REINFORCEMENT LEARNING WITH COMPARATIVE EVIDENCE FOR SOCIAL INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Developing socially intelligent AI remains heavily dependent on human-annotated data, limiting the scale and breadth of social understanding that models can acquire. Methods that derive training signals from unlabeled data offer a path beyond this dependence, but social predictions lack the verification oracles available in domains such as mathematics and coding. Moreover, core social targets such as affect, intent, preference, and pragmatic meaning are often ambiguous. The same observed behavior can support multiple plausible interpretations, making it difficult to verify which is best supported. To address this challenge, we introduce **REINFORCEMENT LEARNING WITH COMPARATIVE EVIDENCE (RLCE)**, a reinforcement learning method that learns social understanding from unlabeled training data without constructing rewards from ground-truth annotations. Given distinct answers in a rollout group, RLCE constructs *evidence tests* that identify observable evidence favoring an answer over another, validates these tests against the input sample, and aggregates test outcomes to determine the best-supported interpretation. Evidence tests are then regenerated as the policy produces new answers during training, enabling tests to evolve with the policy's answers. Across four benchmarks spanning affect, pragmatics, communicative intent, and preference, RLCE attains the strongest performance among seven methods that use no ground-truth training labels for rewards, including consensus, policy LLM-judge verification, multimodal co-evolution, and rubric-based rewards. Gains over the strongest baseline are +3.58 on MOSEI, +4.50 points on UR-FUNNY, +1.15 on IntentQA, and +18.93 on PrefEval. Analyses further show that RLCE exhibits a larger share of reward variation between correct and incorrect predictions than compared rubric methods, can overturn erroneous policy-derived preferences, and benefits from pairwise test construction, compositional test aggregation, and on-policy test evolution. Code and models will be released after the review process; additional implementation details are in the appendix, included in a separate supplementary file.

1 INTRODUCTION

Building machines with social intelligence, the ability to understand people, their affective and cognitive states, preferences, and behavior, has been one of AI's longest-standing and most fundamental ambitions (Picard, 2000; Breazeal, 2000). However, despite the remarkable progress of modern AI, socially intelligent models remains heavily reliant on human-annotated data (Zhang et al., 2026b), which is difficult to obtain at scale because human behavior is highly heterogeneous, context-dependent, and ambiguous to interpret (Zhang et al., 2026b; Ong et al., 2026a). This places a fundamental ceiling on social intelligent AI systems, limiting what models can learn and how well they generalize to new settings (Ong et al., 2026b). Raising this ceiling, we argue, requires learning social understanding from unlabeled training examples, with rewards constructed without ground truth annotations. Yet, this remains challenging as social predictions lack a verification oracle (e.g., a proof checker), unlike solutions in mathematics and coding (Zhao et al., 2026). Consequently, self-generated supervision risks reinforcing errors (Arazo et al., 2020).

To mitigate this problem, prior work has explored proxy verification signals in place of a verification oracle, including LLM judgments (Yuan et al., 2024), self-consistency (Chen et al., 2023; He et al., 2025; Zhang et al., 2026a), and rubric-based rewards (Shao et al., 2025; Gunjal et al., 2026; Wu et al., 2026). However, social understanding poses a distinct verification problem: the social targets to be

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.

modeled - e.g., intent, sentiment, preference, or pragmatic meaning, are ambiguous, and manifest through observable language, expressions, cues, and behavior (Mathur et al., 2024; Vinciarelli et al., 2009). These observations can often be compatible with multiple plausible interpretations, although only one corresponds to the underlying social target (Vinciarelli et al., 2009; 2011). Existing proxy verification signals do not explicitly resolve among these interpretations. Rubric-based methods typically evaluate each response independently against general criteria, which may leave multiple interpretations similarly plausible (Gunjal et al., 2026; Shao et al., 2025). Additionally, consensus-based verification may favor agreement across predictions rather than what is actually observed in the sample (Chen et al., 2023; He et al., 2025; Zhang et al., 2026a).

To address this gap, we introduce **REINFORCEMENT LEARNING WITH COMPARATIVE EVIDENCE (RLCE)**, an RL method for learning social understanding from unlabeled training data. Our core principle is to verify ambiguous answers through comparative evidence: comparing answers pairwise through multiple evidence tests, each of which checks the sample for evidence (i.e., observable support) favoring an answer over another, and then aggregating the test outcomes into a verdict. This allows verification to distinguish among different interpretations of the social target based on the evidence present in the sample. Given answers from a rollout group, RLCE constructs evidence tests that check for evidence favoring an answer over another. For example, if "Nice job" could mean *praise* or *sarcasm*, rather than checking a generic cue such as whether the speaker smiles, which may occur in either case, RLCE constructs a test around a distinction that helps resolve the two interpretations, such as whether the speaker's tone aligns with or contradicts the positive words. Each test is then evaluated against the sample's content to determine whether the specified evidence is present, and the resulting outcomes are aggregated to identify the best-supported answer, which receives the reward for GRPO. As the policy updates, RLCE regenerates evidence tests for its latest answers, keeping them specific to the policy's outputs.

We evaluate RLCE on four established multimodal and text benchmarks spanning core and diverse social understanding tasks: affective understanding (Zadeh et al., 2018), pragmatic reasoning (Hasan et al., 2019), communicative-intent (Li et al., 2023), and preference and personalization (Zhao et al., 2025); comparing against state-of-the-art methods that construct rewards from unlabeled training data, including consensus-based approaches, from self-consistency (Chen et al., 2023) to multimodal evolution (He et al., 2025; Zhang et al., 2026a), policy LLM-Judge (Yuan et al., 2024), and rubric-based rewards (Gunjal et al., 2026; Shao et al., 2025; Wu et al., 2026). Our experiments show: (1) relative to baselines, RLCE achieves the best mean performance across benchmarks, with gains over the strongest baseline of +1.15 (IntentQA), +3.58 (MOSEI), +4.50 (UR-FUNNY), and +18.93 (PrefEval) points; (2) relative to compared rubric-based methods, RLCE exhibits a larger share of reward variation between correct and incorrect predictions, and more often overturns erroneous consensus and self-judgment preferences; (3) ablations of RLCE's components show that pairwise evidence-test construction improves performance, aggregating multiple tests improves verdict accuracy, and evolving tests outperform fixed tests, particularly for free-form policy outputs.

Main Contributions:

1. **RLCE for learning social understanding from unlabeled training data.** We introduce REINFORCEMENT LEARNING WITH COMPARATIVE EVIDENCE (RLCE), an RL method that constructs rewards without human-provided training labels. It verifies answers through pairwise evidence tests that check the input sample for observable evidence favoring an answer over another, and aggregates their outcomes to identify the best-supported answer.
2. **Consistent gains across four social understanding benchmarks.** Across four benchmarks spanning diverse and core forms of social understanding, RLCE achieves the best mean performance among seven methods that construct rewards without ground-truth training labels, exceeding the strongest baseline by +1.15, +3.58, +4.50, and +18.93 points on IntentQA, MOSEI, UR-FUNNY, and PrefEval, respectively.
3. **Analysis of RLCE's learning signal and design.** Relative to the compared rubric-based methods, RLCE exhibits a larger share of reward variation between correct and incorrect predictions and more often overturns erroneous consensus and self-judgment preferences. Ablations further show benefits from pairwise evidence-test construction, compositional test aggregation, and on-policy test evolution.

2 Related Work

Self-generated verification signals. A central challenge in learning from unlabeled training data is constructing reliable learning signals without human-provided labels. Early approaches derive this signal from agreement across the model's sampled predictions. Self-Consistency (Wang et al., 2022) and Universal Self-Consistency (Chen et al., 2023) use agreement across sampled answers, an approach extended to multimodal self-training by VisPlay (He et al., 2025) and Video-Zero (Zhang et al., 2026a). TTRL (Zuo et al., 2026) uses majority-vote pseudo-labels for policy updates, while SRLM (Yuan et al., 2024) relies on direct policy self-judgment. Recent work instead constructs explicit criteria for evaluating responses. Rubrics as Rewards (RaR) (Gunjal et al., 2026) uses fixed criteria, while subsequent methods construct or refine criteria adaptively from model-generated responses (Shao et al., 2025; Guan et al., 2026; Wang et al., 2026; Liu et al., 2026; Wu et al., 2026). By contrast, RLCE constructs pairwise evidence tests for candidate answers, checks whether the corresponding evidence is present in the input sample, and aggregates their outcomes across answers to identify the best-supported answer. Existing verification approaches have been explored less extensively for social understanding, where the underlying target is ambiguous and the same observed behavior can admit multiple plausible interpretations. This motivates RLCE's focus on resolving ambiguity by testing which interpretation is more strongly supported by the observed input.

Socially intelligent AI. Core to social intelligence is inferring latent affective, intentional, preferential, and pragmatic states from observable audiovisual and linguistic behavior. Existing work has largely relied on supervised annotations, ranging from sentiment and emotion labels in MOSEI (Zadeh et al., 2018) to human-authored reasoning traces in Social Genome (Mathur et al., 2025). Recent generalist models similarly learn social understanding through supervised training on annotated data (Ong et al., 2026a;b). Together, these approaches have advanced social understanding but retain a common dependence on human annotations for supervision. This limits how readily learning can scale to naturally occurring social data without training labels.

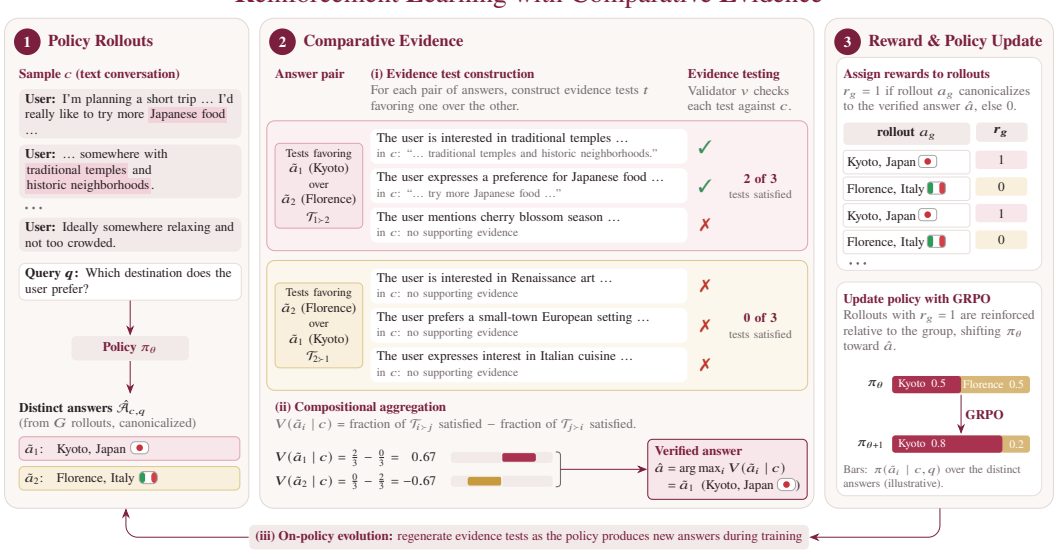


Figure 1: **Method Overview.** Reinforcement Learning with Comparative Evidence (RLCE) verifies candidate answers by comparing them pairwise with multiple tests that check the sample for support favoring an answer over another, and aggregating the test outcomes into a verdict. RLCE instantiates this through three steps: (i) *Evidence-test construction:* constructing pairwise evidence tests that check for observable evidence favoring an answer over another, (ii) *Compositional aggregation:* validating and aggregating these tests across answers to identify the best-supported answer for reward assignment, and (iii) *On-policy evolution:* regenerating evidence tests as the policy is updated and produces new answers during training.

3 REINFORCEMENT LEARNING WITH COMPARATIVE EVIDENCE

3.1 PRELIMINARIES

Problem definition: Learning social understanding from unlabeled training data. We consider critic-free reinforcement learning (e.g., GRPO (Shao et al., 2024)) for social understanding tasks, where policy training and reward construction use no human-provided ground-truth annotations from the training data. Let π_θ denote the policy to be learnt, instantiated as an LLM. Given a multimodal sample c and question $q = (u, \mathcal{A})$, where u denotes the natural-language query and $\mathcal{A}$ the answer space, we sample G rollouts $a_g \sim \pi_\theta(\cdot \mid c, q)$, $a_g \in \mathcal{A}$, $g = 1, \ldots, G$. The goal is to assign rewards using only c and q, without access to the ground-truth answer for the training example. Here a_g represents an interpretation of a social target (e.g., sentiment, intent, preference, or humor) inferred from observable language, expressions, and behavior. Because these cues can be ambiguous, the same observation may support multiple plausible interpretations.

Method overview. REINFORCEMENT LEARNING WITH COMPARATIVE EVIDENCE (RLCE) verifies candidate answers through three components. (i) *Evidence-test construction*: for pairs of distinct answers in a rollout group, RLCE constructs evidence tests that check for observable evidence in c favoring an answer over the other. (ii) *Compositional aggregation*: each test is validated against c, and the resulting test outcomes are aggregated across answer pairs into a verification score that determines the best-supported answer. (iii) *On-policy evolution*: as the policy updates and produces new answers, the evidence tests are regenerated so that verification stays specific to the policy's latest answers. Fig. 1 provides an overview, Sec. 4.2, 4.3, analyzes and ablates (i), (ii) and (iii).

3.2 VERIFICATION WITH COMPARATIVE EVIDENCE

Definition (Evidence test). An *evidence test* t checks whether the sample c contains observable evidence favoring one candidate answer over another. For example, if an answer predicts positive sentiment and another predicts negative sentiment, a test may check whether the clip shows the speaker smiling or laughing, supporting the positive interpretation over the negative one.

Evidence test construction. At training iteration ℓ, we sample a rollout group $\{a_g^{(\ell)}\}_{g=1}^{G} \sim \pi_{\theta_\ell}(\cdot \mid c, q)$ and canonicalize the rollouts into the distinct answer set $\hat{\mathcal{A}}_{c,q}^{(\ell)} = \{\tilde{a}_1^{(\ell)}, \ldots, \tilde{a}_K^{(\ell)}\}$, merging semantically equivalent outputs. Throughout our method, π_f denotes a *frozen* copy of the initial policy π_{θ_0} which is not updated during training. For open-ended outputs, canonicalization is performed by a single greedy call to π_f. For closed-label outputs, no greedy call is required for canonicalization; each rollout is mapped directly to its answer label. RLCE then constructs comparative evidence tests over a distinct candidate answer set $\hat{\mathcal{A}}_{c,q}^{(\ell)} \subseteq \mathcal{A}$. Concretely, for each pair of distinct answers $(\tilde{a}_i^{(\ell)}, \tilde{a}_j^{(\ell)})$, $i < j$, in the set, RLCE constructs one set of evidence tests for observable evidence supporting $\tilde{a}_i^{(\ell)}$ over $\tilde{a}_j^{(\ell)}$, and another supporting $\tilde{a}_j^{(\ell)}$ over $\tilde{a}_i^{(\ell)}$:

$$T_{\tilde{a}_i^{(\ell)} \succ \tilde{a}_j^{(\ell)}} \sim \pi_f\Big(\cdot \mid c, q, \tilde{a}_i^{(\ell)}, \tilde{a}_j^{(\ell)}\Big), \qquad T_{\tilde{a}_j^{(\ell)} \succ \tilde{a}_i^{(\ell)}} \sim \pi_f\Big(\cdot \mid c, q, \tilde{a}_j^{(\ell)}, \tilde{a}_i^{(\ell)}\Big). \tag{1}$$

For example, if two answers interpret a speaker's sentiment as *positive* versus *negative*, the corresponding tests separately check for observable evidence supporting positive over negative sentiment and negative over positive sentiment. Each test in $T_{\tilde{a}_i^{(\ell)} \succ \tilde{a}_j^{(\ell)}}$ is an observable claim whose satisfaction supports $\tilde{a}_i^{(\ell)}$ over $\tilde{a}_j^{(\ell)}$, i.e., it provides supporting evidence for $\tilde{a}_i^{(\ell)}$ and contradicting evidence for $\tilde{a}_j^{(\ell)}$. Each distinct answer is thus associated with tests that support it, $T^{+}_{\tilde{a}_i^{(\ell)}} = \bigcup_{j \neq i} T_{\tilde{a}_i^{(\ell)} \succ \tilde{a}_j^{(\ell)}}$, and tests that contradict it, $T^{-}_{\tilde{a}_i^{(\ell)}} = \bigcup_{j \neq i} T_{\tilde{a}_j^{(\ell)} \succ \tilde{a}_i^{(\ell)}}$.

Evidence testing. Given an evidence test $t \in T^{+}_{\tilde{a}_i^{(\ell)}} \cup T^{-}_{\tilde{a}_i^{(\ell)}}$ associated with an answer $\tilde{a}_i^{(\ell)}$, validation determines whether the claim specified by t is satisfied by the sample c. We use a single validator v to validate whether each constructed test is supported by the input. RLCE is validator-agnostic; in our experiments, we use π_f as v, following prior work demonstrating the effectiveness of policy-based self-judgment (Yuan et al., 2024), while ensuring that performance improvements can be attributed to RLCE's verification procedure rather than a stronger external model. Accordingly, for each evidence

test t, the validator returns a binary satisfaction outcome:

$$z(t) = v(t, c), \qquad z(t) \in \{0, 1\}, \tag{2}$$

where $z(t) = 1$ indicates that t is satisfied by c, and $z(t) = 0$ otherwise. Consistent with the general principle of abstaining rather than forcing a decision under uncertainty (Tomani et al., 2024), the validator may abstain when the sample c is insufficient to determine satisfaction or non-satisfaction of a test. Abstained tests receive no outcome and are excluded from aggregation.

Compositional aggregation. RLCE aggregates the satisfaction outcomes of the supporting and contradicting evidence tests associated with each answer to compute its verification score, $V(\tilde{a}_i^{(\ell)} \mid c)$. Let n_+ and n_- denote the numbers of evidence tests supporting and contradicting $\tilde{a}_i^{(\ell)}$, respectively, which are equal by construction, barring abstentions. The verification score is:

$$V\left(\tilde{a}_i^{(\ell)} \mid c\right) = \frac{1}{n_+} \sum_{t \in T^+_{\tilde{a}_i^{(\ell)}}} z(t) - \frac{1}{n_-} \sum_{t \in T^-_{\tilde{a}_i^{(\ell)}}} z(t). \tag{3}$$

The two terms are the proportions of satisfied supporting and contradicting evidence tests, respectively; their difference measures how strongly the validated evidence-test outcomes support $\tilde{a}_i^{(\ell)}$ relative to the other distinct answers, given the observed sample c. Intuitively, a larger $V(\tilde{a}_i^{(\ell)} \mid c)$ indicates stronger net evidence in favor of $\tilde{a}_i^{(\ell)}$ over alternative answers. We therefore select the distinct answer with the highest verification score as the verified answer:

$$\hat{a}^{(\ell)} = \arg \max_{\tilde{a}_i^{(\ell)} \in \widehat{\mathcal{A}}^{(\ell)}_{c,q}} V\left(\tilde{a}_i^{(\ell)} \mid c\right). \tag{4}$$

The aggregation and selection in Eqs.(3-4) depend on multiple validated evidence tests. As a result, even if some tests are incorrect, ambiguous, or uninformative, the aggregate verification score can remain accurate provided that the remaining tests are reliable. Additionally, verification operates on distinct answers; repeated rollouts are canonicalized before test construction, preventing answer frequency from affecting V.

Ties. If there is no clear winner with the highest verification score V, RLCE requests up to two additional evidence tests for the unresolved comparisons. If the tie remains, or if all rollouts correspond to the same answer (and thus yield no group-relative gradient), no rewards are assigned.

On-policy evolution. As the policy updates and produces new answers, RLCE regenerates evidence tests for the new answers rather than reusing tests from earlier iterations. The constructor and validator π_f remain frozen, so only the evidence tests evolve with the policy's answers, and no gradient computation occurs through π_f. Section 4.3 shows that this is particularly beneficial for free-form outputs, whose answer sets change more substantially during training.

3.3 Policy Update

Reward assignment. Recall that semantically equivalent rollout answers are canonicalized into the distinct answer set $\widehat{\mathcal{A}}^{(\ell)}_{c,q}$, from which the verified answer $\hat{a}^{(\ell)}$ is selected in Eq.(4). We assign a reward of 1 to each rollout whose answer was canonicalized into $\hat{a}^{(\ell)}$, and 0 otherwise:

$$r_g^{(\ell)} = \begin{cases} 1, & \text{if the canonicalized answer of } a_g^{(\ell)} \text{ equals } \hat{a}^{(\ell)}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

In practice, an auxiliary format reward is added before policy optimization; for notational simplicity, we continue to denote the resulting reward by $r_g^{(\ell)}$.

Group Relative Policy Optimization. We optimize the policy with GRPO (Shao et al., 2024). For a group of G rollouts with rewards $\{r_g^{(\ell)}\}_{g=1}^G$, and reward mean $\mu_{\mathcal{G}}^{(\ell)}$ and standard deviation $\sigma_{\mathcal{G}}^{(\ell)}$, the group-relative advantage is $A_g^{(\ell)} = (r_g^{(\ell)} - \mu_{\mathcal{G}}^{(\ell)})/\sigma_{\mathcal{G}}^{(\ell)}$, with $A_g^{(\ell)} = 0$ when $\sigma_{\mathcal{G}}^{(\ell)} = 0$; advantages are used in the standard clipped GRPO objective (all rewards and hyperparameters are in App. C.2.)

Table 1: **Held-out test performance after training on unlabeled training data.** All methods start from the same Qwen3.5-9B base model, and we report the mean over three training runs with different random seeds. For IntentQA, MOSEI, and UR-FUNNY, which are obtained from Human Behavior Atlas (Ong et al., 2026b), we follow its evaluation setup and report LLM-Judge accuracy, binary weighted F1 (F1w2), and weighted F1, respectively. For PrefEval, we follow its standard evaluation protocol and report LLM-Judge accuracy, macro-averaged over its three preference tasks. Δ denotes change from the base model (green/red: improvement/regression), and $\uparrow$ / $\downarrow$ denote higher/lower is better. Overall, RLCE achieves the most consistent performance across the baseline methods.

Method	**IntentQA** ($\uparrow$)	Δ	**MOSEI** ($\uparrow$)	Δ	**UR-FUNNY** ($\uparrow$)	Δ	**PrefEval**† ($\uparrow$)	Δ	**Avg. Rank** ($\downarrow$)
Qwen3.5-9B *(base model)*	55.25	–	64.86	–	63.11	–	44.72	–	–
Universal Self-Consistency (Chen et al., 2023)	56.77	+1.52	66.84	+1.98	59.86	−3.25	41.37	−3.35	5.75
Policy LLM-Judge (Yuan et al., 2024)	57.33	+2.08	67.35	+2.49	60.44	−2.67	46.73	+2.01	4.00
Video-Zero (Zhang et al., 2026a)	55.06	−0.19	64.80	−0.06	62.91	−0.20	43.30	−1.42	6.75
VisPlay (He et al., 2025)	55.34	+0.09	65.00	+0.14	64.96	+1.85	42.55	−2.17	5.75
Rubrics as Rewards (Gunjal et al., 2026)	55.60	+0.35	65.03	+0.17	64.72	+1.61	56.45	+11.73	4.25
RLER (Shao et al., 2025)	56.31	+1.06	65.19	+0.33	64.52	+1.41	56.12	+11.40	4.25
RLAC (Wu et al., 2026)	56.75	+1.50	67.15	+2.29	61.96	−1.15	55.27	+10.55	4.25
RLCE (ours)	**58.48**	+3.23	**70.93**	+6.07	**69.46**	+6.35	**75.38**	+30.66	**1.00**

4 EXPERIMENTS

Benchmarks & baselines. We evaluate RLCE on four core dimensions of social understanding: affect using the sentiment analysis task from MOSEI (Zadeh et al., 2018), pragmatics with UR-FUNNY (Hasan et al., 2019), intent with IntentQA (Li et al., 2023), and preference with PrefEval (Zhao et al., 2025). We compare against seven methods that construct rewards from unlabeled training data: rubric-based methods, including Rubrics as Rewards (RaR) (Gunjal et al., 2026), RLER (Shao et al., 2025), and RLAC (Wu et al., 2026); multimodal evolution methods, Video-Zero (Zhang et al., 2026a) and VisPlay (He et al., 2025); self-agreement via Universal Self-Consistency (USC) (Chen et al., 2023); and a frozen Policy LLM-Judge baseline, adapted from the policy-as-judge formulation of self-rewarding language models (Yuan et al., 2024).

Experimental setup. For MOSEI, UR-FUNNY, and IntentQA, we follow the splits, prompts, and metrics of the Human Behavior Atlas benchmark (Ong et al., 2026b); for PrefEval, we follow the evaluation setup introduced in the original work (Zhao et al., 2025) across its explicit, choice, and persona tasks. All methods use the same Qwen3.5-9B initialization, with no ground-truth training labels used for reward construction. From Table 1, we report the mean over three training runs with different random seeds. Full details of all baseline implementations and the experimental setup are provided in App. D.

4.1 RESULTS

Main results. We first evaluate whether RLCE can learn from unlabeled training data to generalize to a held-out test split. From Tab. 1, RLCE achieves the best mean performance on all four benchmarks across three runs, with an average rank of 1.00. Its gains over the strongest competing baseline are +3.58 points on MOSEI, +4.50 on UR-FUNNY, and +18.93 on PrefEval, each exceeding the seed standard deviation of either method (Tab. 6 provides the standard deviation). On IntentQA, the +1.15-point difference is within one standard deviation (58.48 ± 1.59 vs. 57.33 ± 0.50), indicating comparable performance. Beyond this second-best baseline, RLCE outperforms the remaining baselines on IntentQA by at least 1.71 points beyond its reported standard deviation. Overall, across three runs, RLCE achieves the highest mean performance on the four benchmarks, with gains spanning affect, pragmatics, intent, and preference tasks.

The magnitude of RLCE's gains varies across benchmarks. The largest improvement is observed on PrefEval (+18.93 points), relative to the gains on IntentQA (+1.15), MOSEI (+3.58), UR-FUNNY (+4.50). Several benchmark characteristics may partly account for this difference. Preference-relevant evidence is often explicitly expressed in the preceding interaction (Zhao et al., 2025), PrefEval's text-only format avoids the additional perceptual uncertainty present in the other multimodal benchmarks, and the lower base-model performance (44.72%) leaves greater room for improvement. We further examine the PrefEval result through controlled ablations under the same training setup. Replacing

pairwise with pointwise evidence-test construction reduces performance from 77.77% to 68.51%, while replacing evolving with fixed evidence tests reduces it to 46.57% (Tab. 4, Tab. 3). These matched comparisons highlight both pairwise construction and evolving evidence tests contributing to the PrefEval performance. Overall, benchmark-specific characteristics may partly explain the larger PrefEval gain, while RLCE also improves performance across the three other multimodal benchmarks (IntentQA, MOSEI, UR-FUNNY).

Extension to test-time training. We further evaluate RLCE under a standard setup for test-time training (TTT), where the model adapts directly to the unlabeled test distribution. Following TTRL (Zuo et al., 2026), all methods use the same base initialization and adaptation protocol, with gold annotations withheld until final evaluation. From Tab. 2, TTRL exhibits mixed performance: with rollout group size $G{=}4$, it improves over the base model on IntentQA ($+0.75$) and UR-FUNNY ($+2.70$), but degrades on MOSEI (-1.39) and PrefEval (-8.20). Increasing to $G{=}16$ reverses some of these trends, improving MOSEI ($+3.95$) while degrading the other three tasks. This is consistent with the majority-based pseudo-labeling that TTRL is built on. Increasing G produces a more stable estimate of the model's prevailing prediction, but does not ensure that prediction is correct; when the majority is wrong, a larger G can instead reinforce the error (Arazo et al., 2020). In contrast, RLCE uses $G{=}4$ and improves over the base model on all four benchmarks by leveraging comparative evidence instead of answer frequency.

Table 2: **Test-time training performance.** All methods are initialized from the same Qwen3.5-9B base model. Δ is relative to the base model (green/red denote improvement/regression). Following TTRL (Zuo et al., 2026), we scale test-time adaptation with test-set size, and report performance at the end of the allocated adaptation budget: three passes for IntentQA, UR-FUNNY, and PrefEval, and two for the larger MOSEI test set, using the same dataset-specific budget across methods. G denotes rollout group size; we use $G{=}4$ for RLCE and report TTRL with $G{=}4$ and an expanded-rollout $G{=}16$ setting. †PrefEval is the macro-average over its three task settings (Choice, Persona, Explicit). Overall, RLCE improves over the base model on all four benchmarks, while TTRL exhibits more variable performance changes with both $G{=}4$ and $G{=}16$.

Method	IntentQA (↑)	Δ	MOSEI (↑)	Δ	UR-FUNNY (↑)	Δ	PrefEval† (↑)	Δ
Qwen3.5-9B *(base)*	55.25	–	64.86	–	63.11	–	44.72	–
TTRL ($G{=}4$)	56.00	+0.75	63.47	−1.39	65.81	+2.70	36.52	−8.20
TTRL ($G{=}16$)	53.84	−1.41	68.81	+3.95	60.13	−2.98	18.26	−26.46
RLCE ($G{=}4$) (ours)	**58.34**	+3.09	**70.65**	+5.79	**70.33**	+7.22	**73.70**	+28.98

4.2 Analysis

A larger share of RLCE's reward variation distinguishes correct from incorrect predictions. In GRPO, within-group reward differences determine the relative advantages that drive the policy update (Shao et al., 2024). Reward variation that separates correct from incorrect predictions therefore provides a learning signal more directly aligned with task performance. As shown in Fig. 2, on MOSEI, throughout training, roughly 65-85% of RLCE's reward variation occurs between correct and incorrect predictions, compared with about 30-36% for RLER and RaR. This suggests that RLCE directs a larger share of its learning signal toward separating correct from incorrect predictions.

RLCE more often corrects erroneous consensus preferences. A potential challenge for self-evolution is confirmation bias: when a model uses its own incorrect predictions as supervision, subsequent training can reinforce those errors (Arazo et al., 2020). We examine a related failure mode through USC, which derives supervision from agreement or consensus among the policy's own rollouts. Among cases where USC's consensus selects an incorrect answer, RLCE more often assigns higher reward to the correct answer relative to rubrics methods RaR and RLER. From Fig. 2 (Middle), RLCE does so in 56% of cases on UR-FUNNY, compared with 48% for RaR and 49% for RLER; on MOSEI, the corresponding rates are 51% (RLCE), 41% (RaR), and 37% (RLER), respectively. These results suggest that RLCE provides a stronger corrective signal than the rubric baselines considered when policy consensus is incorrect.

RLCE can revise erroneous Policy-Judge preferences. We next consider the Policy LLM-Judge baseline, where the frozen policy directly judges the candidate answers. This comparison evaluates whether RLCE can overturn an incorrect direct judgment made by the same frozen policy used to validate its evidence tests. This tests whether RLCE's verification procedure can yield a different preference from the policy's own direct assessment. From Fig. 2 (Right), on UR-FUNNY, RLCE

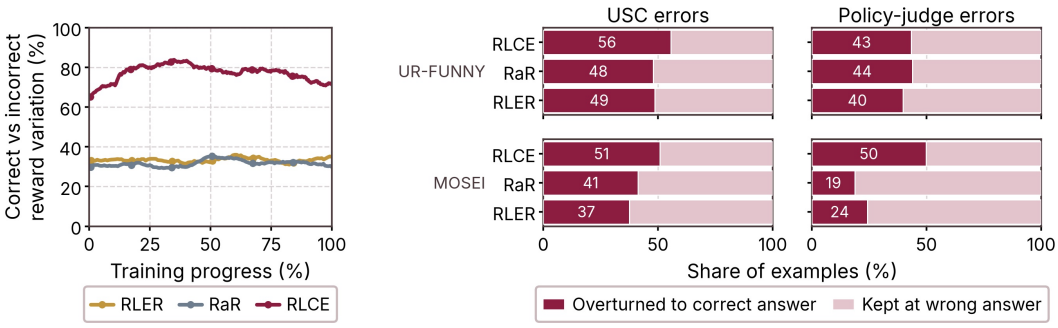


Figure 2: **Analysis of reward variation and correction of erroneous selections.** (Left) On MOSEI, among rollout groups with reward variation, the fraction of total reward variation that distinguishes correct from incorrect predictions, comparing RLCE, Rubrics as Rewards (RaR), and RLER. (Middle, Right) For RLCE, RaR, and RLER methods, among the cases where Universal Self-Consistency (USC) or Policy LLM-Judge selects an incorrect rollout over a correct one, fraction in which each method reverses this ordering by assigning the correct rollout higher reward, on MOSEI and UR-FUNNY.

overturns 43% of Policy-Judge errors, comparable to RaR (44%) and RLER (40%); on MOSEI, RLCE overturns 50%, compared with 19% (RaR) and 24% (RLER), respectively. Despite using the policy LLM as its validator, RLCE overturns at least 43% of erroneous Policy LLM-Judge decisions across both benchmarks, indicating that its verification procedure need not reproduce the policy's direct judgment, with a larger difference from the rubric baselines on MOSEI.

4.3 Ablations

Evolving evidence tests are particularly beneficial for free-form outputs. We compare RLCE's evolving evidence tests with static tests fixed throughout training. From Fig. 3 (Right), evolving tests generally sustain improvement in validation performance, whereas static tests tend to plateau or decline. The separation is substantially larger on the free-form PrefEval tasks than on closed-label MOSEI, where the answer space is fixed and the corresponding evidence tests have less opportunity to change as the policy evolves. From Tab. 3, test performance shows the same pattern: evolving tests improve modestly over static tests on MOSEI (71.73% vs. 69.40%), but yield larger gains on PrefEval (77.77% vs. 46.57%). These results support updating evidence tests around the policy's current answers, particularly in free-form settings where those answers can change substantially during training.

Table 3: **Ablation comparing evolving and static evidence tests across a single training run.** Evolving evidence tests yield stronger performance than static tests, with a larger performance difference on free-form PrefEval compared to closed-label MOSEI. †PrefEval reports the macro average across three task settings (Choice, Persona, Explicit).

Reward	MOSEI	†PrefEval
Static evidence tests	69.40	46.57
Evolving evidence tests	**71.73**	**77.77**

Compositional aggregation reduces sensitivity to errors in individual evidence tests. Across a training run, RLCE validates multiple evidence tests for each example and aggregates their outcomes into a verdict identifying the best-supported answer. Fig. 3 (Left, top panel) examines how this verdict changes as the fraction of incorrect constituent tests increases, i.e., tests whose validated outcomes support an incorrect answer. With 0-20% incorrect tests, the aggregate verdict is correct for 96% of MOSEI examples and 100% of UR-FUNNY examples; with 20-40% incorrect tests, it remains correct for 87% and 86%, respectively. Accuracy declines more substantially once incorrect tests account for roughly 40-60% or more of the constituent tests. These results suggest that aggregating multiple evidence tests can reduce sensitivity to errors in individual tests.

Aggregate verdict accuracy increases as more evidence tests are composed. Fig. 3 (Left, bottom panel) varies the number of evidence tests composed into an aggregate verdict. For each training example, we randomly add tests one at a time without replacement. Verdict accuracy increases from 54% to 69% on MOSEI and from 56% to 66% on UR-FUNNY as the number of composed tests increases from 1 to 12. These results indicate that composing additional evidence tests can improve the reliability of the aggregate verdict.

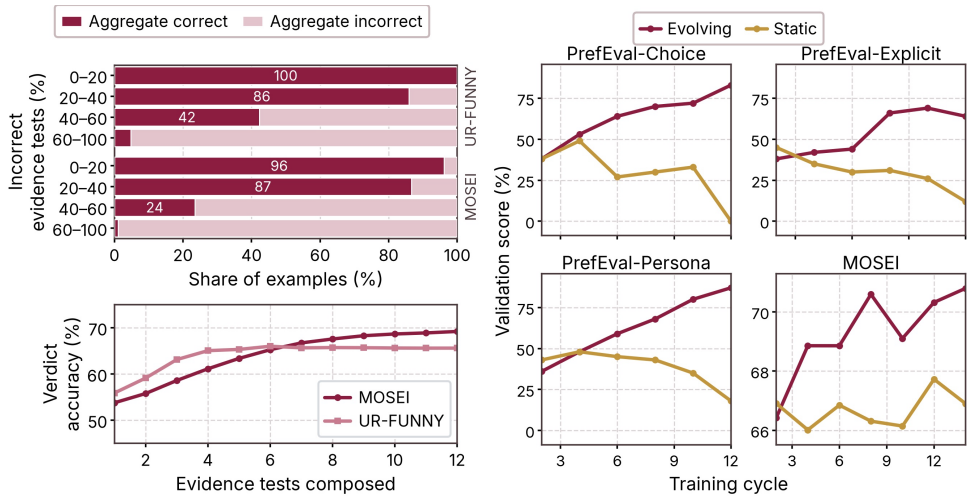


Figure 3: **Analysis of compositional aggregation and evolving evidence tests.** (Left) top panel: On training examples, RLCE's aggregated verdict on the best-supported answer is correct on 86-100% of cases when 0-40% of individual evidence tests are incorrect; bottom panel: verdict accuracy also increases as more evidence tests are composed on MOSEI and UR-FUNNY. For verdict accuracy, MOSEI uses positive and negative polarity, while UR-FUNNY uses binary humor labels, consistent with Ong et al. (2026b). (Right) Regenerating evidence tests as the policy evolves (i.e. evolving tests) yields higher validation performance than keeping them fixed (i.e. static tests), with larger performance gains on the free-form PrefEval tasks than on closed-set MOSEI.

Pairwise evidence tests provide additional signal over pointwise tests. We isolate the pairwise construction component of RLCE by replacing pairwise evidence tests with a pointwise variant that constructs tests independently for each answer and assigns reward when a majority of tests are satisfied, while keeping the remaining components unchanged. We evaluate this ablation on IntentQA and PrefEval, whose answers are not restricted to a small label set, allowing more distinct answers, and hence more pairwise comparisons, to arise within each rollout group. As shown in Tab. 4, RLCE outperforms the pointwise variant on both PrefEval (77.77% vs. 68.51% macro average) and IntentQA (59.61% vs. 56.61%). These results indicate that pairwise construction can provide additional signal over pointwise construction.

Table 4: **Ablation comparing pairwise and pointwise evidence-test construction across a single training run.** Pairwise evidence test construction yields a stronger performance than pointwise construction on IntentQA and PrefEval. †PrefEval reports the macro average across its three task settings (Choice, Persona, Explicit).

Reward	IntentQA	PrefEval†
Pointwise evidence tests	56.61	68.51
Pairwise evidence tests	**59.61**	**77.77**

5 Conclusion

In this work, we introduced Reinforcement Learning with Comparative Evidence, an RL framework that learns social understanding from unlabeled training examples, with rewards constructed without ground-truth annotations. Across four core and diverse benchmarks spanning affect, pragmatics, intent, and preference, RLCE achieves the strongest performance among the baselines. Our analyses further show that a larger share of RLCE's reward variation distinguishes correct from incorrect predictions than the compared rubric methods, and that RLCE can revise erroneous policy-derived preferences. Controlled ablations also highlight the utility of pairwise evidence-test construction, compositional test aggregation, and on-policy test evolution for verification. Overall, these results suggest that RLCE can provide a useful learning signal for improving social understanding from unlabeled training examples across core tasks spanning affect, pragmatics, intent, and preference. More broadly, we hope the same approach may be useful in other settings where correctness is ambiguous and cannot be directly checked by a verification oracle, and may offer a useful direction for learning from unlabeled training data.

AI USE STATEMENT

We used generative AI tools solely to aid and polish the writing of this paper, including improving clarity, grammar, and readability. We did not use generative AI for research ideation or execution, retrieval or discovery of related work, generation of synthetic data, or formulation or proof of mathematical claims. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work trains models to infer affect, humor, intent, and preferences from recordings of people, using existing public research benchmarks (CMU-MOSEI, UR-FUNNY, IntentQA, PrefEval) as released by their authors; no new human-subject data were collected. Automated inference of latent human states carries well-known risks. Such inferences can be wrong in ways that differ across demographic and cultural groups, and can be misused for profiling or surveillance. Label-free self-training adds a specific risk. Because rewards are derived from the model's own judgments about what observable behavior means, the procedure can entrench priors of the base model rather than correct them, and removing annotators removes one safeguard against this. Our results should therefore be read as a study of learning signals on research benchmarks, not as evidence that the resulting models are fit for consequential decisions about individuals. We did not evaluate subgroup-disaggregated performance, which the benchmarks used here do not support well, and consider this necessary before any deployment.

REPRODUCIBILITY STATEMENT

The appendix specifies the datasets, splits, metrics, evaluation harness, and answer-extraction rules (App. A); the full configuration of RLCE, including the reward, GRPO hyperparameters, decoding settings, and prompt design (App. C); and the re-implementation of every baseline with its prompts, hyperparameters, deviations from the original method, and compute (App. D). Main results are reported as means and standard deviations over seeds (Tab. 6). All code and models will be released after the review process.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Cynthia Lynn Breazeal. *Sociable machines: Expressive social exchange between humans and robots*. PhD thesis, Massachusetts Institute of Technology, 2000.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.

Xin Guan, Xiaomeng Hu, Shen Huang, Zhenyi Wang, Bo Zhang, Zijian Li, Pengjun Xie, Bo Liu, and Jiuxin Cao. Evorubric: Self-evolving rubric-driven rl for open-ended generation. *arXiv preprint arXiv:2605.29847*, 2026.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. 2026:127924–127945, 2026.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, and Jianyuan Zhong. Ur-funny: A multimodal language dataset for understanding humor. In *EMNLP-IJCNLP*, pp. 2046–2056, 2019.

Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. Visplay: Self-evolving vision-language models from images. *arXiv preprint arXiv:2511.15661*, 2025.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11963–11974, 2023.

Rui Liu, Dian Yu, Zhenwen Liang, Yucheng Shi, Tong Zheng, Runpeng Dai, Haitao Mi, Pratap Tokekar, et al. Deltarubric: Generative multimodal reward modeling via joint planning and verification. *arXiv preprint arXiv:2605.09269*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. Advancing social intelligence in ai agents: Technical challenges and open questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20541–20560, 2024.

Leena Mathur, Marian Qian, Paul Pu Liang, and Louis-Philippe Morency. Social genome: Grounded social reasoning abilities of multimodal models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24879–24902, 2025.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3470–3487, 2022.

Keane Ong, Sabri Boughorbel, Luwei Xiao, Chanakya Ekbote, Wei Dai, Ao Qu, Jingyao Wu, Rui Mao, Ehsan Hoque, Erik Cambria, et al. Omnisapiens: A foundation model for social behavior processing via heterogeneity-aware relative policy optimization. In *Proceedings of ICML*, 2026a.

Keane Ong, Wei Dai, Carol Li, Dewei Feng, Hengzhi Li, Jingyao Wu, Jiaee Cheong, Rui Mao, Gianmarco Mengaldo, Erik Cambria, et al. Human behavior atlas: Benchmarking unified psychological and social behavior understanding. In *International Conference on Learning Representations*, volume 2026, pp. 44215–44246, 2026b.

Rosalind W Picard. *Affective computing*. MIT press, 2000.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G Finlayson, David Sontag, et al. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 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.

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. Uncertainty-based abstention in llms improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*, 2024.

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE transactions on affective computing*, 3(1):69–87, 2011.

Jianze Wang, Kunwang Zheng, Ying Liu, Yu Cao, Qilong Zhang, Jinlong Chen, Hua Yang, and Qianglong Chen. Serpo: Self-evolving rubric policy optimization for open-ended test-time reinforcement learning. *arXiv preprint arXiv:2607.26873*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Mian Wu, Gavin Zhang, Sewon Min, Sergey Levine, and Aviral Kumar. Rlac: Reinforcement learning with adversarial critic for free-form generation tasks. 2026:139470–139492, 2026.

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

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*, pp. 2236–2246, 2018.

Ruixu Zhang, Deyi Ji, Lanyun Zhu, Xuanyi Liu, Yuxin Meng, Ruihang Chu, and Yujiu Yang. Video-zero: Self-evolution video understanding. *arXiv preprint arXiv:2605.14733*, 2026a.

Yiqun Zhang, Xiaocui Yang, Xingle Xu, Zeran Gao, Yijie Huang, Shiyi Mu, Shi Feng, Daling Wang, Yifei Zhang, Kaisong Song, et al. Affective computing in the era of large language models: A survey from the nlp perspective. *Knowledge-Based Systems*, pp. 115411, 2026b.

Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *Advances in Neural Information Processing Systems*, 38:105816–105879, 2026.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. Do llms recognize your preferences? evaluating personalized preference following in llms. In *International Conference on Learning Representations*, volume 2025, pp. 15888–15931, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.