RITR: RELIABILITY-INFORMED TRIAGE AND CONSTRAINED REWRITING FOR RUBRIC-BASED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rubric-based reinforcement learning (RL) extends training to open-ended tasks, but standard importance-weighted aggregation ignores the reliability of grading evidence. Group normalization can turn judge fluctuations into unit-scale advantages when responses lack genuine quality differences, while uniformly failed criteria offer no successful examples. We propose RITR, a diagnose-before-repair framework that decomposes judge scores into decision direction and confidence. A criterion-level reliability proxy and four-state triage admit only sufficiently confident cross-boundary criteria to the main reward and control their relative weights. For criteria failed by all trusted responses, RITR allocates one sample from a fixed response budget to a targeted rewrite of the best-scoring response. The rewrite is gated on target satisfaction and preservation of previously satisfied criteria, and its scaffold is removed in two stages. Starting from Qwen3-8B, we train separate policies on medical, science, and finance rubric data. RITR improves over uniform-aggregation rubric RL on six of seven benchmarks, by up to 4.95 points on open-ended medical evaluation and 2.60 points on medical multiple-choice tasks. Medical ablations show open-ended gains from triage and further gains from constrained repair.

1 INTRODUCTION

Open-ended responses in medicine, science, or finance can be fluent yet unusable when they omit a contraindication, a unit, or a disclosure an expert would require. Reinforcement learning with verifiable rewards (RLVR) has driven strong progress in mathematics and code, where correctness is checkable (Shao et al., 2024; Guo et al., 2025), but binary feedback cannot express the multidimensional quality that high-stakes open-ended tasks demand (Arora et al., 2025). Rubrics offer a structured alternative: they decompose expert intent into judgeable criteria that an LLM judge can evaluate one at a time, enabling on-policy RL beyond verifiable domains (Gunjal et al., 2026; Huang et al., 2025).

Standard rubric-RL aggregation weights criteria by importance, sums their scores into a response-level reward, and applies a group-relative update (Shao et al., 2024). Importance weights encode how much a criterion *should* matter; they do not measure how reliably the judge *currently* grades it. LLM judges can be biased and unstable (Zheng et al., 2023), and criteria within one rollout group can provide very different evidence: confident pass–fail contrasts, uniformly passing or failing verdicts, or scores near the decision boundary.

Uniform aggregation can turn this heterogeneity into noise-dominated updates. Under the observation model of Section 4, score variation includes both satisfaction-state differences and within-state nuisance; when nuisance dominates, group normalization still produces a nontrivial advantage despite little genuine contrast. Conversely, a criterion failed by every response provides no successful example for a group-relative comparison. These are distinct problems: deciding which observed comparisons to trust, and obtaining better samples where the group has none.

We propose RITR (**R**eliability-**I**nformed **T**riage and constrained **R**ewriting), a diagnose-before-repair framework. RITR separates each score into decision direction and confidence and constructs an intraclass-correlation-inspired reliability proxy $\lambda_k \in [0, 1]$. A four-state diagnosis admits suf-

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.

ficiently confident cross-boundary criteria into a boundary-normalized reward and uses reliability to set their relative contributions. When all trusted responses fail a criterion, RITR allocates one sample from a fixed budget to a targeted rewrite of the best-scoring response, which must satisfy all targets while preserving previously satisfied non-target criteria before receiving a parent-anchored score. A two-stage schedule removes the bonus and then the privileged rewrite prompt, leaving triage active throughout training. These decisions use the judge's expressed confidence; they do not certify the correctness of its verdicts.

Our contributions are as follows.

- We analyze a noise-dominated failure mode of uniform aggregation and introduce boundary-normalized, reliability-weighted four-state triage, computed from a single judgment per response–criterion pair with no repeated grading or added trainable parameters.
- We introduce diagnosis-triggered constrained rewriting within a fixed response budget, with cross-criterion validation and two-stage scaffold removal.
- Across seven benchmarks in three domains, RITR improves over uniform-aggregation rubric RL on six, by up to 4.95 points on open-ended medical evaluation and 2.60 points on medical multiple-choice tasks.

2 Related Work

Rubric-based reward modeling. Rubrics extend RL to open-ended tasks through fine-grained, judgeable criteria (Gunjal et al., 2026; Huang et al., 2025). Subsequent work uses rubrics as rollout scaffolds and for reward shaping (Zhou et al., 2025), evolves them online (Shao et al., 2025), jointly optimizes judge and rubric (Xu et al., 2026), and turns unmet criteria into guidance or privileged trajectories (Zhang et al., 2026b; Hou et al., 2026). These approaches primarily improve rubric construction or exploration guidance; RITR instead focuses on how the evidence from a given rubric and judge is selected and weighted within each rollout group.

Reliability of judge signals. LLM-as-a-judge exhibits position, verbosity, and self-preference biases (Zheng et al., 2023). Generative verifiers use the yes-token probability as a continuous score (Zhang et al., 2025a); conformal methods produce interval estimates (Sheng et al., 2025); policy-aware and focal rewards reweight criteria by saturation or difficulty (Tyagi et al., 2026; Huang et al., 2026); repeated evaluation reduces verifier variance (Kwok et al., 2026). RITR distinguishes *observed separability* (do verdicts cross the boundary?) from *expressed confidence* (how far are scores from it?) and uses both to route criteria and select repair targets. Unlike repeated evaluation, its statistics come from one judgment per response–criterion pair and cannot independently identify confidently wrong judgments.

Guided exploration and scaffold removal. Injecting rubrics or hints into the prompt and later withdrawing them can ease exploration bottlenecks (Zhou et al., 2025; Hou et al., 2026). RITR targets criteria diagnosed as failed by all trusted responses, allocates at most one rewrite per prompt, and validates it against both target and non-target criteria before assigning its reward-channel score.

3 Preliminaries: Rubric-Based Group-Relative RL

For each prompt x, the behavior policy $\pi_{\theta_{\text{old}}}$ samples G responses $\{y_i\}_{i=1}^{G}$. The prompt is paired with a rubric $\{(c_k, w_k)\}_{k=1}^{M}$, where c_k is a criterion description and $w_k \in \mathbb{R} \setminus \{0\}$ a signed importance weight, positive for requirements and negative for pitfalls. An LLM judge evaluates every response–criterion pair with a fixed prompt (Appendix A.1) asking for a single `yes`/`no` token, and we read the normalized probability of `yes` as a continuous score $p_{ik} \in [0, 1]$ (Zhang et al., 2025a). Standard rubric grading thresholds this probability into a verdict $\mathbf{1}[p_{ik} > 0.5]$ and scores the response by the weights of the criteria it meets (Arora et al., 2025). RITR keeps p_{ik} itself, using the threshold only for the decision direction and the distance to it as a confidence margin.

Uniform aggregation. The baseline uses $S_i^{\text{unif}} = \sum_{k=1}^{M} w_k \, \mathbf{1}[p_{ik} > 0.5]$. Here "uniform" means a common aggregation rule without reliability-based selection, not equal importance weights. The group-relative advantage $\widehat{A}_i = (S_i - \bar{S})/(\text{Std}(S) + \varepsilon_{\text{std}})$ is shared by all tokens of y_i (Shao et al.,

2024). For G original-prompt rollouts, the clipped surrogate is

$$\begin{aligned}\mathcal{J}(\theta) &= \mathbb{E}_x\,\mathbb{E}_{\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\ell_{i,t}(\theta)\right] - \beta\, D_{\text{KL}}\big[\pi_\theta \,\|\, \pi_{\text{ref}}\big],\\ \ell_{i,t}(\theta) &= \min\Big(r_{i,t}(\theta)\widehat{A}_i,\ \text{clip}\left(r_{i,t}(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}\right)\widehat{A}_i\Big),\\ r_{i,t}(\theta) &= \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})}.\end{aligned} \tag{1}$$

where π_{ref} is the reference model and ϵ_{clip} the policy-ratio clipping parameter. RITR retains this clipped-loss backbone, changes the construction of S_i, and optionally substitutes a targeted rewrite for one rollout. With a rewrite, the total response count is $K = G + 1$; otherwise all K responses come from the original prompt. This matches the number of responses across the RL comparisons, not necessarily their token count or wall-clock cost.

4 When Uniform Aggregation Becomes Noise-Dominated

For a fixed prompt, consider an observation model relating the aligned judge score g_{ik} (Section 5.1) to a latent satisfaction state $s_{ik} \in \{0, 1\}$:

$$g_{ik} = \mu_{s_{ik}}^{(k)} + \delta_{ik}, \qquad \mathbb{E}[\delta_{ik} \mid s_{ik}] = 0, \qquad \text{Var}(\delta_{ik} \mid s_{ik}) = \tau_{k,s_{ik}}^2. \tag{2}$$

Here Δ_k is the class-mean gap, π_k the satisfaction probability, and $\bar{\tau}_k^2$ the average within-class variance. The law of total variance gives

$$\sigma_k^2 = \text{Var}(g_{ik}) = \underbrace{\Delta_k^2\,\pi_k(1-\pi_k)}_{\text{between-class}} + \underbrace{\bar{\tau}_k^2}_{\text{within-class}}. \tag{3}$$

The first term captures variation in satisfaction; the second captures variation at fixed satisfaction, treated as nuisance for this binary criterion model rather than every notion of response quality.

From criteria to aggregate rewards. Thresholded verdicts b_{ik} obey the same decomposition (Appendix B.4). For $v_{ik} \in \{g_{ik}, b_{ik}\}$ and $\tilde{w}_k = |w_k|$, write

$$\begin{aligned}S_i^{\text{agg}} &= \sum_k \tilde{w}_k v_{ik} = T_i + N_i,\\ T_i &= \sum_k \tilde{w}_k \mathbb{E}[v_{ik} \mid s_{ik}], \qquad N_i = \sum_k \tilde{w}_k \delta_{ik}.\end{aligned} \tag{4}$$

Here δ_{ik} and $\bar{\tau}_k^2$ refer to the chosen score: T_i captures satisfaction-dependent signal and N_i residual nuisance. Binary aggregation matches Base Rubric RL up to a response-independent offset when no pitfall score equals 0.5; Appendix B.1 gives the tie correction.

Proposition 1 (Noise-dominated population normalization)**.** *Assume nuisance terms are uncorrelated across criteria and each is uncorrelated with every satisfaction state. Let Z_X denote population standardization of X. Then*

$$\begin{aligned}\text{Var}(S_i^{\text{agg}}) &= \text{Var}(T_i) + \sum_k \tilde{w}_k^2 \bar{\tau}_k^2,\\ \text{Corr}(Z_{S^{\text{agg}}}, Z_T) &= \sqrt{\frac{\text{Var}(T_i)}{\text{Var}(S_i^{\text{agg}})}} \qquad \big(\text{Var}(T_i) > 0\big).\end{aligned} \tag{5}$$

If all satisfaction states become degenerate while nuisance variance stays bounded away from zero, signal variance vanishes and the standardized aggregate converges in mean square to standardized nuisance.

Appendix B.1 gives the proof. This is a population-level failure mode, not a finite-group identity or a claim that all same-side differences are harmful. In Figure 1, continuous scores reduce the noisy criterion's raw spread $3.4\times$ (0.58 versus 2.00), yet increase its maximum absolute standardized advantage from 1.00 to 1.37: smaller reward spread need not mean a smaller update. The analysis motivates filtering uncertain comparisons and seeking successful samples for uniformly failed criteria; it does not show that our observable proxy identifies the latent variance components.

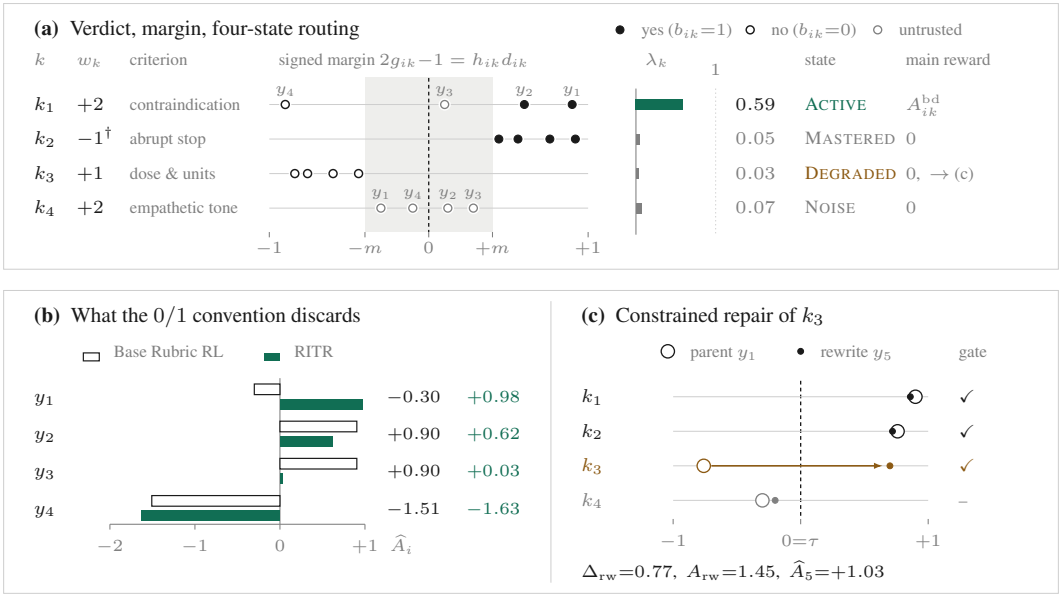

Figure 1: RITR on a constructed instance (G=4, M=4); k_1 and k_4 have equal importance weights. **(a)** Markers show signed margins $2g_{ik} - 1 = h_{ik}d_{ik}$: signs encode verdicts, magnitudes encode confidence, and pale markers with $h_{ik} \leq m$ are untrusted. **(b)** Standardized advantages before repair. Uniform aggregation assigns 2.0 reward units to k_4's untrusted passing verdicts, making y_2 and y_3 tie despite their different k_1 margins (0.60 and 0.10). **(c)** Parent margins (rings) and rewrite margins (dots): k_3 crosses the pass boundary while k_1, k_2 remain passing, with margin changes of 0.04. Illustrative values use Eqs. (6)–(14) with $m = 0.4$, $\theta_c = 0.25$, τ_{pres}=τ_{tar}=0.5, and α_{rw}=0.5; they are not experimental measurements. $^\dagger k_2$ is a pitfall, aligned as $g = 1 - p$.

5 RITR

RITR is a two-phase, diagnose-before-repair procedure (Figure 1). *Diagnosis* computes a reliability proxy and assigns each criterion to ACTIVE, MASTERED, DEGRADED, or NOISE using observed grading evidence; only ACTIVE criteria enter the main reward channel. *Repair* uses DEGRADED criteria as targets for at most one gated rewrite within the fixed response budget, and its scaffold is removed in two stages while triage remains active.

5.1 POLARITY ALIGNMENT AND CONFIDENCE–DIRECTION DECOMPOSITION

Because `yes` means satisfaction for a requirement but violation for a pitfall, we align polarities so that larger scores are always better, and separate each score into a semantic side and a confidence magnitude:

$$g_{ik} = \begin{cases} p_{ik}, & w_k > 0, \\ 1 - p_{ik}, & w_k < 0, \end{cases} \qquad \tilde{w}_k = |w_k|, \qquad \begin{aligned} h_{ik} &= |2g_{ik} - 1|, \\ b_{ik} &= \mathbf{1}[g_{ik} > 0.5], \quad d_{ik} = 2b_{ik} - 1. \end{aligned} \tag{6}$$

The side b_{ik} records whether a score crosses the judge's decision boundary and $h_{ik} \in [0, 1]$ is its confidence margin; both are observed quantities, not verified satisfaction labels.

Lemma 1 (Exact decomposition)**.** *For every $g_{ik} \in [0, 1]$, $2g_{ik} - 1 = h_{ik}d_{ik}$, and $g_{ik} \mapsto (h_{ik}, d_{ik})$ is invertible via $g_{ik} = (1 + h_{ik}d_{ik})/2$. If moreover $s_{ik} \mid g_{ik} \sim \mathrm{Bernoulli}(g_{ik})$, then $h_{ik} = \mathbb{E}[(2s_{ik} - 1)\, d_{ik} \mid g_{ik}]$.*

The algebraic identity requires no calibration assumption; the expected *signed* agreement interpretation does, giving $h_{ik} = 2\Pr(s_{ik} = b_{ik} \mid g_{ik}) - 1$ (Appendix B.2). We retain h_{ik} as a per-response margin in the boundary statistic rather than standardizing it separately, which guarantees neither calibrated confidence in practice nor proportional attenuation of the final, group-standardized advantage.

5.2 CRITERION-LEVEL RELIABILITY

Let $\mathcal{V}_k$ contain the original-prompt responses with a valid judgment for criterion k (a `yes`/`no` token), and let $G_k = |\mathcal{V}_k|$. Using $c_{ik} = h_{ik}^2$, we compute a confidence-weighted score dispersion and an unweighted Bernoulli uncertainty proxy:

$$\begin{aligned} \bar{g}_k &= \frac{\sum_{i\in\mathcal{V}_k} c_{ik}\, g_{ik}}{\sum_{i\in\mathcal{V}_k} c_{ik}}, \qquad V_k^{\text{between}} = \frac{\sum_{i\in\mathcal{V}_k} c_{ik}\,(g_{ik}-\bar{g}_k)^2}{\sum_{i\in\mathcal{V}_k} c_{ik}}, \\ u_k^{\text{within}} &= \frac{1}{G_k}\sum_{i\in\mathcal{V}_k} g_{ik}(1-g_{ik}). \end{aligned} \tag{7}$$

The reliability proxy and its scaling are

$$\lambda_k = \frac{V_k^{\text{between}}}{V_k^{\text{between}} + u_k^{\text{within}}} \in [0,1], \qquad r_k = \sqrt{\lambda_k}. \tag{8}$$

Empty groups or groups with no positive confidence mass are routed to NOISE before evaluating these ratios; we set $\lambda_k = 0$ when both variance terms vanish, since such unanimous endpoint scores provide no boundary contrast. The superscripts describe the observable construction, not identified estimates of the latent components in Eq. equation 3.

Proposition 2 (Calibrated, uniform-weight special case)**.** *Let i be uniform on $\mathcal{V}_k$, assume $s_{ik} \mid g_{ik} \sim \text{Bernoulli}(g_{ik})$, and let $\text{Var}(s_{ik}) > 0$. The uniformly weighted version satisfies*

$$\begin{aligned} \text{Var}(s_{ik}) &= \text{Var}(g_{ik}) + \mathbb{E}[g_{ik}(1-g_{ik})], \\ \lambda_k^{\text{unif}} &= \frac{\text{Var}(g_{ik})}{\text{Var}(g_{ik}) + \mathbb{E}[g_{ik}(1-g_{ik})]} = \frac{\text{Var}(\mathbb{E}[s_{ik} \mid g_{ik}])}{\text{Var}(s_{ik})}. \end{aligned} \tag{9}$$

This special case provides an explained-variance interpretation and motivates an intraclass-correlation-inspired construction (Shrout & Fleiss, 1979). The actual λ_k uses confidence weights only in the between-response term, so it is a *reliability proxy* rather than the exact ratio above, and confidence weighting can increase or decrease the ratio relative to uniform weighting. We use $r_k = \sqrt{\lambda_k}$ as an order-preserving, less aggressive shrinkage than λ_k itself, not as an identified correlation with ground truth. Neither the calibration assumption nor the association between this proxy and actual grading errors is established by our experiments (Appendix B.3).

5.3 BOUNDARY-NORMALIZED ADVANTAGE

Reliability controls relative criterion weights, but each criterion's contrast still needs a scale. RITR uses the weighted spread of binary verdicts rather than raw score dispersion. Define

$$\bar{b}_k = \frac{\sum_{i\in\mathcal{V}_k} c_{ik}\, b_{ik}}{\sum_{i\in\mathcal{V}_k} c_{ik}}, \qquad V_k^{\text{sem}} = \bar{b}_k(1-\bar{b}_k). \tag{10}$$

With stability constant $\epsilon = 10^{-6}$ and clipping threshold $z_{\max} = 2$, the boundary statistic and its centered, scaled form are

$$\begin{aligned} z_{ik}^{\text{bd}} &= \text{clip}\left(h_{ik}\,\frac{b_{ik}-\bar{b}_k}{\sqrt{V_k^{\text{sem}}}+\epsilon}, -z_{\max}, z_{\max} \right), \\ \tilde{z}_{ik}^{\text{bd}} &= z_{ik}^{\text{bd}} - \frac{1}{G_k}\sum_{j\in\mathcal{V}_k} z_{jk}^{\text{bd}}, \qquad A_{ik}^{\text{bd}} = \tilde{w}_k\, r_k\, \tilde{z}_{ik}^{\text{bd}}, \end{aligned} \tag{11}$$

with $A_{ik}^{\text{bd}} = 0$ for $i \notin \mathcal{V}_k$. The denominator reflects observed cross-boundary contrast, while h_{ik} retains a response-specific margin before centering. If all positive-confidence responses lie on one side, $V_k^{\text{sem}} = 0$ and every $z_{ik}^{\text{bd}} = 0$, and zero-confidence responses contribute zero as well, so same-side fluctuations alone cannot activate this channel. When contrast exists, multiplying by h_{ik} and clipping need not preserve zero mean, so we center explicitly: $\sum_{i\in\mathcal{V}_k} A_{ik}^{\text{bd}} = 0$. This does not guarantee that every passing verdict receives a positive centered advantage (Appendices B.4, B.5).

Table 1: Four-state routing based on observed grading evidence. "Quality-valid" includes a nonempty trusted set.

State	Condition	Treatment
NOISE	$G_k < 2$, or $\mathcal{C}_k = \emptyset$, or $\bar{c}_k \le \theta_c$	Main contribution is zero
ACTIVE	Quality-valid; $\{b_{ik} : i \in \mathcal{C}_k\} = \{0, 1\}$	Use A_{ik}^{bd}
MASTERED	Quality-valid; $b_{ik} = 1$ for all $i \in \mathcal{C}_k$	Main contribution is zero
DEGRADED	Quality-valid; $b_{ik} = 0$ for all $i \in \mathcal{C}_k$	Candidate for rewriting

5.4 QUALITY-AWARE FOUR-STATE ROUTING

Let $\bar{c}_k = G_k^{-1} \sum_{i \in \mathcal{V}_k} c_{ik}$, and let $\mathcal{C}_k = \{i \in \mathcal{V}_k : h_{ik} > m\}$ be the *trusted* responses, with $0 \le m < 1$. Evidence is quality-valid when $G_k \ge 2$, $\mathcal{C}_k \ne \emptyset$, and $\bar{c}_k > \theta_c$, where $0 \le \theta_c < 1$. Table 1 assigns states from the trusted responses' observed sides: only ACTIVE criteria contribute to the main reward, and DEGRADED criteria are candidates for repair. Since $g(1-g) = (1-h^2)/4$, the quality gate is equivalent to $u_k^{\mathrm{within}} < (1-\theta_c)/4$ (Appendix B.7), which bounds expressed uncertainty, not grading error. In particular, DEGRADED means that *all trusted responses* receive failing verdicts; it does not assert that every response truly fails. Routing and reliability are separate gates: a criterion whose verdicts all sit near the boundary is sent to NOISE because $\mathcal{C}_k = \emptyset$, irrespective of λ_k, which need not be small there ($\lambda_{k_4} > \lambda_{k_2}$ in Figure 1).

5.5 TRIAGE-GUIDED CONSTRAINED REPAIR

Triage selects observed comparisons; it does not by itself generate better responses. Let $\mathcal{R} = \{k : \mathrm{state}_k = \text{DEGRADED}\}$. When $\mathcal{R} \ne \emptyset$ and the prompt is selected with probability $\rho(t)$, RITR uses its $(G{+}1)$-th sample for a targeted rewrite instead of an undirected rollout.

Parent-anchored rewrite. The parent is the diagnostic response with the highest weighted aligned score, $\mathrm{par} = \arg\max_i \sum_k \tilde{w}_k g_{ik}$. The prompt supplies the original question, this parent, and the criteria in $\mathcal{R}$ as numbered requirements with their polarity, and requests minimal edits that satisfy the targets while preserving correct content, structure, and style (Appendix A.2). Rubric weights and diagnosed states are not revealed.

Validation gate and net-improvement bonus. Given valid aligned parent and rewrite scores g_k^{par} and g_k^{rw}, the gate passes, $\mathbb{I}_{\mathrm{gate}} = 1$, if and only if both conditions hold:

$$\begin{aligned} &\forall k \notin \mathcal{R} :\ g_k^{\mathrm{par}} \ge \tau_{\mathrm{pres}} \Rightarrow g_k^{\mathrm{rw}} \ge \tau_{\mathrm{pres}}, \\ &\forall k \in \mathcal{R} :\ g_k^{\mathrm{rw}} > \tau_{\mathrm{tar}}. \end{aligned} \tag{12}$$

With $\tau_{\mathrm{pres}} = \tau_{\mathrm{tar}} = 0.5$, targets must receive passing verdicts and non-targets must retain any score at or above the preservation threshold. Preservation is an implication, so a non-target the parent already fails is left unconstrained; this is a boundary-level check, not a guarantee that all continuous scores improve. Let $\mathcal{A}_t$ be the ACTIVE criteria and $S_{\mathrm{par}} = \sum_{k \in \mathcal{A}_t} A_{\mathrm{par},k}^{\mathrm{bd}}$ the parent's *pre-standardization boundary score*, distinct from the raw score used to select it. The rewrite score is

$$\begin{aligned} \Delta_{\mathrm{rw}} &= \frac{\sum_{k=1}^{M} \tilde{w}_k \, (g_k^{\mathrm{rw}} - g_k^{\mathrm{par}})}{\sum_{k \in \mathcal{R}} \tilde{w}_k}, \\ A_{\mathrm{rw}} &= S_{\mathrm{par}} + \alpha_{\mathrm{rw}}(t)\ \mathrm{clip}(\Delta_{\mathrm{rw}}, -1, 1)\, |S_{\mathrm{par}}|. \end{aligned} \tag{13}$$

Here $\alpha_{\mathrm{rw}}(t)$ controls the bonus, and Δ_{rw} measures net weighted score change across *all* criteria, normalized by target weight. Passing the gate guarantees neither $\Delta_{\mathrm{rw}} > 0$ nor a positive final advantage: continuous losses on non-targets can offset target gains without crossing a preservation threshold. If $S_{\mathrm{par}} = 0$, then $A_{\mathrm{rw}} = 0$ even for a successful rewrite, but other responses can still have nonzero scores. If $\mathcal{A}_t = \emptyset$, all group scores are zero and there is no reward-driven policy-gradient contribution. We do not introduce an untested nonzero fallback.

Two-stage scaffold decay. Let $\kappa(t; T) = \frac{1}{2}(1 + \cos(\pi t/T))$ for $0 \le t \le T$. In stage one ($0 \le t \le T_1$), $\rho(t) = \rho_0$ and $\alpha_{\mathrm{rw}}(t) = \alpha_{\mathrm{rw}}^0 \kappa(t; T_1)$. In stage two ($T_1 < t \le T_1 + T_2$), $\alpha_{\mathrm{rw}}(t) = 0$ and $\rho(t) = \rho_0 \kappa(t - T_1; T_2)$. Prompts without a selected repair use an original-prompt rollout for the

remaining sample. For $t > T_1 + T_2$, both scaffolds are absent and all K responses are generated without target injection; whether repaired behaviors are retained is an empirical question.

5.6 Unified Training Objective

Let $\mathcal{B}_t$ be the original-prompt responses and $\mathcal{W}_t$ the at most one rewrite, with $|\mathcal{B}_t| + |\mathcal{W}_t| = K$. The first G responses determine whether to request a rewrite. Final routing, main-channel statistics, and centering use $\mathcal{B}_t$: G responses when a rewrite is generated, and all K otherwise. With $\mathcal{A}_t$ denoting the resulting ACTIVE criteria, define

$$
\begin{aligned}
S_i &= \mathbf{1}[i \in \mathcal{B}_t] \sum_{k \in \mathcal{A}_t} A_{ik}^{\mathrm{bd}} + \mathbf{1}[i \in \mathcal{W}_t]\, \mathbb{I}_{\mathrm{gate}}\, A_{\mathrm{rw}}, \\
\widehat{A}_i &= \frac{S_i - \bar{S}}{\mathrm{Std}(S) + \varepsilon_{\mathrm{std}}}.
\end{aligned}
\tag{14}
$$

The mean and standard deviation use all K responses. Since $\sum_{i \in \mathcal{B}_t} S_i = 0$, a rejected rewrite has zero advantage in the reward-driven term, though adding its zero score can still change the common normalization scale and does not mask KL regularization (Appendix B.5). Triage remains active throughout training, while the rewrite term disappears once $\rho(t) = 0$.

Relative weighting, not absolute attenuation. For a positive constant a, group standardization gives $\widehat{A}(aS) = \widehat{A}(S)$ when the stability term is omitted; with one ACTIVE criterion and no rewrite, its common factor r_k cancels. Reliability can change the relative contributions of multiple criteria, and routing can remove a channel, but smaller r_k does not generally imply a smaller final update (Appendix B.6).

The update averages over all K responses. Following privileged-rollout practice (Zhou et al., 2025), both policy probabilities in $r_{i,t}(\theta)$ condition on the original prompt, even for a rewrite generated from an augmented prompt. Applying Eq. equation 1 to this mixed sample set is therefore a privileged-trajectory surrogate, not an unbiased on-policy estimator. Reliability estimation requires no repeated grading or added trainable parameters; equal response counts do not imply equal token or wall-clock costs.

6 Experiments

6.1 Setup

Training data. All RL methods start from Qwen3-8B (Yang et al., 2025), with a separate policy trained for each domain. For healthcare we follow Zhou et al. (2025) and randomly sample 4,500 rubric-annotated instances from HealthBench (Arora et al., 2025); for science we use the Science subset of RubricHub (Li et al., 2026); for finance, the Finance subset of the rubric dataset of Zhang et al. (2026a) (about 1.15k examples).

Evaluation. Medical evaluation uses HealthBench-500 (Arora et al., 2025), a 500-sample held-out subset disjoint from training, and LLMEval-Med (Zhang et al., 2025b), both open-ended and rubric-scored, plus MedQA (Jin et al., 2021) and MedMCQA (Pal et al., 2022). The medical RL data contains no multiple-choice supervision, so the latter two measure transfer to a different task format. Science uses GPQA-Diamond (Rein et al., 2023); finance uses the finance subsets of PRBench (Akyürek et al., 2025) and OneMillion-Bench (Yang et al., 2026). Scores are on percentage scales, with accuracy for multiple-choice tasks.

Baselines. *Initial* is Qwen3-8B before our RL training. *gpt-oss-20b* (OpenAI, 2025) and *Qwen3-14B* are larger reference checkpoints evaluated without additional training, not controlled comparisons of model scale. *RuscaRL* (Zhou et al., 2025) injects rubrics into the rollout prompt and decays them. *Base Rubric RL* uses uniform aggregation S_i^{unif} with the same judge, data, initialization, and response-count budget as RITR.

Implementation. All RL methods use verl (Sheng et al., 2024) with learning rate 10^{-6}, batch size 64, and micro-batch size 32; training lasts 500 steps in healthcare and science and 300 in finance. The training judge (reward) is Qwen3-14B, and the total response budget is $K = 8$ ($G = 7$

Table 2: Main results (%). HealthBench-500, LLMEval-Med, PRBench, and OneMillion-Bench are open-ended and rubric-scored; MedQA, MedMCQA, and GPQA-Diamond are multiple-choice. Best in bold, second best underlined. RL rows are trained from Qwen3-8B.

Model	Medical				Science	Finance	
	HealthBench 500	LLMEval Med	MedQA	MedMCQA	GPQA Diamond	PRBench	OneMillion Bench
Initial (Qwen3-8B)	47.32	58.32	68.11	59.28	47.22	23.99	15.31
gpt-oss-20b	52.73	55.47	**85.54**	**68.58**	43.43	31.58	23.07
Qwen3-14B	48.77	63.57	68.97	<u>64.28</u>	48.74	26.50	19.96
RuscaRL	56.73	61.17	68.10	60.26	52.40	32.15	23.28
Base Rubric RL	<u>60.05</u>	<u>64.32</u>	67.71	58.69	<u>53.53</u>	**34.79**	<u>23.59</u>
RITR (ours)	**64.71**	**69.27**	<u>70.31</u>	61.01	**55.05**	<u>33.57</u>	**24.78**

Table 3: Component ablation on medical benchmarks (%). Parenthesized values are changes relative to Base Rubric RL in points. All RL rows share data, judge, initialization, and sampling budget.

Model	Open-ended		Closed-ended	
	HealthBench-500	LLMEval-Med	MedQA	MedMCQA
Initial (Qwen3-8B)	47.32	58.32	68.11	59.28
Base Rubric RL	60.05	64.32	67.71	58.69
+ reliability triage	63.25 (+3.20)	64.92 (+0.60)	66.30 (−1.41)	58.86 (+0.17)
RITR (triage + repair)	**64.71** (+4.66)	**69.27** (+4.95)	**70.31** (+2.60)	**61.01** (+2.32)

diagnostic responses when rewriting). RITR uses $m = 0.5$, $\theta_c = 0.6$, $\alpha_{\text{rw}}^0 = 0.12$, and $\rho_0 = 1.0$. With $T_1 = T_2 = 100$ steps, the bonus is zero from step 100 and rewrites cease at step 200. Appendix C gives the remaining training settings.

6.2 Main Results

Table 2 reports point estimates from the evaluated checkpoints. RITR ranks first among all listed models on four of seven benchmarks and exceeds Base Rubric RL on six, with the largest gains on the two open-ended medical benchmarks; PRBench is the only exception.

Medical domain. RITR reaches 64.71 on HealthBench-500 and 69.27 on LLMEval-Med, improving over Base Rubric RL by 4.66 and 4.95 points and over the initial model by 17.39 and 10.95. On MedQA and MedMCQA, its gains over Base Rubric RL are 2.60 and 2.32 points, and over the initial model 2.20 and 1.73. Unlike Base Rubric RL, which scores below the initial model on both tasks, the full method shows positive transfer without multiple-choice supervision. This does not establish that triage alone protects medical knowledge: its MedQA score is below the baseline (Table 3).

Science and finance. RITR scores 55.05 on GPQA-Diamond (+1.52 over Base Rubric RL) and 24.78 on OneMillion-Bench (+1.19). PRBench is the only regression: 33.57 versus 34.79 (−1.22 points). Without per-state statistics and repeated runs, we cannot distinguish a systematic limitation from run-to-run variation.

Reference models and RuscaRL. RITR exceeds Qwen3-14B on six benchmarks and gpt-oss-20b on five, including all four open-ended benchmarks, and surpasses RuscaRL on all seven. These comparisons favor the complete procedure; differences in architecture and training preclude causal claims about scale, and the comparisons do not isolate individual components.

6.3 Ablations

Component ablation. In Table 3, + *reliability triage* combines reliability weighting, boundary normalization, and four-state routing without repair. It raises HealthBench-500 by 3.20 points and LLMEval-Med by 0.60, but lowers MedQA by 1.41 and changes MedMCQA by only +0.17. Adding repair raises all four scores above both the baseline and triage-only variant. This supports

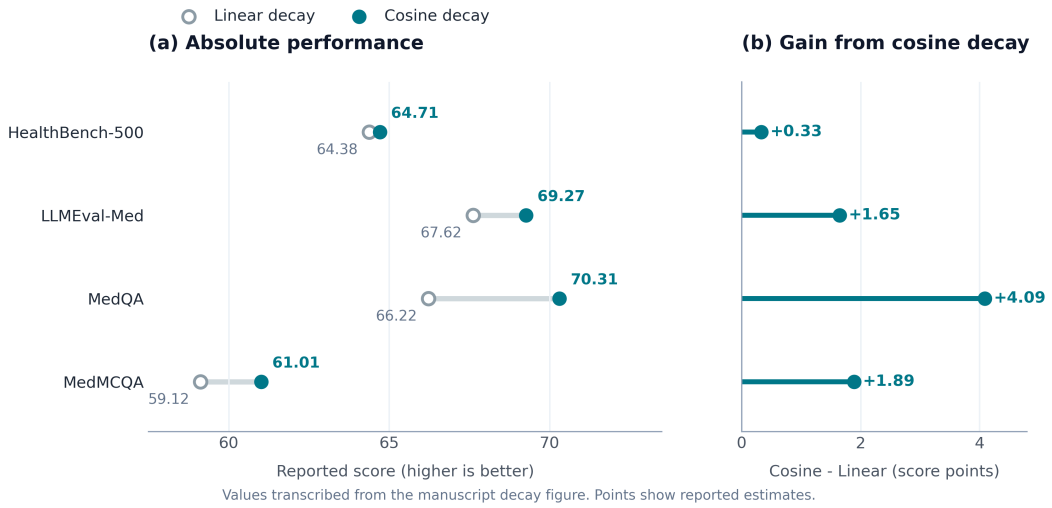

Figure 2: Decay-schedule ablation for the rewrite bonus and rewrite ratio (full RITR, medical domain): cosine versus linear decay of equal length. (a) Absolute scores on the four medical benchmarks; (b) gain of cosine over linear decay.

open-ended gains from the triage bundle and an additional benefit from repair, but does not separate reliability weighting from boundary normalization and routing.

Decay schedule. Cosine decay gives higher point estimates than equal-length linear decay on all four medical benchmarks, by 0.33, 1.65, 4.09, and 1.89 points (Figure 2). Both schedules have continuous-time mean $1/2$ per stage, so they differ only in exposure timing. One plausible reading is that cosine keeps the scaffold strong early, when rewrites are most informative, and removes it quickly late; with one run per configuration this remains untested.

Training dynamics. Single-run HealthBench curves (Appendix C.2) peak during RITR's scaffold phase and then contract, ending with entropy near the baseline and KL above it. They describe optimization dynamics, not behavioral internalization or a causal schedule effect.

7 Conclusion

RITR diagnoses rubric evidence before using it for learning. Boundary-normalized four-state triage selects comparisons and sets relative reliability weights; constrained rewriting seeks improved samples for criteria failed by all trusted responses. Across seven benchmarks in three domains, RITR improves over uniform-aggregation rubric RL on six. Medical ablations show open-ended gains from triage and further gains from repair. Calibrated-judge checks, finer controls, and repeated runs are needed to establish generality.

8 Limitations

We evaluate one policy scale, one training judge, and one run per configuration, without confidence intervals. Ablations do not isolate reliability weighting from boundary normalization and routing, or compare against uniform rewards with identical targeted rewriting. State frequencies, gate acceptance rates, and sensitivity to m, θ_c, and α_{rw}^0 are unreported. RITR uses probability magnitudes whereas Base Rubric RL uses only thresholded verdicts; reward granularity differs, and confidently wrong judgments remain a risk. Group standardization can cancel common scaling, and parent-anchored repair produces no signal without an ACTIVE criterion; the frequency of this case and alternative anchors are untested. Privileged generation differs from original-prompt sampling, and equal response counts need not mean equal computation. Additional judges, controls, and repeated runs are needed to assess generality.

AI USE STATEMENT

We used AI-based tools in several stages of this work. We used them to help draft sections of the paper and to polish the writing. We also used them for retrieval and discovery, such as finding related work. In addition, we used them to support research ideation and execution, such as refining the method design and debugging the code. All AI-generated content was reviewed, verified, and edited by the authors. The authors take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

We are committed to making the results of this work reproducible. The paper provides a complete mathematical specification of RITR: all scoring, reliability, routing, gating, and advantage computations are given as explicit formulas in Sections 4–5.6, with full proofs and worked numerical examples in Appendices B. The judge and rewrite prompts are reproduced verbatim in Appendix A, and the training configuration, including all hyperparameters, is reported in Appendix C. All datasets and base models used in this study are publicly available, and all baselines are re-implementations trained by us under the same framework, data, and budget, as described in Section 6.1.

This work was carried out at an organization. The training code builds on internal infrastructure whose release is subject to intellectual-property and confidentiality review, and we are not permitted to make it public at submission time. We recognize that this limits direct verification, and we have therefore aimed to make the paper self-contained to a level that supports independent re-implementation: the method introduces no learned components beyond the policy itself, every statistic is defined in closed form, and the only stochastic elements are standard sampling and decoding procedures with documented settings. Should the internal review conclude favorably, we intend to release the code publicly; in the meantime, we welcome technical inquiries through the anonymous review channel and will provide any additional implementation details that can be shared without violating these obligations.

REFERENCES

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego Mares, Pavit Singh, Michael Liu, Subodh Chawla, Pete Cline, Lucy Ogaz, Ernesto Hernandez, Zihao Wang, Pavi Bhatter, Marcos Ayestaran, Bing Liu, and Yunzhong He. PRBench: Large-scale expert rubrics for evaluating high-stakes professional reasoning. *arXiv preprint arXiv:2511.11562*, 2025.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. HealthBench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025. doi: 10.48550/arXiv.2505.08775.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Jinkun Hou, Zhuo Liu, Huimin Ren, Hongsheng Xin, Pan Zhou, and Kun Zhan. Rise-rl: Rubric-informed selective exploration for open-ended reinforcement learning. *arXiv preprint arXiv:2608.09123*, 2026.

Yu Huang, Zihua Zhao, Zhaoxin Huan, Wanli Gu, Feng Hong, Xinmu Ge, Lin Yuan, Weichang Wu, Qiang Hu, Xiaolu Zhang, Jun Zhou, and Jiangchao Yao. Focal reward: Balanced reinforcement learning under rubric-based rewards. *arXiv preprint arXiv:2605.26579*, 2026.

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. *arXiv preprint arXiv:2508.12790*, 2025.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Jacky Kwok, Shulu Li, Pranav Atreya, Yuejiang Liu, Yixing Jiang, Chelsea Finn, Marco Pavone, Ion Stoica, and Azalia Mirhoseini. LLM-as-a-verifier: A general-purpose verification framework. *arXiv preprint arXiv:2607.05391*, 2026.

Sunzhu Li, Jiale Zhao, Miteto Wei, Huimin Ren, Yang Zhou, Jingwen Yang, Shunyu Liu, Kaike Zhang, and Wei Chen. Rubrichub: A comprehensive and highly discriminative rubric dataset via automated coarse-to-fine generation. *arXiv preprint arXiv:2601.08430*, 2026.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025. URL https://arxiv.org/abs/2508.10925.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning (CHIL)*, pp. 248–260, 2022.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen-tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. 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, 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. doi: 10.48550/arXiv.2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. Analyzing uncertainty of LLM-as-a-judge: Interval evaluations with conformal prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11286–11328. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.569.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Utkarsh Tyagi, Xingang Guo, MohammadHossein Rezaei, Daniel George, Anas Mahmoud, Jackson Lee, Bing Liu, and Yunzhong He. Not every rubric teaches equally: Policy-aware rubric rewards for RLVR. *arXiv preprint arXiv:2605.20164*, 2026.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tuo Zhao, and Haoyu Wang. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training. *arXiv preprint arXiv:2602.01511*, 2026.

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

Qianyu Yang, Yang Liu, Jiaqi Li, Jun Bai, Hao Chen, Kaiyuan Chen, Tiliang Duan, Jiayun Dong, Xiaobo Hu, Zixia Jia, Yang Liu, Tao Peng, Yixin Ren, Ran Tian, Zaiyuan Wang, Yanglihong Xiao, Gang Yao, Lingyue Yin, Ge Zhang, Chun Zhang, Jianpeng Jiao, Zilong Zheng, and Yuan Gong. Onemillion-bench: How far are language agents from human experts? *arXiv preprint arXiv:2603.07980*, 2026.

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. 2026a. URL https://openreview.net/forum?id=pBjy4ek2QV.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *International Conference on Learning Representations*, 2025a.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. LLMEval-med: A real-world clinical benchmark for medical LLMs with physician validation. *arXiv preprint arXiv:2506.04078*, 2025b.

Wenjian Zhang, Kongcheng Zhang, Jiaxin Qi, Baisheng Lai, and Jianqiang Huang. Experience is the best teacher: Motivating effective exploration in reinforcement learning for llms. *arXiv preprint arXiv:2603.20046*, 2026b.

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, Datasets and Benchmarks Track*, 2023.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, Hengtong Lu, Wei Chen, Yan Xie, and Mingli Song. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. *arXiv preprint arXiv:2508.16949*, 2025.

A PROMPTS

A.1 JUDGE GRADING PROMPT

Every response–criterion pair is evaluated with the following fixed prompt. Let q_{ik}^{yes} and q_{ik}^{no} be the judge's first-token probabilities for the two allowed answers. We use $p_{ik} = q_{ik}^{\text{yes}}/(q_{ik}^{\text{yes}} + q_{ik}^{\text{no}})$, i.e., the probability normalized over `yes` and `no`. The resulting score reflects the judge's expressed confidence, not a verified probability of criterion satisfaction.

> Your job is to look at a conversation and a rubric item, and judge whether the last assistant response in the conversation meets the rubric item.
>
> # Conversation
>
> «conversation»
>
> # Rubric item
>
> «rubric_item»
>
> # Instructions
>
> - Decide whether the response meets the criteria of the rubric item. If a rubric item has multiple criteria, all of them must be met.
>
> - Exception: if the criteria says "such as", "for example", or "including", the response need not include all listed examples.
>
> - Note: some criteria describe something undesirable with a negative point value. In that case answer whether the criteria IS met (i.e. whether the response exhibits the undesirable behavior), not whether the response is good.
>
> # Answer
>
> Answer with a single word, either "yes" (criteria met) or "no" (criteria not met). Do not output anything else.

A.2 TARGETED REWRITE PROMPT

The rewrite of Section 5.5 is generated from a multi-turn prompt: the original user question, the parent response as an `assistant` turn, and the following instruction as a new `user` turn. Each target criterion $k \in \mathcal{R}$ is rendered as one numbered requirement whose polarity prefix is "Include or satisfy" when $w_k > 0$ and "Avoid" when $w_k < 0$.

> Revise your previous answer using the minimum changes necessary to satisfy the requirements below.
>
> Preserve all already-correct content, the overall structure, tone, and approximate length.
>
> Return only the complete revised answer. Do not describe the revision process or mention any rubric.
>
> Requirements:
>
> 1. «polarity»: «criterion_1»
>
> 2. «polarity»: «criterion_2»
>
> ...

The instruction never reveals rubric weights or the diagnosed state of any criterion.

B PROOFS AND DERIVATIONS

B.1 PROOF OF PROPOSITION 1 (POPULATION NORMALIZATION)

Score-specific quantities. For either $v_{ik} \in \{g_{ik}, b_{ik}\}$, define

$$\delta_{ik} = v_{ik} - \mathbb{E}[v_{ik} \mid s_{ik}], \qquad \pi_k = \Pr(s_{ik} = 1),$$
$$\Delta_k = \mathbb{E}[v_{ik} \mid s_{ik} = 1] - \mathbb{E}[v_{ik} \mid s_{ik} = 0],$$
$$\bar{\tau}_k^2 = \mathbb{E}[\mathrm{Var}(v_{ik} \mid s_{ik})].$$

For continuous scores, these definitions agree with Eq. equation 2; all moments are over the response distribution for a fixed prompt.

Decomposition. Let $\tilde{v}_{ik} = p_{ik}$ when $v_{ik} = g_{ik}$ and $\tilde{v}_{ik} = \mathbf{1}[p_{ik} > 0.5]$ when $v_{ik} = b_{ik}$. Polarity alignment gives $\tilde{v}_{ik} = v_{ik}$ for $w_k > 0$ and $\tilde{v}_{ik} = 1 - v_{ik}$ for $w_k < 0$, except at $p_{ik} = 0.5$ in the binary case. Away from those ties, $\sum_k w_k \tilde{v}_{ik} = T_i + N_i - \sum_{k:w_k<0} \tilde{w}_k$, where the final term is response-independent. At exact ties, the binary baseline additionally includes $\sum_{k:w_k<0} \tilde{w}_k \mathbf{1}[p_{ik} = 0.5]$. Thus its equivalence to the aligned aggregate requires the no-tie condition; the proposition for S_i^{agg} itself does not require ties to have zero probability.

Variance. Since $\mathbb{E}[v_{ik} \mid s_{ik}] = \mathbb{E}[v_{ik} \mid s_{ik}{=}0] + \Delta_k s_{ik}$ is affine in s_{ik}, the assumed lack of correlation between every nuisance term and every satisfaction state implies $\text{Cov}(T_i, N_i) = 0$. Cross-criterion nuisance terms are uncorrelated, and $\mathbb{E}[\delta_{ik} \mid s_{ik}] = 0$ gives $\text{Var}(\delta_{ik}) = \bar{\tau}_k^2$. Hence $\text{Var}(N_i) = \sum_k \tilde{w}_k^2 \bar{\tau}_k^2$ and $\text{Var}(S_i^{\text{agg}}) = \text{Var}(T_i) + \text{Var}(N_i)$.

Correlation. For any X with positive variance, define its population standardization as

$$Z_X = \frac{X - \mathbb{E}[X]}{\sqrt{\text{Var}(X)}}.$$

For nonzero $\text{Var}(T_i)$, this is an affine transformation with a deterministic positive scale. Therefore

$$\text{Corr}(Z_{S^{\text{agg}}}, Z_T) = \frac{\text{Cov}(T_i + N_i, T_i)}{\sqrt{\text{Var}(S_i^{\text{agg}})\text{Var}(T_i)}} = \sqrt{\frac{\text{Var}(T_i)}{\text{Var}(S_i^{\text{agg}})}}. \tag{15}$$

Limit. With finitely many bounded-score criteria, degeneracy of all s_{ik} implies $\text{Var}(T_i) \to 0$. If $\text{Var}(N_i)$ is bounded away from zero, then

$$\mathbb{E}\big[(Z_{S^{\text{agg}}} - Z_N)^2\big] = 2\left(1 - \sqrt{\frac{\text{Var}(N_i)}{\text{Var}(T_i) + \text{Var}(N_i)}}\right) \longrightarrow 0. \tag{16}$$

This proves mean-square convergence to a unit-variance nuisance score. The statement concerns population moments, not random finite-group means and standard deviations, and excludes the case in which both signal and nuisance variance vanish. □

B.2 Proof of Lemma 1 (exact decomposition and Bayesian interpretation)

Let $x = 2g_{ik} - 1 \in [-1, 1]$. Then $h_{ik} = |x|$ and, because $g_{ik} > 0.5 \iff x > 0$, $d_{ik} = \text{sign}(x)$ whenever $x \neq 0$, so $h_{ik} d_{ik} = |x| \,\text{sign}(x) = x$. At $x = 0$, $h_{ik} = 0$ and the identity holds regardless of the convention for b_{ik}. Inverting, $g_{ik} = (1 + h_{ik} d_{ik})/2$. The threshold 0.5 is the unique one for which the identity holds: if $b_{ik} = \mathbf{1}[g_{ik} > \theta]$ with $\theta \neq 0.5$, then for g_{ik} strictly between θ and 0.5 the sign of d_{ik} disagrees with that of x and $h_{ik} d_{ik} = -x$.

For the Bayesian interpretation, condition on g_{ik}; then d_{ik} is deterministic with $d_{ik}^2 = 1$ and $\mathbb{E}[2s_{ik} - 1 \mid g_{ik}] = 2g_{ik} - 1$, so

$$\mathbb{E}[(2s_{ik} - 1)d_{ik} \mid g_{ik}] = (2g_{ik} - 1)d_{ik} = h_{ik} d_{ik}^2 = h_{ik}. \tag{17}$$

The product $(2s_{ik} - 1)d_{ik}$ is $+1$ for agreement and -1 for disagreement, so its expectation is $2\Pr(s_{ik} = b_{ik} \mid g_{ik}) - 1 = h_{ik}$. This interpretation requires the stated calibration assumption. Retaining h_{ik} in Eq. equation 11 preserves this input margin before clipping and centering, not an absolute scale for the final standardized advantage. □

B.3 Proof of Proposition 2 and scope of the proxy

Under the proposition's calibration assumption, $\mathbb{E}[s_{ik} \mid g_{ik}] = g_{ik}$ and $\text{Var}(s_{ik} \mid g_{ik}) = g_{ik}(1 - g_{ik})$. Taking moments with i uniform on $\mathcal{V}_k$ gives

$$\text{Var}(s_{ik}) = \text{Var}(\mathbb{E}[s_{ik} \mid g_{ik}]) + \mathbb{E}[\text{Var}(s_{ik} \mid g_{ik})] = \text{Var}(g_{ik}) + \mathbb{E}[g_{ik}(1 - g_{ik})]. \tag{18}$$

Dividing by positive $\text{Var}(s_{ik})$ proves the formula for λ_k^{unif}. If $\text{Var}(g_{ik}) > 0$ as well, $\text{Cov}(g_{ik}, s_{ik}) = \text{Var}(g_{ik})$, so $\text{Corr}(g_{ik}, s_{ik}) = \sqrt{\lambda_k^{\text{unif}}}$. These conclusions apply to the calibrated, uniform-weight special case only. □

For the actual proxy, V_k^{between} uses the measure assigning probability $c_{ik}/\sum_j c_{jk}$ to response i, whereas u_k^{within} uses uniform weights. A total-variance identity under the weighted measure would require weighting *both* terms consistently. Thus Eq. equation 8 is not that identity, nor an estimator of $\Delta_k^2\pi_k(1-\pi_k)$ and $\bar{\tau}_k^2$ from Eq. equation 3. Its nonnegative components place it in $[0,1]$ whenever the denominator is positive; the zero-denominator convention covers unanimous endpoint scores. Confident disagreement can give a high proxy even when the judge is wrong, and confidence weighting need not lower the ratio relative to uniform weights.

B.4 Boundary normalization and constructed examples

Proposition 3 (Weighted boundary variance)**.** *Assume* $\sum_{i\in\mathcal{V}_k} c_{ik} > 0$ *and let* $\omega_i = c_{ik}/\sum_j c_{jk}$. *Then* $\mathrm{Var}_\omega(b_{ik}) = \bar{b}_k(1-\bar{b}_k) = V_k^{\text{sem}}$. *This variance vanishes if and only if all positive-weight responses share the same observed boundary side. Separately, under any fixed joint distribution of* (b,s), *let* $\varepsilon_{k,s} = \Pr(b_{ik} \neq s_{ik} \mid s_{ik} = s)$. *Then* $\mathrm{Var}(b_{ik} \mid s_{ik} = s) = \varepsilon_{k,s}(1-\varepsilon_{k,s})$*; under equal class-conditional error rates,* $\mathbb{E}[b \mid s=1] - \mathbb{E}[b \mid s=0] = 1 - 2\varepsilon_k$.

Proof. A binary variable with weighted mean $\bar{b}_k$ has weighted variance $\bar{b}_k - \bar{b}_k^2$. It is zero exactly when that variable is constant on the positive-weight support. Given $s=0$, b is Bernoulli with parameter $\varepsilon_{k,0}$; given $s=1$, its parameter is $1-\varepsilon_{k,1}$. For thresholded scores $b_{ik} = \mathbf{1}[g_{ik} > 0.5]$, the score-specific moments are therefore

$$\Delta_k = 1 - \varepsilon_{k,1} - \varepsilon_{k,0},$$
$$\bar{\tau}_k^2 = \mathbb{E}[\varepsilon_{k,s_{ik}}(1-\varepsilon_{k,s_{ik}})].$$

The law of total variance yields the same two-term decomposition as Eq. equation 3, with these binary moments. Under equal class-conditional error rates, the mean gap reduces to $1-2\varepsilon_k$. □

These identities do not imply that thresholding always reduces noise: score variance and binary misclassification variance are different quantities. The exact operational property is narrower: same-side score fluctuations cannot activate the boundary channel, although erroneous cross-boundary verdicts can. The following examples illustrate the formulas, not measured training outcomes.

Example A (unanimous observed verdicts). Let $G = 8$, with seven scores at $g = 0.9$ and one at $g = 0.7$, so $b \equiv 1$. Continuous normalization gives $\bar{g} = 0.875$, $\sigma = \sqrt{0.004375} \approx 0.0661$, $z(0.7) \approx -2.65$, and $z(0.9) \approx +0.38$. Boundary normalization instead gives $V_k^{\text{sem}} = 0$ and zero contributions. The continuous difference is nuisance if the responses truly have identical binary satisfaction states; the probabilities alone do not establish that premise.

Example B (cross-boundary contrast). For $g = [0.95, 0.95, 0.95, 0.9, 0.7, 0.05, 0.05, 0.1]$, confidence weighting gives $\bar{b}_k = 3.23/5.49 \approx 0.588$ and $\sqrt{V_k^{\text{sem}}} \approx 0.492$. Before centering, $z^{\text{bd}} \approx +0.753$ at $g = 0.95$, $+0.335$ at $g = 0.7$, and -1.076 at $g = 0.05$; thus the margin retains differences among passing verdicts. The uniform-weight reliability is $\lambda_k^{\text{unif}} \approx 0.678$, whereas the actual weighted proxy is $\lambda_k \approx 0.700$. The same boundary pattern with scores $[0.6, 0.6, 0.6, 0.6, 0.55, 0.4, 0.4, 0.45]$ gives $\lambda_k \approx 0.034$. These are pre-standardization weights, not a guaranteed ratio of final policy-update magnitudes.

Example C (centering can change signs). For $g = [0.6, 0.51, 0.05]$, the uncentered statistic is approximately $[0.896, 0.090, -0.201]$, and the centered statistic is $[0.634, -0.172, -0.462]$. The response at $g = 0.51$ therefore has a negative candidate contribution despite its passing verdict. This concerns the boundary statistic before routing and shows why no general sign guarantee follows from a passing verdict.

B.5 Zero-mean property and the advantage of rejected rewrites

By definition $\sum_{i\in\mathcal{V}_k} \tilde{z}_{ik}^{\text{bd}} = 0$, and since $\tilde{w}_k r_k$ is constant across i, $\sum_{i\in\mathcal{V}_k} A_{ik}^{\text{bd}} = \tilde{w}_k r_k \sum_{i\in\mathcal{V}_k} \tilde{z}_{ik}^{\text{bd}} = 0$. Responses outside $\mathcal{V}_k$ contribute zero, so $\sum_{i\in\mathcal{B}_t} A_{ik}^{\text{bd}} = 0$ for every criterion and, summing over $k \in \mathcal{A}_t$, $\sum_{i\in\mathcal{B}_t} S_i = 0$.

If a rewrite fails the gate, $S_{\text{rw}} = 0$ and $\bar{S} = K^{-1}(\sum_{i\in\mathcal{B}_t} S_i + 0) = 0$, so $\widehat{A}_{\text{rw}} = 0$. Original responses keep their signs and ranking relative to normalization without the rejected rewrite, although

including a zero can change the common scale. This is a statement about the reward-driven term, not a mask for KL regularization. If the gate passes, $\bar{S} = A_{\mathrm{rw}}/K$ and

$$\widehat{A}_{\mathrm{rw}} = \frac{(1 - 1/K)A_{\mathrm{rw}}}{\mathrm{Std}(S) + \varepsilon_{\mathrm{std}}}. \tag{19}$$

For $K > 1$, its sign is therefore the sign of A_{rw}, which need not be positive. If $S_{\mathrm{par}} = 0$, Eq. equation 13 gives $A_{\mathrm{rw}} = 0$ regardless of target improvements. In particular, a group with no ACTIVE criterion has zero scores throughout and no reward-driven update.

B.6 EFFECT OF FINAL GROUP STANDARDIZATION

For any fixed score vector S and positive scalar a,

$$\widehat{A}_i(aS) = \frac{a(S_i - \bar{S})}{a\,\mathrm{Std}(S) + \varepsilon_{\mathrm{std}}} = \frac{S_i - \bar{S}}{\mathrm{Std}(S) + \varepsilon_{\mathrm{std}}/a}. \tag{20}$$

If $\varepsilon_{\mathrm{std}} = 0$ and $\mathrm{Std}(S) > 0$, common positive scaling cancels exactly. With the stability term, cancellation is approximate when $a\,\mathrm{Std}(S) \gg \varepsilon_{\mathrm{std}}$. For one ACTIVE criterion and no rewrite, r_k is such a common factor; it cannot provide absolute attenuation in the scale-free regime. With multiple criteria, changing their reliability weights can change the relative score ordering and thus the update. Routing an entire group to zero remains zero even after standardization. These properties motivate describing reliability as relative weighting and selection rather than a universal gradient-magnitude control.

B.7 QUALITY GATE AND AVERAGE UNCERTAINTY PROXY

For any $g \in [0, 1]$, with $h = |2g - 1|$ and $c = h^2$, we have $c = 1 - 4g(1 - g)$. For a nonempty valid set,

$$u_k^{\mathrm{within}} = \frac{1}{G_k} \sum_{i \in \mathcal{V}_k} g_{ik}(1 - g_{ik}) = \frac{1}{4G_k} \sum_{i \in \mathcal{V}_k} (1 - c_{ik}) = \frac{1 - \bar{c}_k}{4}. \tag{21}$$

Thus $\bar{c}_k > \theta_c$ if and only if $u_k^{\mathrm{within}} < (1 - \theta_c)/4$. The equivalence is algebraic; interpreting it as a bound on actual posterior uncertainty requires calibration, and it is not a bound on the judge's error rate.

C EXPERIMENTAL DETAILS

C.1 HYPERPARAMETERS

Table 4 reports shared training settings and RITR-specific hyperparameters. The total budget is $K = 8$ responses per prompt: RITR uses $G = 7$ original-prompt responses plus one rewrite when repair is selected, and all K responses otherwise come from the original prompt.

C.2 TRAINING DYNAMICS

Figure 3 shows one training run per method over 500 steps. The baseline's smoothed entropy decreases while KL rises gradually. RITR's KL and entropy peak during the first 200 steps and then decline; KL partially rises again later and remains above the baseline at the end. The entropy gap narrows substantially after its early peak. The timing is compatible with a transient effect during scaffold use, but these aggregate metrics cannot determine whether the policy retained particular repaired behaviors or whether the schedule caused the final benchmark gains. We treat them as descriptive diagnostics rather than evidence of behavioral internalization.

C.3 ABLATION PROTOCOL

The triage-only ablation (Table 3) sets $\rho_0 = 0$, so no rewrite is ever generated and all K samples are drawn from the original prompt, and otherwise uses the RITR configuration. The decay-schedule ablation (Figure 2) replaces $\kappa(t; T)$ by $1 - t/T$ in both stages with unchanged $T_1 = T_2 = 100$. Each configuration was trained once.

Table 4: Training configuration. Shared settings apply to Base Rubric RL, the triage-only ablation, and RITR.

Setting	Value
Policy initialization	Qwen3-8B
RL framework	verl
Learning rate	1×10^{-6}
Batch size / micro-batch size	64 / 32
Training steps	500 (healthcare, science), 300 (finance)
Group size G / total budget K	7 / 8
Training judge (reward)	Qwen3-14B
Policy decoding and response length	temperature $= 1.0$, maximum 4096 tokens, thinking disabled
Clip ratio ϵ_{clip}, KL coefficient β	0.2, 10^{-3}
Final score normalization	$\varepsilon_{\text{std}} = 10^{-8}$; population standard deviation over all K responses
Quality margin m, quality threshold θ_c	0.5, 0.6
Boundary stability ϵ and clipping $z_{\max}$	10^{-6}, 2
Gate thresholds τ_{pres}, τ_{tar}	0.5, 0.5
Initial bonus α_{rw}^0 and rewrite ratio ρ_0	0.12, 1.0
Decay lengths T_1 (bonus), T_2 (ratio)	100 steps each, sequential
Decay schedule	Cosine; linear in Figure 2

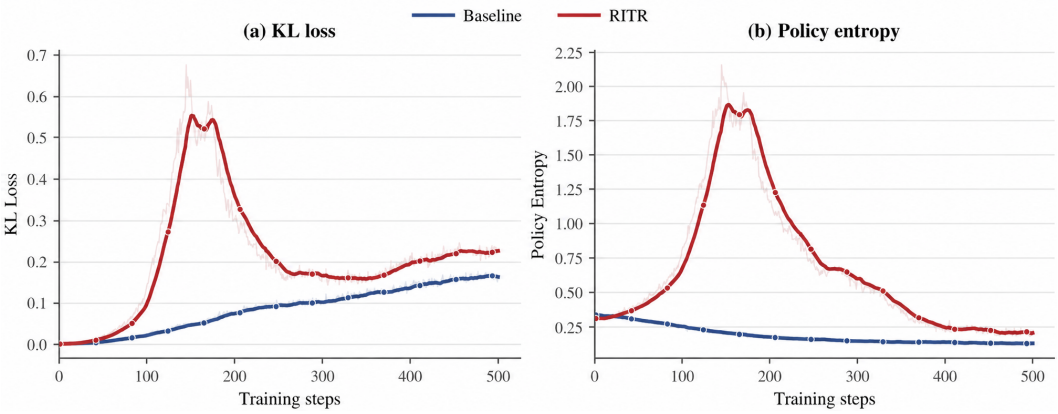

Figure 3: HealthBench training dynamics of Base Rubric RL (blue) and RITR (red): (a) actor KL loss against the reference model and (b) policy entropy. Opaque curves are smoothed; faint curves are raw values.