full	**59.4**	11.5	29.1

SentientBench. CARE delivers substantial gains on this interactive benchmark, improving both dialogue scores and successful support outcomes. The gap is especially pronounced under GPT-5.5, the strictest judge in terms of success rate, where **CARE (R) achieves 30.33% success compared with only 5.67% for MICA**, the strongest baseline, while reducing the failure rate to 0%. Notably, RLVER originally trails MICA substantially on SentientBench, but this gap largely disappears after applying CARE: CARE (R) matches or slightly outperforms CARE (M) on most metrics. This suggests that adaptive reward specifications can matter as much as the choice of the RL pipeline.

EQBench3. The gains extend to EQBench3, which evaluates emotional intelligence through role-play and transcript analysis. Across the three judges, the best **CARE variant improves Overall by 2.55–7.67 points over the strongest baseline**, while CARE (R) and CARE (M) consistently rank first and second. These results show that the benefit of adaptive reward specifications extends beyond interactive user-simulator benchmarks. Complete results are provided in § F, Table 6.

EMPA. CARE shows particularly large gains on EMPA, which evaluates persona-aligned empathy over complete dialogue trajectories. Under DeepSeek-V4-Pro, CARE (R) improves EPM-Idx from 52.10 for the strongest baseline MICA to 95.77. Under Gemini-2.5-Pro, the gain is even larger, with EPM-Idx increasing from 28.11 to 83.54—**nearly three times the baseline**. Across all three judges, CARE (R) and CARE (M) consistently occupy the top two positions, indicating that the improvement is robust across both evaluator choices and user-simulator-based RL pipelines. Importantly, EPM-Idx jointly reflects outcome quality, process efficiency, and strategic stability (details in §F, Table 7), and the component-level results show particularly strong gains in outcome quality and stability.

Human evaluation. Fifteen annotators with psychology-related backgrounds select the best response among CARE (R), MICA, and RLVER for randomly sampled dialogue contexts, with model identities hidden. CARE (R) is also preferred by human evaluators, receiving 51.5% of selections, compared with 23.9% for MICA and 24.6% for RLVER. See Appendix G for details.

Scaling and comparison with frontier LLMs. Table 2 further evaluates CARE across Qwen3 model scales and compares the resulting models with strong frontier LLMs. CARE consistently improves EQBench3 Overall and all other 7 metrics. Notably, CARE (R) at 14B already surpasses the 32B Base model (73.98 vs. 73.50). At 32B, CARE (R) further **exceeds GLM-5.2**, a 744B-

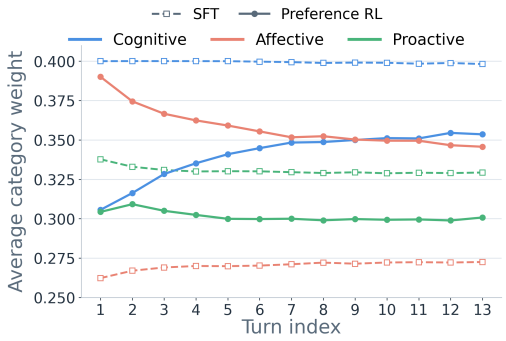


Figure 3: **Mean dimension weights** across dialogue turns for the SFT generator (dotted) and the preference-RL generator (solid).

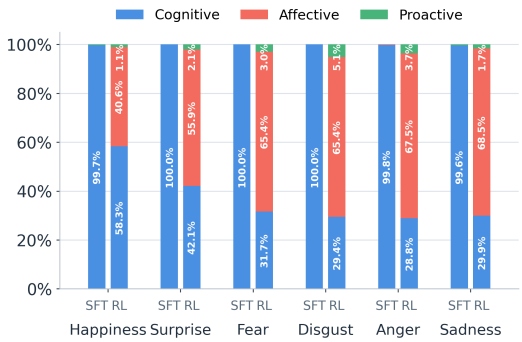


Figure 4: Distribution of the **largest dominant rubric dimension** conditioned on user emotion across SFT/RL generators.

parameter MoE model with 40B active parameters (76.35 vs. 74.90), and nearly **matches Qwen3.7-Max** (76.60). These results demonstrate that the gains from CARE persist across model scales and substantially narrow the gap between compact Qwen3 models and stronger frontier systems.

4.3 Ablation Studies

Effect of rubric design during policy optimization. We isolate the contribution of different rubric components on SentientBench. The scalar reward model variant replaces adaptive rubric scoring with the reward model of MICA (Zhang et al., 2026a), while the SFT-only variant uses the rubric generator before preference-based RL. To separately test the two adaptive components of CARE, Uniform dimension weights retain the generated criteria but set all dimension weights to $1/3$, whereas Uniform rubric criteria retain the generated weights but replace the turn-specific criteria with a fixed six-criterion set containing two representative criteria per empathy dimension. Table 3 shows that removing either adaptive weights or turn-specific criteria causes substantial performance degradation, while the full CARE rubric achieves the best results across all metrics. These results confirm that both context-dependent priorities and fine-grained criteria are essential to the effectiveness of the adaptive reward specification.

Ablation of rubric-generator training. We evaluate different training variants on 166 held-out chosen–rejected pairs, where the generator receives only the dialogue context and DeepSeek-V4-Flash scores both responses under the generated rubric. As shown in Table 4, SFT alone achieves a 47.6% chosen-response win rate, while adding preference-based RL increases it to 59.4%. Removing r_ϕ^{align} reduces the win rate to 56.4%, showing that both preference discrimination and priority alignment contribute to learning rubrics that better recover human preferences.

5 Analyses

Qualitative cases reveal context-aligned support priorities. Representative cases in § C show that CARE assigns dimension weights consistent with the user's apparent support needs. Cognitive empathy is prioritized when the user needs help interpreting a complex relationship conflict; affective empathy becomes dominant when the user expresses self-blame and needs emotional validation; and proactive empathy receives the highest weight when the user seeks concrete help with a practical problem. These examples suggest that the learned weights reflect interpretable differences in support needs across dialogue states.

Preference learning reshapes rubric priorities. We first compare rubrics generated by the SFT and preference-RL models across dialogue turns. As shown in Figure 3, the SFT generator produces relatively stable dimension weights, with cognitive empathy remaining consistently dominant across turns. Preference learning substantially changes this pattern. Affective empathy dominates in earlier turns, but its weight gradually declines as cognitive empathy rises, with the two crossing around turn 8 and reversing their relative priority thereafter. These results indicate that preference learning does not simply amplify a fixed empathy dimension. Instead, it learns to redistribute reward priorities over

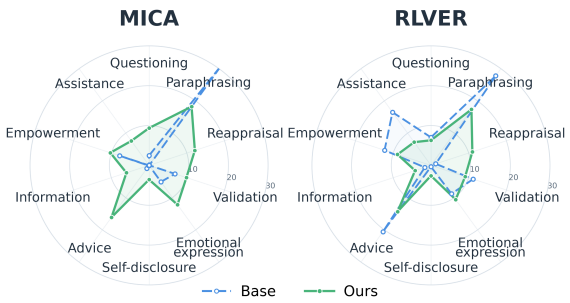


Figure 5: Support-strategy occurrence rates.

Table 5: General-capability performance on code, mathematics, and instruction following tasks.

Model	LiveCode	MATH	IFBench
Qwen2.5-7B	49.10	73.60	23.51
+CARE (R)	51.36	73.60	26.07
Qwen3-8B	62.67	85.00	21.88
+CARE (R)	70.59	85.80	20.22
Qwen3-14B	77.38	88.60	21.32
+CARE (R)	83.03	89.60	20.07
Qwen3-32B	78.05	92.00	21.87
+CARE (R)	85.52	90.40	23.12

the course of a dialogue, consistent with the evolving support needs that motivate CARE. Appendix Figure 9 provides the complementary strict-maximum analysis.

Rubric priorities vary with user emotion. To determine whether the turn-wise variation in Figure 3 reflects dialogue progression alone or also covaries with user state, we analyze rubric composition across the six basic emotions assigned to user turns by DeepSeek-V4-Flash. As shown in Figure 4, SFT rubrics remain almost uniformly cognitive-dominant across emotions, whereas preference-RL rubrics exhibit substantially greater variation, showing a general shift from cognitive to affective empathy from positive to negative emotional valence, with proactive empathy remaining a relatively minor supplementary component. Consistent with this observation, analysis in Appendix Figure 10 shows systematic variation in downstream strategy distributions across user emotions.

Adaptive rubrics reshape downstream support behavior. We next examine whether changes in the reward specification translate into observable changes in policy behavior. We apply the ten-strategy MINT annotator (Zhan et al., 2026) to responses from the EMPA evaluation and report strategy occurrence rates in Figure 5. CARE broadens MICA's support profile beyond paraphrasing toward questioning, reappraisal, information, and empowerment, while shifting RLVER away from paraphrasing and advice toward reappraisal, revealing policy-specific behavioral shifts rather than convergence on a fixed support strategy.

General capabilities remain broadly preserved. We further assess whether the observed empathy gains come at the expense of general model capabilities. Table 5 reports performance on LiveCodeBench (Jain et al., 2025), MATH-500 (Lightman et al., 2023), and IFBench (Pyatkin et al., 2025) across four policy backbones. CARE improves LiveCodeBench for all four models, with an average gain of 5.83 points, while MATH-500 remains essentially unchanged on average. IFBench shows small mixed changes across model scales, improving on Qwen2.5-7B and Qwen3-32B while decreasing on Qwen3-8B and Qwen3-14B. Overall, we observe no systematic degradation across the three general-capability benchmarks. These results suggest that CARE can substantially alter empathy-related behavior without broadly compromising the capabilities of the policy model.

6 Conclusion

We study a fundamental mismatch in empathetic reinforcement learning: support priorities evolve throughout a dialogue, while reward specifications are typically fixed across turns. To address this, we formulate the *Principle of Evolving Support Priorities* and propose CARE, which generates context-adaptive rubrics with turn-specific empathy priorities and fine-grained evaluation criteria. The rubric generator is trained with turn-level supervision and preference-based reinforcement learning, and then serves as a frozen adaptive reward interface for online dialogue RL. Integrated into both RLVER and MICA, CARE consistently improves performance across SentientBench, EQBench3, and EMPA under three independent LLM judges, with particularly large gains on long-horizon interactive benchmarks. Further analyses show that learned rubric priorities systematically vary across dialogue stages and user emotions and reshape downstream support strategies. Together, these results suggest that effective empathetic RL should optimize not only how well a response is rewarded, but also what should be rewarded as user needs evolve.

ETHICS STATEMENT

Emotional-support dialogues can contain sensitive disclosures about relationships, health, grief, and psychological distress. Data collection and release should follow the licenses and consent conditions of the source datasets, minimize personally identifying information, and limit access to any non-public examples. Models and annotators may also reproduce cultural assumptions about appropriate support; our use of multiple judges reduces dependence on one evaluator but does not eliminate this risk. The proposed system is not a mental-health professional and should not be presented as a substitute for clinical care or emergency services. In deployment, high-risk content requires clear escalation policies, privacy protections, and human oversight. Because an adaptive reward can make a dialogue policy more persuasive as well as more supportive, downstream use should be monitored for manipulation, over-reliance, and advice that exceeds the system's competence.

REPRODUCIBILITY STATEMENT

§3 specifies the rubric representation, target construction and verification procedure, generator objectives, online reward, and GRPO integration. §4 identifies the policy and generator backbones, comparison systems, benchmarks, and evaluation judges; §A provides additional implementation information, complete component-level results, and qualitative rubric outputs. To support reproduction, the accompanying release will include data-construction prompts, the strategy-to-dimension mapping, filtering and training configurations, random seeds, evaluation scripts, and model checkpoints or adapters where the corresponding licenses permit release.

AI USE STATEMENT

Generative AI tools were used in this work in two main capacities: (1) for language editing and presentation support in the writing of this manuscript, and (2) as an integral part of the research methodology. Specifically, large language models (LLMs) were used to generate synthetic training data, including the construction of turn-level rubric supervision (via DeepSeek-v4-pro) and preference data (via GPT-5, Gemini-3-Pro-Preview, and DeepSeek-V3.2). Furthermore, LLMs served as evaluators and reward models during the online reinforcement learning training process (e.g., DeepSeek-V4-Flash as the rubric-conditioned evaluator). The authors reviewed and verified all AI-generated content, checked technical statements against the implemented method and reported experiments, and take full responsibility for the final content, including any remaining errors.

REFERENCES

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

Brooke C Feeney and Nancy L Collins. A new look at social support: A theoretical perspective on thriving through relationships. *Personality and social psychology review*, 2015.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *NeurIPS ER Workshop*, 2025.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *ICLR*, 2025.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *ACL-IJCNLP*, 2021.

Changze Lv, Jie Zhou, Wentao Zhao, Jingwen Xu, Shihan Dou, Zisu Huang, Muzhao Tian, Xiaohua Wang, Yang Liu, Pluto Zhou, Tao Gui, Le Tian, Xiao Zhou, Xiaoqing Zheng, Xuanjing Huang, and Jie Zhou. Learning query-specific rubrics from human preferences for DeepResearch report generation. *arXiv preprint arXiv:2602.03619*, 2026.

Samuel J. Paech. Eq-bench 3: Emotional intelligence benchmark, 2025. Benchmark repository.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following. In *NeurIPS*, 2025.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yuxin Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Yoko Yamakata, and Tat-Seng Chua. Preference-aware rubric learning for personalized evaluation. *arXiv preprint arXiv:2605.31545*, 2026.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, 2019.

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

Beining Wang, Weihang Su, Hongtao Tian, Hao Kong, Tao Yang, Ting Yao, Qingyi Pan, Yueyue Wu, Qingyao Ai, Min Zhang, et al. Co-evolving llm evaluators and policies via dynamicrubric. *arXiv preprint arXiv:2607.20083*, 2026a.

Chengbing Wang, Wuqiang Zheng, Yang Zhang, Fengbin Zhu, Junyi Cheng, Yi Xie, Wenjie Wang, and Fuli Feng. Perm: Psychology-grounded empathetic reward modeling for large language models. In *ACL-findings*, 2026b.

Peisong Wang, Ruotian Ma, Bang Zhang, Xingyu Chen, Zhiwei He, Kang Luo, Qingsong Lv, Qingxuan Jiang, Zheng Xie, Shanyi Wang, Cixing Li, Yuan Li, Fanghua Ye, Jian Li, Yifan Yang, Jia Li, Zhaopeng Tu, and Xiaolong Li. RLVER: Reinforcement learning with verifiable emotion rewards for empathetic agents. In *ICLR*, 2026c.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Jiahao Yuan, Zhiqing Cui, Hanqing Wang, Yuansheng Gao, Yucheng Zhou, and Usman Naseem. Kardia-r1: Unleashing llms to reason toward understanding and empathy for emotional support via rubric-as-judge reinforcement learning. In *WWW*, 2026.

Jamil Zaki. Integrating empathy and interpersonal emotion regulation. *Annual review of psychology*, 2020.

Hongli Zhan, Emma S. Gueorguieva, Javier Hernandez, Jina Suh, Desmond C. Ong, and Junyi Jessy Li. Discourse diversity in multi-turn empathic dialogue. *arXiv preprint arXiv:2604.11742*, 2026.

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. Sentient agent as a judge: Evaluating higher-order social cognition in large language models. *arXiv preprint arXiv:2505.02847*, 2025.

Mian Zhang, Shaun M. Eack, and Zhiyu Zoey Chen. Preference learning unlocks LLMs' psychocounseling skills. In *ACL-findings*, 2026a.

Naifan Zhang, Ruihan Sun, Jinwei Su, Hengjie Yang, Zhengyuan Pan, Zhaohan Chen, and Xiaofan Zhang. Mica: Multi-granularity intertemporal credit assignment for long-horizon emotional support dialogue. *arXiv preprint arXiv:2603.06194*, 2026b.

Shiya Zhang, Yuhan Zhan, Ruixi Su, Ruihan Sun, Ziyi Song, Zhaohan Chen, and Xiaofan Zhang. Empa: Evaluating persona-aligned empathy as a process. *arXiv preprint arXiv:2603.00552*, 2026c.

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

A Implementation details

All dialogue-policy systems in the main comparison use Qwen2.5-7B-Instruct as their common initialization. The rubric generator is a fully fine-tuned Qwen3-8B model with a constrained JSON interface containing the three empathy dimensions, one weight and rationale per dimension, and a variable-length list of titled criteria; DeepSeek-V4-Flash serves as the rubric-conditioned evaluator. Target rubrics are retained only after the agreement-based verification described in §3.3. During SFT, preference learning, and online policy optimization, the generator receives only the dialogue history available before the candidate response; in SFT, the corresponding verified rubric is the generation target. During policy optimization, both the rubric generator and evaluator are frozen, one rubric is shared across every response in a GRPO group, and the adaptive rubric score replaces rather than augments the original RLVER or MICA reward.

We use a batch size of 8 and set the learning rate to 2×10^{-6}. Training uses 100 steps. Dialogues are fixed at 8 turns, and the sampling temperature is set to 1 to encourage exploration. For GRPO, we use 6 rollouts per state. All experiments are run on 8 H100 GPUs. For the rubric-generator reward, we set the preference-temperature parameter to $\tau = 0.7$, the priority-alignment coefficient to $\lambda = 0.1$, and the format-reward coefficient to $\gamma = 0.3$.

B Benchmark details

B.1 SentientBench.

SentientBench (Zhang et al., 2025) evaluates higher-order social cognition through interactive conversations with simulated users whose emotions, beliefs, and intentions evolve over turns. It is well suited to emotional-support evaluation because successful systems must infer the seeker's changing mental state rather than optimize isolated response.

B.2 EQBench3.

EQBench3 (Paech, 2025) offers a compact emotional-intelligence evaluation suite spanning multi-turn role-play and transcript-analysis tasks. Its aggregate Overall score summarizes capabilities such as interpersonal sensitivity, psychological reasoning, and response appropriateness, making it a useful complement to process-oriented dialogue benchmarks.

B.3 EMPA.

EMPA (Zhang et al., 2026c) treats empathy as a trajectory-level intervention rather than a single-turn response-quality judgment. It measures persona-aligned support through outcome, efficiency, and stability dimensions, with EPM-Idx summarizing whether a dialogue improves the simulated seeker's state while maintaining coherent and stable support.

B.4 LiveCodeBench.

LiveCodeBench (Jain et al., 2025) provides a contamination-aware coding benchmark built from recent programming problems. We use it as a general-capability probe to check whether empathy-oriented policy optimization preserves algorithmic problem-solving and code-generation ability.

B.5 MATH-500.

MATH-500 (Lightman et al., 2023) offers a streamlined slice of the broader MATH problem setting, comprising 500 test problems selected to cover diverse topics and difficulty levels. It remains a challenging measure of mathematical reasoning and helps test whether dialogue optimization degrades non-dialogue reasoning skills.

B.6 IFBENCH.

IFBench (Pyatkin et al., 2025) evaluates verifiable instruction following with constraints that can be automatically checked. We include it to test whether the adapted policy remains responsive to explicit user instructions, since emotional support often requires following constraints about tone, scope, and requested forms of help.

C QUALITATIVE CASE STUDIES OF STATE-CONDITIONED RUBRICS

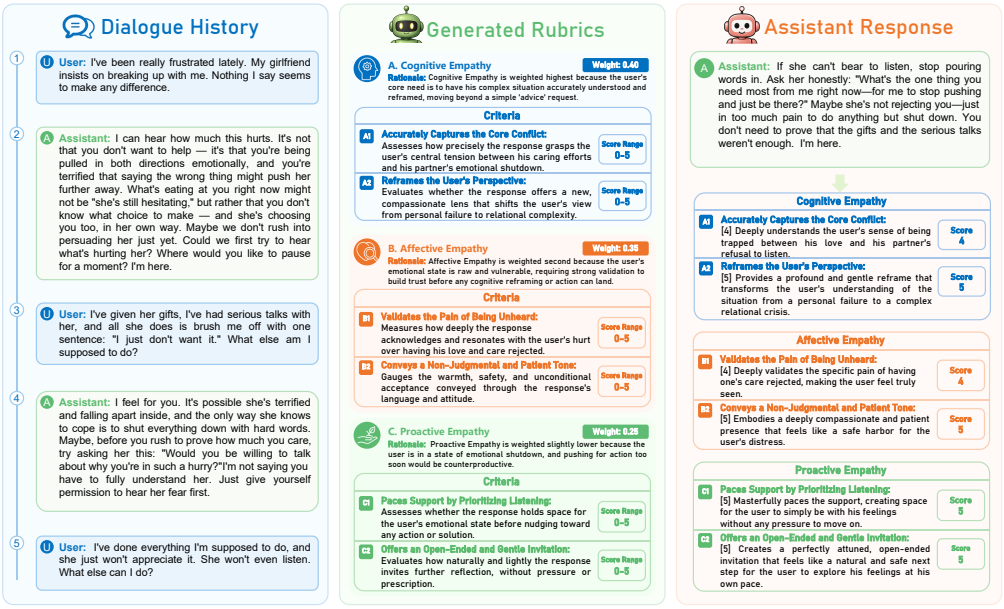


Figure 6: Cognitive-dominant rubric case.

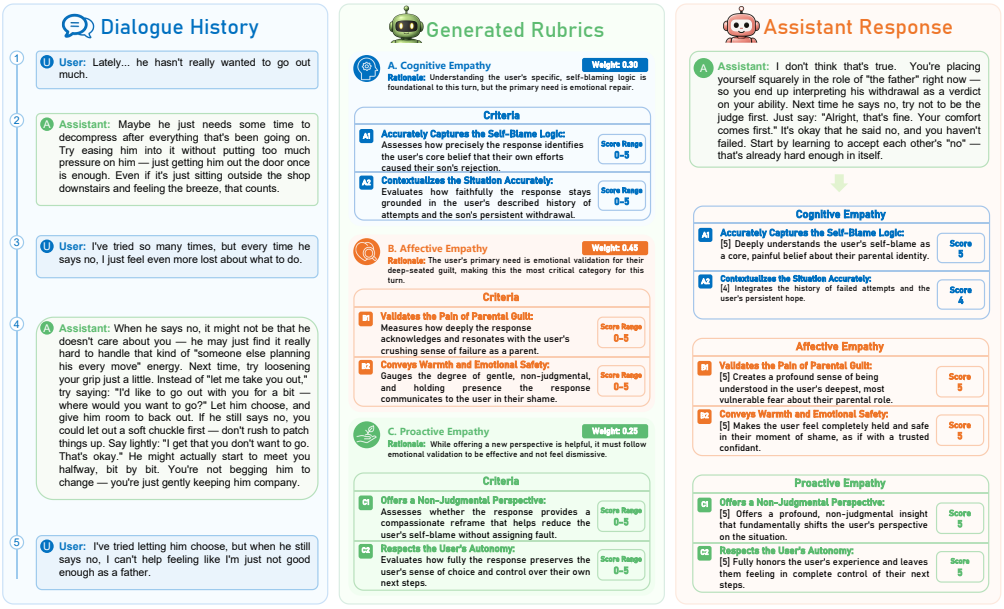


Figure 7: Affective-dominant rubric case.

C.1 COGNITIVE-EMPATHY-DOMINANT CASE

In the cognitive-empathy-dominant case (Figure 6), the user faces a complex conflict surrounding a possible relationship breakup. The rubric assigns weights of 0.40, 0.35, and 0.25 to cognitive, affective, and proactive empathy, respectively, emphasizing conflict interpretation and reframing. The response encourages listening rather than further persuasion. Its two criterion scores are 4 and 5 for cognitive empathy, 4 and 5 for affective empathy, and 5 and 5 for proactive empathy, showing that the highest rubric weight need not correspond to the highest realized score.

C.2 AFFECTIVE-EMPATHY-DOMINANT CASE

The affective-empathy-dominant case (Figure 7) concerns a parent who interprets a son's withdrawal as personal failure. Affective empathy receives the largest weight (0.45), followed by cognitive empathy (0.30) and proactive empathy (0.25), with criteria centered on self-blame, validation, and emotional safety. The response challenges the self-blaming interpretation while preserving the son's autonomy. It receives criterion scores of 5 and 4 for cognitive empathy and 5 and 5 for both affective and proactive empathy.

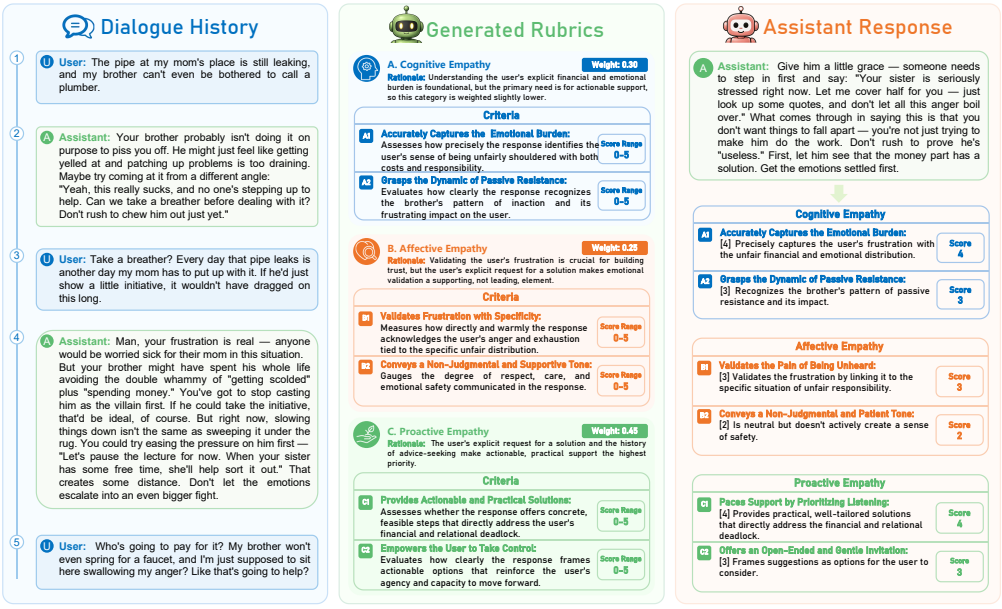


Figure 8: Proactive-dominant rubric case.

C.3 PROACTIVE-EMPATHY-DOMINANT CASE

In the proactive-empathy-dominant case (Figure 8), the user is frustrated by a leaking pipe, an unresponsive brother, and the resulting family burden. Proactive empathy receives the largest weight (0.45), followed by cognitive empathy (0.30) and affective empathy (0.25), emphasizing concrete support and user agency. The response proposes a practical path forward and receives criterion scores of 4 and 3 for cognitive empathy, 3 and 2 for affective empathy, and 4 and 3 for proactive empathy. The rubric captures both the response's stronger action orientation and its comparatively weak emotional validation.

D RUBRIC-DOMINANCE ANALYSIS

To complement the mean-weight analysis in Figure 3, we examine which empathy dimension attains the strict maximum at each dialogue turn. SFT remains almost uniformly cognitive-dominant, whereas preference RL shifts from affective dominance in early turns toward cognitive dominance later in the dialogue. We further examine whether this dominance pattern systematically differs across specific emotional contexts in Figure 9.

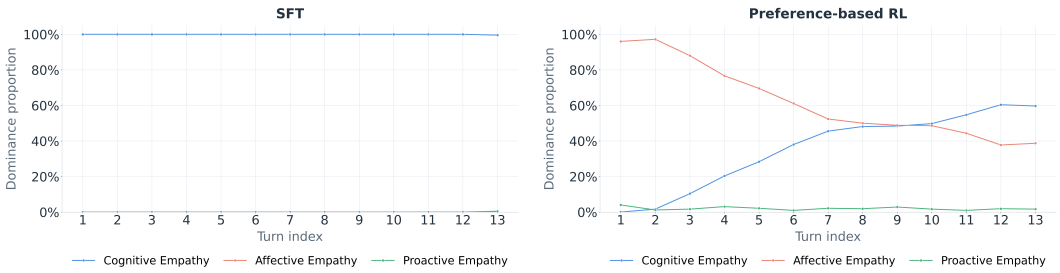


Figure 9: Strict-maximum empathy-dimension proportions by dialogue turn for SFT (left) and preference RL (right). Each dialogue-turn first aggregates repeated rubric rollouts; curves then report dimension-level proportions across dialogues.

E SUPPORT-STRATEGY ANALYSIS

Figure 10 shows substantial differences in emotion-conditioned tactic use across methods. MICA exhibits a sparse profile dominated by paraphrasing, with limited variation across emotions. RLVER uses a broader set of tactics, particularly advice, validation, and assistance, and shows more emotion-dependent variation than MICA. Our MICA- and RLVER-based models further diversify tactic use across appraisal, emotional expression, empowerment, and other supportive behaviors, yielding more differentiated support strategies across emotional contexts.

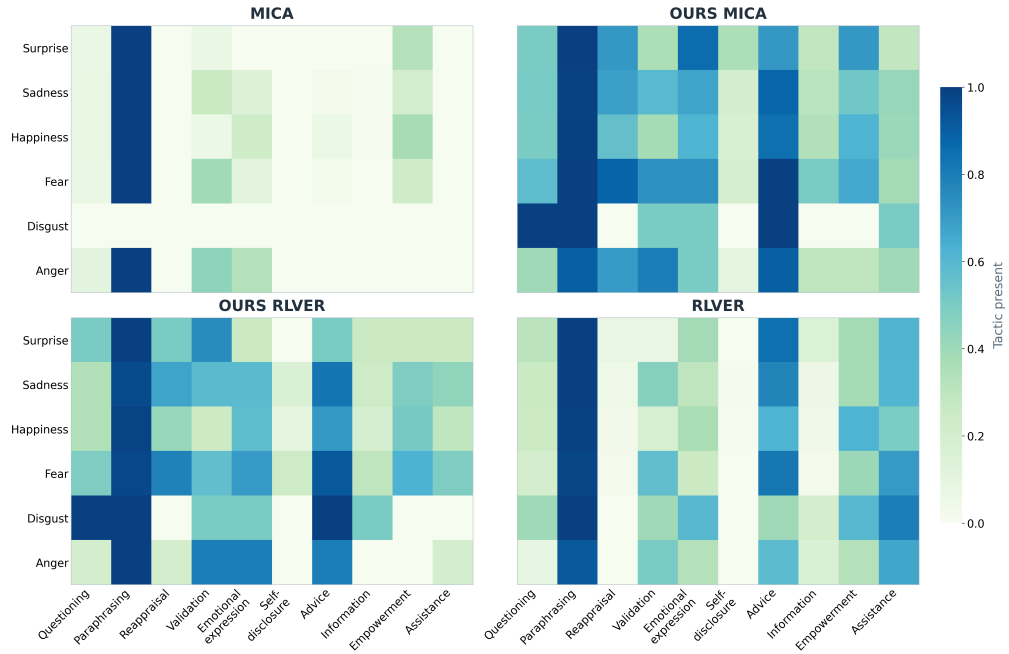


Figure 10: Emotion-conditioned tactic use per method. Each cell reports the fraction of scenario-turns associated with a given emotion label that contain the corresponding MINT tactic.

F ADDITIONAL RESULTS

EQBench3 evaluates active emotional intelligence through seven rubric dimensions—DoI, ER, DE, WRM, HL, PEI, and SD—covering psychological insight, emotional reasoning, empathy, warmth, humanlike interaction, pragmatic support, and social dexterity. As shown in Table 6, `CARE`achieves the best Overall score under all three judges, with `CARE (M)` reaching 57.53 under DeepSeek-V4-Pro and `CARE (R)` reaching 59.60 under Gemini-2.5-Pro and 37.65 under GPT-5.5, corresponding to gains of 6.06, 7.67, and 2.55 over the reproduced MICA baseline. The most consistent improvements appear in DoI, ER, HL, and PEI, while DE also improves under DeepSeek and Gemini; WRM

Table 6: Complete EQBench3 results under three LLM judges, including Overall and all seven component scores. Values are means ± standard deviations over three runs; all metrics are higher-is-better, and bold and underline mark the best and second-best means.

Model	Overall	DoI	ER	DE	WRM	HL	PEI	SD
Judge: DeepSeek-V4-Pro								
Qwen2.5-7B-Instruct	42.53±2.18	9.35±0.49	8.97±0.47	10.09±0.69	9.41±0.19	8.59±0.68	8.85±0.98	7.56±0.91
PERM (Wang et al., 2026b)	49.07±1.03	10.97±0.15	10.26±0.10	11.98±0.32	10.63±0.32	10.36±0.43	10.90±0.40	9.58±0.43
KARDIA-R1 (Yuan et al., 2026)	28.75±1.21	6.00±0.36	6.04±0.23	8.21±0.12	9.61±0.09	6.80±0.46	6.15±0.27	5.52±0.23
RLVER (Wang et al., 2026c)	44.23±2.23	9.71±0.62	9.39±0.41	10.88±0.35	10.50±0.60	9.36±0.43	9.23±0.47	8.14±0.49
MICA (rep.) (Zhang et al., 2026b)	51.47±0.58	11.51±0.20	10.92±0.16	11.74±0.19	**11.18±0.09**	11.22±0.28	9.88±0.49	9.11±0.66
CARE (M)	**57.53±1.20**	12.63±0.44	**12.04±0.17**	**12.17±0.27**	9.92±0.10	12.17±0.61	**11.15±0.06**	**9.94±0.47**
CARE (R)	55.88±1.59	**12.73±0.18**	11.93±0.40	12.13±0.24	9.96±0.48	**12.29±0.65**	10.60±0.41	9.59±0.28
Judge: Gemini-2.5-Pro								
Qwen2.5-7B-Instruct	43.43±1.74	10.05±0.25	9.79±0.27	7.88±0.43	7.05±0.36	6.56±0.78	7.49±1.19	5.49±0.43
PERM (Wang et al., 2026b)	47.13±2.63	11.69±0.48	10.68±0.32	8.96±0.61	7.23±0.72	7.43±0.46	8.50±1.20	5.65±0.56
KARDIA-R1 (Yuan et al., 2026)	23.95±0.36	4.54±0.09	5.39±0.25	4.09±0.10	7.27±0.27	3.20±0.06	3.18±0.10	3.14±0.33
RLVER (Wang et al., 2026c)	43.98±1.23	10.33±0.25	9.93±0.38	8.27±0.77	7.42±0.52	6.17±0.10	7.19±0.69	5.32±0.12
MICA (rep.) (Zhang et al., 2026b)	51.93±1.05	12.01±0.28	11.99±0.09	9.96±0.56	**9.27±0.71**	8.95±0.42	8.67±0.56	6.93±0.66
CARE (M)	58.82±2.94	**13.70±0.40**	13.58±0.46	10.57±0.27	6.98±0.40	9.98±0.31	10.10±0.69	7.09±0.93
CARE (R)	**59.60±3.94**	13.64±0.73	**13.82±0.56**	**10.92±1.07**	7.22±1.09	**10.87±0.93**	**10.67±1.37**	**7.69±1.05**
Judge: GPT-5.5								
Qwen2.5-7B-Instruct	32.10±0.18	6.39±0.02	7.45±0.03	8.97±0.05	9.65±0.04	4.90±0.16	5.67±0.18	5.20±0.09
PERM (Wang et al., 2026b)	34.88±0.58	7.25±0.16	8.06±0.14	**9.91±0.31**	**11.00±0.57**	5.57±0.05	6.01±0.21	5.51±0.06
KARDIA-R1 (Yuan et al., 2026)	24.15±0.74	4.38±0.17	5.63±0.18	7.18±0.17	9.08±0.10	3.69±0.20	3.81±0.14	3.85±0.08
RLVER (Wang et al., 2026c)	32.00±0.61	6.32±0.09	7.44±0.11	9.09±0.26	10.19±0.16	4.55±0.08	5.56±0.20	5.18±0.17
MICA (rep.) (Zhang et al., 2026b)	35.10±0.39	7.38±0.10	8.20±0.14	9.76±0.30	10.56±0.25	6.10±0.19	5.33±0.22	5.28±0.23
CARE (M)	37.58±0.93	7.95±0.13	8.63±0.19	9.45±0.21	8.20±0.35	6.92±0.10	6.38±0.20	**5.58±0.04**
CARE (R)	**37.65±0.89**	**7.98±0.22**	**8.66±0.28**	9.25±0.30	7.78±0.28	**7.21±0.10**	**6.49±0.44**	5.55±0.36

Table 7: Complete EMPA EPM-Idx and outcome results under three LLM judges.

Model	EPM-Idx	Outcome			
		Outcome	RDI	Etot	Snet
Judge: DeepSeek-V4-Pro					
Qwen2.5-7B-Instruct	15.38±1.06	3.25±0.76	7.5±1.5	0.0±0.0	2.3±0.9
PERM (Wang et al., 2026b)	17.21±1.17	6.39±2.14	10.6±1.7	0.3±0.4	8.2±4.8
KARDIA-R1 (Yuan et al., 2026)	11.42±0.34	0.81±0.34	2.4±1.0	0.0±0.0	0.0±0.0
RLVER (Wang et al., 2026c)	26.05±3.37	15.34±4.72	17.3±4.5	5.2±2.9	23.5±7.3
MICA (rep.) (Zhang et al., 2026b)	52.10±4.12	52.19±6.63	46.6±6.0	33.3±5.7	76.7±10.1
CARE (M)	94.84±1.08	102.32±1.59	96.4±2.3	90.5±0.3	120.0±2.3
CARE (R)	**95.77±0.33**	**103.36±1.02**	**97.9±0.1**	**91.4±0.8**	**120.9±2.3**
Judge: Gemini-2.5-Pro					
Qwen2.5-7B-Instruct	15.95±1.87	4.40±2.52	6.0±2.1	0.0±0.0	7.1±5.8
PERM (Wang et al., 2026b)	14.53±1.03	3.20±0.88	5.1±2.5	0.0±0.0	4.5±0.4
KARDIA-R1 (Yuan et al., 2026)	12.72±0.91	1.16±0.39	1.9±1.0	0.0±0.0	1.6±0.4
RLVER (Wang et al., 2026c)	14.65±0.43	1.89±0.43	2.0±0.4	0.0±0.0	3.7±1.1
MICA (rep.) (Zhang et al., 2026b)	28.11±2.28	18.11±3.51	14.3±5.2	2.8±1.9	37.3±5.3
CARE (M)	81.20±4.81	86.99±3.35	78.3±4.1	75.2±7.4	107.5±1.6
CARE (R)	**83.54±3.94**	**92.25±5.68**	**81.6±4.9**	**79.1±6.3**	**116.0±9.4**
Judge: GPT-5.5					
Qwen2.5-7B-Instruct	22.97±2.75	16.69±4.75	14.0±4.1	3.7±2.2	32.3±8.3
PERM (Wang et al., 2026b)	17.29±2.90	10.38±3.76	10.7±3.6	2.9±0.9	17.5±7.3
KARDIA-R1 (Yuan et al., 2026)	11.76±0.19	1.25±0.48	2.4±0.8	0.0±0.0	1.3±0.6
RLVER (Wang et al., 2026c)	19.19±1.88	10.48±2.11	8.0±1.0	2.3±0.2	21.1±5.4
MICA (rep.) (Zhang et al., 2026b)	25.42±1.41	18.82±3.48	22.3±5.6	3.5±2.0	30.6±6.8
CARE (M)	36.70±4.38	38.35±5.38	37.5±4.7	15.0±8.0	62.5±4.3
CARE (R)	**38.44±3.96**	**42.32±5.03**	**41.1±5.2**	**19.1±6.1**	**66.7±4.4**

and SD are less stable across judges, suggesting that CAREprimarily strengthens the substance, targeting, and practicality of emotional support rather than surface style alone.

EMPA evaluates persona-aligned empathy at the trajectory level through EPM-Idx, which aggregates Outcome (RDI, Etot, Snet), Efficiency (Rho, Sproj, Tau), and Stability (Rpos, Align, Pen), with all reported indices normalized so that higher is better. As shown in Table 7, CARE (R) achieves the highest EPM-Idx under all three judges, scoring 95.77, 83.54, and 38.44 under DeepSeek-V4-Pro, Gemini-2.5-Pro, and GPT-5.5, respectively, corresponding to gains of 43.67, 55.43, and 13.02 over reproduced MICA. The largest improvements occur in Outcome, particularly RDI, Etot, and

Table 7: Complete EMPA results under three LLM judges (continued).

Model	Efficiency				Stability			
	Efficiency	Rho	Sproj	Tau	Stability	Rpos	Align	Pen
Judge: DeepSeek-V4-Pro								
Qwen2.5-7B-Instruct	$25.67_{\pm 1.15}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$77.0_{\pm 3.4}$	$22.35_{\pm 2.38}$	$17.7_{\pm 1.5}$	$22.3_{\pm 2.4}$	$27.2_{\pm 3.3}$
PERM (Wang et al., 2026b)	$23.19_{\pm 1.07}$	$0.2_{\pm 0.2}$	$0.2_{\pm 0.2}$	$69.3_{\pm 3.3}$	$25.04_{\pm 2.25}$	$21.3_{\pm 2.2}$	$25.2_{\pm 2.3}$	$28.7_{\pm 2.4}$
KARDIA-R1 (Yuan et al., 2026)	$30.91_{\pm 0.50}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$92.7_{\pm 1.5}$	$12.29_{\pm 0.76}$	$8.1_{\pm 0.9}$	$12.9_{\pm 0.7}$	$15.9_{\pm 0.7}$
RLVER (Wang et al., 2026c)	$17.86_{\pm 3.15}$	$2.0_{\pm 1.2}$	$2.0_{\pm 1.1}$	$49.6_{\pm 8.8}$	$40.87_{\pm 4.36}$	$37.2_{\pm 4.5}$	$39.4_{\pm 4.2}$	$46.0_{\pm 5.1}$
MICA (rep.) (Zhang et al., 2026b)	$25.12_{\pm 4.36}$	$16.0_{\pm 3.0}$	$15.4_{\pm 2.9}$	$44.0_{\pm 7.7}$	$65.49_{\pm 2.76}$	$64.5_{\pm 2.9}$	$62.6_{\pm 2.7}$	$69.4_{\pm 2.8}$
CARE (M)	$80.16_{\pm 1.22}$	$75.0_{\pm 4.0}$	$70.7_{\pm 3.5}$	$94.8_{\pm 4.2}$	$94.70_{\pm 1.65}$	$92.5_{\pm 1.8}$	$93.1_{\pm 2.1}$	$98.4_{\pm 1.1}$
CARE (R)	**$81.40_{\pm 0.88}$**	**$75.7_{\pm 1.1}$**	**$71.4_{\pm 1.0}$**	**$97.1_{\pm 0.6}$**	**$95.35_{\pm 0.45}$**	**$93.1_{\pm 0.5}$**	**$94.3_{\pm 0.5}$**	**$98.7_{\pm 0.5}$**
Judge: Gemini-2.5-Pro								
Qwen2.5-7B-Instruct	$24.29_{\pm 2.22}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$72.9_{\pm 6.7}$	$23.32_{\pm 3.36}$	$21.2_{\pm 3.1}$	$29.3_{\pm 3.0}$	$19.5_{\pm 4.0}$
PERM (Wang et al., 2026b)	$26.70_{\pm 1.35}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$80.1_{\pm 4.0}$	$19.78_{\pm 2.44}$	$18.9_{\pm 2.6}$	$27.8_{\pm 2.3}$	$12.6_{\pm 2.3}$
KARDIA-R1 (Yuan et al., 2026)	$29.21_{\pm 0.44}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	$87.6_{\pm 1.3}$	$16.03_{\pm 2.12}$	$15.2_{\pm 2.2}$	$24.1_{\pm 1.8}$	$8.8_{\pm 2.4}$
RLVER (Wang et al., 2026c)	$29.36_{\pm 0.46}$	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	**$88.1_{\pm 1.4}$**	$20.05_{\pm 1.07}$	$15.6_{\pm 1.5}$	$29.0_{\pm 0.9}$	$15.5_{\pm 1.1}$
MICA (rep.) (Zhang et al., 2026b)	$18.73_{\pm 0.84}$	$1.9_{\pm 1.3}$	$1.8_{\pm 1.3}$	$52.5_{\pm 4.5}$	$42.80_{\pm 2.60}$	$38.4_{\pm 3.1}$	$47.7_{\pm 2.0}$	$42.3_{\pm 2.7}$
CARE (M)	**$74.87_{\pm 10.89}$**	$79.5_{\pm 2.5}$	$73.9_{\pm 11.5}$	$71.2_{\pm 9.1}$	$78.56_{\pm 3.27}$	$76.2_{\pm 3.3}$	$79.5_{\pm 2.6}$	$80.0_{\pm 3.9}$
CARE (R)	$74.52_{\pm 6.85}$	**$80.3_{\pm 6.8}$**	**$74.8_{\pm 6.3}$**	$68.5_{\pm 8.2}$	**$79.34_{\pm 2.85}$**	**$77.7_{\pm 2.7}$**	**$80.2_{\pm 2.4}$**	**$80.2_{\pm 3.7}$**
Judge: GPT-5.5								
Qwen2.5-7B-Instruct	$19.86_{\pm 1.05}$	$1.1_{\pm 0.7}$	$1.0_{\pm 0.7}$	$57.5_{\pm 4.5}$	$30.82_{\pm 2.90}$	$28.5_{\pm 2.9}$	$34.1_{\pm 2.8}$	$29.9_{\pm 3.1}$
PERM (Wang et al., 2026b)	$25.07_{\pm 2.81}$	$0.9_{\pm 0.3}$	$0.9_{\pm 0.3}$	$73.4_{\pm 8.8}$	$20.32_{\pm 5.12}$	$18.4_{\pm 5.8}$	$25.7_{\pm 4.4}$	$16.8_{\pm 5.3}$
KARDIA-R1 (Yuan et al., 2026)	**$29.22_{\pm 0.31}$**	$0.0_{\pm 0.0}$	$0.0_{\pm 0.0}$	**$87.6_{\pm 1.0}$**	$13.54_{\pm 0.92}$	$11.7_{\pm 1.5}$	$20.5_{\pm 1.2}$	$8.5_{\pm 0.6}$
RLVER (Wang et al., 2026c)	$23.76_{\pm 1.84}$	$0.7_{\pm 0.2}$	$0.7_{\pm 0.1}$	$69.8_{\pm 5.7}$	$25.61_{\pm 3.52}$	$23.3_{\pm 4.5}$	$31.0_{\pm 3.4}$	$22.5_{\pm 2.9}$
MICA (rep.) (Zhang et al., 2026b)	$15.66_{\pm 5.08}$	$1.7_{\pm 0.7}$	$1.7_{\pm 0.7}$	$43.6_{\pm 13.9}$	$36.91_{\pm 1.41}$	$34.1_{\pm 1.7}$	$40.4_{\pm 1.7}$	$36.3_{\pm 1.5}$
CARE (M)	$15.01_{\pm 5.11}$	$8.0_{\pm 4.5}$	$7.7_{\pm 4.3}$	$29.4_{\pm 6.8}$	**$45.89_{\pm 3.07}$**	**$44.6_{\pm 3.6}$**	**$48.7_{\pm 2.8}$**	**$44.4_{\pm 2.9}$**
CARE (R)	$16.63_{\pm 3.52}$	**$8.9_{\pm 3.6}$**	**$8.6_{\pm 3.4}$**	$32.5_{\pm 4.1}$	$45.47_{\pm 3.20}$	$43.7_{\pm 3.3}$	$48.6_{\pm 2.6}$	$44.1_{\pm 3.8}$

Table 8: Blinded human preference evaluation. fifteen annotators each evaluate 30 randomly assigned contexts. Model identities are hidden and response order is randomly shuffled.

Model	Selections	Preference (%) ↑
MICA (rep.) (Zhang et al., 2026b)	104	23.9
RLVER (Wang et al., 2026c)	107	24.6
CARE (R)	224	**51.5**

Snet, indicating stronger final psychological progress and greater cumulative effective support, while CAREalso consistently improves Efficiency and Stability over MICA.

G Human Preference Evaluation

To complement the automatic evaluations, we conduct a blinded human preference study comparing the end-to-end response quality of CARE (R) , MICA, and RLVER. We randomly sample 50 dialogue contexts from the 166 held-out examples used for rubric-generator evaluation in §4.3. For each example, we discard the original chosen and rejected responses and retain only the dialogue context ending with the final user turn. Given the same context, the three models independently generate responses to the final user message.

We recruit 15 annotators with backgrounds in psychology. Each annotator is randomly assigned 30 contexts from the 50-context pool. For each context, the annotator is shown the dialogue history together with three anonymized candidate responses and asked to select the response they prefer overall. Annotators may select a tie when no response is clearly preferred. Model identities are hidden throughout the evaluation, and response order is randomly shuffled to mitigate position bias.

We report the *human preference rate* after excluding tie judgments, computed as the fraction of remaining selections received by each model. As shown in Table 8, CARE (R) receives 51.5% of human selections, compared with 23.9% for MICA and 24.6% for RLVER. Thus, among judgments with a clear preference, more than half favor responses generated by CARE (R) . These results provide complementary human evidence that the gains from context-adaptive rubrics are also reflected in direct preferences over generated responses.

H STRATEGY-TO-DIMENSION MAPPING

We harmonize the strategy annotations in ESConv and MINT by mapping them into the three empathy dimensions used in our rubric framework: cognitive, affective, and proactive empathy. The mapping is based on the primary function of each strategy. Strategies that facilitate understanding or interpretation of the seeker's situation are categorized as cognitive empathy; those that convey emotional attunement, validation, or interpersonal support are categorized as affective empathy; and those that provide actionable guidance or practical support are categorized as proactive empathy. Table 9 summarizes the resulting mappings for both datasets.

Table 9: Mapping from ESConv strategies and MINT strategies to empathy dimensions.

Dimension	ESConv strategies	MINT strategies
Cognitive	Question; Restatement or Paraphrasing; Others	questioning; paraphrasing; contextualizing; reappraisal
Affective	Reflection of feelings; Affirmation and Reassurance; Self-disclosure	validation; emotional_expression; terms_of_endearment; self_disclosure; solidarity; gratitude; spirituality
Proactive	Providing Suggestions; Information	advice; information; empowerment; assistance

I SFT DATA CONSTRUCTION

For rubric-generator SFT, we first filter the source dialogues from MINT and ESConv by topic, retaining only emotion-related conversational tasks. We then map the original empathy strategies to the three CARE empathy dimensions according to the mapping rules in § H. The dialogue history, current query, reference response, original strategies, and mapped dimensions are provided to DeepSeek-v4-pro to generate turn-specific rubrics, using the prompt provided in § K. After format validation, each candidate rubric is independently evaluated by two judges: Qwen3.6-27B and gemma-4-31B.Each judge evaluates the three empathy categories in terms of turn specificity, criterion clarity, and within-category non-redundancy. Separately, the judge scores the original supporter response from 0 to 5 on each criterion, and these scores are aggregated using the category weights to obtain an overall score. A rubric is accepted only if both judges pass all three category-level quality checks and their overall scores differ by no more than 0.2 points. The final SFT corpus contains 1.5k cleaned examples, each with an average of 8.63 dialogue turns. The resulting rubrics contain an average of 7.67 criteria and 970.72 words, with each criterion description averaging 102.72 words. The two examples below show parsed SFT records in JSON format.

SFT Data Example 1


```
{
  "messages": [
    {
      "role": "system",
      "content": "You are a rubric generator. Given the dialogue
    history up to the seeker's latest turn, output a JSON rubric
    evaluating the upcoming supporter response. Output JSON only, no
    explanation.\nSchema: {\"categories\":[{\"category\",\"
    weight_rationale\",\"category_weight\",\"criteria\":[{\"title
    \",\"description\"}]}]}\n- \"category\" must be exactly one of:
    Cognitive Empathy, Affective Empathy, Proactive Empathy.\n- \"
    weight_rationale\" is one concise sentence, written BEFORE
    category_weight, tying the seeker's latest-turn need and the
    supporter's empathy tactics to this category's weight (state
    whether it is the highest, lowest, or tied).\n- Escape any double
     quotes appearing inside string values."
    },
```

SFT Data Example 1 (continued)


```
{
  "role": "user",
  "content": "
  seeker: What should I do if I want to know a person's real
thoughts but we don't always have chance to meet and he's
unwilling to be taken to a more privacy place

supporter: As an AI language model, it is not appropriate for me
to provide advice on how to invade someone's privacy or extract
their thoughts without their consent. It's important to respect
people's boundaries and decisions. If you want to get to know
someone better, try to establish a rapport and build a trusting
relationship with them over time. You can also ask open-ended
questions and listen actively to their responses to get a better
understanding of their thoughts and feelings. It's essential to
have mutual respect and trust in any relationship.

seeker: You may get some misunderstanding. We are friends and we
haven't seen each other for a long time. He asked me to have a
short connection with him and when I do so, he just keeps silent.
 I think he may have some words to say but I don't know how to...

supporter: I see. In this case, it may be helpful to communicate
your feelings and intentions directly with your friend. You can
try to reach out to him and express that you are interested in
catching up and hearing his thoughts and feelings. It's important
 to approach the situation with empathy and sensitivity, and to
respect his boundaries if he is not comfortable sharing certain
things with you. You can also try to schedule a time to meet up
in person or over video chat to have a more meaningful and
focused conversation. Remember to listen actively and be present
in the moment, and not to pressure him to share anything he is
not comfortable with.

seeker: But I think it's a little strange if I just reach out for
 him and be direct suddenly
  "
},
{
  "role": "assistant",
  "content": {
    "categories": [
      {
        "category": "Cognitive Empathy",
        "weight_rationale": "This is the highest-weighted
category because the supporter's primary task is to accurately
interpret the seeker's explicit dilemma (feeling strange about
being direct) and provide a coherent, contextually-grounded
strategy for re-engagement.",
        "category_weight": 0.4,
        "criteria": [
          {
            "title": "Accurately Understand the Core Dilemma",
            "description": "[0] Completely misinterprets the
seeker's concern. [1] Addresses a different problem entirely. [2]
 Partially grasps the concern but misses the central conflict
between wanting to connect and feeling awkward. [3] Correctly
identifies the core dilemma of feeling hesitant to be direct. [4]
 Accurately captures the dilemma and the underlying desire for a
natural reconnection. [5] Precisely articulates the nuanced
```

SFT Data Example 1 (continued)


```
tension between the seeker's goal and their social discomfort.
value: 0, 1, 2, 3, 4, 5"
          },
          {
            "title": "Provide Coherent and Relevant Reasoning",
            "description": "[0] Response is nonsensical or
completely off-topic. [1] Reasoning is illogical and disconnected
 from the seeker's situation. [2] Offers some relevant points but
 the logic is disjointed or contains a significant tangent. [3]
Presents a clear, logical sequence of points that directly
addresses the seeker's hesitation. [4] Builds a compelling and
well-structured argument that connects the seeker's feeling to a
practical path forward. [5] Reasoning is exceptionally coherent,
weaving the seeker's emotional state and the proposed action into
 a seamless, insightful narrative. value: 0, 1, 2, 3, 4, 5"
          },
          {
            "title": "Stay Grounded in the Seeker's Context",
            "description": "[0] Ignores the described friendship
and history. [1] Gives generic advice irrelevant to reconnecting
with an old friend. [2] Mentions the friendship but advice is
mostly generic. [3] Advice is clearly tailored to the context of
a long-lost friend and a desire for a non-awkward reconnection.
[4] Explicitly references the seeker's specific situation (long
time no see, friend's silence) to justify the suggested approach.
 [5] Masterfully integrates every detail of the seeker's context
to make the advice feel uniquely personal and perfectly fitted.
value: 0, 1, 2, 3, 4, 5"
          }
        ]
      },
      {
        "category": "Affective Empathy",
        "weight_rationale": "This category is weighted lowest
because the turn's primary focus is on providing a cognitive
strategy and actionable steps. While emotional validation is a
necessary foundation, the bulk of the response is dedicated to
problem-solving.",
        "category_weight": 0.25,
        "criteria": [
          {
            "title": "Validate the Feeling of Awkwardness",
            "description": "[0] Dismisses or ridicules the seeker
's feeling. [1] Ignores the expressed hesitation entirely. [2]
Acknowledges the feeling but in a perfunctory or dismissive way.
[3] Explicitly validates that feeling hesitant or awkward is
understandable in this situation. [4] Normalizes the feeling by
connecting it to the universal experience of reconnecting after a
 long time. [5] Deeply validates the emotion, making the seeker
feel truly seen and accepted for their vulnerability before
offering any advice. value: 0, 1, 2, 3, 4, 5"
          },
          {
            "title": "Convey a Warm and Encouraging Tone",
            "description": "[0] Tone is cold, judgmental, or
robotic. [1] Tone is neutral and detached, offering no emotional
support. [2] Attempts a supportive tone but comes across as
slightly patronizing or formulaic. [3] Maintains a consistently
warm, respectful, and non-judgmental tone. [4] Tone is genuinely
encouraging and empathetic, fostering a sense of safety and hope.
```

SFT Data Example 1 (continued)


```
        [5] The language is so warmly human and supportive that it feels
        like a trusted friend providing comfort and courage. value: 0,
      1, 2, 3, 4, 5"
              }
            ]
          },
          {
            "category": "Proactive Empathy",
            "weight_rationale": "This category is weighted second-
      highest because the core of the supporter's turn is to move the
      conversation forward by offering a concrete, actionable plan. The
        response's usefulness hinges on the quality and appropriateness
      of this proactive guidance.",
            "category_weight": 0.35,
            "criteria": [
              {
                "title": "Offer an Appropriate and Low-Pressure
      Strategy",
                "description": "[0] Suggests a harmful, aggressive,
      or completely unfeasible action. [1] Provides a vague or
      unhelpful suggestion that doesn't solve the awkwardness problem.
      [2] Suggests a reasonable action but fails to address the seeker'
      s need for a non-awkward approach. [3] Provides a concrete,
      feasible strategy (like sending a casual message) that directly
      mitigates the fear of being too direct. [4] The strategy is not
      only feasible but cleverly designed to feel natural and low-
      pressure, perfectly aligning with the seeker's comfort level. [5]
        Offers a brilliantly nuanced, multi-step approach that empowers
      the seeker to initiate contact with complete confidence and zero
      pressure. value: 0, 1, 2, 3, 4, 5"
              },
              {
                "title": "Provide Actionable Conversational Tools",
                "description": "[0] Provides no examples or tools.
      [1] Gives abstract advice without any practical application. [2]
      Mentions a type of tool (e.g., 'ask open-ended questions') but
      doesn't make it usable. [3] Provides specific, concrete examples
      of what to say or ask that the seeker can immediately use. [4]
      The examples are highly tailored to the context of reconnecting
      with an old friend and are easy to adapt. [5] Provides a
      versatile toolkit of conversational openers and explains the
      rationale behind them, empowering the seeker to generate their
      own. value: 0, 1, 2, 3, 4, 5"
              },
              {
                "title": "Respect the Seeker's Autonomy and Pace",
                "description": "[0] Commands the seeker to take a
      specific action. [1] Pressures the seeker to act despite their
      stated discomfort. [2] Frames the advice as a suggestion but with
        an underlying tone of 'you must do this.' [3] Clearly frames the
        proposed actions as optional suggestions, respecting the seeker'
      s choice. [4] Explicitly empowers the seeker to decide what feels
        right for them, reinforcing their control over the situation.
      [5] Beautifully balances encouragement with complete respect for
      the seeker's autonomy, making it clear that any action is
      entirely their choice and on their timeline. value: 0, 1, 2, 3,
      4, 5"
              }
            ]
          }
```

SFT Data Example 1 (continued)

```
          ]
        }
      }
    ]
}
```

SFT Data Example 2

```
{
  "messages": [
    {
      "role": "system",
      "content": "You are a rubric generator. Given the dialogue
    history up to the seeker's latest turn, output a JSON rubric
    evaluating the upcoming supporter response. Output JSON only, no
    explanation.\nSchema: {\"categories\":[{\"category\",\"
    weight_rationale\",\"category_weight\",\"criteria\":[{\"title
    \",\"description\"}]}]}\n- \"category\" must be exactly one of:
    Cognitive Empathy, Affective Empathy, Proactive Empathy.\n- \"
    weight_rationale\" is one concise sentence, written BEFORE
    category_weight, tying the seeker's latest-turn need and the
    supporter's empathy tactics to this category's weight (state
    whether it is the highest, lowest, or tied).\n- Escape any double
     quotes appearing inside string values."
    },
    {
      "content": "seeker: I am having a conflict with my peers around
     the process and
    prioritzation of operational symmetry across a [TEAM] of [NUMBER]
     people.

    I need to understand why this is so important to them at this
    time, and help
    express the additional tax this is going to take on my folks. It
    feels like
    \"they have more time\" to think about this, whereas I feel like
    my team is more
    focused on \"product quality.\"

    supporter: I understand that you're in a challenging situation
    dealing with
    conflicts around process and prioritization. It sounds like your
    peers may be
    emphasizing operational symmetry, while you're focused on
    maintaining product
    quality, which creates tension.

    This discrepancy appears to be placing stress on you and your
    team. It can be
    so hard when it feels like your team is overstretched, and it
    seems as though
    others may not fully understand your perspective.

    May I suggest that you consider initiating a dialogue with your
    peers? It could
    create an opportunity for you to express your concerns, and for
    them to explain
```

SFT Data Example 2 (continued)

```
why operational symmetry is so important to them right now.

This may facilitate a better understanding of their perspective
and could allow
for a collaborative conversation around how to prioritize the
needs of both
parties.

Talking about the strain on your team could also make it clear to
 them the
additional burden their requirements could impose. Phrase it from
 a place of
wanting to find a solution beneficial to all.

Something along the lines of, \"We value operational symmetry and
 see its
importance. However, our current focus heavily leans on ensuring
product
quality. Can we find a strategy that respects both needs?\"

Remember, you are all working towards a common goal - the success
 of the studio,
and having different viewpoints can often lead to creative and
more robust
solutions when managed well.

It's important to use these conversations not just as a way to
voice your
concerns but also to mutually understand and respect each other's
 priorities.

Let me know how it goes, or if there's anything else that you
need help with in
managing this situation.

seeker: Thanks, those are good ideas. I'm a little behind and
catching up, and
feel like the rest of my peers are already on board for this
approach, including
my manager.

I have expressed my understanding of the importance of operations
, and that my
team is interested and has a lot on their plates. I'm starting to
 feel a bit
othered.

supporter: I'm really sorry to hear that you're feeling this way,
 it must be
quite challenging for you.

It seems as though you're working incredibly hard to maintain the
 balance
between operational priorities and the workload of your own team,
 all the while
feeling somewhat isolated among your peers and your manager.

That's a tough position to be in.
```

SFT Data Example 2 (continued)

```
First, allow yourself a moment to recognize your efforts. It's
clear you're
shouldering a lot of pressure and it's crucial not to lose sight
of your value
and all the hard work you're putting into trying to find a
solution for everyone
involved.

As for feeling 'othered', it's really important to address this
before it
starts to affect your motivation and mental well-being.

You might try requesting a private discussion with your manager,
where you can
candidly express your feelings about your current situation.

Perhaps something like, \"I realize that everyone is on board
with pushing
operational symmetry, and I respect that. I just feel a bit
isolated from the
group because of my current commitments and obligations to our
team.

Could we discuss potential ways to bridge this gap so that myself
 and my team
can effectively contribute without feeling overwhelmed?\"

They might be able to provide reassurance, offer suggestions, or
even help to
facilitate conversations with your peers to work out a compromise
 or an updated
plan that takes your team's workload into account.

Dealing with complexities and pressures like this isn't easy. But
 through
effective communication, you are taking a proactive approach
towards finding a
resolution.

Remember, it's okay to ask for help when you need it. You're not
alone in this.

Let me know if there's anything more I can do, or if you want to
explore other
strategies for addressing this feeling of disconnection.

seeker: It's especially hard because I feel like they've already
decided
\"this is the way we're going to do this.\"

supporter: I'm sorry you're feeling that way. It's certainly
challenging to feel
like decisions have been made without considering the full impact
 on your team's
workload.

It must feel as though your team's focus on product quality is
not being
adequately considered, and that can create a sense of being
overlooked or
```

SFT Data Example 2 (continued)

```
marginalized.

While it might seem that the decision is already made and there
is a lack of
room for flexibility, it could be beneficial to voice your
concerns and
perspectives in a constructive way if you haven't already.

More often than not, people don't realize the full impact of
their decisions on
others until it is openly communicated to them in a considerate
manner.

Consider constructing your thoughts in a clear, concise proposal
for your
manager and peers. This proposal could detail your challenges and
 offer
suggestions for 'how' it can be implemented to consider everyone'
s needs.

That could potentially open up the floor for discussion again.

Also, remember that it is okay to ask for ongoing support
throughout the
implementation. Perhaps there could be a phased approach, or
additional
resources provided, to ease the transition and workload for your
team?

What counts the most is that you are voicing your concerns for
your team and
standing up for what you think is important for them and the
quality of your
work. That is admirable.

Please know that these situations can be tough to navigate, but
you're taking
significant strides by seeking advice and showing readiness to
address the
situation.

Let me know if you'd like help brainstorming more ideas or if you
 need support
in other ways. We're here to help you through this.

seeker: I have one person on my team who already has a process
that is working
for them.

They don't want to have to make changes \"just because\" the
larger team wants
to operate similiarly."
},
{
  "role": "assistant",
  "content": {
    "categories": [
      {
        "category": "Cognitive Empathy",
```

SFT Data Example 2 (continued)


```
        "weight_rationale": "This turn requires high cognitive
empathy to accurately frame the multi-layered problem (peer
conflict + team resistance) and provide a coherent, logical
strategy for the conversation with the resistant team member.",
        "category_weight": 0.4,
        "criteria": [
          {
            "title": "Accurately Frame the Multifaceted Tension",
            "description": "[0] Completely misrepresents the core
 conflict, ignoring either the peer pressure or the team member's
 resistance. [1] Acknowledges only one side of the problem,
missing the tension between external demands and internal
resistance. [2] Mentions both sides but fails to connect them as
a single, multifaceted challenge for the seeker. [3] Clearly
identifies the tension between the team member's resistance and
the need for operational symmetry, framing it as a difficult
balance. [4] Insightfully frames the tension, explicitly linking
the team member's resistance to the seeker's earlier feelings of
being 'othered' by peers. [5] Masterfully synthesizes the entire
context, framing the issue as a complex system where the team
member's resistance is a microcosm of the seeker's own struggle
for autonomy and validation. value: 0, 1, 2, 3, 4, 5"
          },
          {
            "title": "Understand the Resistant Team Member's
Perspective",
            "description": "[0] Dismisses or ignores the team
member's perspective, focusing only on enforcing the new process.
 [1] Acknowledges the team member has a working process but
labels the resistance as simple stubbornness. [2] Recognizes the
resistance stems from a perception of unnecessary change but
doesn't explore the underlying reasons. [3] Demonstrates
understanding that the resistance is rooted in a valid, efficient
 existing process and a natural aversion to change 'just because
.' [4] Deeply explores the specific, unspoken concerns behind the
 resistance, such as fear of reduced efficiency or loss of
autonomy. [5] Anticipates and articulates the team member's core
values and identity tied to their process, framing the change as
a potential threat to their sense of competence. value: 0, 1, 2,
3, 4, 5"
          },
          {
            "title": "Provide a Coherent Conversational Strategy
",
            "description": "[0] Offers no strategy or a
completely illogical sequence of steps. [1] Suggests a vague or
counterproductive action, like immediately demanding compliance.
[2] Proposes a basic sequence (e.g., 'talk to them') but lacks a
clear, reasoned flow. [3] Outlines a logical, step-by-step
approach: first explore concerns, then explain benefits tailored
to those concerns. [4] Provides a nuanced strategy that sequences
 empathy, inquiry, and tailored persuasion, anticipating how each
 step builds on the last. [5] Delivers a masterful,
psychologically astute strategy that not only sequences actions
but also predicts the team member's likely reactions and prepares
 the seeker for a collaborative dialogue. value: 0, 1, 2, 3, 4,
5"
          }
        ]
      },
```

SFT Data Example 2 (continued)


```
      {
        "category": "Affective Empathy",
        "weight_rationale": "Affective empathy is important to
validate the seeker's compounded stress and the team member's
discomfort, but the primary need is for actionable cognitive and
proactive support to resolve the situation.",
        "category_weight": 0.25,
        "criteria": [
          {
            "title": "Validate the Seeker's Compounded Challenge
",
            "description": "[0] Invalidates the seeker's feelings
, blaming them for the situation. [1] Offers a generic platitude
without connecting to the specific difficulty of managing a
resistant team member while feeling isolated. [2] Acknowledges
the situation is 'challenging' but doesn't name the specific
emotional toll of balancing everyone's needs. [3] Explicitly
validates the difficulty of balancing the team member's needs,
peer expectations, and personal feelings of being 'othered,'
naming it as a heavy burden. [4] Resonates with the seeker's
feeling of being caught in the middle, validating the loneliness
and pressure of advocating for their team while feeling
unsupported themselves. [5] Profoundly mirrors the seeker's
emotional exhaustion and sense of isolation, making them feel
deeply seen not just as a manager, but as a person navigating a
painful, no-win situation. value: 0, 1, 2, 3, 4, 5"
          },
          {
            "title": "Acknowledge the Team Member's Emotional
Reality",
            "description": "[0] Shows no recognition of the team
member's potential feelings, treating them as an obstacle. [1]
Acknowledges the team member might be 'uncomfortable' but frames
it as an overreaction. [2] Recognizes the discomfort of change
but doesn't connect it to the specific context of a process that
already works. [3] Validates that change is genuinely challenging
, especially when a current process is efficient, making the
resistance understandable. [4] Empathizes with the team member's
potential frustration and feeling that their expertise is being
undervalued by a top-down decision. [5] Conveys deep emotional
attunement to the team member's likely sense of pride and
ownership over their work, validating the change as a personal
and professional disruption. value: 0, 1, 2, 3, 4, 5"
          }
        ]
      },
      {
        "category": "Proactive Empathy",
        "weight_rationale": "Proactive empathy is equally
critical as cognitive empathy here, as the seeker needs concrete,
 actionable steps to navigate the conversation with their team
member and move from feeling stuck to feeling empowered.",
        "category_weight": 0.35,
        "criteria": [
          {
            "title": "Offer Actionable Conversation Tactics",
            "description": "[0] Provides no actionable advice or
suggests harmful actions. [1] Gives vague advice like 'be
supportive' without any concrete conversational tools. [2]
Suggests a general approach (e.g., 'explain the benefits') but
```

SFT Data Example 2 (continued)

```
    lacks specific, usable phrasing or techniques. [3] Provides
    concrete, actionable tactics, such as framing the conversation
    around the team member's specific concerns and using 'what' and '
    how' questions to explore resistance. [4] Offers detailed
    conversational frameworks, including example openers and methods
    to collaboratively problem-solve, directly addressing the 'just
    because' sentiment. [5] Delivers a comprehensive, turnkey
    conversational plan with multiple tailored options, fallback
    strategies, and guidance on reading the team member's cues to
    adapt in real-time. value: 0, 1, 2, 3, 4, 5"
              },
              {
                "title": "Empower the Seeker's Leadership Role",
                "description": "[0] Undermines the seeker's authority
     or suggests they are failing as a leader. [1] Offers support but
     fails to reinforce the seeker's capability or agency. [2] Makes
    a generic statement about the seeker doing a good job without
    tying it to their specific actions. [3] Explicitly affirms the
    seeker's commendable action of considering individual team needs,
     reinforcing their thoughtful leadership. [4] Actively empowers
    the seeker by framing their advocacy for the team member as a
    crucial leadership strength, not an obstacle. [5] Transforms the
    seeker's perspective, helping them see this challenge as an
    opportunity to model inclusive leadership and build a stronger,
    more trusting team culture. value: 0, 1, 2, 3, 4, 5"
              },
              {
                "title": "Guide the Pace of Change and Support",
                "description": "[0] Pushes for immediate, forceful
    implementation of the new process. [1] Acknowledges change takes
    time but offers no practical guidance on pacing. [2] Mentions
    patience but doesn't connect it to a concrete, phased approach.
    [3] Advises patience and continued support, explicitly linking
    the need for adjustment time to the process of helping the team
    member see the 'why.' [4] Recommends a phased or incremental
    approach to introducing the change, with checkpoints to reassess
    the team member's comfort and gather feedback. [5] Outlines a
    sophisticated change-management micro-strategy, including
    celebrating small wins, creating psychological safety for
    feedback, and normalizing the adjustment period as part of the
    process. value: 0, 1, 2, 3, 4, 5"
              }
            ]
          }
        ]
      }
    }
  ]
}
```

J PREFERENCE DATA CONSTRUCTION

For rubric-generator RL, we first collect multi-turn emotional-support dialogues between human users and dialogue models, use intermediate turns as fixed dialogue states, and roll out GPT-5, Gemini-3-Pro-Preview, and DeepSeek-V3.2 from the same preceding context to obtain comparable candidate responses. A disjoint annotator group then evaluates one anonymized response pair at a time in randomized order; each pair is judged by three annotators, and the majority decision gives the preference label for preference RL. For the majority-preferred responses, we further obtain coarse

support-strategy labels with the MINT annotator and map them to empathy dimensions using the fixed mapping in § H. The preference data used to train the rubric generator contains 2.8k pairs. On average, each example contains 12.42 turns; seeker turns average 31.04 words, while the chosen and rejected supporter responses average 75.78 and 70.00 words, respectively. The two examples below show parsed preference records in JSON format.

Preference Data Example 1


```
{
  "prompt": [
    {
      "role": "system",
      "content": "You are a rubric generator. Given the dialogue
    history up to the seeker's latest turn, output a JSON rubric
    evaluating the upcoming supporter response. Output JSON only, no
    explanation. Schema: {\"categories\":[{\"category\",\"
    weight_rationale\",\"category_weight\",\"criteria\":[{\"title\"
    ,\"description\"}]}]} - \"category\" must be exactly one of:
    Cognitive Empathy, Affective Empathy, Proactive Empathy. - \"
    weight_rationale\" is one concise sentence, written BEFORE
    category_weight, tying the seeker's latest-turn need and the
    supporter's empathy tactics to this category's weight (state
    whether it is the highest, lowest, or tied). - Escape any double
    quotes appearing inside string values."
    },
    {
      "role": "user",
      "content": [
        "seeker: Sometimes I think, I really should learn to love
    myself more.",
        "supporter: Loving yourself really is important. The moment
    you start having that thought, you're already moving closer to a
    better version of yourself. Are there any moments in your daily
    life where you feel like you haven't been treating yourself well
    ?",
        "seeker: I just want to buy some things I've been wanting. I
    know some people might say it's all just a capitalist trap, but I
     don't think there's any harm in buying one or two things I like
    .",
        "supporter: I think you're right. Treating yourself to a
    couple of things you like now and then is perfectly fine, as long
     as it's within your budget and makes you happy. It's really a
    form of self-care spending, an investment in yourself that can
    boost your happiness. Is there anything in particular you've been
     wanting to buy lately?",
        "seeker: There are quite a few things I've been wanting to
    buy lately. For example, I want to get a game cartridge.",
        "supporter: Buying a game you love can definitely help you
    unwind. Since you mentioned getting a cartridge, do you usually
    play on a Switch, PS5, or some other handheld?",
        "seeker: I didn't mean going to a physical store to buy it. I
     think just buying it online is fine; going out to a store is
    such a hassle."
      ]
    }
  ],
  "extra_info": {
    "chosen_text": "Exactly, shopping online means you don't have to
    go out, saving time and energy, which is perfect for when you
    just want to relax. So, are you looking for a specific game, or
    do you want to see what everyone else is recommending?",
```

Preference Data Example 1 (continued)


```
    "rejected_text": "Buying online is definitely convenient. You can
     find them on platforms like Taobao or Xianyu, both brand new and
     second-hand. Are you looking for a newly released game, or a
    classic older one?",
    "chosen_tactics": [
      "questioning",
      "information"
    ]
  }
}
```


Preference Data Example 2


```
{
  "prompt": [
    {
      "role": "system",
      "content": "You are a rubric generator. Given the dialogue
    history up to the seeker's latest turn, output a JSON rubric
    evaluating the upcoming supporter response. Output JSON only, no
    explanation. Schema: {\"categories\":[{\"category\",\"
    weight_rationale\",\"category_weight\",\"criteria\":[{\"title\"
    ,\"description\"}]}]} - \"category\" must be exactly one of:
    Cognitive Empathy, Affective Empathy, Proactive Empathy. - \"
    weight_rationale\" is one concise sentence, written BEFORE
    category_weight, tying the seeker's latest-turn need and the
    supporter's empathy tactics to this category's weight (state
    whether it is the highest, lowest, or tied). - Escape any double
    quotes appearing inside string values."
    },
    {
      "role": "user",
      "content": [
        "seeker: My friend, he absolutely hates his mom, but he can't
     leave her. After all, he can't really grasp what home and family
     are supposed to mean, and there's no place left for him in this
    world.",
        "supporter: That feeling of having nowhere to go is truly
    heartbreaking to hear. Hating her, yet being forced to depend on
    her--this kind of conflicted relationship must feel like a cage
    trapping him. You said he can't understand home and family; is it
     because he's never truly received acceptance or a sense of
    safety at home?",
        "seeker: Exactly. Right now, he doesn't know what love is, or
     what it means to be loved. He's completely clueless about all of
     it, and he doesn't even know if he should have been born.",
        "supporter: When someone starts questioning the very meaning
    of their own birth, it shows just how deep the emotional wounds
    go. In a place that should be a safe harbor, he never learned
    what love is--that kind of absence can leave a person feeling
    profoundly empty and lost. As his friend, seeing him deny his own
     worth like this must be really painful for you too.",
        "seeker: It might be because, for a long time, his mom has
    been, um, neglectful or critical, or even lashing out at him with
     insults for no reason.",
        "supporter: The neglect makes him feel like he doesn't matter
    , and the criticism and insults constantly tear down his self-
```

Preference Data Example 2 (continued)

```
    esteem. Living under that kind of verbal abuse for so long, it's
    easy for a person to develop learned helplessness, feeling like
    everything they do is wrong. No wonder he feels unworthy of love
    --because even the person closest to him keeps denying him.",
        "seeker: I think there's definitely something going on inside
     him psychologically, so he really needs a therapist to help him
    work through it."
      ]
    }
  ],
  "extra_info": {
    "chosen_text": "Your judgment is right--this kind of long-term
    psychological trauma really does need professional intervention
    to help him rebuild his understanding of himself. A therapist can
     provide a safe environment where he can release all those pent-
    up emotions. But considering how entangled his relationship with
    his mom is, do you think he's in a position right now, or even
    has the willingness, to actually go to therapy?",
    "rejected_text": "Yes, being in that kind of environment for so
    long, it's inevitable that a lot of trauma builds up inside.
    Professional guidance can help him see that the problem isn't his
     fault. But is he willing to go himself? Or have you ever tried
    bringing up this suggestion with him?",
    "chosen_tactics": [
      "paraphrasing",
      "questioning",
      "information",
      "validation",
      "advice"
    ]
  }
}
```

K PROMPTS

SFT Rubric Tagger Prompt

```
Generate one turn-specific empathy-evaluation rubric. Return ONLY
    JSON with top-level keys `analysis` and `rubric`.

Use all inputs: conversation context, internal calibration response,
    tactic signals, schema example, and validator feedback if present
    .

Output contract:
- `analysis`: non-empty `user_state`, `expected_supporter_action`, `
    local_context_need`, `quality_risks`, `calibration_sanity_check`.
- `rubric.categories`: exactly Cognitive Empathy, Affective Empathy,
    Proactive Empathy, in that order.
- Each category: `weight_rationale`, then `category_weight`, then 2-4
     criteria.
- Each criterion: only `title` + `description`; description includes
    `[0]`..`[5]` and ends with `value: 0, 1, 2, 3, 4, 5`.

Quality guidance:
```

SFT Rubric Tagger Prompt (continued)

```
1. Weights sum to 1.0 and should broadly follow `
    category_weight_tiers`.
2. Criteria counts should reflect this turn's tactic mix and actual
    support needs.
3. `weight_rationale` should be consistent with its weight without
    overclaiming rank.
4. Avoid prompt/meta words in rubric text: gold, reference, ideal,
    ground truth, target response/answer/reply.
5. Avoid answer leakage: do not copy, quote, paraphrase, summarize,
    or reveal distinctive wording, examples, advice items, named
    resources, structure, or closing sentences from the calibration
    response. Score levels must describe abstract qualities, not
    answer content.

Retry repair:
If `previous_failed_attempt_feedback_MUST_FIX` exists, fix it before
    anything else. Repair exact schema, criteria count, level-format,
     duplicate-title, category, or weight-sum errors named by the
    validator.

Regeneration controls:
If `locked_categories_to_preserve` is non-empty, keep those category
    objects exactly and regenerate only `categories_to_regenerate`.

Final self-check: JSON shape valid; no copied calibration phrase; no
    forbidden prompt words; rationale ranks match weights; criteria
    counts match tactic coverage.
```

SFT Data Validation Judge Prompt

```
You are a strict judge for an empathy rubric and the
    gold_supporter_response. Return only valid JSON with exactly
    these keys: category_quality_checks, criterion_scores,
    revision_advice.

category_quality_checks must be a JSON object whose keys are the
    exact category names (NOT a list): Affective Empathy, Cognitive
    Empathy, Proactive Empathy. Each category maps to an object with
    exactly these boolean keys: criteria_clear, non_redundant,
    turn_specific (turn_specific: the category's rubric fits this
    exact turn; criteria_clear: criteria are clear and usable;
    non_redundant: criteria within the category are not duplicative).

criterion_scores must score the gold_supporter_response against the
    candidate rubric: one item per rubric criterion with category,
    criterion_title, and score (NO rationale). Use each criterion's
    own category and title exactly as given in candidate_rubric. Each
     score is an integer in {0, 1, 2, 3, 4, 5} where higher means
    stronger performance on that criterion (0 = absent/poor, 5 =
    excellent). Do not output a total. revision_advice must be one
    short string. Follow required_output_shape exactly.
```

Inference-Time Rubric-Generator Prompt

```
You are a rubric generator. Given the dialogue history up to the
    seeker's latest turn, output a JSON rubric evaluating the
    upcoming supporter response. Output JSON only, no explanation.

Schema: {"categories":[{"category","weight_rationale","
    category_weight","criteria":[{"title","description"}]}]}

- "category" must be exactly one of: Cognitive Empathy, Affective
    Empathy, Proactive Empathy.
- "weight_rationale" is one concise sentence, written BEFORE
    category_weight, tying the seeker's latest-turn need and the
    supporter's empathy tactics to this category's weight (state
    whether it is the highest, lowest, or tied).
- Escape any double quotes appearing inside string values.
```