NOT JUST SATURATION: EVENT-STRUCTURED REWARDS FOR OPEN-ENDED POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Judge scores for responses to the same prompt can concentrate as an open-ended generation policy improves. The remaining differences may poorly reflect candidate quality and misdirect group-relative updates. Fine-grained evaluation exposes defects overlooked by holistic scores, but direct aggregation can distort rewards: one defect may be reported under several criteria, multiple independent defects may collapse into one criterion score, and excessive penalties can reverse a reasonable preference. We introduce CARMA, which separates the identification of independent defects from their reward costs. A verifier groups evidence-linked diagnoses resolved by one shared semantic repair into an event. Each event contributes its largest supported member penalty, and same-prompt preferences calibrate 24 shared severity increments. Event penalties adjust a separately obtained holistic score, preserving complementary quality information. Across private Chinese and public English writing tracks on Qwen3-8B and Qwen3.5-27B, CARMA improves policy preference over holistic-reward group-centered training. Offline audits show more accurate near-tie rankings and closer agreement with reference advantages than holistic scoring and ungrouped diagnostic aggregation. Gains in pairwise and feedback-driven settings support reusing events across training interfaces.

1 Introduction

The writing-quality gap. A strong writing model can produce fluent same-prompt responses with almost identical judge scores. Creative writing requires coherence, style, and instruction-following judgments beyond correctness checks. Post-training uses human-preference rewards (Christiano et al., 2017; Ouyang et al., 2022) or language-model judgments (Zheng et al., 2023). GRPO converts rewards into policy advantages (Shao et al., 2024); meaningful quality differences must survive.

Reward reliability. Homogeneous rewards can eliminate group-relative advantages (He et al., 2026), but concentration alone does not show that useful information is lost. Positive affine compression preserves order and standardized direction while shrinking mean-centered training weights. Reward models can misjudge subtle content differences (Liu et al., 2025a); optimizing imperfect rewards can improve scores without improving quality (Gao et al., 2023). Chasing the Tail emphasizes accuracy in high-reward regions (Zhang et al., 2026a). We study evidence-linked diagnoses for closely matched candidate responses under a shared writing task.

Independent defects. Rubric rewards specify criteria; evidence-grounded judges connect assessments to text (Gunjal et al., 2026; Hong et al., 2026). Yet a diagnosis need not identify an independent defect. An unexplained goal change can trigger weak-motivation, characterization, and causal-transition diagnoses. One clarification can resolve all three, so repeated charges overcount. Conversely, an age inconsistency and an unrelated key-state inconsistency require separate repairs, although a binary continuity criterion records one violation (Figure 1). Overpenalizing a minor flaw can reverse a reasonable preference. Rewards need independent penalty units and calibrated costs.

Aggregation and learning. Aggregation errors shift the group mean and every response's advantage. Prior work studies multi-reward failures (Yuan et al., 2026), recursive rubric refinement (Shen et al., 2026), and preference-learned criteria (Xie et al., 2025). Tournament comparisons and tex-

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.

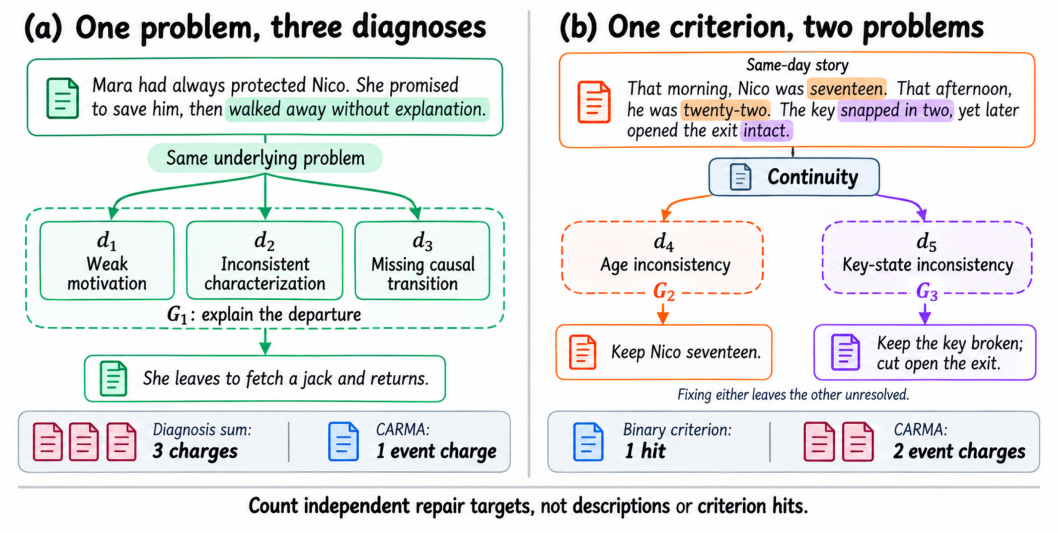


Figure 1: **Independent repair targets determine penalty counts.** (a) Three diagnoses of Mara's unexplained departure share one causal repair and incur one event charge. (b) The age and key-state inconsistencies fall under one continuity criterion but require independent repairs and incur two event charges. Figure 2 develops CARMA's reward construction for the same example.

tual feedback offer complementary interfaces (Zhang et al., 2026b; Song et al., 2026). We identify independent problems and calibrate their shared costs from same-prompt preferences.

Event-structured rewards. We propose CARMA, separating what the problem is from how much it should cost. A frozen proposer supplies evidence-linked diagnoses; a verifier groups those resolved by one shared semantic repair. Each event incurs its largest supported member penalty; independent penalties adjust a separate holistic score. Same-prompt preferences calibrate 24 nonnegative severity increments across eight families. With fixed diagnoses and events, copying preserves rewards and centered advantages for the entire same-prompt rollout group.

Evaluation and contributions. We evaluate CARMA through Qwen3-8B and Qwen3.5-27B writing policy training, fixed-candidate audits, WritingBench, and LitBench (Wu et al., 2025; Fein et al., 2026). Internal-track preference reaches 60.4% and 58.3% against the same-backbone holistic reference. Audits and matched-supervision ablations test counting, calibration, and near-tie ordering; ArenaRL and RLTF-SD integrations test comparison and feedback interfaces. Our contributions are:

- **Analysis.** We show how diagnostic aggregation distorts local reward rankings and group-relative advantages, distinguishing score dispersion from learning-signal reliability.
- **Method.** We group diagnoses by shared repair before learning defect costs, obtaining invariance to exact within-event replication under a fixed representation.
- **Evaluation.** We test reward construction through controlled audits, matched-supervision ablations, two-backbone policy training, and pairwise and feedback integrations.

2 Reward Scale and Group-Relative Advantages

Training advantages. For K responses to one prompt, let $r \in \mathbb{R}^K$ be their rewards, $\bar{r} = K^{-1}\sum_i r_i$ their mean, and $\sigma(r)^2 = K^{-1}\sum_i (r_i - \bar{r})^2$ their variance. Magnitude-preserving group centering (MPGC) uses $A_i^{\text{train}} = r_i - \bar{r}$, without reward-STD normalization, following the mean-centering choice in Dr. GRPO (Liu et al., 2025b).

Offline direction audit. For geometric analysis and the offline cosine audit, we also define the standardized direction

$$A_i^{\text{std}}(r) = \frac{r_i - \bar{r}}{\sigma(r)}, \qquad \sigma(r) > 0. \tag{1}$$

We use this direction for offline analysis; training uses unnormalized centered rewards.

Observation (Affine compression preserves direction, not magnitude). *For $a > 0$ and $\sigma(r) > 0$, $A^{\mathrm{train}}(ar + b\mathbf{1}) = aA^{\mathrm{train}}(r)$ and $A^{\mathrm{std}}(ar + b\mathbf{1}) = A^{\mathrm{std}}(r)$.*

Thus affine score compression shrinks the actual reward-derived update weights, while preserving their direction and ranking. It does not establish that the remaining differences are valid quality signals. Quantization, clipping, and newly introduced ties require separate treatment.

Effects of aggregation errors. A response-specific aggregation error changes its reward and, through the group mean, every centered advantage. Standardization adds sensitivity through the denominator. Even an order-preserving nonlinear transform can change both the relative update weights and the positive or negative signs of individual advantages (Appendix B).

We audit local rankings and agreement with reference advantages from separate GRM evaluation calls. Updates also depend on score-function gradients, importance ratios, clipping, and KL regularization; identical rewards yield zero reward-derived advantages. Related studies examine GRPO length and difficulty biases (Liu et al., 2025b; Yang et al., 2026), DAPO clipping and sampling (Yu et al., 2025a), and GSPO sequence-level ratios (Zheng et al., 2025). We examine their reward inputs.

3 Learning Rewards from Independent Defects

3.1 From Criteria to Independent Defects

A criterion names an aspect of quality; a diagnosis reports a violation. Neither defines an independent penalty unit. Binary hits $h(x, y) \in \{0, 1\}^C$ and holistic score $S(x, y)$ can be identical for responses containing one shared defect or several independent defects. A scorer using only (h, S) cannot distinguish them. Diagnosis addition retains detail but can repeat charges. CARMA groups reports before learning costs from same-prompt preference comparisons.

3.2 Grouping Diagnoses by Shared Repair

A frozen judge proposes $\mathcal{D}(x, y) = \{d_i\}_{i=1}^N$ under a fixed domain specification. Diagnoses contain family, severity, confidence, claim, relation type, literal evidence, and canonical tags. Contradictions may cite multiple spans; omissions may cite task requirements and response context. The proposer receives neither preference labels nor holistic scores during evidence extraction.

Shared repair. The partition $\mathcal{P}(x, y)$ groups diagnoses of the same violated requirement or relation. Repair targets identify the affected entity or action, relation, and contextual scope. The verifier supplies a repair resolving every member and preserving requirements and unrelated facts.

Separate repair targets. Diagnoses remain separate when they concern distinct validated repair targets. A shared event requires one scoped, requirement-consistent repair of the same underlying violation that resolves every member. Changing "two red cups" to "one blue cup" fixes quantity and color in one edit, but they remain distinct semantic targets. Several diagnoses of an unexplained goal change can share one causal clarification. All members need a common repair; pairwise similarity chains are insufficient.

Verification. Evidence and repair targets justify grouping before cost fitting. Shared preprocessing validates evidence and removes exact duplicates for CARMA and matched diagnostic controls. Every accepted diagnosis is assigned to an event once; invalid partitions use singletons (Appendix F).

3.3 Computing Event Penalties

The canonical tags define a cached binary support map $q_{ci} \in \{0, 1\}$ for each diagnosis. For a collection of criterion descriptions $\mathcal{C}$, let

$$u_i(\mathcal{C}) = p_i \max_{c \in \mathcal{C}} q_{ci}, \tag{2}$$

Here $p_i \in [0, 1]$ is evidence confidence; an empty criterion collection gives zero. This map gates canonical-tag coverage, rather than rechecking evidence or assigning another confidence. With a covered tag, $u_i = p_i$. Criterion views use cached support, so repetition cannot multiply support.

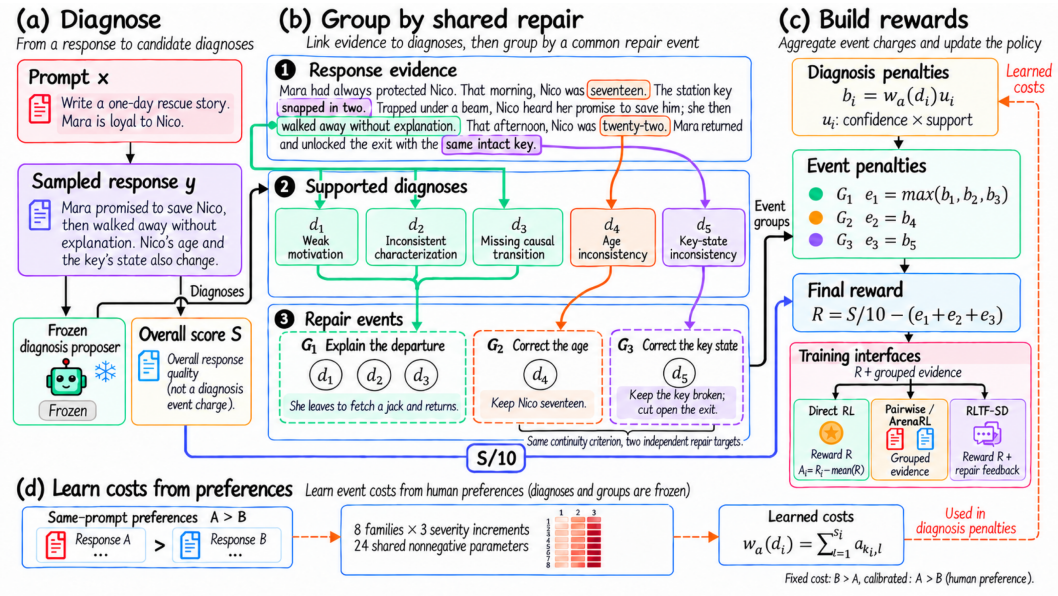


Figure 2: **CARMA reward construction, calibration, and interfaces.** (a) Frozen diagnoses and a separate overall score; the response is abbreviated (full input: Appendix J). (b) Shared repair groups d_1–d_3; age and key state remain independent. (c) Event maxima adjust $S/10$ and supply scalar, pairwise, and feedback interfaces. (d) Preferences fit 24 shared increments with other inputs fixed. The heatmap and human-preference example are schematic; public-track calibration uses GRM preferences (Appendix I).

Diagnosis, event, and response penalties. We compute rewards in three steps:

$$b_i(a;\mathcal{C}) = w_a(d_i)u_i(\mathcal{C}), \qquad e_G(a;\mathcal{C}) = \max_{i\in G} b_i(a;\mathcal{C}),$$
$$R_a(x,y;\mathcal{C}) = \frac{S(x,y)}{10} - \sum_{G\in\mathcal{P}(x,y)} e_G(a;\mathcal{C}). \tag{3}$$

Here $\mathcal{C}$ contains the response's tagged criteria. Each b_i combines one complete diagnosis's cost, confidence, and support. The maximum selects one complete member penalty. The sum charges independent events; singleton events recover learned diagnostic addition without grouping.

Overall quality. The separate score $S \in [0, 10]$ captures quality beyond diagnosed defects. Its fixed coefficient anchors costs to the normalized scale; Section 5.5 tests its removal. Confidence is a fixed input, not a calibrated probability. Cross-event additivity is a modeling assumption.

Worked example. Figure 2 follows the story in Figure 1. Explaining that Mara leaves to fetch a jack resolves diagnoses d_1, d_2, d_3, forming event G_1. The age and key-state inconsistencies require independent repairs, forming G_2 and G_3 despite sharing the continuity criterion. Their combined penalty is $\max(b_1, b_2, b_3) + b_4 + b_5$.

3.4 Calibrating Shared Costs from Preferences

We share nonnegative severity increments across diagnoses of the same family and severity; individual events and diagnoses have no free parameters. For family k_i and severity $s_i \in \{1, 2, 3\}$,

$$w_a(d_i) = \sum_{\ell=1}^{s_i} a_{k_i,\ell}, \qquad a_{k,\ell} \geq 0. \tag{4}$$

Eight families $k \in \{1, \ldots, 8\}$ and three increments $\ell \in \{1, 2, 3\}$ give 24 coefficients. Higher severity cannot reduce cost with family, confidence, and support fixed; severity levels need not have comparable costs across families.

Let $\mathcal{Q}$ contain strictly ordered same-prompt preference pairs (y_i, y_j, z_{ij}), with $z_{ij} \in \{-1, +1\}$ marking the preferred response. Tie annotations are outside this binary calibration set. We fit

$$\begin{aligned} \Delta R_{ij}(a) &= R_a(x, y_i; \mathcal{C}_i) - R_a(x, y_j; \mathcal{C}_j), \\ \ell_{ij}(a) &= \log(1 + \exp\left[-z_{ij}\Delta R_{ij}(a)/\tau\right]), \\ \mathcal{L}_\eta(a) &= \frac{1}{|\mathcal{Q}|} \sum_{(i,j)\in\mathcal{Q}} \ell_{ij}(a) + \frac{\eta}{2}\|a\|_2^2, \qquad a \geq 0. \end{aligned} \tag{5}$$

Here $\tau > 0$ is the margin temperature and $\eta \geq 0$ the regularization weight. Each pair shares prompt x. Scores, diagnoses, supports, and partitions are fixed; only 24 costs are fitted. Including S calibrates penalties against the training reward. Fitting penalties before adding S optimizes a different margin. This is a binary preference objective (Ouyang et al., 2022; Rafailov et al., 2023).

Costs can change an event's maximal member, making the objective nonconvex and piecewise smooth. Bound-constrained L-BFGS (Byrd et al., 1995) uses two fixed starts per regularization value. We retain the lower training objective and select on disjoint validation prompts.

CARMA's scalar reward drives group-relative training; events also provide ArenaRL comparator evidence (Zhang et al., 2026b) and RLTF-SD critiques (Song et al., 2026) while preserving their optimizers (Figure 2c). We evaluate reward-only, feedback-only, and joint variants (Appendix E).

4 Invariance and Sensitivity of Event Rewards

With diagnoses, event partitions, supports, holistic scores, and costs fixed, reordering reports or events, exactly copying a diagnosis within its event, and repeating cached criterion support preserve Eq. 3 (Proposition 1, Appendix A). Rewards, MPGC advantages, and standardized audit directions remain unchanged under a fixed zero-variance convention. Semantic redescription is tested separately in Section 5.2; exact deduplication also preserves rewards under copying.

Support changes are bounded by the greatest affected member penalty in each event (Proposition 2). Centering cannot amplify reward errors in ℓ_2 norm, whereas standardized directions are sensitive to small group variance (Eq. 8). Even without STD normalization, aggregation errors can dominate meaningful quality differences between candidates in a narrow same-prompt reward group.

Falsely merging independent penalties $u, v \geq 0$ raises reward by $\min(u, v)$; false splits can multiply a defect's cost. Separate defects may interact in human preference, so cross-event additivity remains an assumption. Section 5.5 tests the training effect of injected merges.

5 Experiments

5.1 Experimental Setup

Data. We use two tracks: private Chinese creative writing and public English writing. Public data mix WritingPrompts (Fan et al., 2018) and REER's DeepWriting-20K (Wang et al., 2025) at 70:30. Each track uses 4,000 reward-training, 500 development, 8,000 policy-training, and 400 evaluation prompts. Appendix I gives sources; audits use separate internal sets.

Policies and optimization. Qwen3-8B and Qwen3.5-27B (Yang et al., 2025; Qwen Team, 2026) sample eight responses per prompt over 200 rollout iterations. Methods share a checkpoint within each backbone and track after task SFT and prior post-training. Generation caps are 1,024 tokens internally and 8,192 publicly, without extra length-based filtering or penalties. 8B uses three seeds; 27B uses one run with prompt-bootstrap uncertainty. Appendix D gives optimization settings. WritingBench evaluates generation (Wu et al., 2025); LitBench evaluates fixed-pair reward ordering (Fein et al., 2026).

Comparisons. Table 2 compares controlled rewards, full pipelines, and training interfaces. Controlled rewards share MPGC. Holistic uses a separate GRM call; the learned diagnostic sum $S/10 - \sum_i w_a(d_i)u_i(\mathcal{C})$ shares CARMA's diagnoses, 24 costs, and preferences. Multi-aspect scoring averages six scores from one separate call (Appendix J). RaR (Gunjal et al., 2026) and RRD-WU (Shen et al., 2026) retain their rubrics; RRD-WU uses MPGC here. Interfaces include pairwise rank rewards, ArenaRL (Zhang et al., 2026b), and RLTF-SD (Song et al., 2026).

Table 1: **Controlled reward-counting tests** (%). Rephrasing flips use 150 pairs per condition; added-defect response and minor-flaw ranking use 300 each. Scoring is rerun for added defects. Exact-copy controls are described in the text.

Reward	Same label ↓	Cross label ↓	Added defect ↑	Minor flaw ↑
Criterion hits	**0.0**	47.3	53.3	58.0
Learned diagnostic sum	51.3	58.7	87.0	61.0
Semantic deduplication	<u>0.7</u>	6.7	85.3	<u>65.3</u>
Fixed event costs	**0.0**	<u>5.3</u>	<u>90.0</u>	62.7
CARMA	1.3	**4.0**	**90.3**	**74.0**

Evaluation design. Fixed-candidate audits apply each reward to the same saved responses, isolating reward construction. Policy comparisons train each method separately and evaluate generated responses. Pale pink marks CARMA configurations; bold and underlining identify the best and second-best results over the comparison scope stated in each table caption.

5.2 Controlled Tests of Reward Counting

Interventions. Repeated-diagnosis tests fix the response and defect while rephrasing its diagnosis under the same criterion tag or distinct tags. The independent-defect test adds an unexplained age change to a response with another continuity error; the two require separate repairs. Minor-flaw pairs have adjudicated overall preferences, including flaws that preserve or change the preferred response. Fixed and learned costs use the same responses and partitions.

The exact-copy control duplicates a complete diagnosis within its existing event after extraction, fixing all other scoring inputs. Grouping methods flip no pairs in either tag condition; criterion hits flip no same-label pairs, while diagnostic addition can charge the copy twice. Table 1 instead reports semantic redescription: fields and event assignment may change when the same defect is described again. Appendix D.1 separates this empirical robustness test from the exact-copy control.

Results. Same-label rephrasing flips 51.3% of diagnostic-sum rankings, versus at most 1.3% for criterion hits and grouping. Cross-label rephrasing affects criterion hits (47.3%), versus at most 6.7% for grouping. Adding an independent defect changes criterion-hit rewards on 53.3% of items and diagnosis-based rewards on 85.3–90.3%; rerunning the holistic judge can change rewards even with fixed hits. Learned costs preserve more minor-flaw preferences (74.0%) than fixed costs (62.7%).

5.3 Ranking and Advantage Agreement

Table 7 and Appendix B audit aggregation with shared diagnosis precision/recall (84.0%/76.0%). Holistic scoring, diagnostic addition, and CARMA correctly order 35/60, 31/60, and 44/60 near-tie pairs. Across 200 eight-response groups, advantage cosines are 0.79, 0.71, and 0.91. Diagnostic addition has fewer dead groups but poorer agreement with reference advantages from separate GRM evaluation calls.

On 120 held-out pairs, human agreement is 101/120, judge–human accuracy 62/120, and judge repeatability 85/120. The 37 stable errors separate repeatability from validity (Appendix C).

5.4 Policy Training on Private and Public Data

Each column of Table 2 uses its own holistic-reward policy as the 50% reference.

Private-data training. CARMA reaches 60.4% on 8B and 58.3% on 27B; diagnostic addition and multi-aspect scoring stay near the reference. Section 5.5 tests grouping, evidence, and costs.

Public-data training. Public costs are fitted on English preferences. The public comparison places CARMA at 59.8% on 8B and 57.9% on 27B. Table 2 reports the full comparisons; rates across tracks have different reference policies and are not direct cross-track comparisons.

Human preference. Blinded raters compare private-track 27B CARMA with each opponent on 300 prompts. Against overall-score training, wins/ties/losses are 170/31/99: preference 61.8%, with prompt-bootstrap 95% CI $[56.7, 67.0]$. Against RRD-WU, counts are 140/47/113: preference 54.5%,

Table 2: **Policy preference on two training tracks** (%; 400 prompts each). Each column uses its own overall-score policy as the 50% reference. 8B: three-run mean ± SD; 27B: one run ± prompt-bootstrap SE. Best and second-best are marked across all rows.

Method	Private Chinese		Public English	
	8B	27B	8B	27B
Initial policy	36.5 ± 1.5	41.6 ± 2.3	39.1 ± 1.4	43.7 ± 2.1
A. Controlled rewards				
Overall score (reference)	50.0	50.0	50.0	50.0
Learned diagnostic sum	51.2 ± 1.2	50.6 ± 1.8	52.3 ± 1.6	51.6 ± 2.0
Multi-aspect scoring	52.4 ± 1.7	51.1 ± 1.4	53.4 ± 1.6	52.5 ± 1.9
CARMA	$\underline{60.4 \pm 1.3}$	$\underline{58.3 \pm 2.2}$	$\underline{59.8 \pm 1.4}$	$\underline{57.9 \pm 2.2}$
B. Full reward pipelines				
RaR-Explicit	53.0 ± 2.2	52.7 ± 1.6	54.2 ± 1.3	53.1 ± 2.0
RaR-Implicit	56.1 ± 0.9	55.9 ± 1.9	56.0 ± 1.7	55.2 ± 2.1
RRD-WU	57.4 ± 2.4	56.5 ± 2.0	56.9 ± 1.5	55.8 ± 2.2
C. Training interfaces				
Pairwise rank reward	56.7 ± 1.6	55.7 ± 2.4	56.2 ± 1.6	55.3 ± 2.1
ArenaRL	58.0 ± 1.9	56.4 ± 2.1	57.4 ± 1.6	56.1 ± 2.1
RLTF-SD	57.1 ± 1.4	55.4 ± 1.2	56.8 ± 1.8	55.7 ± 2.0
CARMA + RLTF-SD	$\mathbf{62.1 \pm 1.1}$	$\mathbf{60.3 \pm 1.1}$	$\mathbf{61.5 \pm 1.4}$	$\mathbf{59.7 \pm 2.1}$

Table 3: **Component ablations on private-track Qwen3-8B.** Percent; preference: three-run mean ± SD against overall-score training; ranking: 60 near-tie pairs. All variants share diagnoses and preference supervision.

Variant	Reward rule	Preference ↑	Ranking ↑
Full CARMA	Event max	$\mathbf{60.4 \pm 1.3}$	**73.3**
Within-event mean	Event mean	$\underline{56.7 \pm 1.0}$	65.0
Semantic deduplication	Semantic groups	55.9 ± 1.2	58.3
Fixed event costs	Fixed costs	55.1 ± 1.4	61.7
No overall score	Remove $S/10$	54.6 ± 1.5	$\underline{66.7}$
No event partition	Diagnosis sum	51.2 ± 1.2	51.7
Criterion interactions	Criterion pairs	53.4 ± 1.1	56.7

CI $[49.3, 59.7]$. The latter interval includes 50%. Appendix E.2 adds interface comparisons; all counts refer to distinct opponents in the blinded human comparisons.

5.5 Component Ablations and Grouping Accuracy

Table 3 compares grouping and cost choices on private-track 8B with diagnoses and calibration supervision fixed. Removing the event partition gives diagnostic addition with the same 24 shared costs. Semantic deduplication groups similar descriptions; CARMA groups shared repairs. Separate problems can sound alike, while one problem can cross criteria (Appendix K.1). The remaining rows compare event means, fixed costs, and removal of the holistic branch.

On 150 prompts and 218 diagnosis pairs, the verifier makes 190/218 correct decisions: 81/94 same-event merges (86.2% recall) and 109/124 independent-pair separations (87.9% specificity). False merges (15) and splits (13) are 6.9% and 6.0%, with pairs clustered by prompt.

With costs fixed, injecting false merges and retraining from a common initialization reduces preference from 60.4% to 54.9% as corruption rises from 0% to 20%. Appendix H details the intervention.

Table 4: **Integration with ArenaRL and RLTF-SD** (%; 400 private-track prompts). Each panel uses its original method as the 50% reference. Bold/underline: best/second within a panel.

ArenaRL: comparator evidence			**RLTF-SD: reward and feedback**		
Comparator input	8B	27B	Training signals	8B	27B
Original	50.0	50.0	Original	50.0	50.0
Raw diagnoses	<u>54.6</u>	<u>53.5</u>	Event reward	<u>57.3</u>	<u>56.1</u>
Repair events	**59.1**	**57.4**	Event feedback	55.1	54.3
			Reward + feedback	**60.2**	**58.6**

5.6 Integration with ArenaRL and RLTF-SD

Table 4 compares each added variant with its original method on private data. ArenaRL keeps its comparator, tournament, and objective. Raw diagnoses add evidence; events additionally identify shared semantic repairs, giving 59.1%/57.4% versus 54.6%/53.5% in the comparison.

RLTF-SD variants replace only the reward, only the feedback, or both; the joint variant has the highest preference. Evaluation uses first-round responses without feedback at test time. Appendix E describes the interfaces and feedback quality.

5.7 Continuing from Training Plateaus

Late writing-policy runs show flattening reward and evaluation loss, with occasional mild overfitting. From holistic, RRD-WU, and ArenaRL plateau checkpoints, equal-budget continuations retain the method or add CARMA reward (first two) or event evidence (ArenaRL). Relative to each starting policy, the continuation comparison gives 58.8–60.1% with CARMA and 50.7–51.3% otherwise (Appendix G).

5.8 WritingBench and LitBench Evaluation

WritingBench evaluates new responses from both training tracks. Internal CARMA improves over holistic training and RRD-WU on both backbones; public scores are 7.34/7.56 versus holistic's 6.84/7.21 (Table 5). Its writing tasks differ from the training mixture.

LitBench evaluates fixed story pairs. With the same LitBench-Train preferences and diagnoses, CARMA reaches 76.1% versus 69.5% for diagnostic addition; domain-external frozen costs with English-adapted prompts give 69.8%. Both protocols appear in Appendix D.3.

6 Related Work

Rubric-based rewards. Rubrics specify task criteria for policy training (Chen et al., 2026; Viswanathan et al., 2025; Gunjal et al., 2026; Huang et al., 2025a). RRD removes redundant or downweights correlated criteria (Shen et al., 2026); Auto-Rubric learns criteria from preferences (Xie et al., 2025). RULERS grounds judgments in evidence (Hong et al., 2026), while SARA studies rubric interference (Yu et al., 2026). CARMA defines response-specific penalty units by shared repair before calibrating shared costs.

Fine-grained feedback and writing. Localized feedback, error attribution, and claim-level rewards provide finer supervision (Wu et al., 2023; Xu et al., 2025; Gao et al., 2026; Gandhi et al., 2026). Preference learning and generative rewards support storytelling and open-ended writing (Castricato et al., 2022; Jia et al., 2025; Liao et al., 2026). CARMA studies how multiple defect reports contribute to one reward using matched-supervision ablations and fixed-response audits.

Judge reliability. Judges remain sensitive to length, preference leakage, position, and prompt variation (Huang et al., 2025b; Li et al., 2026; Xu et al., 2026; Bhat & Varma, 2026). RLTF-SD addresses feedback-conditioned advantage collapse with a first-turn reward baseline (Song et al., 2026); CARMA targets defect aggregation. Exact-copy invariance assumes fixed diagnoses and partitions; semantic-redescription robustness is tested empirically. Appendix N gives the broader comparison.

Table 5: **WritingBench generation and LitBench reward ranking.** WritingBench uses a ten-point scale; LitBench reports accuracy (%). Best and second-best are marked within each panel. Benchmark protocols are in Appendix D.3.

WritingBench: generation quality ↑				
Policy	Private Chinese		Public English	
	8B	27B	8B	27B
Overall score	6.62	7.07	6.84	7.21
RaR-Implicit	6.81	7.28	7.01	7.35
RRD-WU	6.94	7.26	<u>7.09</u>	7.39
ArenaRL	6.87	<u>7.30</u>	7.05	<u>7.40</u>
RLTF-SD	<u>6.95</u>	7.24	7.03	7.33
CARMA	**7.16**	**7.43**	**7.34**	**7.56**

LitBench: reward-ranking accuracy		
Reward	Cost fitting	Accuracy ↑
Overall score	—	64.0
RaR-Implicit	Original pipeline	70.0
RRD-WU	Original pipeline	<u>73.8</u>
Learned diagnostic sum	LitBench-Train	69.5
CARMA	Frozen; adapted prompts	69.8
CARMA	LitBench-Train	**76.1**

7 Discussion and Limitations

CARMA separates independently repairable problems from the number of descriptions they receive, then calibrates costs from preferences. Table 3 compares this principle with diagnostic addition and semantic similarity grouping using the same diagnoses and supervision.

Atomic repair is approximate: verification can merge independent targets, split shared targets, or retain invalid diagnoses. The natural audit splits 13/94 same-event pairs (13/218 overall); their training effect is unmeasured. Independent defects can interact in preference, so additivity remains an assumption. Calibration cannot recover positive qualities missed by all scoring inputs.

Evaluation and cost. Reward construction and machine evaluation share the GRM and may share systematic errors, even with separate prompts and calls. Human comparisons provide an additional check, but the interval against RRD-WU includes 50%. Diagnostics, verification, and training-time revision also add computation.

8 Conclusion

CARMA groups diagnoses by shared repair and learns category and severity costs from preferences. One problem incurs one event charge; independent problems incur separate charges. Exact copying preserves rewards and centered advantages under fixed inputs. Audits and policy comparisons test this construction, while comparison and feedback interfaces demonstrate reuse. Plateau continuations examine further improvement after existing training signals flatten.

AI Use Statement

Generative AI tools assisted with manuscript review, interpretation of reported results, consistency checks, and figure preparation. A locally deployed language model assisted with transferring experiment values into LaTeX. Language-model judges form part of the research pipeline (Appendix D). The authors are responsible for the final content and verification against experiment records.

References

Savita Bhat and Vasudeva Varma. All prompts are created equal? evaluating robustness of LLM judges against non-adversarial prompt variations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38730–38745. Association for Computational Linguistics, July 2026. doi: 10.18653/v1/2026.findings-acl.1929. URL https://aclanthology.org/2026.findings-acl.1929/.

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Louis Castricato, Alexander Havrilla, Shahbuland Matiana, Michael Pieler, Anbang Ye, Ian Yang, Spencer Frazier, and Mark Riedl. Robust preference learning for storytelling via contrastive reinforcement learning. *arXiv preprint arXiv:2210.07792*, 2022. URL https://arxiv.org/abs/2210.07792.

Angelica Chen, Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. Learning from Natural Language Feedback. *Transactions on Machine Learning Research*, 2024a. ISSN 2835-8856. URL https://openreview.net/forum?id=xo3hI5MwvU.

Hao Chen, Ziyu Han, Yukun Yan, Qingfu Zhu, Maosong Sun, and Wanxiang Che. From Holistic Evaluation to Structured Criteria: Rubrics Across the Evolving LLM Landscape, 2026. URL https://arxiv.org/abs/2606.08625.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024b. doi: 10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30*, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html.

Ganqu Cui, Lifan Yuan, Zefan Wang, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Litbench: A benchmark and dataset for reliable evaluation of creative writing. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 7740–7755, 2026. doi: 10.18653/v1/2026.eacl-long.362. URL https://aclanthology.org/2026.eacl-long.362/. arXiv:2507.00769.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 1066–1083, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.100. URL https://aclanthology.org/2023.wmt-1.100/.

Shubham Gandhi, Saurabh Goyal, Kiran Kate, and Yara Rizk. DRACO: Fine-grained credit assignment with dynamic rubrics for long-horizon agent training. *arXiv preprint arXiv:2609.04094*, 2026. URL https://arxiv.org/abs/2609.04094.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023. PMLR 202.

Mingqi Gao, Hongyuan Dong, Yifei Chen, Zhisheng Zhong, Zheng Ruan, Wenjin Hou, Yu Chen, Han Hu, and Yansong Tang. Claim-level rubric rewards for video caption reinforcement learning, July 2026. URL https://arxiv.org/abs/2607.05150.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54/.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2507.17746.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2605.21125.

Yihan Hong, Huaiyuan Yao, Bolin Shen, Wanpeng Xu, Hua Wei, and Yushun Dong. From rubrics to reliable scores: Evidence-grounded text evaluation with LLM judges. *arXiv preprint arXiv:2601.08654*, 2026. URL https://arxiv.org/abs/2601.08654. Accepted to EMNLP 2026 Main Conference.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiaxin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, Jianguo Li, and Junbo Zhao. Reinforcement Learning with Rubric Anchors, 2025a. URL https://arxiv.org/abs/2508.12790.

Zeyu Huang, Zihan Qiu, Zili Wang, Edoardo M. Ponti, and Ivan Titov. Post-hoc Reward Calibration: A Case Study on Length Bias. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Iu8RytBaji.

Ruipeng Jia, Yunyi Yang, Yongbo Gai, et al. Writing-zero: Bridge the gap between non-verifiable tasks and verifiable rewards. *arXiv preprint arXiv:2506.00103*, 2025.

Yoonho Lee, Joseph Boen, and Chelsea Finn. Feedback Descent: Open-Ended Text Optimization via Pairwise Comparison, 2025. URL https://arxiv.org/abs/2511.07919.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791. Association for Computational Linguistics, November 2025a. doi: 10.18653/v1/2025.emnlp-main.138. URL https://aclanthology.org/2025.emnlp-main.138/.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference Leakage: A Contamination Problem in LLM-as-a-judge. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2502.01534.

Zongxia Li, Yapei Chang, Yuhang Zhou, et al. Semantically-aware rewards for open-ended R1 training in free-form generation. *arXiv preprint arXiv:2506.15068*, 2025b.

Jianxing Liao, Tian Zhang, Xiao Feng, et al. RLMR: Reinforcement learning with mixed rewards for creative writing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2305.20050.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. RM-Bench: Benchmarking reward models of language models with subtlety and style. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, 2025a. arXiv:2410.16184.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like Training: A Critical Perspective. In *Proceedings of the Second Conference on Language Modeling (COLM)*, 2025b. URL https://arxiv.org/abs/2503.20783.

Renjie Luo, Zichen Liu, Xiangyan Liu, Chao Du, Min Lin, Wenhu Chen, Wei Lu, and Tianyu Pang. Language Models Can Learn from Verbal Feedback Without Scalar Rewards, 2025. URL https://arxiv.org/abs/2509.22638.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. arXiv:2203.02155.

Qwen Team. Qwen3.5-27B model card. Hugging Face, 2026. URL https://huggingface.co/Qwen/Qwen3.5-27B.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Qingyu Ren, Tianjun Pan, Xingzhou Chen, and Xuhong Wang. From coarse to fine: Benchmarking and reward modeling for writing-centric generation tasks. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2796–2810, 2026. doi: 10.18653/v1/2026.findings-acl.134. URL https://aclanthology.org/2026.findings-acl.134/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. Spurious rewards: Rethinking training signals in RLVR. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2506.10947.

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

William F. Shen, Xinchi Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. Rethinking rubric generation for improving LLM judge and reward modeling for open-ended tasks. *arXiv preprint arXiv:2602.05125*, 2026. URL https://arxiv.org/abs/2602.05125.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.acl-long.1230. URL https://aclanthology.org/2025.acl-long.1230/.

Yuda Song, Lili Chen, Fahim Tajwar, Remi Munos, Deepak Pathak, J. Andrew Bagnell, Aarti Singh, and Andrea Zanette. Expanding the capabilities of reinforcement learning via text feedback. *arXiv preprint arXiv:2602.02482*, 2026. URL https://arxiv.org/abs/2602.02482.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2507.18624.

Hanyang Wang, Lu Wang, Chaoyun Zhang, Tianjun Mao, Si Qin, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Text2grad: Reinforcement learning from natural language feedback. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/pdf/a9821c721f942c1437ced59f36bcffafdc8774f1.pdf.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 10582–10592, 2024a. arXiv:2406.12845.

Haozhe Wang, Haoran Que, Qixin Xu, Minghao Liu, Wangchunshu Zhou, Jiazhan Feng, Wanjun Zhong, Wei Ye, Tong Yang, Wenhao Huang, Ge Zhang, and Fangzhen Lin. Reverse-engineered reasoning for open-ended generation. *arXiv preprint arXiv:2509.06160*, 2025. URL https://arxiv.org/abs/2509.06160.

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, August 2024b. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/.

Zijun Weng, Xiaohui Hu, Shuangyong Song, Yongxiang Li, Kaidong Yu, and Xuanjing Huang. Prompt-level reward specifications for open-ended post-training. *arXiv preprint arXiv:2605.29275*, 2026. URL https://arxiv.org/abs/2605.29275.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. arXiv:2503.05244.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In *Advances in Neural Information Processing Systems*, 2023.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, Zhaoyang Liu, and Bolin Ding. Auto-rubric: Learning from implicit weights to explicit rubrics for reward modeling. *arXiv preprint arXiv:2510.17314*, 2025. URL https://arxiv.org/abs/2510.17314.

Yuzheng Xu, Tosho Hirasawa, Tadashi Kozuno, and Yoshitaka Ushiku. Am I more pointwise or pairwise? Revealing position bias in rubric-based LLM-as-a-judge. In *Findings of the Association for Computational Linguistics: EMNLP*, 2026. URL https://arxiv.org/abs/2602.02219.

Zishan Xu, Shuyi Xie, Qingsong Lv, Shupei Xiao, Linlin Song, Sui Wenjuan, and Fan Lin. Diagnosing failures in large language models' answers: Integrating error attribution into evaluation framework. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21148–21165. Association for Computational Linguistics, July 2025. doi: 10.18653/v1/2025.findings-acl.1089. URL https://aclanthology.org/2025.findings-acl.1089/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. Your Group-Relative Advantage Is Biased, 2026. URL https://arxiv.org/abs/2601.08521.

Dingyao Yu, Tong Zhang, Yutao Mou, Yunxiao Zhang, Wei Ye, and Shikun Zhang. Mitigating rubric interference in LLM judges via on-policy self-distillation. *arXiv preprint arXiv:2608.14684*, 2026. URL https://arxiv.org/abs/2608.14684.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025a. URL https://arxiv.org/abs/2503.14476.

Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, Melanie Kambadur, Dhruv Mahajan, and Rui Hou. Self-generated critiques boost reward modeling for language models. 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. 11499–11514. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.naacl-long.573. URL https://aclanthology.org/2025.naacl-long.573/.

Yu Yuan, Yaoyou Fan, Lili Zhao, Guangting Zheng, Kai Zhang, Lu Pan, Ke Zeng, and Qi Liu. Uncovering and mitigating aggregation-induced reward hacking in multi-reward reinforcement learning, 2026. URL https://arxiv.org/abs/2609.00213.

Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. Chasing the tail: Effective rubric-based reward modeling for large language model post-training. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026a. arXiv:2509.21500.

Qiang Zhang, Boli Chen, Fanrui Zhang, Ruixue Ding, Shihang Wang, Qiuchen Wang, Yinfeng Huang, Haonan Zhang, Rongxiang Zhu, Pengyong Wang, Ailin Ren, Xin Li, Pengjun Xie, Jiawei Liu, Ning Guo, Jingren Zhou, and Zheng-Jun Zha. Arenarl: Scaling RL for open-ended agents via tournament-based relative ranking. *arXiv preprint arXiv:2601.06487*, 2026b. URL https://arxiv.org/abs/2601.06487.

Xiaoying Zhang, Yipeng Zhang, Hao Sun, Kaituo Feng, Chaochao Lu, Chao Yang, and Helen Meng. Critique-GRPO: Advancing LLM Reasoning with Natural Language and Numerical Feedback, 2025. URL https://arxiv.org/abs/2506.03106.

Xinran Zhang. Rethinking atomic decomposition for LLM judges: A prompt-controlled study of reference-grounded QA evaluation, March 2026. URL https://arxiv.org/abs/2603.28005.

Zheng Zhang, Ao Lu, Yuanhao Zeng, Ziwei Shan, Jinjin Guo, Lufei Li, Yexin Li, and Kan Ren. Grad2reward: From sparse judgment to dense rewards for improving open-ended LLM reasoning. *arXiv preprint arXiv:2602.01791*, 2026c. URL https://arxiv.org/abs/2602.01791.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization, 2025. URL https://arxiv.org/abs/2507.18071.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track*, 2023. arXiv:2306.05685.

Mingchen Zhuge, Changsheng Zhao, Dylan R. Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, Yangyang Shi, Vikas Chandra, and Jürgen Schmidhuber. Agent-as-a-Judge: Evaluate Agents with Agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 80569–80611. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhuge25a.html.

A Proofs and Grouping Errors

A.1 Formal Properties

Proposition 1 (Conditional representation invariance)**.** *Fix diagnoses, partition, supports, holistic score, and learned parameters for each response. Reordering diagnoses or events, exactly copying a diagnosis within its event, and repeating cached criterion support leave Eq. 3 unchanged.*

A new supported independent event decreases reward by its effective cost, strictly if positive.

Advantage invariance. Applying these transformations to a rollout group preserves its reward vector, hence both training advantages and standardized audit directions, including under a fixed zero-variance convention. The result concerns fixed diagnoses and partitions. Semantic rephrasing is evaluated separately in Section 5.2; exact deduplication can share this replication invariance.

Proposition 2 (Support perturbation)**.** *For a fixed partition, define* $\Delta_i = |\max_{c\in\mathcal{C}} q_{ci} - \max_{c\in\mathcal{C}'} q_{ci}|$. *Then*

$$|R_a(\mathcal{C}) - R_a(\mathcal{C}')| \le \sum_{G\in\mathcal{P}} \max_{i\in G} w_a(d_i) p_i \Delta_i. \tag{6}$$

Let $H = I - \mathbf{1}\mathbf{1}^\top/K$. Since $A^{\text{train}}(r) = Hr$ and $\|H\|_2 = 1$,

$$\|A^{\text{train}}(r) - A^{\text{train}}(r')\|_2 \le \|r - r'\|_2. \tag{7}$$

For two nondegenerate reward vectors, standardized audit geometry separately gives

$$\|A^{\text{std}}(r) - A^{\text{std}}(r')\|_2 \le \frac{2\|r - r'\|_2}{\min\{\sigma(r), \sigma(r')\}}. \tag{8}$$

The first bound controls changes in the training weights; the second controls changes in the standardized audit direction. In a narrow reward group, centered weights are small, so a response-specific reward error can be large relative to the intended signal even without STD normalization.

A.2 Proofs

Proof of Proposition 1. With inputs fixed, reordering or copying members preserves each event maximum. Repeated supports preserve Eq. 2; event order preserves the sum in Eq. 3. □

Proof of Proposition 2. For two criterion views, write $b_i = w_a(d_i)p_i \max_{c\in\mathcal{C}} q_{ci}$ and $b'_i = w_a(d_i)p_i \max_{c\in\mathcal{C}'} q_{ci}$. For every fixed event G,

$$\left|\max_{i\in G} b_i - \max_{i\in G} b'_i\right| \le \max_{i\in G} |b_i - b'_i| = \max_{i\in G} w_a(d_i)p_i\Delta_i.$$

Sum this inequality over all fixed events; the fixed overall score cancels. □

From reward differences to advantage differences. Let $H = I - \mathbf{1}\mathbf{1}^\top/K$ be the centering projection. Then

$$A^{\text{train}}(r) = Hr, \qquad A^{\text{std}}(r) = \sqrt{K}\,\frac{Hr}{\|Hr\|_2}.$$

For nonzero vectors z, z', triangle inequality and the reverse triangle inequality give

$$\left\|\frac{z}{\|z\|_2} - \frac{z'}{\|z'\|_2}\right\|_2 \le \frac{2\|z - z'\|_2}{\min(\|z\|_2, \|z'\|_2)}. \tag{9}$$

Applying $z = Hr$, using $\|H\|_2 = 1$ and $\|Hr\|_2 = \sqrt{K}\sigma(r)$ proves Eq. 8.

False split and merge. For a false merge of two fixed events, the reward increase is $u + v - \max(u, v) = \min(u, v)$. If one event of cost u is split into m groups retaining the same maximal diagnosis cost, the penalty increases by $(m-1)u$. Fresh extraction can change the fields and costs as well, so these identities describe the isolated effect of changing the partition alone.

B Reward Scale and Advantage Geometry

For $a > 0$, $A^{\text{train}}(ar + b\mathbf{1}) = aA^{\text{train}}(r)$ while $A^{\text{std}}(ar + b\mathbf{1}) = A^{\text{std}}(r)$. Affine compression shrinks the training weights without changing their direction; the standardized vector is an offline diagnostic. Quantization can erase distinctions, and misordering can attach positive advantage to the worse response. Table 6 audits five constructions; Table 7 adds detection metrics for the three-reward comparison.

B.1 Worked examples

Preserving order does not preserve advantages. Consider two score vectors related by an increasing mapping on these three values:

$$\begin{aligned} u &= (-1, 0.9, 1), \quad A^{\text{std}}(u) \simeq (-1.41, 0.65, 0.76), \\ s &= (-1, -0.9, 1), \quad A^{\text{std}}(s) \simeq (-0.76, -0.65, 1.41). \end{aligned} \tag{10}$$

Every pairwise order is preserved, but the middle candidate's advantage changes sign. Ranking alone therefore does not determine the centered weights used in policy training.

Table 6: **Group-relative advantage audit.** 200 prompts, eight Qwen3-8B responses per prompt. Reference utilities come from separate GRM evaluation calls using task-specific criteria.

Reward	Adv. cosine ↑	Sign (%) ↑	Dead (%) ↓
Overall score	0.79	72.0	4.0
Learned diagnostic sum	0.71	65.0	**1.0**
Criterion interactions	0.80	74.0	<u>1.5</u>
Fixed event costs	<u>0.86</u>	<u>80.0</u>	2.0
CARMA	**0.91**	**86.0**	2.0

Table 7: **Detection and reward reliability.** Precision, recall, ranking, and dead-group rates are percentages. Ranking uses 60 near-tie pairs; advantage cosine and dead groups use 200 groups of eight responses. Diagnosis-based rows share detection outputs.

Reward	Precision ↑	Recall ↑	Ranking ↑	Cosine ↑	Dead ↓
Overall score	—	—	<u>58.3</u>	<u>0.79</u>	4.0
Diagnostic sum	**84.0**	**76.0**	51.7	0.71	**1.0**
CARMA	**84.0**	**76.0**	**73.3**	**0.91**	<u>2.0</u>

C Judge Agreement and Repeatability

Table 8: **Repeated judge decisions and human-reference agreement** on 120 pairs.

Judge decision	Correct	Incorrect	Total
Same decision	48	37	85
Different decision	14	21	35
Total	**62**	**58**	**120**

Human annotators agree on 101/120 pairs. The judge repeats its winner on 85/120 but matches the adjudicated human label on only 62/120; 37 pairs are stable and incorrect. Repeatability therefore does not imply preference validity. Both candidate orders are mapped back to response identity.

Agreement. For N pairs, let $h_j^{(1)}, h_j^{(2)}$ be the two independent human decisions, h_j^* the adjudicated reference, and $J_j^{(1)}, J_j^{(2)}$ repeated judge decisions. All labels include A, B, and tie, mapped

back to response identity:

$$\text{HumanAgreement} = \frac{1}{N}\sum_j \mathbf{1}[h_j^{(1)} = h_j^{(2)}],$$

$$\text{JudgeHumanAgreement} = \frac{1}{N}\sum_j \mathbf{1}[J_j^{(1)} = h_j^{*}],$$

$$\text{JudgeRepeatability} = \frac{1}{N}\sum_j \mathbf{1}[J_j^{(1)} = J_j^{(2)}].$$

D Training and Evaluation Details

Data and initialization. The private Chinese and public English tracks use the splits described in Section 5, with data sources in Appendix I. Within each backbone and training track, all methods share the same starting checkpoint after task SFT and prior post-training.

GRM supervision and machine evaluation. We trained an in-house GRM on task–response examples with scalar ratings and textual critiques, using human assessments and mean ratings from GPT, Claude Opus, and Gemini as supervision. The frozen GRM supplies reward scores, diagnoses, event verification, revision feedback, and machine evaluation through separate calls. Evaluation calls receive neither constructed rewards nor diagnostic penalties.

Policy optimization. Each rollout batch contains 16 prompts and eight responses per prompt. We train for 200 rollout iterations at learning rate 2×10^{-6}, temperature 0.8, and top-p 0.95. Generation limits are 1,024 tokens on the private track and 8,192 on the public track. The PPO-style likelihood-ratio clip (Schulman et al., 2017) is 0.2, the reference-KL coefficient is 0.02, and gradient clipping is 1.0. MPGC uses $A_i = r_i - \bar{r}$ without group-STD normalization. RLTF-SD processes each rollout in two eight-prompt minibatches; other interfaces use one full-batch update.

Preference calibration. The scorer outputs and event partitions are fixed before fitting the 24 nonnegative severity increments. Ties are retained for evaluation but excluded from binary cost fitting. We minimize Eq. 5 with bound-constrained L-BFGS, using $\tau = 0.2$ and selecting $\eta \in \{0.001, 0.01, 0.1, 1\}$ on development preferences. Matched diagnosis-based controls use the same strict preferences and parameterization; the unit-cost sum is a separate control.

Evaluation. The GRM judges each prompt in both response orders. After restoring response identity, a win, tie, and loss receive 1, 0.5, and 0; we average within prompts before averaging across prompts. Qwen3-8B reports the mean and sample SD over three training seeds. Fixed-checkpoint results use 100,000 prompt-bootstrap resamples, retaining both orders together; SE is the SD of resampled means, and 95% intervals use their 2.5th and 97.5th percentiles.

D.1 Controlled-Test Protocols

Exact copies. Copy a complete diagnosis within its event, keeping the response, overall score, support, and partition fixed. Event maxima are unchanged; diagnosis addition charges twice.

Rephrasing. Table 1 keeps the response and defect fixed but changes their diagnostic description. A rewrite may change the reported label, severity, confidence, or event assignment; the flip columns quantify sensitivity to these field and assignment changes while responses are held fixed.

Independent and minor defects. Adding a defect triggers fresh scoring, so the overall score can change even if the criterion-hit vector stays fixed. Minor-flaw pairs instead use adjudicated overall preferences to test whether the assigned penalties preserve their adjudicated overall quality ordering.

D.2 Evaluation Metrics

Preference and ranking. Tie-adjusted preference is $(W + T/2)/N$. Policy comparisons use 400 prompts, and human comparisons use 300. Near-tie ordering uses 60 independently annotated response pairs selected by $|\Delta S| \leq 0.2$ on the ten-point scale. Duplication tests use 150 pairs per label condition; added-defect and minor-flaw tests use 300 pairs each.

Reward diagnostics. Advantage cosine compares centered candidate rewards with reference utilities from separate task-specific GRM evaluation calls on 200 groups of eight responses. Cosine and

sign agreement use the common groups where all required vectors are nonzero. Dead-group rate counts zero-variance rewards. Detection precision and recall match diagnoses to annotated defects by repair target; partition accuracy counts merged same-event and separated independent pairs.

D.3 Public Benchmark Evaluation

LitBench. We score the 2,480 fixed test story pairs in LitBench (Fein et al., 2026) against its official preference labels. Frozen costs are calibrated outside LitBench with English-adapted prompts. The fitted condition learns the 24 increments on LitBench-Train and evaluates on the official test split. The learned diagnostic sum shares its training preferences and diagnoses. This evaluates reward ordering, rather than training a policy on LitBench.

WritingBench. Policies from both tracks generate responses to the 1,000 WritingBench queries (Wu et al., 2025). We preserve query language and requirements, using its five query-specific criteria and ten-point scale with the GRM as a common scorer. Criteria are fixed across methods for each query; scores are averaged within queries and then across queries. This evaluates policy generation, separately from fixed-pair reward ordering.

E ArenaRL and RLTF-SD Integration

Text feedback and training interfaces. Language-feedback imitation, Text2Grad, and Critique-GRPO use critiques to improve model training (Chen et al., 2024a; Wang et al., 2026; Zhang et al., 2025); feedback-conditioned policy learning also operates without scalar rewards (Luo et al., 2025). Feedback Descent uses pairwise textual feedback for inference-time text optimization (Lee et al., 2025). ArenaRL and RLTF-SD provide comparison and feedback-driven policy interfaces (Zhang et al., 2026b; Song et al., 2026). CARMA supplies event-organized evidence, repair feedback, or scalar rewards with the original training procedure preserved at each interface.

ArenaRL. The original comparator receives the task and two responses. The raw-diagnosis variant additionally shows each response's diagnoses and evidence. The event variant organizes these same diagnoses by shared repair target, keeping independently repairable problems separate. The comparator, tournament, and policy objective are held fixed throughout policy training.

RLTF-SD. Reward-only replaces the scalar reward with CARMA while retaining the original feedback. Feedback-only replaces feedback with one repair instruction per event (Appendix J) while retaining the original reward. Joint replaces both. These match Event reward, Event feedback, and Reward + feedback in Table 4; CARMA + RLTF-SD in Table 2 is joint. GRM-guided revisions serve as self-distillation targets under the original objective. Evaluation uses first-round responses without test-time feedback.

Comparison references. Table 2 compares each interface against the corresponding holistic-reward policy; Table 4 compares each variant against its original interface policy. These reference policies differ across the two tables. Interface cells use single runs; the 8B policy table averages three seeds. Each interface retains its optimizer while using the specified reward, evidence, or feedback.

E.1 Feedback Quality

Table 9: **RLTF-SD feedback quality** (%). Validity is measured per suggestion; repair and damage per attempted revision.

Feedback	Validity ↑	Repair ↑	Damage ↓
Original critique	73.0	61.0	12.0
Deduplicated feedback	<u>78.0</u>	<u>63.3</u>	<u>10.7</u>
Event-guided feedback	**86.0**	**71.3**	**8.0**

Table 9 evaluates 300 suggestions and 300 attempted revisions per condition: evidence validity, target repair, and damage to unaffected content.

E.2 Human Comparisons

Table 10: **Blinded human preferences** on 300 prompts per comparison. Preference is $(W + T/2)/300$ (%). RLTF-SD + CARMA uses event reward and feedback.

Comparison	Wins	Ties	Losses	Preference
CARMA vs. RRD-WU	140	47	113	54.5
CARMA vs. overall score	170	31	99	61.8
ArenaRL + events vs. ArenaRL	157	39	104	58.8
RLTF-SD + CARMA vs. RLTF-SD	155	45	100	59.2

Responses are anonymized and presented in randomized order. Raters judge independently and may choose a tie. Preference is calculated from the displayed counts as $(W + T/2)/300$; prompt-level bootstrap intervals use the same resampling procedure described in Appendix D.

F CARMA Algorithm

Algorithm 1 separates frozen evidence extraction from preference calibration. The active member of each event is recomputed during fitting because its identity can change with the cost vector.

Algorithm 1 Event-structured rewards with preference-calibrated costs

Require: Responses $\mathcal{Y}$; strict-preference sets $\mathcal{Q}_{\mathrm{tr}}, \mathcal{Q}_{\mathrm{dev}}$
Require: Frozen judges J_S, J_D, J_V; regularization grid Λ; two initial cost vectors
Ensure: Shared costs $\hat{a} \in \mathbb{R}_+^{8\times 3}$ and reward $R_{\hat{a}}$

Evidence extraction and event partitioning

for $(x, y) \in \mathcal{Y}$ **do**
  $S \leftarrow J_S(x, y)$; $\mathcal{D} \leftarrow J_D(x, y)$; validate evidence and remove exact duplicates
  $\mathcal{P} \leftarrow J_V(x, y, \mathcal{D})$
  Resolve evidence issues in shared $\mathcal{D}$; validate $\mathcal{P}$, falling back to singletons
  Cache $(S, \mathcal{D}, \mathcal{P}, u)$, with $u_i = p_i \max_c q_{ci}$
end for

Preference calibration

for $\eta \in \Lambda$ **do**
  for $j \in \{1, 2\}$ **do**
    $a_{\eta,j} \leftarrow \mathrm{LBFGSB}\big(\mathcal{L}_\eta, a^{(j)}, a \geq 0\big)$
  end for
  $a_\eta \leftarrow \arg\min_{a\in\{a_{\eta,1},a_{\eta,2}\}} \mathcal{L}_\eta(a; \mathcal{Q}_{\mathrm{tr}})$
end for
$\hat{a} \leftarrow a_\eta$ with highest preference agreement on $\mathcal{Q}_{\mathrm{dev}}$
function REWARD$(a, S, \mathcal{D}, \mathcal{P}, u)$
  $b_i \leftarrow u_i \sum_{\ell=1}^{s_i} a_{k_i,\ell}$ for every $d_i \in \mathcal{D}$
  $e_G \leftarrow \max_{i\in G} b_i$ for every $G \in \mathcal{P}$
  return $S/10 - \sum_{G\in\mathcal{P}} e_G$
end function

G Continuation from Plateau Checkpoints

For each original method, select a late checkpoint where reward and evaluation loss have stopped improving. Fork two continuations from this checkpoint with the same prompts and additional budget: retain the original method, or add the CARMA component specified in Section 5.7. Evaluate both against their own starting policy using the same frozen GRM comparator. The midpoint and endpoint correspond to half and all of the allocated additional training budget.

Table 11: **Continuation from plateau checkpoints** (% preference against each block's starting policy). Start is 50%; mid and final use half and all of the additional budget.

Starting method	Continuation	Start	Mid	Final
Overall score	Original	50.0	51.0	50.7
	CARMA reward	50.0	**55.3**	**60.1**
RRD-WU	Original	50.0	51.1	51.3
	CARMA reward	50.0	**54.8**	**59.4**
ArenaRL	Original	50.0	50.9	51.1
	Event evidence	50.0	**54.6**	**58.8**

We monitor reward before checkpoint selection. After switching, common pairwise evaluation tracks quality; each run logs training reward and evaluation loss. The paired continuations measure further improvement from the same selected plateau checkpoints.

H Training Effects of False Event Merges

We corrupt a fixed fraction of partitions by merging two verifier-separated events independently annotated as distinct repair targets before reward computation. The affected prompts are sampled independently for each seed, and learned costs remain frozen throughout policy training at every corruption level. Each corruption rate trains a fresh policy from the same initialization with the same budget; the nominal rate is the fraction of these annotated distinct-target event pairs within a response that are merged. At 0%, 5%, 10%, 15%, and 20% false-merge rates, policy preference is 60.4%, 59.2%, 58.1%, 56.6%, and 54.9%, respectively.

I Training Data

Private data. The private track uses Chinese creative-writing tasks for reward and policy training. Held-out prompts evaluate policy quality; separate internal sets audit reward construction.

Public data. We mix WritingPrompts-train (Fan et al., 2018) and the English writing subset of REER's `m-a-p/DeepWriting-20K` (Wang et al., 2025) at 70:30, with permission from DeepWriting's maintainers. The sources supply story premises and writing instructions with reference texts. Both tasks and writing content are English. We add no uniform length target and retain original task constraints.

Reference texts help identify writing tasks; public preference labels compare fresh responses to the same prompt from task-adapted policies. The GRM assesses task fulfillment and writing quality, retaining ties for evaluation. CARMA and the learned diagnostic sum use the same preferences to fit their 24 costs. Table 2 separates policy results by track; both tracks are evaluated on WritingBench (Appendix D.3). LitBench instead tests reward ordering on fixed public story pairs.

J Judge Prompts and a Complete Example

These compact templates specify each stage's role, input boundaries, rules, and output contract. Within a domain, category, tag, and scoring guides stay fixed; tasks and responses are variable input data.

J.1 Core Judging Templates

Diagnosis prompt

Role: You are an evidence-grounded writing evaluator.
Input: Task, response, and fixed category and tag guide; no score or preference label.

1. **Find supported defects.** Quote task or response evidence; omissions cite the requirement and response context. Do not require optional genre elements. Do not use length as a quality proxy; violations of explicit length requirements, supported repetition, or missing required content may be defects.
2. **Describe each diagnosis.** Use one of the eight categories and three severity levels in the guide. Confidence in $[0, 1]$ expresses evidence support, independently of severity. Preserve supported perspectives; do not group them or assign penalty costs.
3. **Output.** Use the supplied JSON schema: id, category, severity, confidence, claim, relation type, evidence, and canonical tags. Return an empty list for no supported defects.

Event-verification prompt

Role: You verify evidence and identify shared semantic repairs.
Input: Task, response, and validated diagnoses with stable ids.

1. **Check support.** Report unsupported evidence for the shared validation stage. Preserve all original diagnosis fields; do not invent evidence or re-grade diagnoses.
2. **Test a shared repair.** Group members only if one specific repair resolves their common target while preserving requirements and unrelated facts. Shared words, categories, similarity chains, or a broad rewrite are insufficient. Independent targets remain separate even within one sentence.
3. **Output.** Return events with id, member ids, shared problem, repair, and content to preserve, plus any validation issues. Assign every accepted diagnosis exactly once. Use singleton events when no common target and repair can be established.

All matched diagnostic aggregations share validated evidence, including verifier-reported issues, and exact-duplicate removal. Invalid partitions fall back to singletons (Algorithm 1).

Overall-score prompt (separate call)

Role: You assess overall task fulfillment and writing quality.
Input: Task, response, and fixed scoring guide; no diagnoses, events, or preference labels.

1. **Assess quality.** Consider task-relevant strengths, coherence, content, development, and effective expression. Weigh flaws by their overall effect; fluency does not excuse a central requirement violation.
2. **Avoid shortcuts.** Do not favor length, brevity, or optional genre elements by themselves. Useful detail and creativity may improve quality without becoming mandatory.
3. **Output.** Return one score from 0 to 10 and a brief explanation.

Multi-aspect scoring prompt (one call)

Role: You assess the six fixed writing-quality aspects.
Input: Task, response, and the aspect definitions and scoring guide.

1. **Aspects.** Task fulfillment; coherence and continuity; development of subjects or characters; organization and progression; voice and register; expression and readability.
2. **Apply the guide.** Apply all six aspects to the requested writing form. Character motivation and dialogue are relevant only when present or required; other forms use subject development and an appropriate voice. Do not penalize absent optional genre elements or omit an aspect.
3. **Output.** Return six scores from 0 to 10, each with a short explanation, in one call.

The multi-aspect baseline averages all six scores on the normalized scale, $\sum_m s_m/60$.

Pairwise evaluation prompt

Role: You compare anonymous responses to the same task.
Input: Task and candidates A and B, with their origins hidden.

1. **Compare quality.** Apply the same task-fulfillment and writing-quality criteria to both candidates. Use decisive evidence; do not favor position, source, or length alone.
2. **Weigh differences.** A minor flaw can break a near tie, but should not outweigh a clear overall quality advantage disproportionately. Choose tie when no meaningful preference is supported.
3. **Output.** Return A, B, or tie and a concise reason.

Map decisions from both candidate orders to response identity (Appendix D).

J.2 Training-Interface Supplements

ArenaRL comparator: event-evidence supplement

Role: You are the original ArenaRL comparator with auxiliary evidence.
Input: Original task and candidates, plus candidate-aligned events and diagnoses.

1. **Use evidence.** Check reported problems against the responses. Event and diagnosis counts are not quality scores; several diagnoses may describe one problem.
2. **Preserve the interface.** Follow the original comparison instructions and scoring rules. Keep the tournament and rank-to-advantage procedure unchanged.
3. **Output.** Return the original comparator output schema; add no new ranking or scoring fields.

Event-guided revision feedback

Role: You provide focused repair feedback for a writing response.
Input: Task, response, and validated events in stable order.

1. **Address each event.** Give one instruction stating its problem, evidence, shared repair, affected scope, and requirements and unrelated content to preserve.
2. **Keep repairs compatible.** Separate events may touch the same sentence. Make necessary structural changes, but avoid unrelated embellishment. Seek an alternative if a repair conflicts with a requirement; report unresolved conflicts through the supplied interface.
3. **Output.** Return event ids and repair instructions in the supplied feedback schema. Do not reassign diagnosis fields or penalty weights.

RLTF-SD uses this feedback for training-time revision and self-distillation. Policy evaluation uses only first-round responses without test-time feedback (Appendix E).

J.3 Worked Diagnosis and Repair Example

The example follows Figures 1 and 2 and supports the illustrative calculation in Appendix K.1.

Writing task and response

Task: Write a one-day rescue story. Mara is loyal to Nico.
Response: Mara had always protected Nico. That morning, Nico was seventeen. The station key snapped in two. Trapped under a beam, Nico heard her promise to save him; she then walked away without explanation. That afternoon, Nico was twenty-two. Mara returned and unlocked the exit with the same intact key.

Supported diagnoses. The following fields illustrate the fixed category and severity guide; confidences are inputs to the worked calculation, not evaluation results.

1. d_1**: motivation; severity 2; confidence .90.** "Walked away without explanation" supplies no reason for leaving Nico trapped.

2. d_2**: characterization; severity 2; confidence .80.** Her unexplained departure is not reconciled with "had always protected Nico."
3. d_3**: causal coherence; severity 2; confidence .85.** The transition from "promise to save him" to departure lacks a connecting decision or circumstance.
4. d_4**: continuity; severity 3; confidence .95.** "Seventeen" that morning and "twenty-two" that afternoon violate the one-day setting.
5. d_5**: continuity; severity 2; confidence .90.** The key "snapped in two" but later opens the exit as the "same intact key," with no explanation.

Event decisions and feedback.

- $G_1 = \{d_1, d_2, d_3\}$: explain that Mara leaves to fetch a jack and returns. This shared repair restores motivation, loyalty, and the causal transition.
- $G_2 = \{d_4\}$: keep Nico seventeen throughout that day.
- $G_3 = \{d_5\}$: keep the key broken and obtain a cutter to open the exit. Age and key state remain distinct targets despite sharing the continuity category.

Example repaired response

Mara had always protected Nico. That morning, seventeen-year-old Nico became trapped under a beam in the station. The station key snapped in two. Mara promised to save him, then explained that she needed a jack and ran to the maintenance shed. She returned with a worker carrying a jack and a cutter. Together they lifted the beam. With the key still broken, the worker cut open the exit. That afternoon, seventeen-year-old Nico stepped outside beside Mara.

The repair resolves all five diagnoses and preserves the one-day rescue and Mara's loyalty.

K Defect Guide and Reward Examples

The proposer uses the eight categories in Table 12. **Low severity** denotes a local flaw with limited effect; **medium severity** disrupts a passage or supporting requirement; **high severity** undermines a central requirement, fact, or causal structure. The LLM identifies category and severity, while preferences determine shared penalty costs used in the final response reward.

Table 12: **Eight diagnostic categories.** Different categories may describe one repair event.

Category	Diagnostic focus
Motivation	Missing reason for an action
Characterization	Unreconciled behavior relative to established traits
Causal coherence	Missing action–outcome link
Continuity	Contradictory facts or states
Instruction adherence	Violated explicit requirement
Narrative development	Missing setup or resolution
Dialogue and voice	Inconsistent speaker voice
Expression/readability	Repetition or unclear wording

K.1 Worked Reward Calculations

Several diagnoses, one repair. All diagnoses in the following calculations have canonical tags covered by $\mathcal{C}$, so $u_i = p_i$. For the rescue story in Appendix J, use illustrative medium-severity costs .060, .050, .070, and .080 for motivation, characterization, causal coherence, and continuity, and high-severity continuity cost .120. The first three member penalties are $.060 \times .90 = .0540$, $.050 \times .80 = .0400$, and $.070 \times .85 = .0595$. Their shared event charges .0595. The separate age and key-state events charge $.120 \times .95 = .1140$ and $.080 \times .90 = .0720$. At illustrative score $S = 8.8$,

$$R_{\text{CARMA}} = .88 - (.0595 + .1140 + .0720) = .6345,$$
$$R_{\text{sum}} = .88 - (.0540 + .0400 + .0595 + .1140 + .0720) = .5405.$$

A copied causal-transition diagnosis preserves CARMA's .6345 but lowers the sum to .4810. These numbers illustrate aggregation, rather than fitted experimental costs.

Similar wording, separate repairs. A guard opens a locked door without explanation, and a witness leaves without explanation. Explaining the guard's action leaves the witness's departure unresolved. Medium causal-coherence cost .070 with confidences .90 and .80 gives separate charges .063 and .056. Separate events charge .119; merging charges .063. At $S = 8.8$, rewards are .761 and .817: the merged version undercounts the two independently repairable problems.

A minor flaw and overall preference. Suppose coherent story A is preferred to generic story B, with overall scores 8.9 and 8.7. A has one low-severity expression flaw and B has none. At confidence 1, cost .005 preserves $R_A = .885 > R_B = .870$; fixed cost .040 reverses the order, $R_A = .850 < R_B = .870$. The diagnosis stays fixed; its cost determines the preference.

L Aggregation Baselines

Let $b_i = w_a(d_i)u_i(\mathcal{C})$. CARMA charges $\sum_G \max_{i\in G} b_i$; the mean variant charges $\sum_G |G|^{-1} \sum_{i\in G} b_i$. The learned diagnostic sum and the no-partition ablation both charge $\sum_i b_i$ and fit the same 24 nonnegative increments. Fixed costs replace learned category costs with severity levels $(1, 2, 3)$ times a development-selected scale. Removing the holistic branch gives $R = -\sum_G \max_{i\in G} b_i$. Table 3 reports these changes together.

Semantic deduplication groups diagnoses within each response by similarity of their descriptions and quoted evidence. The design uses normalized BGE-M3 embeddings (Chen et al., 2024b) and cosine similarity, with complete-link clustering and a development-selected threshold from $\{.75, .80, .85, .90, .95\}$. Category labels and repair decisions are excluded from its similarity input. Each cluster is charged its largest member penalty, with the 24 costs fitted on the same preferences. Exact duplicates are removed for both semantic grouping and CARMA. This isolates the grouping rule with identical diagnostic evidence and preference supervision.

L.1 Criterion Interactions

A criterion-interaction model discounts co-occurring criteria and remains monotone in violations:

$$g(h) = -\sum_c \alpha_c h_c + \sum_{(c,d)\in\mathcal{G}} \gamma_{cd} h_c h_d, \qquad \gamma_{cd} \geq 0, \qquad \alpha_c \geq \sum_{e\ni c} \gamma_e. \tag{11}$$

We parameterize $\alpha_c = a_c + \sum_{e\ni c} \gamma_e$, $a_c \geq 0$. This comparator captures recurring interactions but cannot distinguish two responses with identical hit vectors and different event structures.

M Extension beyond Writing

Fine-grained evaluation also appears in mathematical reasoning: process supervision supplies step judgments, and PRMBench tests nuanced error detection (Lightman et al., 2024; Wang et al., 2024b; Song et al., 2025); PRIME learns implicit process rewards (Cui et al., 2025). Agent-as-a-Judge provides intermediate feedback on agent execution (Zhuge et al., 2025). Grouping reports that share a repair offers a natural extension to these settings, with domain-specific repair definitions and empirical evaluation left to future work.

N Extended Related Work

Rubric-based rewards. Rubrics turn broad judgments into task-specific criteria (Chen et al., 2026). Checklists, RaR, and Rubric Anchors use such criteria for policy training (Viswanathan et al., 2025; Gunjal et al., 2026; Huang et al., 2025a); prompt-level specifications combine requirements, global quality, and executable constraints (Weng et al., 2026). RRD decomposes criteria and removes redundant or downweights correlated criteria estimated from sampled responses; AutoRubric learns explicit criteria from preferences (Shen et al., 2026; Xie et al., 2025). RULERS grounds judgments in evidence and calibrates them to human score boundaries (Hong et al., 2026), while SARA studies interference between jointly evaluated rubrics (Yu et al., 2026). Claim-level

rewards and DRACO localize judgments to caption claims and agent steps (Gao et al., 2026; Gandhi et al., 2026). CARMA groups each response's diagnoses by shared repair and calibrates shared costs from same-prompt preferences.

Fine-grained feedback and aggregation. Localized human feedback and translation-error evaluation provide finer supervision than overall scores (Wu et al., 2023; Fernandes et al., 2023; Guerreiro et al., 2024). Error-attribution judges categorize failures, while prompt-controlled studies examine when atomic decomposition helps (Xu et al., 2025; Zhang, 2026). Critic-RM learns critiques and rewards; W^2Bench measures requirement-specific writing rankings (Yu et al., 2025b; Ren et al., 2026). ArmoRM combines interpretable quality objectives through contextual weighting, and Grad2Reward derives token-level credit from judge gradients (Wang et al., 2024a; Zhang et al., 2026c). CARMA aggregates defect events: a shared repair defines the penalty unit, an event maximum limits repeated charges, and learned costs adjust a complementary holistic score.

Rewards for creative writing. Preference learning has been applied to storytelling (Castricato et al., 2022). Writing-Zero uses pairwise generative rewards, RLMR combines subjective writing rewards with objective constraint verification, and PrefBERT provides semantically aware rewards for free-form generation (Jia et al., 2025; Liao et al., 2026; Li et al., 2025b). We study how multiple reports of writing defects should contribute to one reward, using matched-supervision ablations and fixed-response audits alongside policy evaluation.

Judge and reward reliability. LLM judges support scoring and ranking but remain sensitive to evaluation design (Li et al., 2025a). Relevant failures include length bias, preference leakage, rubric-option position bias, and equivalent prompt variations (Huang et al., 2025b; Li et al., 2026; Xu et al., 2026; Bhat & Varma, 2026). Spurious-reward results further show that policy gains alone do not explain reward informativeness (Shao et al., 2026). RLTF-SD addresses feedback-conditioned advantage collapse using a first-turn reward baseline (Song et al., 2026). CARMA instead targets defect aggregation, assessed through near-tie ranking and advantage audits. Exact-copy invariance assumes fixed diagnoses and partitions; robustness to semantic redescription is evaluated empirically.