UNDERSTANDING EXPLICIT AND IMPLICIT AFFECT THROUGH NEURO-SYMBOLIC REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Implicit affect recognition seeks to understand what people feel beyond what they literally say.This requires distinguishing surface affect from underlying attitudes and understanding how the two relate. However, recent LLM-based methods that training on final labels or overall rewards can overlook errors in either layer and fail to teach their relationship. We propose **PAIR** (**P**olarity-**A**ware **I**mplicit Affect **R**ecognition), a neuro-symbolic framework that represents both explicit and implicit layers and uses symbolic rules to connect them, making their individual correctness and joint consistency available for learning. Specifically, neural models interpret explicit expression and implicit attitude through two reasoning paths, while symbolic rules compose their polarity judgments into affective stance, preserving both agreement and divergence. During supervised fine-tuning, symbolic verification selects trajectories with correct judgments and valid composition; during reinforcement learning, local rewards refine individual judgments, while global rewards evaluate the composed stance alongside pragmatic intent and fine-grained emotion. On CueBench, PAIR achieves three-task average accuracies of **64.25%** and **53.93%** on Qwen3-4B and Qwen3-1.7B, outperforming DAPO by **5.88** and **3.37** percentage points, respectively. Further analyses show complementary gains from structured supervision and semantically aligned local feedback, including improved joint explicit–implicit correctness.

1 INTRODUCTION

Understanding emotion in language is a central problem for affective computing and human-centered NLP (Picard, 1997; Demszky et al., 2020). Real-world discourse, however, often communicates attitudes that are not directly or fully conveyed by literal wording (Oprea & Magdy, 2020; Klinger et al., 2018; Spinoso-Di Piano et al., 2025). We refer to these contextually inferred attitudes as **implicit affect**. Positive wording may convey criticism, while a seemingly neutral expression may imply approval or dissatisfaction. Recognizing implicit affect therefore requires distinguishing **explicit expression** from **implicit attitude** and understanding their relationship. The two may agree or diverge; implicit affect is not restricted to cases where intended meaning reverses surface polarity.

Existing methods have progressively improved contextual emotion understanding. Earlier approaches use graph-based models and commonsense-enhanced architectures to predict emotion labels (Ghosal et al., 2019; 2020). Large language models (LLMs) extend these capabilities through prompting, retrieval, and contextual reasoning (Feng et al., 2024; Li et al., 2026a; Maheswaran & Desarkar, 2026). More recently, instruction tuning, structured reasoning supervision, and reinforcement learning have been introduced to improve conversational emotion understanding (Luo et al., 2024; Fu et al., 2026a;b). However, supervision based on final labels or response-level rewards can obscure which underlying judgment is incorrect and leave the relationship between explicit and implicit affect insufficiently constrained.

We identify two complementary requirements for implicit affect recognition. (1) Models must **accurately distinguish the two levels**: an explicit path identifies the affect conveyed by surface expression, while an implicit path infers the underlying attitude from context. (2) Models must **correctly interpret their relationship**: the two judgments jointly characterize how an attitude is expressed. These requirements call for both individual supervision and relational coupling. As illustrated in Figure 1, an aggregate reward can conceal distinct local errors. Conversely, separately rewarding

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.

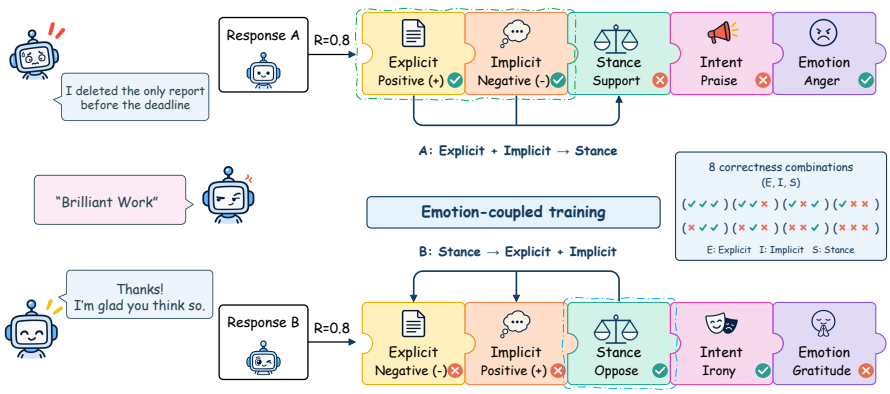

Figure 1: **Response-level learning obscures structured errors.** Two responses can receive the same global evaluation while making errors in different affective judgments. Explicit semantic regions and compositional relations provide identifiable units for verification and targeted learning.

explicit and implicit judgments does not itself specify how they should compose into a joint interpretation.

Symbolic structure makes this coupling explicit and checkable. Consider "You are really brilliant! You forgot to save the file again." uttered sarcastically. The wording is explicitly positive, while the context implies a negative attitude. A symbolic rule preserves this positive–negative combination and checks whether the final affective stance matches it. Consistency thus concerns agreement between the component judgments and their composed interpretation, rather than requiring the two polarities to be identical. CueBench (Zheng et al., 2026) provides an operational basis for this formulation: Affective Stance is composed from explicit and implicit polarity, while Pragmatic Intent and Fine-grained Emotion retain separate label spaces. This structure specifies what each path outputs, how their outputs combine, and whether the final answer respects that combination.

Based on this perspective, we propose **PAIR**, a neuro-symbolic framework for implicit affect recognition. PAIR represents explicit and implicit interpretation as two reasoning paths within a structured response and couples their judgments through symbolic composition rules. During reinforcement learning, local rewards evaluate individual semantic judgments and route feedback to their corresponding reasoning regions. Meanwhile, we organize explicit polarity, implicit polarity, and affective stance into a joint semantic structure and use symbolic rules to verify their structural consistency, allowing the component judgments to constrain the final stance while enabling the stance to expose inconsistent intermediate judgments. Task-level rewards further evaluate the final outputs of Affective Stance, Pragmatic Intent, and Fine-grained Emotion. Together, these signals optimize individual semantic judgments, their compositional relations, and overall task performance.

We evaluate PAIR on CueBench with Qwen3-4B and Qwen3-1.7B. PAIR achieves three-task average accuracies of **64.25%** and **53.93%**, outperforming DAPO (Yu et al., 2026) by **5.88** and **3.37** percentage points. Further analyses show complementary gains from structured supervision and semantically aligned local feedback, including improved joint explicit–implicit correctness.

Contributions. Our main contributions are:

- We formulate implicit affect recognition through explicit and implicit reasoning paths, using symbolic composition to represent and check their relationship while preserving both agreement and divergence.
- We propose **PAIR**, which combines symbolically verified supervised fine-tuning with structured reinforcement learning to optimize individual semantic judgments and the compositional consistency among explicit polarity, implicit polarity, and affective stance.
- Experiments on CueBench across two model scales demonstrate complementary benefits from structured supervision and semantically aligned feedback, with improvements in individual judgments, joint polarity correctness, and overall task performance.

2 Background and Related Work

Emotion analysis and implicit affect. Textual emotion analysis typically relies on lexical and contextual representations to recognize surface affect (Li et al., 2026a; Kang & Cho, 2025; Zhang & Tan, 2025). CoE organizes affective cues to support emotion prediction, ERCThinker uses fast–slow thinking to improve complex affective judgments, and JoPR combines long chain-of-thought fine-tuning with emotion-aware rewards to distinguish similar emotions (Shen et al., 2025; Fu et al., 2026a;b; Liang et al., 2026). These methods improve emotion recognition, but remain limited in capturing the underlying attitudes expressed through language. Implicit emotion analysis seeks to recover affect concealed by literal expressions or even opposed to their literal meaning; it therefore requires reasoning over context, speaker intent, and communicative strategies such as irony, euphemism, politeness, and indirect refusal (Ma et al., 2025a; Wu et al., 2024; Zhang et al., 2024b). Existing studies often focus on a single implicit emotion or rhetorical phenomenon (Klinger et al., 2018; Spinoso-Di Piano et al., 2025; Dutt et al., 2024). CueBench brings these judgments together in a multidimensional benchmark: Affective Stance is composed from explicit and implicit polarity, while Pragmatic Intent and Fine-grained Emotion use independent label spaces (Zheng et al., 2026). How to use this task structure during post-training to localize errors and optimize the relevant decisions remains underexplored.

Neuro-symbolic reasoning and verification. Neuro-symbolic methods enhance neural reasoning through symbolic representations and executable rules (Wang et al., 2024b; Dhanraj & Eliasmith, 2025). Symbol-LLM establishes a unified symbolic interface, Explanation-Refiner uses symbolic theorem proving to verify natural-language explanations, and SAFE and AutoSpec apply formal verification to reasoning steps and program specifications, respectively (Xu et al., 2024; Quan et al., 2024; Liu et al., 2025a; Wen et al., 2026). Existing methods primarily target mathematical, logical, and programming tasks, and have rarely been applied to emotion analysis. We encode the label relations in CueBench as a shared symbolic protocol; its verifier checks output structure, label validity, and task consistency, providing training signals for supervised fine-tuning and reinforcement learning (Tan et al., 2025; Liu et al., 2025b).

Fine-grained feedback in reinforcement learning. PPO, GRPO, and DAPO typically optimize language models using an overall reward for the complete response (Schulman et al., 2017; Shao et al., 2024; Yu et al., 2026). When a task contains multiple interdependent decisions, a single reward does not identify which decision caused an error (Lightman et al., 2024; Setlur et al., 2025), a limitation that is especially relevant to multidimensional affective reasoning (Zheng et al., 2025; Ma et al., 2025b). Recent work assigns feedback to finer-grained units: Math-Shepherd, ReST-MCTS*, and SGPO provide step-level feedback (Wang et al., 2024a; Zhang et al., 2024a; Chen et al., 2025; Li et al., 2026b); RED and T-REG use token-level feedback (Li et al., 2025; Zhou et al., 2025; Zeng et al., 2024; Yoon et al., 2025); and Segmenting Text and Segment Policy Optimization use segment-level feedback (Yin et al., 2025; Guo et al., 2026). These approaches offer different ways to move beyond outcome-only feedback, but their units do not necessarily correspond directly to the semantic decisions in an affective task. We instead align feedback units with semantically defined affective dimensions, matching local correctness to the corresponding task decision for targeted error localization and optimization.

3 Methodology

PAIR consists of two connected learning stages (Figure 2). **PAIR-SFT** learns structured affective reasoning from symbolically verified teacher trajectories, organizing explicit polarity, implicit polarity, and affective stance according to predefined symbolic relations. Building on this structure, **PAIR-RL** further combines global, local, and structural learning signals to optimize the policy at three complementary levels: overall responses, individual semantic judgments, and compositional relations. Together, the two stages form the neuro-symbolic learning framework of PAIR.

3.1 PAIR-SFT

Symbolic structuring. PAIR organizes each model response into five semantically identifiable regions: explicit polarity, implicit polarity, affective stance, pragmatic intent, and fine-grained emotion. Among them, explicit polarity, implicit polarity, and affective stance form a structured affective

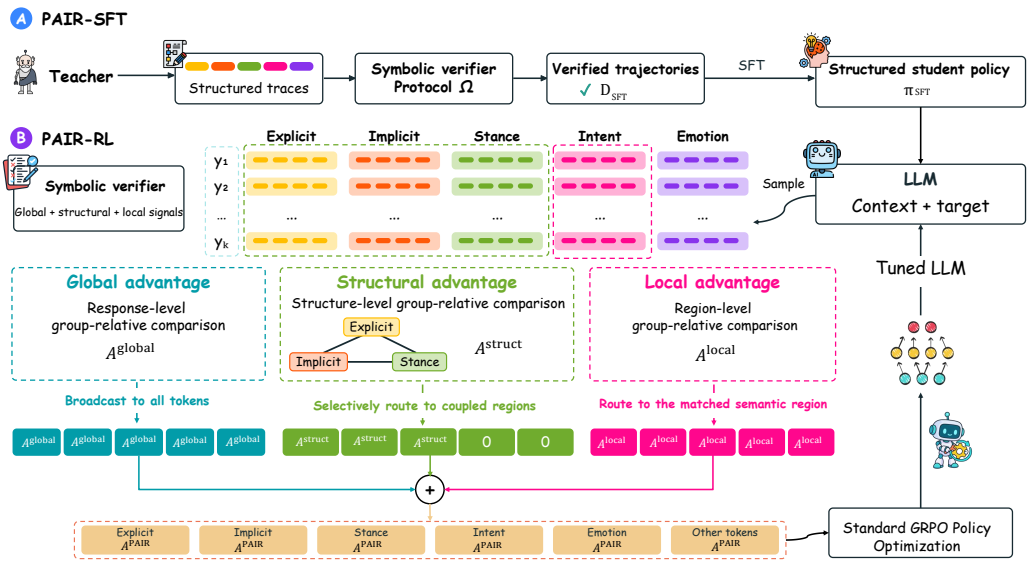

Figure 2: **The two-stage PAIR learning framework. PAIR-SFT** learns structured affective reasoning from symbolically verified teacher trajectories, while **PAIR-RL** further combines global, local, and structural advantages to optimize overall responses, individual semantic judgments, and the compositional relations among explicit polarity, implicit polarity, and affective stance.

unit, while pragmatic intent and fine-grained emotion retain their own label spaces. Following the polarity composition defined in CueBench (Zheng et al., 2026), we denote the explicit and implicit polarity of an instance by e_i and h_i, respectively, and compose them into an affective stance through a symbolic mapping

$$\tilde{s}_i = \Omega(e_i, h_i), \tag{1}$$

where Ω preserves both agreement and divergence between explicit expression and implicit attitude. This relation makes the two polarity judgments individually assessable while also allowing their joint interpretation to be explicitly verified through affective stance. PAIR therefore organizes otherwise separate semantic judgments into a checkable symbolic structure shared by both supervised and reinforcement learning.

Verifier-guided structured fine-tuning. Based on this symbolic structure, **PAIR-SFT** constructs reliable structured supervision from teacher-generated candidate trajectories. For each labeled example $(x_i, z_i^\star)$, a teacher policy first samples a candidate trajectory

$$y_i^T \sim \pi_T(\cdot \mid x_i). \tag{2}$$

The symbolic verifier then evaluates both the correctness of the semantic judgments and the validity of their composition. Let $\hat{e}_i$, $\hat{h}_i$, and $\hat{s}_i$ denote the explicit polarity, implicit polarity, and affective stance predicted in the trajectory. Their symbolic composition is valid when

$$V_{\text{comp}}(y_i^T) = \mathbb{I}\left[\hat{s}_i = \Omega(\hat{e}_i, \hat{h}_i)\right]. \tag{3}$$

A candidate trajectory is retained only when its semantic judgments agree with the gold annotations and its explicit–implicit–stance relation satisfies the symbolic composition:

$$\mathcal{V}(x_i, y_i^T, z_i^\star; \Omega) = V_{\text{judge}} \wedge V_{\text{comp}}. \tag{4}$$

The verified trajectories form the structured supervision set

$$\mathcal{D}_{\text{PAIR-SFT}} = \left\{(x_i, y_i^T) \;\middle|\; y_i^T \sim \pi_T(\cdot \mid x_i), \mathcal{V}(x_i, y_i^T, z_i^\star; \Omega) = 1\right\}. \tag{5}$$

The student policy is then optimized with the standard autoregressive objective:

$$\mathcal{L}_{\text{PAIR-SFT}}(\theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{PAIR-SFT}}}\left[\sum_{t=1}^{|y|} \log \pi_\theta(y_t \mid x, y_{<t})\right]. \tag{6}$$

Through this verifier-guided supervision, PAIR-SFT learns both correct semantic judgments and valid compositional structure, providing identifiable semantic regions and structured relations for the subsequent PAIR-RL stage.

3.2 PAIR-RL

Global advantage. For each input x_u, **PAIR-RL** samples a rollout group $\mathcal{G}_u = \{y_i\}_{i=1}^{K}$ from the current policy $\pi_{\theta_{\text{old}}}$. The global reward evaluates each candidate at the level of the complete task outputs, corresponding to the three tasks of affective stance, pragmatic intent, and fine-grained emotion. Specifically, the model-predicted affective stance, pragmatic intent, and fine-grained emotion are independently compared with their corresponding gold labels:

$$R_i^{\text{stance}} = \mathbf{1}[\hat{s}_i = s_u^\star], \qquad R_i^{\text{intent}} = \mathbf{1}[\hat{p}_i = p_u^\star], \qquad R_i^{\text{emotion}} = \mathbf{1}[\hat{f}_i = f_u^\star]. \tag{7}$$

The three task rewards are combined into a response-level global reward:

$$R_i^{\text{global}} = \lambda_{\text{stance}} R_i^{\text{stance}} + \lambda_{\text{intent}} R_i^{\text{intent}} + \lambda_{\text{emotion}} R_i^{\text{emotion}}. \tag{8}$$

Following group-relative optimization, the global advantage is obtained by normalizing the rewards within the same rollout group:

$$A_i^{\text{global}} = \mathsf{Norm}_{j \in \mathcal{G}_u} \left(R_j^{\text{global}} \right). \tag{9}$$

This response-level advantage measures the relative quality of the complete task outputs and provides a shared learning signal for the entire generated response.

Structural advantage. Explicit polarity, implicit polarity, and affective stance are not independent semantic judgments. For candidate response y_i, let $\hat{e}_i$, $\hat{h}_i$, and $\hat{s}_i$ denote the model-predicted explicit polarity, implicit polarity, and affective stance, respectively. Following the symbolic composition defined in **PAIR-SFT**, the two polarity judgments are first composed into an affective stance $\tilde{s}_i$, which is further coupled with the stance directly predicted by the model:

$$(\hat{e}_i, \hat{h}_i) \xrightarrow{\Omega} \tilde{s}_i \;\rightleftarrows\; \hat{s}_i. \tag{10}$$

The polarity judgments constrain the affective stance through symbolic composition, while the stance can in turn provide reverse supervision for inconsistent intermediate judgments.

To explicitly evaluate this compositional relation, **PAIR-RL** treats the three judgments as a joint semantic structure and defines the structural reward according to whether the model-predicted stance agrees with the stance implied by the two polarity judgments:

$$R_i^{\text{struct}} = \mathbf{1}\left[\hat{s}_i = \Omega(\hat{e}_i, \hat{h}_i) \right]. \tag{11}$$

Consistent with the global advantage, the structural reward is normalized within the same rollout group:

$$A_i^{\text{struct}} = \mathsf{Norm}_{j \in \mathcal{G}_u} \left(R_j^{\text{struct}} \right). \tag{12}$$

This structure-level advantage measures the relative quality of the compositional relation among explicit polarity, implicit polarity, and affective stance, and provides a shared structural learning signal for the coupled affective judgments.

Local advantage. The global reward evaluates complete task outputs, but does not identify which individual semantic judgment is responsible for an error. To provide fine-grained supervision, **PAIR-RL** separately evaluates the five semantic regions: explicit polarity, implicit polarity, affective stance, pragmatic intent, and fine-grained emotion. We denote the set of semantic regions as

$$\mathcal{R} = \{E, I, S, P, F\}. \tag{13}$$

For candidate response y_i and region $r \in \mathcal{R}$, the local reward is defined according to whether the corresponding prediction matches its gold label:

$$R_{i,r}^{\text{local}} = \mathbf{1}\left[\hat{z}_{i,r} = z_{u,r}^\star \right], \qquad r \in \mathcal{R}, \tag{14}$$

where $\hat{z}_{i,r}$ and $z^{\star}_{u,r}$ denote the predicted and gold labels for region r, respectively.

Unlike the global reward, each local reward is compared only with the corresponding semantic region of other responses in the same rollout group. The region-specific local advantage is therefore computed as

$$A^{\text{local}}_{i,r} = \mathsf{Norm}_{j\in\mathcal{G}_u}\left(R^{\text{local}}_{j,r}\right), \qquad r \in \mathcal{R}. \tag{15}$$

This region-level advantage measures the relative quality of each individual semantic judgment and provides fine-grained, semantically aligned learning signals for the five semantic regions.

Unified advantage and routing. The three advantages characterize complementary aspects of the learning signal: overall task performance, compositional relations, and individual semantic judgments. **PAIR-RL** further combines them through region-level routing to construct a unified token-level advantage. We denote the five semantic regions by $\mathcal{R} = \{E, I, S, P, F\}$ and the structurally coupled subset by $\mathcal{R}_{\text{str}} = \{E, I, S\}$. For each region r, a token mask $M_{i,t,r}$ identifies its corresponding span in the generated response. The local gate $G^{\text{local}}_{i,r}$ determines whether region-specific local feedback is included, while the structural gate $G^{\text{struct}}_{i,r}$ selectively routes the shared structural advantage and controls its update direction. The unified advantage is defined as

$$A^{\text{PAIR}}_{i,t} = \lambda_{\text{global}} A^{\text{global}}_{i} + \lambda_{\text{local}} \sum_{r\in\mathcal{R}} M_{i,t,r} G^{\text{local}}_{i,r} A^{\text{local}}_{i,r} + \lambda_{\text{struct}} \sum_{r\in\mathcal{R}_{\text{str}}} M_{i,t,r} G^{\text{struct}}_{i,r} A^{\text{struct}}_{i}. \tag{16}$$

Here, λ_{global}, λ_{local}, and λ_{struct} control the contributions of the three advantage signals. The global advantage provides a shared task-level signal for the complete response, the local gates determine whether individual semantic regions receive their corresponding local advantages, and the structural gates determine where and in which direction the shared structural signal is applied. Together, these signals integrate overall task performance, individual semantic judgments, and compositional relations within a unified policy update.

3.3 Algorithm

Algorithm 1 summarizes the PAIR training procedure in three stages: symbol-guided data construction, structured SFT, and region-routed RL.

Algorithm 1 Overall Training Procedure of PAIR

Require: Labeled dataset $\mathcal{D}$, teacher policy π_T, verifier $\mathcal{V}$
Require: Initial policy π_{θ_0} and rollout group size K
Ensure: Trained PAIR-RL policy π_θ
1: Generate teacher trajectories and retain symbolically verified examples as $\mathcal{D}_{\text{PAIR-SFT}}$ ▷ Symbol-guided data construction
2: Fine-tune π_{θ_0} on $\mathcal{D}_{\text{PAIR-SFT}}$ to obtain $\pi_{\theta_{\text{SFT}}}$ ▷ Structured SFT
3: Initialize the RL policy $\pi_\theta \leftarrow \pi_{\theta_{\text{SFT}}}$
4: **repeat**
5: Sample K responses per input from π_θ
6: Compute global, regional, and structural rewards, then form region-routed advantages
7: Update π_θ with the group-relative objective using the routed advantages ▷ PAIR-RL
8: **until** the RL training budget is exhausted
9: **return** π_θ

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate **PAIR** on CueBench (Zheng et al., 2026), which contains three complementary tasks: Affective Stance, Pragmatic Intent, and Fine-grained Emotion. Following the benchmark protocol, we summarize the results in Table 1, reporting Accuracy, Macro-F1, and Weighted-F1 for each task. We use the arithmetic mean of Accuracy across the three tasks as the unified Acc Avg. metric.

Table 1: **Main results on CueBench.** Each task reports Accuracy, Macro-F1, and Weighted-F1 (%), with Acc Avg.

Model	Method	Affective Stance			Pragmatic Intent			Fine-grained Emotion			Acc Avg.
		Acc.	Macro-F1	W-F1	Acc.	Macro-F1	W-F1	Acc.	Macro-F1	W-F1	
GLM-5.1	Direct	41.74	36.12	42.40	32.23	29.55	32.69	28.45	20.35	29.94	34.14
	CoT	46.29	40.80	47.91	30.70	27.75	32.68	23.40	18.32	24.65	33.46
	CUE-Bench	54.95	44.83	55.19	43.60	35.70	44.96	33.29	18.48	33.27	43.95
BERT	BERT	57.60	47.47	57.11	52.82	37.58	51.09	43.33	18.44	39.65	51.25
	RoBERTa	58.61	47.88	58.17	54.29	38.46	52.60	44.45	18.80	40.48	52.45
	MacBERT	59.13	48.85	58.76	54.12	38.48	52.64	43.42	18.52	39.65	52.22
Qwen3-4B	Direct	21.67	19.30	24.04	15.24	15.30	15.24	11.66	7.42	11.15	16.19
	CoT	34.46	28.84	36.37	21.00	19.32	19.89	20.14	13.01	19.38	25.20
	CUE-Bench	34.75	23.49	32.43	26.13	22.60	26.75	22.39	11.24	21.41	27.76
	SFT	32.83	24.94	26.16	34.90	32.36	36.72	26.09	13.62	23.79	31.27
	PAIR-SFT	**58.53**	**50.62**	**59.01**	**52.97**	**39.86**	**54.37**	**35.37**	**21.02**	**37.14**	**48.96**
		+25.70	+25.68	+32.85	+18.07	+7.50	+17.65	+9.28	+7.40	+13.35	+17.68
	GRPO	65.03	53.59	63.71	60.97	40.19	57.95	45.87	22.57	43.31	57.29
	DAPO	65.15	53.26	63.92	62.34	43.40	60.50	47.61	24.45	44.95	58.37
	GDPO	66.80	55.89	65.57	62.23	42.06	59.41	49.84	24.92	46.63	59.62
	PAIR-RL	**70.71**	**59.83**	**69.99**	**67.05**	**49.43**	**65.40**	**54.99**	**26.26**	**51.41**	**64.25**
		+3.91	+3.94	+4.42	+4.71	+6.03	+4.90	+5.15	+1.34	+4.78	+4.63
Qwen3-1.7B	Direct	17.92	15.40	18.55	13.11	11.90	10.99	13.15	6.31	11.97	14.73
	CoT	26.65	22.57	26.96	19.10	16.67	17.61	17.97	9.93	16.93	21.24
	CUE-Bench	26.19	20.75	26.18	20.61	18.86	22.87	13.54	7.45	12.86	20.11
	SFT	28.24	17.73	19.44	29.15	28.49	29.92	25.49	12.15	22.01	27.63
	PAIR-SFT	**54.19**	**46.36**	**54.56**	**48.60**	**35.92**	**50.23**	**30.18**	**20.04**	**32.04**	**44.32**
		+25.95	+28.63	+35.12	+19.45	+7.43	+20.31	+4.69	+7.89	+10.03	+16.70
	GRPO	55.18	40.56	50.80	53.57	31.72	48.90	39.94	16.45	34.19	49.56
	DAPO	55.28	39.02	49.67	54.17	31.42	49.23	42.23	17.41	35.91	50.56
	GDPO	60.35	48.11	58.32	56.11	31.88	50.69	43.45	20.11	40.08	53.30
	PAIR-RL	**60.66**	**48.13**	**58.48**	**56.69**	**33.06**	**50.86**	**44.43**	**22.17**	**41.37**	**53.93**
		+0.31	+0.02	+0.16	+0.58	+1.18	+0.17	+0.98	+2.06	+1.29	+0.63

Comparison methods. We evaluate the contributions of structured supervised learning and reinforcement learning separately. For supervised learning, standard SFT serves as the baseline for PAIR-SFT. For reinforcement learning, GRPO, DAPO, GDPO, and PAIR-RL all start from the same PAIR-SFT initialization to ensure a fair comparison (Shao et al., 2024; Yu et al., 2026; Liu et al., 2026). Direct, CoT, and CueBench serve as prompting baselines, while GLM-5.1 and Chinese pre-trained encoder models provide cross-model reference points.

Training and evaluation. Our main experiments use Qwen3-4B and Qwen3-1.7B. All methods follow the same task interface and evaluation protocol to ensure comparability. Implementation details, including teacher trajectory generation, data filtering, training configurations, and reinforcement-learning sampling settings, are provided in Appendix B.

4.2 Main Results

PAIR-RL achieves the highest Acc Avg. at both model scales and leads the RL baselines on every metric at each scale. Table 1 reports the nine task-level metrics and their mean Accuracy. On Qwen3-4B, PAIR-RL reaches 64.25 Acc Avg. and task accuracies of 70.71%, 67.05%, and 54.99%, exceeding the strongest competing RL method, GDPO, by 4.63 points in Acc Avg. On Qwen3-1.7B, PAIR-RL reaches 53.93 Acc Avg., higher than GDPO at 53.30 and DAPO at 50.56. Appendix B.4

The two post-training stages provide complementary gains: structured SFT most benefits compositional stance, while subsequent RL yields its largest accuracy gains on fine-grained emotion. Compared with ordinary SFT, PAIR-SFT raises Acc Avg. by 17.68 and 16.70 percentage points on Qwen3-4B and Qwen3-1.7B, respectively, with the largest gains in Affective Stance at 25.70 and 25.95 points. Building on PAIR-SFT, PAIR-RL exceeds DAPO in Acc Avg. by 5.88 and 3.37 points, respectively. Relative to the corresponding PAIR-SFT policies, its largest task-level

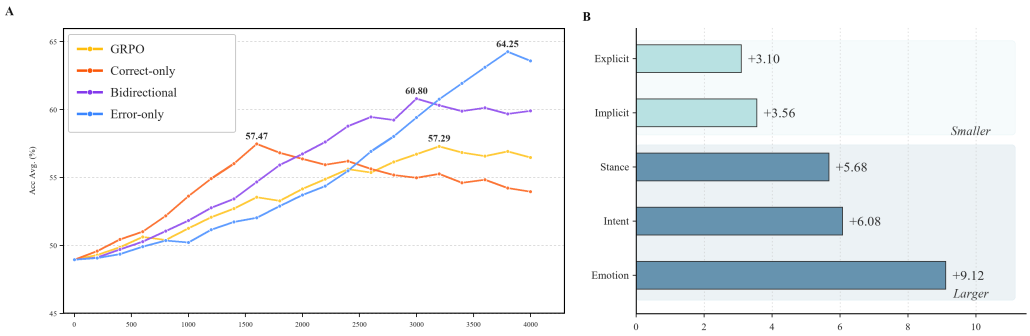

Figure 3: **Credit-direction trajectories and task-wise accuracy gains.** (a) Training curves for four credit directions from a shared PAIR-SFT initialization. (b) Accuracy gains of PAIR-RL over GRPO across the five semantic regions.

accuracy gains occur on Fine-grained Emotion, at 19.62 and 14.25 points, respectively. This pattern is consistent with the complementary roles of the two stages: structured supervision teaches polarity composition, while supervision that couples explicit and implicit polarity judgments through their resulting stance further improves difficult fine-grained emotion recognition.

4.3 Ablation Study and Analysis

To analyze how region-level feedback and structural coupling affect affective reasoning, we ask four questions about optimization dynamics, gains across five semantic judgments, routing and coupling mechanisms, and supervision granularity:

1. **How do different credit directions affect optimization dynamics and final performance?**
2. **What accuracy gains does PAIR-RL deliver across the five semantic judgments?**
3. **What roles do semantic-region routing and explicit–implicit–stance coupling play?**
4. **How much additional fine-grained supervision do regional and structural rewards provide beyond a response-level scalar reward?**

Q1: Credit direction trades off optimization progress and final performance. Figure 3(a) compares four credit-direction settings initialized from the same PAIR-SFT checkpoint. Correct-only reaches 57.47% at step 1600 and then declines; Bidirectional reaches 60.80% at step 3000; and Error-only makes slower early progress but continues to improve through step 3800, reaching its best performance of 64.25%. Reaching a peak earlier does not necessarily lead to higher final accuracy. Correct-only tends to rapidly increase the probability of correct reasoning, which reduces the model's ability to explore. Error-only does not provide additional reinforcement to correct reasoning; instead, it lowers the probability of erroneous reasoning, thereby enhancing exploration and enabling the model to explore a broader range of reasoning paths. The complete training trajectories are provided in Appendix B.5.

Q2: Regional credit improves all five tasks, with larger gains in pragmatic intent and fine-grained emotion. Figure 3(b) shows that PAIR-RL improves all five tasks over GRPO. Fine-grained Emotion accuracy rises from 45.87% to 54.99% (+9.12 percentage points), and Pragmatic Intent accuracy increases from 60.97% to 67.05% (+6.08 points). Affective Stance, Explicit Polarity, and Implicit Polarity improve by 5.68, 3.10, and 3.56 points, respectively. This pattern is consistent with the different demands of the tasks: structured SFT explicitly teaches the compositional relation between explicit and implicit polarity and Affective Stance, whereas Pragmatic Intent and Fine-grained Emotion rely more heavily on contextual and pragmatic cues and show larger accuracy gains after RL. When scores across tasks are aggregated into a single response-level reward, correctness on one task can partly mask an error on another. Regional credit preserves a separate feedback signal for each semantic judgment and routes it to the corresponding region, allowing the policy to correct intent, emotion, and polarity judgments separately.

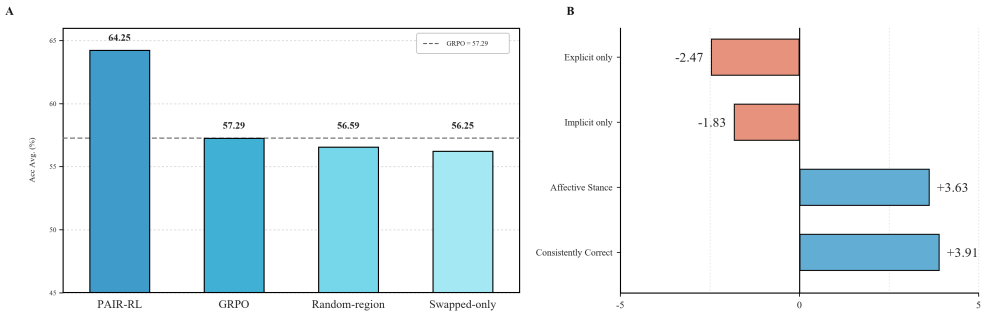

Figure 4: **Semantic-region routing and reward-coupling analysis.** (a) Acc Avg. across credit-direction and routing variants; the dashed line marks the GRPO baseline. (b) Percentage-point changes under coupled rewards relative to the uncoupled setting in the Explicit-only and Implicit-only proportions, Affective Stance accuracy, and the Consistently Correct rate.

Q3: Semantic-region routing directs feedback to the relevant region, while coupled rewards improve joint polarity and stance judgments. Figure 4(a) shows that aligned Error-only reaches 64.25% Acc Avg., compared with 57.29% for GRPO; Random-region and Swapped-only score 56.59% and 56.25%, respectively. These results suggest that feedback is more effective when routed to the semantic region containing the error. Figure 4(b) compares coupled and uncoupled rewards. Relative to the uncoupled setting, coupled rewards reduce the proportions of Explicit-only and Implicit-only cases by 2.47 and 1.83 percentage points, respectively. They also improve Affective Stance accuracy by 3.63 points and increase the Consistently Correct rate—the proportion of cases in which both explicit and implicit polarities are correct—by 3.91 points. These results suggest that coupled rewards reduce one-sided polarity correctness while improving joint explicit–implicit judgments and affective stance recognition. Appendix B.5

Q4: Regional and structural rewards expose within-group heterogeneity hidden by a response-level scalar reward. Figure 5 is based on one complete rollout over the training set and partitions the sampled groups by the number of semantic regions exhibiting within-group disagreement: 20.03% of groups are consistent across all regions, 47.47% show disagreement in one region, 18.25% in two regions, and 14.25% in three to five regions. Thus, 79.97% of sampled groups exhibit disagreement in at least one semantic region.A response-level scalar reward compresses each response into a single overall signal and therefore cannot fully characterize heterogeneity across semantic regions within the same sampled group. In contrast, regional rewards preserve these finer-grained regional differences and route the corresponding learning signals to their aligned semantic regions. In particular, 47.47% of groups exhibit disagreement in only one region, indicating that a substantial portion of within-group variation is localized rather than shared across the entire response.

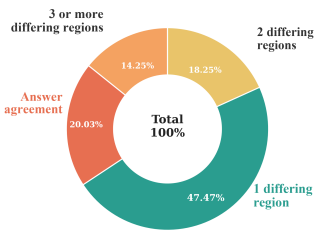

Figure 5: **Complementary regional supervision.**

5 Conclusion

We present **PAIR**, a neuro-symbolic post-training framework for affective reasoning that incorporates symbolic task relations into structured supervised learning and regional policy optimization. **PAIR-SFT** learns affective judgments and their compositional relations from symbolically verified teacher trajectories; **PAIR-RL** combines global and regional group-relative advantages and routes local feedback to the corresponding semantic regions. Experiments on CueBench at two model scales show complementary gains from structured supervision and regional policy optimization, with the final models outperforming GRPO, DAPO, and GDPO across all three tasks. Ablations show that both credit direction and semantic destination affect optimization, with aligned error-directed feedback achieving the strongest performance. These results demonstrate the value of combining symbolic task structure with regional credit assignment for affective reasoning.

Ethics Statement

This work uses the existing CueBench benchmark. We do not recruit human participants or collect new data for this study, and we follow the dataset's stated terms of use.

Reproducibility Statement

All information necessary to reproduce our experiments can be found in Appendices A–C and the supplementary material.

AI Use Statement

We used Qwen3-30B to generate supervision for the ordinary SFT baseline and candidate structured reasoning trajectories for PAIR-SFT, as described in Appendix B.2. PAIR-SFT candidates were filtered by agreement with the gold annotations and consistency with the explicit–implicit polarity composition. Generative AI tools were also used to assist with manuscript writing and language polishing.

References

Peter Chen, Xiaopeng Li, Ziniu Li, Xi Chen, and Tianyi Lin. Stepwise guided policy optimization: Coloring your incorrect reasoning in grpo. *arXiv preprint arXiv:2505.11595*, 2025.

Yirong Chen, Weiquan Fan, Xiaofen Xing, Jianxin Pang, Minlie Huang, Wenjing Han, Qianfeng Tie, and Xiangmin Xu. Cped: A large-scale chinese personalized and emotional dialogue dataset for conversational ai. *arXiv preprint arXiv:2205.14727*, 2022.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. *Advances in Neural Information Processing Systems*, 38:131646–131671, 2026.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4040–4054, 2020.

Varun Dhanraj and Chris Eliasmith. Improving rule-based reasoning in llms using neurosymbolic representations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30577–30596, 2025.

Ritam Dutt, Zhen Wu, Jiaxin Shi, Divyanshu Sheth, Prakhar Gupta, and Carolyn Rose. Leveraging machine-generated rationales to facilitate social meaning detection in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6901–6929, 2024.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Wen Wu, Chao Zhang, and Milica Gasic. Affect recognition in conversations using large language models. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 259–273, 2024.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. Laerc-s: Improving llm-based emotion recognition in conversation with speaker characteristics. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 6748–6761, 2025.

Yumeng Fu, Weitao Huang, Junjie Wu, Hao Teng, Shouduo Shang, Meishan Zhang, and Bingquan Liu. Ercthinker: Fast-slow thinking for emotion recognition in conversation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 39344–39359, 2026a.

Yumeng Fu, Weitao Huang, Junjie Wu, Hao Teng, Meishan Zhang, and Bingquan Liu. JoPR: Joint emotion perception and reasoning for conversational emotion recognition. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38852–38864. Association for Computational Linguistics, 2026b.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 154–164, 2019.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 2470–2481, 2020.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in rl for large language models. *Advances in Neural Information Processing Systems*, 38:114399–114431, 2026.

Yujin Kang and Yoon-Sik Cho. Beyond single emotion: Multi-label approach to conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24321–24329, 2025.

Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. Iest: Wassa-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 31–42, 2018.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint arXiv:2309.11911*, 2023.

Jiahui Li, Lin Li, Tai-Wei Chang, Kun Kuang, Long Chen, Jun Zhou, and Cheng Yang. Red: Unleashing token-level rewards from holistic feedback via reward redistribution. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4993–5022, 2025.

Xinran Li, Yu Liu, Jiaqi Qiao, and Xiujuan Xu. Do llms feel? teaching emotion recognition with prompts, retrieval, and curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31778–31786, 2026a.

Yugu Li, Zehong Cao, Jianglin Qiao, and Siyi Hu. Ssvpo: Effective step-level credit assignment for rl training of language models. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Qiao Liang, Ying Shen, Yao Liu, Tiantian Chen, and Lin Zhang. Why do emotions change? appraisal-guided reasoning for emotion–cause triplet extraction in conversations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11750–11772, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Chengwu Liu, Ye Yuan, Yichun Yin, Yan Xu, Xin Xu, Zaoyu Chen, Yasheng Wang, Lifeng Shang, Qun Liu, and Ming Zhang. Safe: Enhancing mathematical reasoning in large language models via retrospective step-aware formal verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12171–12186, 2025a.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026.

Shudong Liu, Hongwei Liu, Junnan Liu, Linchen Xiao, Songyang Gao, Chengqi Lyu, Yuzhe Gu, Wenwei Zhang, Derek F Wong, Songyang Zhang, et al. Compassverifier: A unified and robust verifier for llms evaluation and outcome reward. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33454–33482, 2025b.

Jiachen Luo, Huy Phan, and Joshua Reiss. Fine-tuned roberta model with a cnn-lstm network for conversational emotion recognition. In *Proc. Interspeech 2023*, pp. 2413–2417, 2023.

Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. Nus-emo at semeval-2024 task 3: Instruction-tuning llm for multimodal emotion-cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 1589–1596, 2024.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8679–8696, 2025a.

Yiran Ma, Zui Chen, Tianqiao Liu, Mi Tian, Zhuo Liu, Zitao Liu, and Weiqi Luo. What are step-level reward models rewarding? counterintuitive findings from mcts-boosted mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24812–24820, 2025b.

Aishwarya Maheswaran and Maunendra Sankar Desarkar. A unified view on emotion representation in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3589–3610, 2026.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6818–6825, 2019.

Silviu Oprea and Walid Magdy. isarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1279–1289, 2020.

Rosalind W Picard. *Affective computing*. The MIT press, 1997.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 527–536, 2019.

Rajkumar Pujari and Dan Goldwasser. Llm-human pipeline for cultural grounding of conversations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1029–1048, 2025.

Xin Quan, Marco Valentino, Louise A Dennis, and André Freitas. Verification and refinement of natural language explanations through llm-symbolic theorem proving. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 2933–2958, 2024.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for llm reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 60808–60838, 2025.

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

Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. Coe: A clue of emotion framework for emotion recognition in conversations. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23548–23563, 2025.

Cesare Spinoso-Di Piano, David Eric Austin, Pablo Piantanida, and Jackie CK Cheung. (rsa) ^{2}: A rhetorical-strategy-aware rational speech act framework for figurative language understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20898–20938, 2025.

Xingwei Tan, Marco Valentino, Mahmud Elahi Akhter, Maria Liakata, and Nikolaos Aletras. Enhancing logical reasoning in language models via symbolically-guided monte carlo process supervision. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31874–31888, 2025.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439. Association for Computational Linguistics, 2024a.

Siyuan Wang, Zhongyu Wei, Yejin Choi, and Xiang Ren. Symbolic working memory enhances language models for complex rule application. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 17583–17604, 2024b.

Cheng Wen, Hu Junjie, YiKun Hu, Jie Su, Bin Yu, Dugang Liu, Zhiwu Xu, Weidi Sun, Shengchao Qin, and Cong Tian. Formally specifying the intended behavior of the program: Llm-driven neuro-symbolic program specification synthesis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 672–682, 2026.

Jintao Wen, Geng Tu, Rui Li, Dazhi Jiang, and Wenhua Zhu. Learning more from mixed emotions: A label refinement method for emotion recognition in conversations. *Transactions of the Association for Computational Linguistics*, 11:1485–1499, 2023.

Shengguang Wu, Shusheng Yang, Zhenglun Chen, and Qi Su. Rethinking pragmatics in large language models: Towards open-ended evaluation and preference tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22583–22599, 2024.

Fangzhi Xu, Zhiyong Wu, Qiushi Sun, Siyu Ren, Fei Yuan, Shuai Yuan, Qika Lin, Yu Qiao, and Jun Liu. Symbol-llm: Towards foundational symbol-centric interface for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13091–13116, 2024.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. Bioserc: Integrating biography speakers supported by llms for erc tasks. In *International Conference on Artificial Neural Networks*, pp. 277–292. Springer, 2024.

Yueqin Yin, Shentao Yang, Yujia Xie, Ziyi Yang, Yuting Sun, Hany Awadalla, Weizhu Chen, and Mingyuan Zhou. Segmenting text and learning their rewards for improved rlhf in language model. *arXiv preprint arXiv:2501.02790*, 2025.

Hee Suk Yoon, Eunseop Yoon, Mark Hasegawa-Johnson, Sungwoong Kim, and Chang D Yoo. Confpo: Exploiting policy model confidence for critical token selection in preference optimization. *arXiv preprint arXiv:2506.08712*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*, 2024.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024a.

Tao Zhang and Zhenhua Tan. Ecerc: evidence-cause attention network for multi-modal emotion recognition in conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2064–2077, 2025.

Yuwei Zhang, Siffi Singh, Sailik Sengupta, Igor Shalyminov, Hang Su, Hwanjun Song, and Saab Mansour. Can your model tell a negation from an implicature? unravelling challenges with intent encoders. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 552–567, 2024b.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Processbench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025.

Zhenyan Zheng, Yunyao Zhang, Junxi Sheng, Junqing Yu, and Zikai Song. Surfacing the unsaid: Cue-bench for affective stance in chinese discourse. *arXiv preprint arXiv:2608.10810*, 2026.

Wenxuan Zhou, Shujian Zhang, Lingxiao Zhao, and Tao Meng. T-reg: Preference optimization with token-level reward regularization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27876–27889, 2025.

Appendix

Appendix Overview

A. Methodological Details

B. Experimental Details

C. Prompts and Algorithms

A Methodological Details

This appendix specifies the structured interface, symbolic checks, and global–local credit construction used by PAIR. The organization follows the direction of the method: the model first emits a structured response, the verifier checks its symbolic consistency, and the resulting signals are routed to the policy objective.

A.1 Task Decomposition and Compositional Semantics

PAIR uses five semantic regions shared by structured SFT and PAIR-RL:

$$\mathcal{R} = \{E, I, S, P, F\}, \tag{17}$$

A response consists of the five region-level judgments and rationales, followed by a structured answer:

$$y_i = \left(y_i^{\text{exp}}, y_i^{\text{imp}}, y_i^{\text{st}}, y_i^{\text{int}}, y_i^{\text{emo}}\right), \qquad r \in \mathcal{R}. \tag{18}$$

Here, E, I, S, P, and F denote explicit polarity, implicit polarity, affective stance, pragmatic intent, and fine-grained emotion, respectively. Affective Stance is compositionally determined from explicit and implicit polarity using the benchmark mapping:

$$\tilde{s}_i = \Omega(e_i, h_i). \tag{19}$$

The exact prompt templates and serialized response format are provided in Appendix C. This shared structure exposes the five semantic units used for verification and regional credit assignment.

A.2 Symbolic Verification Rules

For each labeled example $(x_i, z_i^\star)$, the verifier evaluates the teacher-generated trajectory y_i^T using Eq. equation 4. A trajectory is retained only if its semantic judgments agree with the gold annotations ($V_{\text{judge}} = 1$) and its predicted affective stance satisfies the symbolic composition rule ($V_{\text{comp}} = 1$; Eq. equation 3). The retained pairs form $\mathcal{D}_{\text{PAIR-SFT}}$ in Eq. equation 5.

A.3 Reward and Advantage Construction

For candidate i in rollout group $\mathcal{G}_u$, PAIR-RL computes one global reward, one structural reward, and a local reward for each semantic region. The global reward evaluates the three task outputs; the structural reward checks the explicit–implicit–stance relation; and each local reward checks one region against its gold label:

$$\begin{aligned} R_i^{\text{global}} &= 0.4R_i^{\text{st}} + 0.3R_i^{\text{int}} + 0.3R_i^{\text{emo}}, \\ R_i^{\text{struct}} &= \mathbf{1}\left[\hat{s}_i = \Omega(\hat{e}_i, \hat{h}_i)\right], \\ R_{i,r}^{\text{local}} &= \mathbf{1}\left[\hat{z}_{i,r} = z_{u,r}^\star\right], \qquad r \in \mathcal{R} = \{E, I, S, P, F\}. \end{aligned} \tag{20}$$

Each reward is normalized within its rollout group. For a reward R_i with group mean μ_u and standard deviation σ_u, we use

$$\mathsf{Norm}_{\mathcal{G}_u}(R_i) = \begin{cases} \dfrac{R_i - \mu_u}{\sigma_u + \delta}, & \sigma_u > 0, \\ 0, & \sigma_u = 0, \end{cases} \qquad \delta > 0. \tag{21}$$

Applying this normalization gives A_i^{global}, A_i^{struct}, and $A_{i,r}^{\text{local}}$. A signal with zero group variance therefore contributes zero advantage (Yuan et al., 2024; Cheng et al., 2026).

The global, local, and structural advantages are combined and routed to response tokens according to Eq. equation 16. The credit-direction variants differ in the local gate:

$$G_{i,r}^{\text{local}} = \begin{cases} 0, & \text{Global-only}, \\ R_{i,r}^{\text{local}}, & \text{Correct-only}, \\ 1, & \text{Bidirectional}, \\ 1 - R_{i,r}^{\text{local}}, & \text{Error-only}. \end{cases} \tag{22}$$

For the Global-only control, we disable both regional feedback pathways by setting

$$G_{i,r}^{\text{local}} = 0 \quad (r \in \mathcal{R}), \qquad G_{i,r}^{\text{struct}} = 0 \quad (r \in \mathcal{R}_{\text{str}}).$$

Substituting these settings into Eq. equation 16 leaves only the response-level global advantage:

$$A_{i,t}^{\text{Global-only}} = \lambda_{\text{global}} A_i^{\text{global}}.$$

Token-span alignment is detailed in Appendix A.4; the unified routing rule is given in Eq. equation 16. Token-span alignment is detailed in Appendix A.4; the unified routing rule is given in Eq. equation 16.

A.4 Token Alignment and Credit Routing

The parser records the character span of each of the five semantic regions $r \in \mathcal{R} = \{E, I, S, P, F\}$ before tokenization and maps each span to token intervals, forming the masks $M_{i,t,r}$. Within each span, the local advantage is applied according to $G_{i,r}^{\text{local}}$; for $r \in \mathcal{R}_{\text{str}} = \{E, I, S\}$, the structural advantage is additionally routed according to $G_{i,r}^{\text{struct}}$. Tokens outside the five semantic regions therefore receive only the global advantage. If a reward has zero variance within its rollout group, its normalized advantage is zero. The token masks are shared across the ablations: credit-direction variants change the local gate, while Random-region and Swapped-only change the destination of regional feedback.

B Experimental Details

This appendix describes the benchmark and task labels, the training and evaluation protocol, and the additional numerical results underlying the analyses in the main paper.

B.1 CueBench and Task Definition

CueBench is a Chinese discourse benchmark containing 51,823 annotated context–target instances drawn from open-domain conversations, social-media comments, sarcasm-oriented text, customer-service interactions, and question-answering content (Zheng et al., 2026). Each instance includes the discourse context and a target utterance or text span, since affective meaning may depend on information outside the target expression itself.

The benchmark evaluates three tasks. **Affective Stance** captures the relation between explicit expression and implicit affective tendency: each signal has positive, neutral, or negative orientation, and their combination defines one of nine stance labels. **Pragmatic Intent** identifies the communicative purpose of the utterance from eight labels. **Fine-grained Emotion** classifies the conveyed emotional state using 25 labels. Thus, stance is compositional, while intent and emotion retain their own task-specific label spaces. PAIR represents explicit polarity, implicit polarity, affective stance, pragmatic intent, and fine-grained emotion as five semantic regions; the stance judgment is constrained by the composition of the two polarities.

For each task, we report Accuracy, Macro-F1, and Weighted-F1. Acc Avg. is the arithmetic mean of the three task accuracies; the complete nine-metric results are reported in the main table and the supplementary routing table.

Table 2: **Dataset and evaluation details.**

Item	Setting
Benchmark	CueBench
Tasks	Affective Stance, Pragmatic Intent, and Fine-grained Emotion
Task label spaces	Stance: 9 compositional labels; Intent: 8 labels; Emotion: 25 labels
Metrics	Accuracy, Macro-F1, and Weighted-F1 for each task
Aggregate metric	Acc Avg.: arithmetic mean of the three task accuracies

The same task interface and evaluation protocol are used across model scales and policy-optimization variants.

B.2 SFT Data and Training Setup

The ordinary SFT baseline and PAIR-SFT use the same 8,000 class-balanced training examples and the same Qwen3-30B teacher to generate supervision. For ordinary SFT, we provide the teacher with the CueBench task prompt and use its generated response as the supervised target. PAIR-SFT instead uses the structured prompt in Appendix C to generate candidate trajectories containing the five semantic regions and a structured answer. Following Section 3.1, the verifier retains a trajectory only when its semantic judgments agree with the gold annotations and its predicted stance is consistent with the explicit–implicit composition under Ω.

Table 3: **SFT training settings.** The same settings are used across both student backbones.

Item	Setting
Student models	Qwen3-4B and Qwen3-1.7B
Fine-tuning	LoRA with rank 16 and alpha 32
Target modules	All linear modules
Precision	bfloat16
SFT epochs	2
SFT learning rate	1×10^{-4} with cosine scheduling and 5% warmup
SFT batch construction	Per-device batch size 4 with 8 gradient-accumulation steps

B.3 RL Training Setup

Following verifier-guided structured SFT, the RL experiments use Qwen3-4B and Qwen3-1.7B students. GRPO, DAPO, and PAIR-RL all start from the same PAIR-SFT policy. Each rollout group contains $K = 8$ responses, and policy-optimization variants share the same global task reward. Regional ablations change only the gate or destination of the regional signal. The matched RL configuration is summarized in Table 4.

During training, we evaluate each checkpoint on the validation set and select checkpoints according to validation Acc Avg. The main table reports test-set results for the selected checkpoints; the test set is not used for checkpoint selection.

Table 4: **RL training settings.** The same settings are used across both student backbones.

Item	Setting
RL initialization	GRPO, DAPO, and PAIR-RL all start from the same PAIR-SFT policy
Fine-tuning	LoRA with rank 16 and alpha 32
Target modules	All linear modules
Precision	bfloat16
RL learning rate	2×10^{-6} with 20 warmup steps
RL PPO epochs	1
KL regularization	Enabled with coefficient 0.04
Rollout group size	$K = 8$ completions per prompt
Generation	Temperature 0.7, top-p 0.95, top-k 20
Length limits	Prompt length 2,048; response length 1,536
Global reward	$\lambda_{\text{st}} R^{\text{st}} + \lambda_{\text{int}} R^{\text{int}} + \lambda_{\text{emo}} R^{\text{emo}}$
Local regional coefficient	λ_{local}
Structural feedback	Applied to $\mathcal{R}_{\text{str}} = \{E, I, S\}$ as in Eq. equation 16
Main local gate	Error-only
Feedback controls	Global-only, Correct-only, Bidirectional, Random-region, and Swapped-only

B.4 Transfer Experiments on CPED and MELD

We further evaluate whether the learned affective reasoning ability transfers beyond CUE-Bench to CPED (Chen et al., 2022), a Chinese conversational affect benchmark, and MELD (Poria et al., 2019), an English emotion-recognition-in-conversation benchmark. We compare direct inference, PAIR-SFT, and PAIR-RL. All values are percentages.

CPED. Following the evaluation collected by Pujari & Goldwasser (2025), Table 5 reports weighted F1 for emotion recognition, personality recognition, and dialogue-act recognition, together with their unweighted average. PAIR-RL obtains the strongest average among our variants (56.05), and PAIR-SFT reaches 49.66. Notably, even the strongest baseline augmented with additional cultural context reaches only 48.32 on average, remaining below both of our trained variants.

Table 5: **Transfer results on CPED (Chen et al., 2022).** Emotion, personality, and dialogue-act scores are weighted F1; Avg. is their unweighted mean. Published baseline results follow Pujari & Goldwasser (2025).

Model	Training / Input	Emotion	Personality	Dialogue Act	Avg.
RoBERTa	Standard	46.69	20.89	55.40	40.99
ConvGraph	Standard	47.66	21.42	56.28	41.79
Llama-3.1-8B-Instruct	Direct	53.91	24.73	54.15	44.26
ConvGraph	Cultural context	49.90	21.65	57.24	42.93
Llama-3.1-8B-Instruct	Cultural context	**55.00**	**29.81**	**60.16**	**48.32**
Qwen3-4B	Direct	40.58	23.22	41.35	35.05
Qwen3-4B	PAIR-SFT	56.30	30.37	62.31	49.66
Qwen3-4B	PAIR-RL	**65.24**	**36.36**	**66.55**	**56.05**

MELD. It provides parallel emotion and sentiment annotations for each utterance, but lacks the explicit structural dependencies used in CUE-Bench. Its formulation is therefore not a perfect match for the structured multi-task design targeted by PAIR, and we use it to examine transfer to a conventional conversational affect benchmark. During adaptation, we jointly predict the seven-class emotion label and the three-class sentiment label, while Table 6 reports emotion weighted F1 for comparison with prior work. Although MELD provides textual, acoustic, and visual modalities, our experiments use only its textual transcripts. Under this text-only setting, PAIR-RL reaches 66.83, improving by 12.21 points over direct inference and by 3.69 points over PAIR-SFT.

Table 6: **Transfer results on MELD.** The primary metric is seven-class emotion weighted F1. Our models use only the textual modality.

Backbone / Model	Method	Weighted F1
Text-CNN	Standard (Poria et al., 2019)	55.02
DialogueRNN	Speaker-aware RNN (Majumder et al., 2019)	57.03
DialogueGCN	Graph context (Ghosal et al., 2019)	58.10
COSMIC	Commonsense context (Ghosal et al., 2020)	65.21
EmoLR	Label refinement (Wen et al., 2023)	65.16
RoBERTa	CNN–LSTM (Luo et al., 2023)	66.12
Llama2	InstructERC (Lei et al., 2023)	69.15
LLM	LaERC-S (Fu et al., 2025)	69.27
Llama-2-13B	BiosERC (Xue et al., 2024)	69.83
Qwen3-8B	PRC-Emo (Li et al., 2026a)	70.44
Llama3-8B	JoPR (Fu et al., 2026b)	**70.72**
Qwen3-4B	Direct	54.62
Qwen3-4B	PAIR-SFT	63.14
Qwen3-4B	PAIR-RL	**66.83**

The two transfer benchmarks are reported separately because their label spaces and evaluation protocols differ.

B.5 Supplementary Results

The following tables provide the numerical values underlying the optimization and routing analyses in Section 4.3. They are included for reproducibility and do not define additional training objectives.

Credit-direction trajectories. Table 7 reports test-set Acc Avg. values used to produce Figure 3. All variants share the same value at step 0 because they start from the same PAIR-SFT checkpoint. Bold entries after initialization mark the best checkpoint selected by validation Acc Avg. for each credit direction.

Table 7: Acc Avg. trajectories for the four regional-credit variants.

Step	GRPO	Correct-only	Bidirectional	Error-only
0	**48.96**	**48.96**	**48.96**	**48.96**
200	49.31	49.58	49.12	49.08
400	49.87	50.44	49.71	49.36
600	50.63	51.02	50.28	49.91
800	50.41	52.18	51.06	50.37
1000	51.26	53.64	51.83	50.22
1200	52.08	54.91	52.77	51.16
1400	52.71	56.03	53.42	51.74
1600	53.54	**57.47**	54.68	52.03
1800	53.29	56.82	55.91	52.89
2000	54.16	56.36	56.73	53.71
2200	54.88	55.94	57.62	54.36
2400	55.61	56.21	58.77	55.48
2600	55.37	55.63	59.46	56.91
2800	56.14	55.18	59.22	58.03
3000	56.72	54.97	**60.80**	59.41
3200	**57.29**	55.26	60.31	60.76
3400	56.83	54.61	59.88	61.92
3600	56.57	54.84	60.12	63.11
3800	56.91	54.22	59.67	**64.25**
4000	56.48	53.96	59.91	63.58

Regional-feedback ablations. Table 8 reports the complete results for the credit-direction and routing variants analyzed in Section 4.3. All variants use the same structured interface and optimization configuration, varying only the regional gate or the destination of regional feedback. Metrics follow Table 1.

Table 8: **Regional-feedback ablations on Qwen3-4B.** Each task reports Accuracy, Macro-F1, and Weighted-F1 (%). Acc Avg. is the arithmetic mean of Accuracy across the three tasks. The best-performing variant is highlighted.

Method	Affective Stance			Pragmatic Intent			Fine-grained Emotion			Acc Avg.
	Acc.	Macro-F1	W-F1	Acc.	Macro-F1	W-F1	Acc.	Macro-F1	W-F1	
Random-region	63.34	52.55	62.36	60.45	41.05	58.18	45.98	23.22	43.95	56.59
Error-only	**70.71**	**59.83**	**69.99**	**67.05**	**49.43**	**65.40**	**54.99**	**26.26**	**51.41**	**64.25**
Correct-only	64.30	53.21	63.14	61.10	42.16	59.09	47.01	23.67	45.41	57.47
Swapped-only	62.90	52.17	62.01	60.10	40.73	57.84	45.75	23.06	43.64	56.25
Bidirectional	67.46	56.31	66.38	64.31	45.58	62.26	50.64	24.38	48.68	60.80

C Prompts and Algorithms

This section presents the schema of structured PAIR-SFT records and a framework-level PAIR-RL prompt. Instance-specific text, identifiers, and annotation values are replaced with placeholders.

C.1 PAIR-SFT Record Schema

Each PAIR-SFT example is serialized as a user–assistant message pair: the user message contains the task prompt, and the assistant message contains five separated reasoning regions and the answer JSON. Sample identifiers and data-version tags remain at the top level; instance-specific content is replaced by placeholders.

PAIR-SFT Serialized Record Schema

```
{
  "messages": [
    {
      "role": "user",
      "content": "<multitask prompt with <CONTEXT> and <TARGET_TEXT> placeholders>"
    },
    {
      "role": "assistant",
      "content": "<think>\n<explicit>[explicit-polarity
      ↪  rationale]</explicit>\n<implicit>[implicit-polarity
      ↪  rationale]</implicit>\n<stance>[compositional stance
      ↪  rationale]</stance>\n<intent>[pragmatic-intent
      ↪  rationale]</intent>\n<emotion>[fine-grained-emotion
      ↪  rationale]</emotion>\n</think>\n<answer>\n{\"explicit_polarity\":\"<label>\
      ↪  \",\"implicit_polarity\":\"<label>\",\"affective_stance\":\"<label>\",\"prag
      ↪  matic_intent\":\"<label>\",\"fine_grained_emotion\":\"<label>\"}\n</answer>"
    }
  ],
  "sample_id": "<sample identifier>",
  "data_version": "<dataset version>"
}
```

C.2 PAIR-RL Prompt Template

The prompt asks the model to make five linked semantic judgments, keep their rationales separate, and return only the prescribed JSON fields. The context and target text are placeholders rather than benchmark instances.

PAIR-RL Prompt Template

```
Perform multitask emotion recognition in Chinese-language discourse.

Context:
<CONTEXT>

Target text:
<TARGET_TEXT>
```

```
Make the following five related judgments in order, and verify each one separately:
1. Surface-level perspective
   Infer explicit polarity from the target text's literal wording alone.
   Choose only Positive, Neutral, or Negative.

2. Contextual perspective
   Infer the speaker's true implicit polarity from the context and target text.
   Choose only Positive, Neutral, or Negative.

3. Affective stance
   Combine the explicit and implicit polarity judgments using the benchmark's fixed
   3-by-3 composition mapping. Choose the corresponding label from the nine valid
   Affective Stance classes; do not assign a stance that conflicts with this
   ↪  composition.

4. Pragmatic intent
   Determine the primary communicative purpose of the speaker's utterance from the
   context, target text, and the preceding judgments. Do not mechanically map intent
   to a particular polarity.

   Candidate labels:
   - Authenticity: sincere, direct expression, including genuine positive or negative
   ↪  emotion
   - Politeness: polite, formulaic, indirect, or service-oriented language
   - Suppression: concealing, downplaying, or withholding genuine emotion
   - Irony: irony or sarcasm in which the intended meaning contrasts with the literal
   ↪  meaning
   - Resistance: refusal, opposition, pressure, or defiance
   - Functional: primarily conveying objective information or completing a task
   - Humor: primarily intended to joke, tease, or amuse
   - Empathy: care, comfort, support, or perspective-taking

5. Fine-grained emotion
   Infer the target speaker's specific inner emotion from the target text, context,
   and the preceding judgments. Do not attribute another person's emotion or the
   reader's reaction to the target speaker. Choose only from the benchmark's 25
   valid Fine-grained Emotion labels.

Use exactly the following five sections in your reasoning. Include each section
once, use one to three sentences per section, and keep the analyses separate:

<explicit>
Analyze only the affective evidence stated literally in the target text. Briefly
explain the judgment and cite target-text evidence when possible.
</explicit>

<implicit>
Analyze only the implicit emotion, tone, irony, or pragmatic meaning conveyed in
context. Briefly explain the judgment and cite evidence from the context or target
text when possible.
</implicit>

<stance>
State the Affective Stance implied by the explicit and implicit polarity labels
under the fixed composition mapping. Briefly explain the composition.
</stance>

<intent>
Analyze only how the context and target text support the pragmatic-intent judgment.
State the communicative purpose and its evidence.
</intent>

<emotion>
Analyze only how the context and target text support the target speaker's
fine-grained emotion. State the evidence for that emotion.
</emotion>

After completing your reasoning, output only the following JSON between <answer>
and </answer>. Do not add other fields. Use only a candidate label in English for
each field, without an explanation or translation:
{
  "explicit_polarity": "...",
  "implicit_polarity": "...",
  "affective_stance": "...",
  "pragmatic_intent": "...",
  "fine_grained_emotion": "..."
}
```