RISE-RL: RUBRIC-INFORMED SELECTIVE EXPLORATION FOR OPEN-ENDED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning Large Language Models (LLMs) for open-ended tasks is challenging because responses must satisfy multidimensional criteria without following a single correct generation trajectory. Existing rubric-based reinforcement learning (RL) methods compress fine-grained criterion-level feedback into scalar rewards, making persistent capability gaps difficult to target under limited on-policy exploration. We propose **RISE-RL** (Rubric-Informed Selective Exploration), which uses frequently missed rubric criteria under natural rollouts to elicit privileged trajectories that are difficult to discover through unguided exploration alone. RISE-RL retains only trajectories whose complete-rubric reward exceeds the mean reward of natural rollouts, and then re-evaluates them under the original prompt to emphasize behaviors that remain weakly supported by the natural policy. The resulting guidance signal is optimized through a separate auxiliary objective and removed once its additional benefit diminishes. Experiments with 4B and 14B models across writing, chat, health, and science show that RISE-RL achieves the highest mean score on every evaluated benchmark under guidance-free evaluation. Compared with standard Rubric-RL, it improves the average score by 1.3 points at the 4B scale and **3.3 points at the 14B scale**, including a **6.0-point** gain on CreativeWriting-V3. It also improves creative-writing diversity and yields gains on objectively scored medical and scientific benchmarks. These results indicate that selective internalization through reward filtering and policy support shaping is effective for open-ended reinforcement learning.

1 INTRODUCTION

Open-ended question answering is a core capability of large language models, spanning open-domain dialogue (Zhang et al., 2020; Roller et al., 2021; Thoppilan et al., 2022), creative writing (Franceschelli & Musolesi, 2023; Gómez-Rodríguez & Williams, 2023), health assistance (Lin et al., 2025; Singhal et al., 2023; 2025), and scientific explanation (Wan et al., 2024; Besrour et al., 2025). Unlike mathematical reasoning (Ren et al., 2025; Chen et al., 2025) or code generation (Le et al., 2022; Cao et al., 2026), open-ended tasks admit no single correct trajectory and require responses to satisfy diverse criteria, including factuality, completeness, style, structure, safety, and human preferences. Moreover, substantially different responses may be equally valid, giving rise to a multi-peaked policy distribution with multiple desirable generation modes. Reinforcement learning for open-ended tasks therefore faces the challenge of improving rubric alignment without sacrificing exploration or response diversity. Balancing these objectives remains a central challenge.

Existing methods still struggle to improve open-ended capabilities while preserving exploration and diversity. Rubric-based RL provides fine-grained criterion-level feedback but aggregates criterion-wise scores into a scalar sequence-level reward for policy optimization (Gunjal et al., 2025; Huang et al., 2026b). Consequently, criterion-specific failures may not be directly translated into targeted policy updates. OPD provides dense token-level supervision by distilling a teacher distribution on states visited by the student policy (Li et al., 2026b; Zhao et al., 2026b; Hübotter et al., 2026). While such supervision facilitates knowledge transfer, it offers limited control over which guided behaviors are selectively internalized. Mixed-policy optimization further combines privileged trajectories with on-policy samples in the same group-relative objective (Bi et al., 2025; Huang et al., 2026a). This

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.

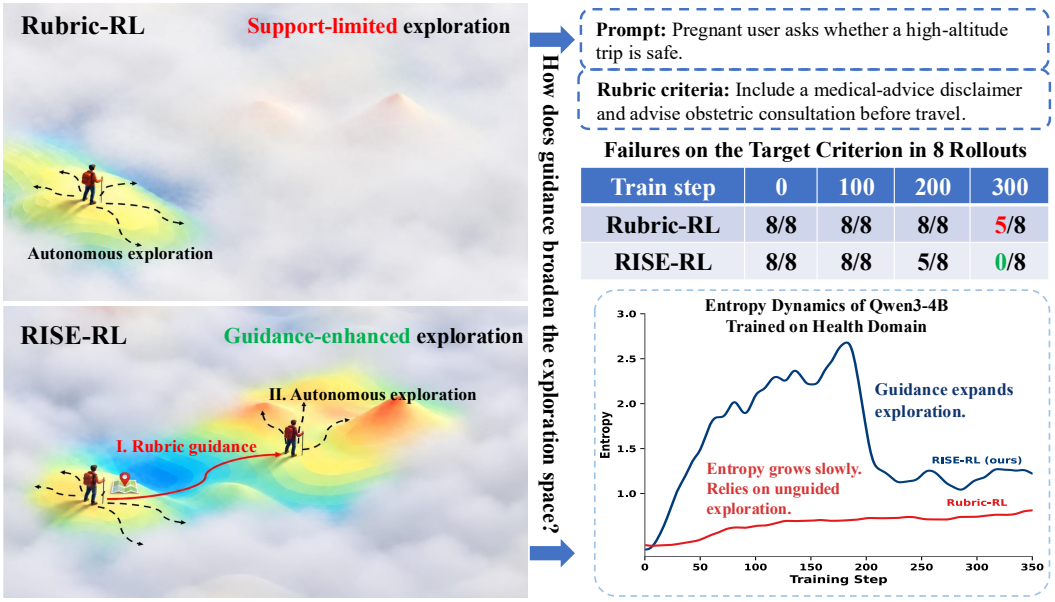

Figure 1: Overview of the guidance–exploration mechanism in RISE-RL. Selective rubric guidance addresses criterion failures and expands the accessible policy space. Once these behaviors become accessible, guidance is removed and training continues with autonomous on-policy exploration.

coupling alters the relative advantages of natural rollouts and can cause scaffold-removal collapse, a performance drop when generation returns to natural prompts.

To address these limitations, we propose RISE-RL, a selective-guidance paradigm for open-ended tasks. RISE-RL uses rubric feedback to identify capability gaps, filters guided trajectories by complete-rubric reward, and estimates their natural-policy support—the token probabilities assigned under the original prompt—to emphasize beneficial behaviors that remain unlikely without guidance. The resulting signal is optimized through a separate auxiliary objective, keeping guided learning outside natural group-relative optimization. As its benefit diminishes, guidance is removed and training continues with unguided on-policy exploration, facilitating selective internalization while retaining autonomous exploration and response diversity.

We evaluate RISE-RL on Qwen3-4B and Qwen3-14B across four RubricHub domains and eight downstream benchmarks. RISE-RL achieves higher mean scores than standard Rubric-RL on every evaluated benchmark at both model scales, with average gains of **1.3** and **3.3** points at the 4B and 14B scales, respectively. On Qwen3-14B, the gains reach **8.0 points** on Arena-Hard-v2 and **6.0 points** on CreativeWriting-V3. In the creative-writing evaluation, RISE-RL also improves output diversity by **5.8%**. These gains further extend to objectively scored tasks, including improvements of **3.3 points** on MedQA and **3.6 points** on GPQA-Diamond. Overall, these results suggest that RISE-RL improves complex rubric alignment while supporting generation diversity and facilitating the internalization of domain knowledge and reasoning capabilities.

Our main contributions, which directly address the limitations of prior work, are threefold:

- **Criterion-Level Selective Guidance for Open-Ended RL:** We introduce RISE-RL, which retains the fine-grained diagnostic information that is normally collapsed into a scalar rubric reward. By identifying high-value criteria that remain repeatedly unsatisfied across natural rollouts, RISE-RL constructs targeted guidance for behaviors that are rarely elicited through unguided exploration.
- **Selective Internalization by Gain and Policy Support:** RISE-RL retains only privileged trajectories that improve the complete-rubric reward over the natural-rollout baseline. It then removes the privileged criteria and re-evaluates the retained trajectories under the original prompt, using token probabilities to concentrate learning on beneficial behaviors that remain weakly supported by the natural policy. The resulting guidance signal is optimized

through a separate auxiliary objective, preventing privileged trajectories from affecting the reward normalization and relative-advantage estimation of natural on-policy rollouts.

- **Broad Empirical Evaluation across Models and Domains:** Experiments with Qwen3-4B and Qwen3-14B across writing, chat, health, and science show that RISE-RL achieves higher mean scores than Rubric-RL on every evaluated benchmark at both scales under guidance-free evaluation. The gains are particularly pronounced on challenging open-ended benchmarks at the 14B scale, while improvements on objectively scored tasks further suggest within-domain transfer to verifiable tasks. Training-dynamics analyses and ablations additionally support the benefits of selective early guidance followed by unguided on-policy exploration.

2 Related Work

Rubric-Based RL and OPD. Rubric-based RL (Gunjal et al., 2025) characterizes the multidimensional quality of open-ended responses through fine-grained evaluation criteria and aggregates criterion-level scores into a scalar reward for policy optimization. However, this aggregation can discard part of the diagnostic information provided by the judge, making failures on specific criteria difficult to translate into targeted policy updates. A complementary line of work instead provides dense supervision through policy distillation. On-policy distillation (OPD) queries a teacher distribution on states visited by the student policy (Li et al., 2026b), while on-policy self-distillation (OPSD) constructs the teacher using the same model under privileged contexts (Zhao et al., 2026a). Rubric-Guided Self-Distillation further supplies instance-specific rubrics as privileged information to the teacher (Rezaei et al., 2026). Although these methods offer token-level guidance, their objectives primarily emphasize matching the teacher distribution, without explicitly coordinating dense supervision with the preservation of autonomous policy exploration.

Guided Trajectory Learning for Policy Optimization. Recent work has begun to combine natural on-policy rollouts with externally guided trajectories. LUFFY integrates on-policy rollouts with off-policy expert demonstrations and introduces policy shaping to balance imitation and exploration (Yan et al., 2026). Critique-GRPO (Zhang et al., 2025b) and RGR-GRPO (Bi et al., 2025) further construct refined responses from critiques and rubrics, respectively, and jointly optimize them with the original rollouts. RuscaRL instead injects rubric-based scaffolds into rollout prompts, producing trajectories conditioned on additional guidance (Zhou et al., 2025). These methods demonstrate that guided trajectories can expose the policy to high-quality behaviors that are difficult to discover through autonomous exploration alone. However, guided and natural trajectories often share the same optimization objective, coupling the learning signal from privileged trajectories with the relative advantages of natural samples. This coupling may reduce control over their respective contributions and exacerbate distribution shift when the scaffold is removed. RISE-RL differs by filtering guided trajectories with the complete rubric, estimating their policy support after removing the privileged context, and selectively internalizing them through a separate auxiliary objective.

3 Method

3.1 Group Relative Policy Optimization

We build on Group Relative Policy Optimization (GRPO) (Shao et al., 2024) for policy optimization under the unguided prompt. We refer to this as the natural-rollout branch. Given a prompt q, the old policy $\pi_{\theta_{\text{old}}}$ samples a group of G responses $\{o_i\}_{i=1}^{G}$, each assigned a sequence-level reward r_i. GRPO estimates the advantage of each response by normalizing its reward within the group:

$$\hat{A}_i = \frac{r_i - \text{mean}\big(\{r_j\}_{j=1}^{G}\big)}{\text{std}\big(\{r_j\}_{j=1}^{G}\big) + \epsilon_A}, \tag{1}$$

where ϵ_A is a small constant for numerical stability. The resulting sequence-level advantage is shared by all tokens in o_i. We instantiate the natural rollout branch with GRPO, following standard implementation practice.

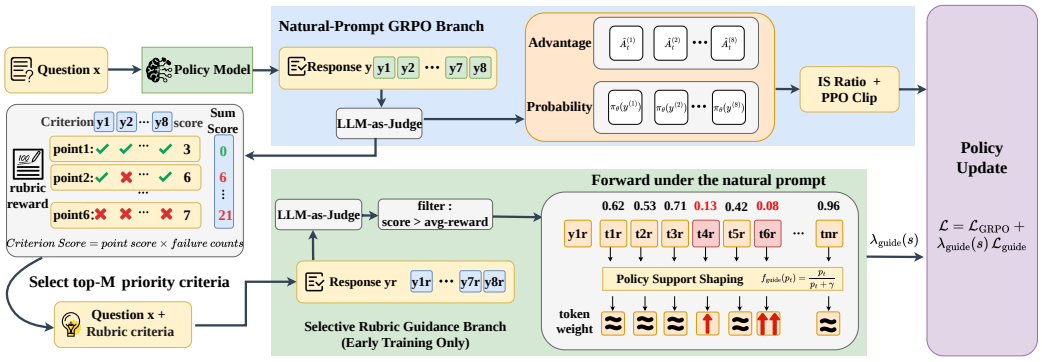

Figure 2: Overview of RISE-RL. Alongside natural-prompt GRPO, an early selective-guidance branch targets unmet rubric criteria and reinforces low-support tokens in reward-improving trajectories. Guidance is removed once its reward advantage saturates, after which training continues with pure GRPO.

3.2 Criterion-Level Selective Feedback

For each prompt q, we associate a rubric $\mathcal{C}(q) = \{c_k\}_{k=1}^{K}$, where criterion c_k has weight w_k. For response o_i, the rubric grader produces a binary judgment

$$z_{i,k} = \mathbb{I}[o_i \text{ satisfies } c_k], \qquad z_{i,k} \in \{0, 1\}, \tag{2}$$

and GRPO uses the aggregated sequence-level reward

$$r_i = \frac{\sum_{k=1}^{K} w_k z_{i,k}}{\sum_{k=1}^{K} w_k}. \tag{3}$$

Although suitable for policy optimization, this scalar reward obscures which important criteria are repeatedly missed across natural rollouts. Given $\mathcal{O}^{\text{nat}} = \{o_i^{\text{nat}}\}_{i=1}^{G}$, we therefore define the priority of criterion c_k as

$$p_k = w_k \sum_{i=1}^{G} \left(1 - z_{i,k}^{\text{nat}}\right), \tag{4}$$

which jointly captures criterion importance and failure frequency within the current rollout group. We then select the M highest-priority criteria as targeted feedback:

$$\mathcal{C}^{\text{fb}}(q) = \text{TopM}\left(\mathcal{C}(q); \{p_k\}_{k=1}^{K}\right). \tag{5}$$

We select at most M criteria with positive priority; if fewer than M have $p_k > 0$, we use only those criteria, and skip guidance when none exists. This concentrates guidance on high-value capabilities that remain insufficiently covered by the current policy.

3.3 Quality-Filtered Selective Guidance

We append the selected criteria to the original prompt and sample a group of privileged trajectories:

$$\begin{aligned} q^{\text{priv}} &= q \oplus \mathcal{C}^{\text{fb}}(q), \\ o_j^{\text{priv}} &\sim \pi_{\theta_{\text{old}}}\left(\cdot \mid q^{\text{priv}}\right), \qquad j = 1, \ldots, G. \end{aligned} \tag{6}$$

Because privileged conditioning may cause cross-criterion degradation or prompt mismatch, we re-evaluate each candidate using the complete rubric; this filtering reduces, but does not eliminate, these risks. Let $\bar{r}^{\text{nat}}$ and σ^{nat} denote the mean and standard deviation of the natural-rollout rewards $\{r_i^{\text{nat}}\}_{i=1}^{G}$, respectively. We define the normalized reward gain of each privileged trajectory as

$$A_j^{\text{ref}} = \frac{r_j^{\text{priv}} - \bar{r}^{\text{nat}}}{\sigma^{\text{nat}} + \epsilon_A}, \qquad \epsilon_A = 10^{-6}. \tag{7}$$

To identify important yet under-supported tokens in reward-improving trajectories, we remove the selected criteria and teacher-force each reference trajectory under the original prompt:

$$p_{j,t} = \pi_\theta \left(o_{j,t}^{\text{ref}} \mid q, o_{j,<t}^{\text{ref}} \right). \tag{8}$$

Re-evaluating the retained trajectory under the original prompt estimates its support without privileged criteria: a low $p_{j,t}$ indicates that the corresponding token, despite appearing in a higher-reward trajectory, remains weakly supported by the natural policy. Following the saturating transformation in LUFFY (Yan et al., 2026), we define the policy support factor

$$\rho_{j,t} = \frac{p_{j,t}}{p_{j,t} + \gamma}, \tag{9}$$

where $\gamma > 0$ controls the sensitivity to low-probability tokens. As $p_{j,t}$ decreases, $-\log \rho_{j,t}$ increases, assigning stronger guidance to weakly supported behavior while suppressing updates to already-supported tokens.

For each retained trajectory $o_j^{\text{ref}} \in \mathcal{O}^{\text{ref}}$, let T_j denote its token length. We define the trajectory-level weight as $\widetilde{A}_j = \min\big(A_j^{\text{ref}}, A_{\max}\big)$, where $A_{\max} = 5$ limits large normalized gains when the natural-rollout reward variance is small. The selective guidance loss is then

$$\mathcal{L}_{\text{guide}} = -\frac{1}{|\mathcal{O}^{\text{ref}}|} \sum_{o_j^{\text{ref}} \in \mathcal{O}^{\text{ref}}} \frac{\widetilde{A}_j}{T_j} \sum_{t=1}^{T_j} \log \rho_{j,t}. \tag{10}$$

This objective combines two forms of selectivity: reward filtering removes non-improving privileged trajectories, while policy support weighting focuses learning on higher-reward behaviors weakly supported by the current policy. Together, reward-based trajectory selection and probability-adaptive token weighting enable selective internalization of guided responses. We recompute $p_{j,t}$ under the current policy π_θ and backpropagate only through these token probabilities, treating trajectories, rewards, and trajectory weights as fixed. If $\mathcal{O}^{\text{ref}} = \emptyset$, we set $\mathcal{L}_{\text{guide}} = 0$.

3.4 Decoupled Optimization and Guidance Removal

We incorporate the filtered guidance signal through a separate auxiliary objective:

$$\mathcal{L}(\theta, s) = \mathcal{L}_{\text{GRPO}}(\theta) + \lambda_{\text{guide}}(s)\mathcal{L}_{\text{guide}}(\theta), \tag{11}$$

where s is the training step. Natural rollouts are optimized by GRPO, while filtered privileged trajectories contribute only through the auxiliary loss. This keeps them outside the reward normalization and relative-advantage estimation of natural rollouts, while allowing independent control over guidance strength and token-level weighting.

As the policy internalizes the guided behaviors, the reward advantage of privileged rollouts over natural rollouts gradually decreases. We measure this additional benefit by

$$\begin{aligned} \Delta_r(s) &= \bar{r}^{\text{priv}}(s) - \bar{r}^{\text{nat}}(s), \\ \bar{r}^{\text{priv}}(s) = \frac{1}{G}\sum_{j=1}^{G} r_j^{\text{priv}}(s), &\qquad \bar{r}^{\text{nat}}(s) = \frac{1}{G}\sum_{i=1}^{G} r_i^{\text{nat}}(s). \end{aligned} \tag{12}$$

A large $\Delta r(s)$ indicates that privileged feedback still elicits behaviors less accessible to the natural policy, whereas a narrowing gap suggests diminishing marginal benefit.

Based on a preliminary run, we set s_{switch} near the onset of the plateau in the smoothed reward-gap curve and use $\lambda_{\text{guide}}(s) = \lambda_0 \mathbb{I}[s < s_{\text{switch}}]$, removing guidance thereafter to continue training with pure GRPO. Before s_{switch}, RISE-RL adds G privileged rollouts/evaluations per prompt and teacher-forces the retained trajectories, yielding a rollout/evaluation multiplier of $1 + s_{\text{switch}}/S$; wall-clock costs are reported in Appendix B.2.

4 Experiments

4.1 Experimental Setup

Datasets. To evaluate the effectiveness of our method across diverse open-ended domains, we construct the training corpus from RubricHub (Li et al., 2026a), a large-scale, multi-domain dataset

Table 1: Main results across four open-ended domains using Qwen3-4B and Qwen3-14B. Each selected checkpoint is independently evaluated three times, and the mean performance is reported. Colored subscripts indicate absolute changes relative to the corresponding Initial model, with improvements shown in green and degradations in red. The best result within each model scale is shown in bold.

Method	Writing		Chat	Health			Science		Avg.
	Writing Bench	Creative Writing-V3	Arena-Hard V2	Health Bench	LLMEval-Med	MedQA	GPQA	Research QA	
Qwen3-4B (Non-Thinking)									
Initial	56.14	40.28	7.91	37.43	64.76	64.79	42.34	64.29	47.24
SFT	66.43$_{+10.29}$	32.96$_{-7.32}$	11.39$_{+3.48}$	40.73$_{+3.30}$	64.87$_{+0.11}$	59.77$_{-5.02}$	36.95$_{-5.39}$	69.19$_{+4.90}$	47.79$_{+0.55}$
OPD	54.55$_{-1.59}$	27.78$_{-12.50}$	6.32$_{-1.59}$	35.46$_{-1.97}$	63.07$_{-1.69}$	61.63$_{-3.16}$	40.24$_{-2.10}$	60.35$_{-3.94}$	43.68$_{-3.56}$
Rubric-RL	70.92$_{+14.78}$	43.30$_{+3.02}$	20.76$_{+12.85}$	52.00$_{+14.57}$	71.85$_{+7.09}$	68.44$_{+3.65}$	47.73$_{+5.39}$	74.52$_{+10.23}$	56.19$_{+8.95}$
RuscaRL	56.75$_{+0.61}$	38.67$_{-1.61}$	9.41$_{+1.50}$	43.54$_{+6.11}$	65.81$_{+1.05}$	63.39$_{-1.40}$	41.41$_{-0.93}$	66.31$_{+2.02}$	48.16$_{+0.92}$
RISE-RL (Ours)	**72.15**$_{+16.01}$	**46.73**$_{+6.45}$	**20.89**$_{+12.98}$	**52.32**$_{+14.89}$	**72.22**$_{+7.46}$	**69.40**$_{+4.61}$	**49.12**$_{+6.78}$	**77.08**$_{+12.79}$	**57.49**$_{+10.25}$
Qwen3-14B (Non-Thinking)									
Initial	63.24	64.94	21.09	43.85	69.46	73.51	52.61	68.27	57.12
SFT	73.31$_{+10.07}$	60.60$_{-4.34}$	33.74$_{+12.65}$	48.66$_{+4.81}$	71.63$_{+2.17}$	74.15$_{+0.64}$	39.90$_{-12.71}$	73.54$_{+5.27}$	59.44$_{+2.32}$
OPD	66.05$_{+2.81}$	51.53$_{-13.41}$	20.07$_{-1.02}$	45.39$_{+1.54}$	71.56$_{+2.10}$	74.91$_{+1.40}$	48.82$_{-3.79}$	69.69$_{+1.42}$	56.00$_{-1.12}$
Rubric-RL	76.89$_{+13.65}$	64.93$_{-0.01}$	49.94$_{+28.85}$	57.92$_{+14.07}$	76.40$_{+6.94}$	78.22$_{+4.71}$	55.98$_{+3.37}$	77.26$_{+8.99}$	67.19$_{+10.07}$
RuscaRL	62.45$_{-0.79}$	42.78$_{-22.16}$	45.96$_{+24.87}$	50.92$_{+7.07}$	71.49$_{+2.03}$	77.39$_{+3.88}$	46.13$_{-6.48}$	74.10$_{+5.83}$	58.90$_{+1.78}$
RISE-RL (Ours)	**77.37**$_{+14.13}$	**70.96**$_{+6.02}$	**57.97**$_{+36.88}$	**59.44**$_{+15.59}$	**78.31**$_{+8.85}$	**81.57**$_{+8.06}$	**59.60**$_{+6.99}$	**78.96**$_{+10.69}$	**70.52**$_{+13.40}$

equipped with fine-grained rubric-based evaluation criteria. RubricHub covers five domains: writing, health, chat, science, and instruction following. We select the writing, health, chat, and science subsets for reinforcement learning training and validation, thereby enabling a broad assessment of the proposed method across heterogeneous open-ended tasks.

Benchmarks. We evaluate RISE-RL across four domains using open-ended generation and verifiable reasoning tasks: (1) **Writing**: WritingBench (Wu et al., 2026) and CreativeWriting-V3; (2) **Chat**: Arena-Hard-v2 (Li et al., 2024); (3) **Health**: HealthBench (Arora et al., 2025), LLMEval-Med (Zhang et al., 2025a), and MedQA (Jin et al., 2021); (4) **Science**: ResearchQA (Yifei et al., 2026) and GPQA-Diamond (Rein et al., 2023). This suite assesses the model's ability to satisfy multidimensional rubrics in creative tasks and maintain factual rigor in knowledge-dense domains. All evaluations follow official protocols.

Baselines. We compare RISE-RL against imitation and RL-based baselines using Qwen3-4B/14B models: (1) **SFT**: Fine-tuned on high-quality responses from RubricHub. (2) **OPD** (Li et al., 2026b): On-policy distillation using Qwen3-14B as a teacher for the 4B model, and Qwen3-32B for the 14B model. (3) **Rubric-RL** (Gunjal et al., 2025): Standard RL using scalar rewards aggregated from fine-grained rubric criteria. (4) **RuscaRL** (Zhou et al., 2025): A scaffold-mixed baseline that jointly optimizes rubric-conditioned and natural responses within a single group-relative objective.

Implementation Details. We use **gpt-oss-120b** as the rubric grader, which has demonstrated high agreement with human judgment (Li et al., 2026a). SFT and OPD are implemented using SWIFT (Zhao et al., 2025), while all RL methods are implemented using verl (Sheng et al., 2025). During the early guidance stage, RISE-RL additionally uses quality-filtered privileged candidates through a separate auxiliary objective. These candidates are excluded from the natural GRPO group and are no longer generated after s_{switch}. All RL methods are trained separately for each domain.

For policy optimization, we use GRPO for the natural rollout branch. In RISE-RL, the shaping parameter is $\gamma = 0.1$, and the guidance weight λ_{guide} is set to 0.05 for writing and chat, and 0.01 for health and science domains. Baselines, including RuscaRL, follow their best-reported configurations (Zhou et al., 2025). Further implementation details, full training configurations, and evaluation protocols are provided in Appendix A.

4.2 Main Results

Table 1 reports the downstream performance of RISE-RL. Overall, RISE-RL consistently achieves the strongest results across both Qwen3-4B and Qwen3-14B scales, effectively balancing complex alignment with autonomous exploration.

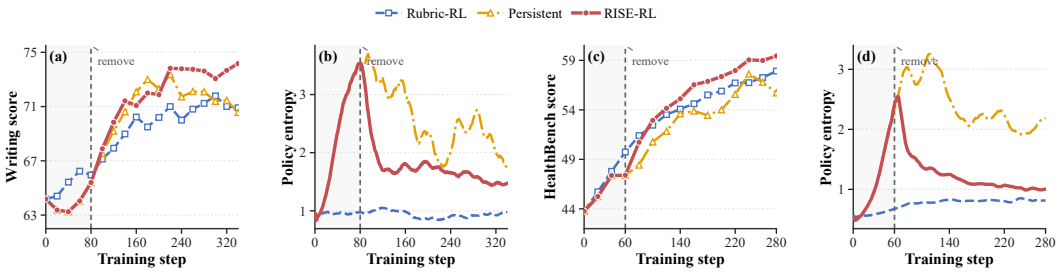

Figure 3: Training dynamics of Qwen3-14B on the writing and health domains. We compare the benchmark performance and policy entropy of Rubric-RL, persistent guidance, and RISE-RL on (a–b) the writing domain and (c–d) the health domain.

Creative Domains: Enhancing Quality and Diversity. In domains requiring high creativity and output diversity, such as **Writing** and **Chat**, RISE-RL shows substantial gains (up to **+6.0** and **+8.0** points on the 14B scale, respectively). On challenging benchmarks like **CreativeWriting-V3**, standard RL often struggles to discover high-reward trajectories or suffers from diversity collapse. RISE-RL's ability to target failed criteria exposes the model to rare creative modes (e.g., narrative coherence), while its *guidance removal* mechanism ensures that the policy returns to unguided exploration, preserving the creative openness essential for these tasks.

Knowledge-Dense Domains: Internalizing Logical Requirements. In domains governed by strict factual and logical constraints, such as **Health** and **Science**, RISE-RL demonstrates that aligning with open-ended rubrics simultaneously reinforces performance on objective benchmarks. For instance, training on Health and Science subsets leads to a **+3.3** point gain on **MedQA** and a **+3.6** point gain on **GPQA-Diamond**. These results indicate that selective rubric guidance helps the model internalize domain-specific factual requirements and logical rigor, which are essential for both nuanced generation and verifiable reasoning.

Comparison with Joint Optimization and Distillation. Notably, RISE-RL maintains a substantial performance advantage over **OPD** and **RuscaRL**. OPD relies on teacher distillation, whose effectiveness depends on both the incremental value and the learnability of teacher trajectories; when the teacher–student gap is large, the additional supervision may be difficult to absorb, consistent with Rethinking OPD (Li et al., 2026b). RuscaRL directly optimizes scaffold-conditioned samples, but the resulting behaviors do not always transfer reliably once the scaffold is removed. Consistent with the concerns raised in RGR-GRPO, we observe that this limitation varies across domains: in some cases, scaffold removal causes substantial performance degradation that subsequent on-policy training fails to fully recover (Bi et al., 2025). We provide further analysis of RuscaRL's training dynamics in Appendix B.8. In contrast, RISE-RL decouples privileged guidance from natural exploration, enabling more stable autonomous on-policy optimization.

Compute-Budget Comparison. Under the same hardware allocation, RISE-RL achieves a higher average score across WritingBench and CreativeWriting-V3 than Rubric-RL with approximately matched training time on Qwen3-4B (59.44 vs. 58.22; 21.48 vs. 21.08 hours). It also outperforms Rubric-RL with a matched rollout budget (59.44 vs. 57.68). Extending Rubric-RL training to 24.80 hours, exceeding RISE-RL in both training time and rollout count, does not close the gap. These results indicate that, within the examined setting and budget range, the gains cannot be attributed solely to additional training time or a larger rollout budget. Details are provided in Appendix B.2.

4.3 Effect of Guidance Removal on Training Dynamics

To analyze the role of dynamic guidance, we compare the training dynamics of RISE-RL with Rubric-RL and persistent guidance on **Writing** and **Health** domains (Fig. 3).

Training Dynamics. Early guidance increases policy entropy from 0.9 to 3.5 in Writing and from 0.5 to 2.5 in Health, alongside performance gains. After guidance removal, entropy declines from its peak, while benchmark performance continues to improve. In contrast, persistent guidance maintains elevated, fluctuating entropy and exhibits late-stage performance plateaus or declines, includ-

ing a 2.8-point drop from the peak Writing score. These observations support early guidance followed by autonomous on-policy optimization in the evaluated settings.

Guidance-removal sensitivity experiments further show that removal at steps 40 and 100 retains average improvements over Rubric-RL. This suggests some flexibility in removal timing; see Appendix B.6 for details. A reward-based automatic removal rule also achieves comparable or higher average scores than manual scheduling on Qwen3-4B in Writing and Science (Appendix B.7).

4.4 Case Studies

Creative Writing: Quality and Diversity. As shown in Fig. 4, we report relative gains over Qwen3-4B in creative quality and output diversity. Quality is evaluated across five aggregated dimensions, with negative criteria reverse-scored before averaging. Diversity is assessed jointly over six responses per prompt across five set-level dimensions, using three random response permutations. All gains are computed as $(S_{\text{method}} - S_{\text{base}})/S_{\text{base}} \times 100\%$. RISE-RL improves all quality dimensions and achieves a **+5.8%** diversity gain, whereas Rubric-RL degrades Coherence and Style. These results suggest that RISE-RL improves creative-writing quality while also enhancing output diversity. Details of the creative-writing evaluation and additional analyses are provided in Appendix E.1.

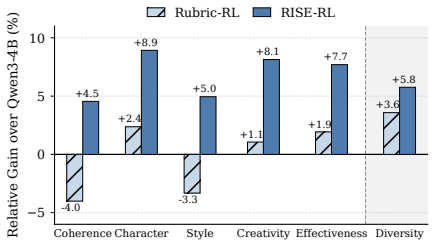

Figure 4: Relative gains over Qwen3-4B in overall creative-writing quality and output diversity.

Health: Acquisition of Persistent Criteria. To examine how RISE-RL resolves capability gaps, we analyze criterion-level failures on HealthBench (see Figure 8 in Appendix E.3). The initial policy exhibits persistent failures on critical requirements (e.g., *Medical-advice disclaimer and pre-travel obstetric consultation*, 8/8 failures). While Rubric-RL struggles to recover these behaviors through scalar rewards, RISE-RL leverages selective guidance to reduce failures. By step 300, RISE-RL achieves zero failures on both the combined disclaimer-and-consultation criterion and the pregnancy-history criterion, whereas Rubric-RL still fails on these criteria in five and one of eight responses, respectively.

4.5 Ablation Study

Sensitivity to Guidance Strength. We select $\lambda_{\text{guide}} = 0.05$ based on the largest training-reward increase among four coefficients in 140-step preliminary Writing runs with Qwen3-4B. A subsequent post-hoc evaluation also finds that this setting achieves the highest final CreativeWriting-V3 score among the tested values, providing additional support for the reward-based selection; see Appendix B.5 for details.

Effect of the Number of Injected Criteria. We vary the number of violated criteria injected into the privileged prompt over $\{1, 5, 10, \text{All}\}$. All variants use Qwen3-4B in the Writing domain under identical settings (Figure 5(a)). Injecting five criteria performs best on both WritingBench and CreativeWriting-V3, indicating a trade-off among robustness, coverage, and specificity. A single criterion may be overly sensitive to judge noise or an unrepresentative failure, whereas too many criteria reduce selectivity and introduce lower-priority, overlapping, or competing requirements that dilute the task signal. Overall, the results favor targeted over exhaustive criterion guidance.

Effect of Policy Support Shaping. We evaluate the contribution of policy support shaping by replacing the token-level auxiliary loss $-\log[p_{j,t}/(p_{j,t} + \gamma)]$ with the standard NLL loss $-\log p_{j,t}$. The ablated variant retains online privileged-trajectory generation, reward filtering, trajectory-level gain weights, probability evaluation under the original prompt, and joint optimization with the natural GRPO branch. As shown in Figure 5(b), support shaping achieves higher scores on HealthBench and LLMEval-Med under the same guidance-loss coefficient and matched short-run budget. These results support the effectiveness of the policy support shaping component in the evaluated setting.

Effect of Reward Filtering. We remove the positive-gain restriction and use all privileged trajectories with signed gains in the auxiliary loss. On Qwen3-4B in the Science domain, this reduces the

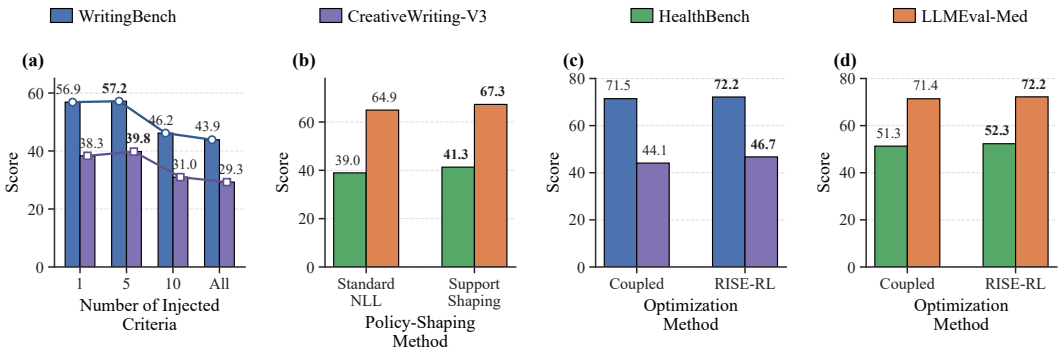

Figure 5: Ablations on Qwen3-4B. (a) Injected criterion count for Writing ("All": all identified violations). (b) Policy support shaping for Health. (c–d) Coupled vs. decoupled optimization for Writing and Health, respectively. Panels (a–b) use matched short-run budgets.

average score across ResearchQA and GPQA-Diamond from 63.10 to 60.44, with decreases on both benchmarks. These results support restricting auxiliary learning to reward-improving trajectories in this setting. Details are provided in Appendix B.3.

Effect of Criterion Selection. We compare our criterion-selection strategy with two alternatives on Qwen3-4B in Science: randomly selecting five criteria and selecting the five criteria with the highest rubric weights. Our strategy achieves an average score of 63.10 across ResearchQA and GPQA-Diamond, compared with 59.53 and 60.46, respectively. These results support jointly considering criterion importance and rollout failure frequency when selecting guidance. Details are provided in Appendix B.4.

Coupled versus Decoupled Optimization. We compare our decoupled design with a coupled variant using the same criterion-selection and guidance-generation procedures. The coupled variant jointly optimizes privileged trajectories and natural rollouts within a shared group-relative objective, whereas RISE-RL incorporates privileged trajectories through a separate auxiliary objective. As shown in Figure 5(c–d), the decoupled design achieves higher scores on all four evaluated benchmarks across Writing and Health. These results support the effectiveness of the separate auxiliary-learning design relative to the evaluated coupled optimization variant.

5 Conclusion and Discussion

In this work, we introduced **RISE-RL**, a training paradigm that balances selective rubric-informed guidance with autonomous exploration for open-ended reinforcement learning. By identifying persistent criterion failures and providing decoupled, quality-filtered guidance during early training, RISE-RL effectively internalizes weakly supported behaviors from reward-improving guided trajectories. Extensive experiments across four diverse domains with 4B and 14B models demonstrate that RISE-RL consistently outperforms standard Rubric-RL and strong baselines while maintaining output diversity. Our results suggest that transitioning from targeted guidance to unguided exploration provides an effective approach for aligning large language models with complex, multidimensional human preferences.

Despite these gains, RISE-RL has several limitations. It introduces additional early-stage computation and depends on the coverage and reliability of explicit rubrics and their graders. Although preliminary automatic-removal results show that guidance can transition online based on normalized reward progress, the current shared threshold is empirically chosen and requires further validation across model scales and domains. More broadly, RISE-RL provides a natural starting point toward **recursive self-improvement (RSI)**. Future work could maintain a persistent feedback loop that identifies emerging capability gaps and adaptively adjusts guidance and the transition back to autonomous exploration, enabling continual **feedback–exploration–internalization** in open-ended environments.

AI USE STATEMENT

In this work, we used generative AI tools to assist with Chinese-English translation, language polishing, and LaTeX formatting and troubleshooting. We also used generative AI tools for retrieval and discovery, including searching for potentially relevant literature and related work. We did not use generative AI tools to generate research ideas, formulate mathematical claims, develop theoretical arguments, design the research methodology or experiments, implement the proposed method, generate or process experimental data, analyze data, or interpret experimental results. All AI-assisted writing and formatting suggestions were manually reviewed, verified, and revised by the authors. For literature suggested or identified with AI assistance, the authors independently consulted the original sources and verified their relevance and bibliographic information before including them in the paper. We take responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work evaluates a general-purpose reinforcement learning method using existing health-domain research benchmarks, including HealthBench, LLMEval-Med, and MedQA. The health case study is included solely to analyze criterion-level model behavior and does not constitute medical advice or clinical validation. The trained models are research artifacts and are not intended for diagnosis, treatment, clinical decision support, or deployment in healthcare settings. Their outputs may be incomplete, incorrect, or overconfident, particularly in safety-critical situations, and should not be relied upon without review by qualified healthcare professionals.

Our study does not recruit human participants or collect new patient data. We use existing benchmark data for research purposes and follow the applicable licenses and terms of use. Improvements on these benchmarks should not be interpreted as evidence of clinical safety, effectiveness, or deployment readiness.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the datasets, training setup, hyperparameters, and baseline implementations in the appendix. The complete prompt and rubric templates used in our experiments are also included in the appendix. In addition, we submit anonymized source code, configuration files, and training scripts as supplementary materials. Together, these materials provide the information necessary to reproduce the main experimental results reported in this paper.

REFERENCES

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.

Ines Besrour, Jingbo He, Tobias Schreieder, and Michael Färber. Squai: Scientific question-answering with multi-agent retrieval-augmented generation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 6603–6608, 2025.

Baolong Bi, Shenghua Liu, Yiwei Wang, Siqian Tong, Lingrui Mei, Yuyao Ge, Yilong Xu, Jiafeng Guo, and Xueqi Cheng. Reward and guidance through rubrics: Promoting exploration to improve multi-domain reasoning. *arXiv preprint arXiv:2511.12344*, 2025.

Ruisheng Cao, Mouxiang Chen, Jiawei Chen, Zeyu Cui, Yunlong Feng, Binyuan Hui, Yuheng Jing, Kaixin Li, Mingze Li, Junyang Lin, et al. Qwen3-coder-next technical report. *arXiv preprint arXiv:2603.00729*, 2026.

Luoxin Chen, Jinming Gu, Liankai Huang, Wenhao Huang, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Kaijing Ma, et al. Seed-prover: Deep and broad reasoning for automated theorem proving. *arXiv preprint arXiv:2507.23726*, 2025.

Giorgio Franceschelli and Mirco Musolesi. On the creativity of large language models. *arXiv preprint arXiv:2304.00008*, 2023.

Carlos Gómez-Rodríguez and Paul Williams. A confederacy of models: a comprehensive evaluation of llms on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14504–14528, 2023.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.

Hsiu-Yuan Huang, Chenming Tang, Weijie Liu, Clive Bai, Saiyong Yang, and Yunfang Wu. Think outside the policy: In-context steered policy optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2758–2776, 2026a.

Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. *Advances in Neural Information Processing Systems*, 38:138362–138383, 2026b.

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, et al. Reinforcement learning via self-distillation. *arXiv preprint arXiv:2601.20802*, 2026.

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.

Hung Le, Yue Wang, Akhilesh Deepak Gotmare, Silvio Savarese, and Steven Chu Hong Hoi. Coderl: Mastering code generation through pretrained models and deep reinforcement learning. *Advances in Neural Information Processing Systems*, 35:21314–21328, 2022.

Sunzhu Li, Jiale Zhao, Huimin Ren, Zhenlin Wei, Yang Zhou, Jingwen Yang, Shunyu Liu, Kaike Zhang, and Chen Wei. Rubrichub: A comprehensive and highly discriminative rubric dataset via automated coarse-to-fine generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31320–31344, 2026a.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, et al. Rethinking on-policy distillation of large language models: Phenomenology, mechanism, and recipe. *arXiv preprint arXiv:2604.13016*, 2026b.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

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.

ZZ Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*, 2025.

MohammadHossein Rezaei, Anas Mahmoud, Zihao Wang, Utkarsh Tyagi, Advait Gosai, Razvan-Gabriel Dumitru, Aakash Sabharwal, Bing Liu, and Yunzhong He. Rubric-guided self-distillation: Post-training without rubric verifiers. *arXiv preprint arXiv:2606.12507*, 2026.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 300–325, 2021.

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

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. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Publisher correction: large language models encode clinical knowledge. *Nature*, 620(7973):E19, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation. *arXiv preprint arXiv:2405.09939*, 2024.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, et al. Writingbench: A comprehensive benchmark for generative writing. *Advances in Neural Information Processing Systems*, 38, 2026.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *Advances in Neural Information Processing Systems*, 38:117157–117186, 2026.

Li S Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *Transactions of the Association for Computational Linguistics*, 14:1344–1368, 2026.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, et al. Llmeval-med: a real-world clinical benchmark for medical llms with physician validation. *arXiv preprint arXiv:2506.04078*, 2025a.

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. *arXiv preprint arXiv:2506.03106*, 2025b.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pp. 270–278, 2020.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*, 2026a.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 29733–29735, 2025.

Ziqi Zhao, Xinyu Ma, Liu Yang, Yujie Feng, Daiting Shi, Jingzhou He, Xin Xin, Zhaochun Ren, and Xiao-Ming Wu. Rosd: Reflective on-policy self-distillation for language model reasoning across domains. *arXiv preprint arXiv:2605.28014*, 2026b.

Yang Zhou, Sunzhu Li, Shunyu Liu, Wenkai Fang, Kongcheng Zhang, Jiale Zhao, Jingwen Yang, Yihe Zhou, Jianwei Lv, Tongya Zheng, et al. Breaking the exploration bottleneck: Rubric-scaffolded reinforcement learning for general llm reasoning. *arXiv preprint arXiv:2508.16949*, 2025.

APPENDIX

TABLE OF CONTENTS

A Dataset and Implementation Details **15**

A.1 Training Data 15

A.2 RISE-RL Training Procedure 16

A.3 Training Configuration 16

A.4 Evaluation Implementations 17

B Additional Experimental Results **20**

B.1 Main Results with Standard Deviations 20

B.2 Compute-Budget Comparison 20

B.3 Effect of Reward Filtering 21

B.4 Effect of Criterion Selection 21

B.5 Sensitivity to Guidance Strength 22

B.6 Sensitivity to Guidance Removal Timing 23

B.7 Automatic Guidance Removal 23

B.8 Training Dynamics of RuscaRL 24

C Theoretical Analysis **25**

C.1 A Hybrid View of SFT and On-Policy Distillation 25

D Prompt and Rubric Templates **28**

D.1 Selective Guidance Prompt 28

D.2 Diversity Evaluation Prompt 29

E Case Studies **31**

E.1 Creative-Writing Case Study 31

E.2 Criterion-Level Analysis across Health Questions 38

E.3 Health Case Study 40

Table 2: Statistics of the reinforcement learning training data.

Domain	Prompts	Share
Science	29,418	34.07%
Health	29,681	34.37%
Writing	17,444	20.20%
Chat	9,812	11.36%
Total	**86,355**	**100.00%**

Table 3: Summary of the training data used in our experiments.

Training Paradigm	Domain Organization	Examples (Aggregate)
SFT	One mixed-domain run	26,194
RL	Four separate domain-specific runs	86,355

A Dataset and Implementation Details

A.1 Training Data

We use the RubricHub (Li et al., 2026a) dataset to construct the training data for both the reinforcement learning methods and the supervised fine-tuning baseline. We focus on four domains: Writing, Chat, Health, and Science, where Health corresponds to the Medical domain in the original RubricHub dataset. RubricHub provides two forms of training data used in our experiments: prompt–rubric pairs for reinforcement learning and high-quality prompt–response pairs for supervised fine-tuning.

Reinforcement Learning Data. For all reinforcement learning methods, including Rubric-RL, RuscaRL, and RISE-RL, we use the prompt–rubric pairs from the four selected RubricHub domains. We train each domain independently, resulting in a separate domain-specific model for Writing, Chat, Health, and Science. The complete reinforcement learning corpus contains 86,355 prompts, including 29,418 Science examples, 29,681 Health examples, 17,444 Writing examples, and 9,812 Chat examples. The detailed domain distribution is reported in Table 2.

Each reinforcement learning instance consists of an open-ended prompt and a set of weighted, instance-specific evaluation criteria. For each generated response, `gpt-oss-120b` independently determines whether each rubric criterion is satisfied. The resulting criterion-level judgments are aggregated into a weight-normalized scalar reward for policy optimization. RISE-RL additionally uses these criterion-level judgments to identify repeatedly missed criteria and construct selective rubric guidance. No reference response is required during reinforcement learning.

Supervised Fine-Tuning Data. For the SFT baseline, we use the official supervised fine-tuning dataset released with RubricHub. It contains 26,194 high-quality prompt–response pairs mixed across the Writing, Chat, Health, and Science domains. The target responses are obtained through the multi-stage response refinement procedure used in RubricHub. We train the SFT baseline on the combined four-domain corpus using the standard next-token prediction objective.

The SFT and reinforcement learning datasets correspond to different data components provided by RubricHub. SFT uses a single mixed-domain corpus of refined responses, whereas each RL model is trained separately on the prompt–rubric pairs of one domain. SFT therefore serves as a supervised reference rather than a data-matched control. Our primary controlled comparison is between RISE-RL and Rubric-RL, which use the same backbone, domain-specific training data, rubric grader, and optimization configuration. The evaluation datasets are drawn from separate benchmark sources and are not used for training.

Algorithm 1: Training Procedure of RISE-RL

Require: Training set $\mathcal{D}$, policy π_θ, total training steps S, group size G, feedback budget M, guidance weight λ_0, gain cap $A_{\max}$, and removal step s_{switch}

for $s = 1, \ldots, S$ **do**
  Sample $q \sim \mathcal{D}$ and generate natural rollouts

$$\mathcal{O}^{\text{nat}} \leftarrow \text{Rollout}(\pi_{\theta_{\text{old}}}, q, G)$$

  Evaluate the natural rollouts:

$$(Z^{\text{nat}}, R^{\text{nat}}) \leftarrow \text{Evaluate}(\mathcal{O}^{\text{nat}}, \mathcal{C}(q))$$

  Compute the GRPO loss:

$$\mathcal{L}_{\text{GRPO}} \leftarrow \text{GRPO}(\mathcal{O}^{\text{nat}}, R^{\text{nat}})$$

  Set $\mathcal{L} \leftarrow \mathcal{L}_{\text{GRPO}}$
  if $s < s_{\text{switch}}$ **then**
    Select up to M positive-priority criteria $\mathcal{C}^{\text{fb}}$ from Z^{nat}; if none exists, skip to the policy update
    Construct the privileged prompt

$$q^{\text{priv}} \leftarrow q \oplus \mathcal{C}^{\text{fb}}$$

    and generate privileged rollouts $\mathcal{O}^{\text{priv}}$
    Evaluate $\mathcal{O}^{\text{priv}}$ using the complete rubric $\mathcal{C}(q)$
    Retain trajectories $\mathcal{O}^{\text{ref}}$ with positive normalized reward gains
    Teacher-force $\mathcal{O}^{\text{ref}}$ under the original prompt q to estimate natural-policy token support $p_{j,t}$
    Compute policy-support weights, clipped normalized trajectory gains, and $\mathcal{L}_{\text{guide}}$
    Update the total loss:

$$\mathcal{L} \leftarrow \mathcal{L}_{\text{GRPO}} + \lambda_0 \mathcal{L}_{\text{guide}}$$

  end if
  Update the policy parameters:

$$\theta \leftarrow \text{Update}(\theta, \mathcal{L})$$

end for
return π_θ

A.2 RISE-RL Training Procedure

Algorithm 1 summarizes the complete RISE-RL training procedure, including natural-prompt policy optimization, selective privileged-trajectory generation, quality filtering, policy-support shaping, and the transition to guidance-free training.

A.3 Training Configuration

Rubric-RL and RISE-RL. We implement Rubric-RL and RISE-RL using the `verl` framework. For a controlled comparison, both methods use the same backbone model, training data, sampling configuration and optimization hyperparameters. RISE-RL additionally introduces a selective guidance branch with several method-specific hyperparameters. The complete training configuration is summarized in Table 4.

RuscaRL. We implement RuscaRL (Zhou et al., 2025) following its original scaffolding strategy and training configuration. Specifically, RuscaRL applies linear intra-group scaffolding differentiation and gradually reduces the rubric scaffolding using a step-wise sigmoid schedule. The main training configuration is summarized in Table 5. We train RuscaRL for 500 optimization steps in all domains.

Checkpoint and Guidance Schedule Selection. For RISE-RL, Rubric-RL, and our reproduction of RuscaRL, checkpoints are selected based on training reward convergence. Downstream benchmark scores are not used for checkpoint selection. For the main RISE-RL experiments, the early guidance-removal step is determined once for each model–domain setting based on the observed training-reward dynamics, rather than selected through a downstream-benchmark sweep over removal steps. Timing sensitivity and an automatic reward-based alternative are evaluated in Appendices B.6 and B.7, respectively. Separately, the Writing guidance coefficient is selected based on the

Table 4: Training configurations for Rubric-RL and RISE-RL.

Category	Configuration
Shared	Training batch size: `64` Maximum prompt length: `4096` Maximum response length: `8192` Number of rollouts per prompt: `8` Overlong length: `4096` Overlong penalty factor: `0.5` Learning rate: `1e-6` Warmup steps: `10` Weight decay: `0.1` Entropy coefficient: `0.0` Reward range: `[0,1]` KL-loss coefficient: `0.0` Asymmetric PPO clipping ϵ_{low}: `0.2` Asymmetric PPO clipping ϵ_{high}: `0.28`
RISE-RL	Number of guided re-rollouts: `8` Guidance support coefficient γ: `0.1` Guidance-loss coefficient λ_{guide}: `0.05 (Writing/Chat), 0.01 (Health/Science)` Number of injected criteria M: `5` Trajectory-gain cap $A_{\max}$: `5`
Hardware	8 × NVIDIA H200 GPUs

largest increase in training reward over 140-step preliminary runs on Qwen3-4B and then fixed for the full Writing training runs at both model scales. The subsequent CreativeWriting-V3 evaluation in Appendix B.5 is a post-hoc sensitivity analysis and is not used to select or readjust the coefficient.

Representative Removal Schedules. We illustrate the schedule selection using the Qwen3-14B Writing and Health settings shown in Figure 3. Based only on preliminary training-reward dynamics, we set $s_{\text{switch}} = 80$ for Writing and $s_{\text{switch}} = 60$ for Health, near the onset of a plateau in the smoothed privileged–natural reward gap. These values were fixed before downstream evaluation. The vertical "remove" markers in Figure 3 indicate the corresponding transitions. After guidance removal, policy entropy decreases from its guidance-induced peak while benchmark performance continues to improve, suggesting that the acquired behaviors are consolidated through autonomous on-policy optimization. In contrast, persistent guidance maintains elevated or fluctuating entropy and leads to late-stage performance plateaus or degradation. These observations motivate removing guidance once its marginal training benefit diminishes.

Supervised Fine-Tuning. We implement the SFT baseline using `SWIFT` with full-parameter fine-tuning. The model is trained on the mixed-domain RubricHub SFT corpus for three epochs using bfloat16 precision and DeepSpeed ZeRO-3, and we evaluate the final checkpoint. The main training hyperparameters are summarized in Table 6.

On-Policy Distillation. We implement OPD (Li et al., 2026b) using the `SWIFT` framework with full-parameter fine-tuning. Following the standard on-policy distillation setup, the student model generates trajectories from its current policy, while a larger teacher model provides token-level supervision on the student-generated states. For the 4B student, we use Qwen3-14B as the teacher; for the 14B student, we use Qwen3-32B as the teacher. All models are trained with bfloat16 precision and DeepSpeed ZeRO-2. The main training hyperparameters are summarized in Table 7.

A.4 Evaluation Implementations

We evaluate all models on eight benchmarks spanning four domains: Writing, Chat, Health, and Science. The evaluation suite includes both open-ended generation tasks and multiple-choice question answering tasks. For open-ended benchmarks, we follow the official rubric-based or pairwise-comparison protocols. For multiple-choice benchmarks, we report answer accuracy based on the

Table 5: Training configuration for RuscaRL.

Category	Configuration
RuscaRL	RL algorithm: GRPO Inter-step scaffolding decay: step sigmoid ($\alpha = 125$, $t_0 = 0.2$) Intra-group scaffolding differentiation: linear Rubric grader: `gpt-oss-120b`
Sampling	Temperature: 0.7 Top-p: 0.8; Top-k: 20 Rollouts per prompt: 8 Maximum response length: 8192
Training	Optimizer: Adam Learning rate: 1×10^{-6} (constant) Training batch size: 64 Mini-batch size: 32 KL-loss coefficient: 1×10^{-3} Entropy coefficient: 0 Total training steps: 500
Hardware	8 $\times$ NVIDIA H200 GPUs

Table 6: Training configuration for the SFT baseline.

Hyperparameter	Configuration
Training framework	`SWIFT`
Fine-tuning type	Full-parameter fine-tuning
Training epochs	3
Learning rate	1×10^{-5}
Per-device batch size	2
Gradient accumulation steps	4
Maximum sequence length	20,000
Warmup ratio	0.05
Validation split	1%
Precision	bfloat16
Distributed training	DeepSpeed ZeRO-3
Hardware	8 $\times$ NVIDIA H200 GPUs

extracted final choice. Unless otherwise specified, evaluation is performed without rubric guidance or any other privileged information.We evaluate on the complete official test sets without subsampling and follow the official evaluation implementations and default generation settings.

For each selected checkpoint, we independently repeat the complete evaluation pipeline three times, including response generation and LLM-based judging, and report the mean and standard deviation. This protocol quantifies the end-to-end evaluation variability for a fixed checkpoint. The question types and evaluation metrics are summarized in Table 8.

Writing. We evaluate writing ability using WritingBench and CreativeWriting-V3.

WritingBench (Wu et al., 2026) evaluates long-form responses to diverse real-world writing instructions using instance-specific criteria covering multiple dimensions of writing quality. We follow the official rubric-based evaluation protocol and report the aggregated rubric score.

CreativeWriting-V3 evaluates open-ended creative writing, including narrative coherence, characterization, style, originality, and overall effectiveness. We follow the official rubric-based evaluation protocol and report the aggregate rubric score.

Chat. We use Arena-Hard-v2 (Li et al., 2024) to evaluate general open-ended instruction-following and conversational ability. Candidate responses are compared against reference-model

Table 7: Training configuration for the OPD baseline.

Hyperparameter	Configuration
Training framework	`SWIFT`
Training objective	On-policy distillation (GKD)
Student–teacher pairs	4B–14B; 14B–32B
Fine-tuning type	Full-parameter fine-tuning
Training steps	50
Learning rate	1×10^{-5}
Per-device batch size	4
Gradient accumulation steps	16
Maximum sequence length	16,000
Maximum completion length	8,192
Warmup ratio	0.05
Precision	bfloat16
Distributed training	DeepSpeed ZeRO-2
Hardware	8 $\times$ NVIDIA H200 GPUs

Table 8: Summary of the evaluation protocols.

Domain	Benchmark	Question Type	Metric
Writing	WritingBench	Open-ended writing	Rubric score
	CreativeWriting-V3	Creative writing	Rubric score
Chat	Arena-Hard-v2	Open-ended dialogue	Style-controlled win rate
Health	HealthBench	Open-ended healthcare	Weighted rubric score
	LLMEval-Med	Open-ended clinical QA	Overall usability
	MedQA	Multiple choice	Exact-match accuracy
Science	ResearchQA	Long-form scholarly QA	Rubric coverage
	GPQA-Diamond	Multiple choice	Exact-match accuracy

responses under the official style-controlled pairwise evaluation protocol, and we report the resulting win rate.

Health. We evaluate health-domain capabilities on HealthBench (Arora et al., 2025), LLMEval-Med (Zhang et al., 2025a), and MedQA (Jin et al., 2021).

HealthBench consists of open-ended healthcare conversations evaluated using physician-written, conversation-specific rubric criteria with importance weights. The final score is computed from the weighted satisfaction of these criteria.

LLMEval-Med evaluates open-ended responses to real-world clinical questions using expert-written reference answers and checklists. We follow the official checklist-based evaluation pipeline and report the overall usability score.

MedQA is a multiple-choice medical question-answering benchmark. We report exact-match accuracy of the predicted answer choice.

Science. We evaluate scientific knowledge and reasoning using ResearchQA (Yifei et al., 2026) and GPQA-Diamond (Rein et al., 2023).

ResearchQA evaluates long-form scholarly question answering using query-specific rubric items. We report the normalized rubric-coverage score.

GPQA-Diamond is a graduate-level multiple-choice benchmark covering biology, physics, and chemistry. We report exact-match accuracy of the predicted answer choice.

Table 9: Full results corresponding to Table 1, with evaluation standard deviations, for Qwen3-4B and Qwen3-14B across four domains. Each benchmark entry reports the mean ± standard deviation over three independent evaluations of the same selected checkpoint. Avg. denotes the unweighted average of the eight benchmark means. The highest mean score for each benchmark and the highest Avg. within each model scale are shown in bold.

Method	Writing		Chat	Health			Science		Avg.
	Writing Bench	Creative Writing-V3	Arena-Hard V2	Health Bench	LLMEval -Med	MedQA	GPQA	Research QA	
Qwen3-4B (Non-Thinking)									
Initial	56.14 ± 0.10	40.28 ± 0.77	7.91 ± 0.07	37.43 ± 0.12	64.76 ± 0.28	64.79 ± 0.12	42.34 ± 1.98	64.29 ± 0.22	47.24
SFT	66.43 ± 0.30	32.96 ± 1.00	11.39 ± 0.13	40.73 ± 0.11	64.87 ± 0.82	59.77 ± 0.85	36.95 ± 2.68	69.19 ± 0.43	47.79
OPD	54.55 ± 0.22	27.78 ± 0.52	6.32 ± 0.11	35.46 ± 0.19	63.07 ± 0.15	61.63 ± 0.17	40.24 ± 2.01	60.35 ± 0.36	43.68
Rubric-RL	70.92 ± 0.21	43.30 ± 0.36	20.76 ± 0.62	52.00 ± 0.01	71.85 ± 0.34	68.44 ± 0.39	47.73 ± 1.20	74.52 ± 0.71	56.19
RuscaRL	56.75 ± 0.39	38.67 ± 0.52	9.41 ± 0.49	43.54 ± 0.13	65.81 ± 0.50	63.39 ± 0.36	41.41 ± 2.92	66.31 ± 0.92	48.16
RISE-RL (Ours)	$\mathbf{72.15 \pm 0.12}$	$\mathbf{46.73 \pm 0.56}$	$\mathbf{20.89 \pm 0.25}$	$\mathbf{52.32 \pm 0.25}$	$\mathbf{72.22 \pm 0.73}$	$\mathbf{69.40 \pm 0.25}$	$\mathbf{49.12 \pm 2.52}$	$\mathbf{77.08 \pm 0.29}$	**57.49**
Qwen3-14B (Non-Thinking)									
Initial	63.24 ± 0.15	64.94 ± 0.40	21.09 ± 0.31	43.85 ± 0.22	69.46 ± 0.52	73.51 ± 0.44	52.61 ± 1.98	68.27 ± 0.27	57.12
SFT	73.31 ± 0.12	60.60 ± 0.88	33.74 ± 0.39	48.66 ± 0.43	71.63 ± 0.39	74.15 ± 0.98	39.90 ± 1.69	73.54 ± 0.32	59.44
OPD	66.05 ± 0.08	51.53 ± 0.60	20.07 ± 0.29	45.39 ± 0.18	71.56 ± 0.35	74.91 ± 1.12	48.82 ± 2.43	69.69 ± 0.33	56.00
Rubric-RL	76.89 ± 0.20	64.93 ± 0.71	49.94 ± 0.42	57.92 ± 0.06	76.40 ± 0.41	78.22 ± 0.68	55.98 ± 2.05	77.26 ± 0.08	67.19
RuscaRL	62.45 ± 0.13	42.78 ± 1.08	45.96 ± 0.48	50.92 ± 0.17	71.49 ± 0.65	77.39 ± 0.49	46.13 ± 2.74	74.10 ± 0.34	58.90
RISE-RL (Ours)	$\mathbf{77.37 \pm 0.10}$	$\mathbf{70.96 \pm 0.53}$	$\mathbf{57.97 \pm 0.31}$	$\mathbf{59.44 \pm 0.11}$	$\mathbf{78.31 \pm 0.20}$	$\mathbf{81.57 \pm 0.91}$	$\mathbf{59.60 \pm 1.09}$	$\mathbf{78.96 \pm 0.15}$	**70.52**

B Additional Experimental Results

B.1 Main Results with Standard Deviations

Table 9 reports the full benchmark results with standard deviations. For each method and model scale, we independently evaluate the same selected checkpoint three times and report the mean and standard deviation across these evaluations. These standard deviations quantify variability from repeated response generation and evaluation for a fixed checkpoint.

Table 9 indicates generally limited variability across the three independent evaluations: 62 of the 96 benchmark entries have standard deviations at or below 0.5 points. Evaluation variability differs across benchmarks, with WritingBench and HealthBench consistently exhibiting standard deviations below 0.5 points across all methods and both model scales. GPQA exhibits the largest standard deviation in every method–scale setting, ranging from 1.09 to 2.92 points. This greater variability may partly reflect its relatively small evaluation set, which makes the aggregate score more sensitive to changes in individual predictions. Overall, these results indicate that most reported benchmark scores are reasonably consistent across repeated evaluations, with greater variability observed on GPQA.

B.2 Compute-Budget Comparison

We compare RISE-RL and Rubric-RL using Qwen3-4B in the Writing domain under the same hardware allocation. Rubric-RL requires approximately 3.72 minutes per optimization step. RISE-RL incurs additional computation only during its initial guidance phase, which lasts for 80 steps in this setting and requires approximately 6.87 minutes per step because of the additional privileged rollouts, complete-rubric evaluation, and teacher-forcing computation. After guidance is removed, RISE-RL follows the same natural-prompt training procedure as Rubric-RL and therefore has approximately the same per-step cost.

We consider complementary controls for wall-clock time and rollout budget. First, we extend Rubric-RL to 340 steps, resulting in a training time approximately matched to that of RISE-RL (21.08 versus 21.48 hours). Second, we train Rubric-RL for 360 steps to match the total number of generated and rubric-evaluated trajectories. Specifically, with a training batch size of $B = 64$ and a rollout group size of $G = 8$, RISE-RL generates

$$280 \times B \times G + 80 \times B \times G = 184{,}320$$

trajectories: $280 \times B \times G$ natural trajectories over the full training run and an additional $80 \times B \times G$ privileged trajectories during the first 80 steps. Rubric-RL at 360 steps generates

$$360 \times B \times G = 184{,}320$$

Table 10: Compute-budget comparison on Qwen3-4B in Writing. The 340-step Rubric-RL run approximately matches the wall-clock training time of RISE-RL, while the 360-step run matches its total number of generated and rubric-evaluated trajectories. The 400-step run exceeds RISE-RL in both budgets. Avg. denotes the unweighted mean of the two benchmark scores. Training times are approximate.

	Training budget		Writing performance		
Method	Steps	Time (h)	WritingBench	CreativeWriting-V3	Avg.
Rubric-RL	280	17.36	70.92	43.30	57.11
Rubric-RL	340	21.08	71.86	44.57	58.22
Rubric-RL	360	22.32	71.38	43.97	57.68
Rubric-RL	400	24.80	71.92	42.53	57.23
RISE-RL (Ours)	280	21.48	**72.15**	**46.73**	**59.44**

Table 11: Reward-filtering ablation on Qwen3-4B in the Science domain. Avg. denotes the unweighted mean of ResearchQA and GPQA-Diamond scores. Best scores are bold.

Method	ResearchQA	GPQA-Diamond	Avg.
RISE-RL w/o reward filtering	75.93	44.95	60.44
RISE-RL	**77.08**	**49.12**	**63.10**

natural trajectories, exactly matching this rollout budget. We additionally report Rubric-RL at 400 steps, which exceeds RISE-RL in both wall-clock time and the number of generated trajectories.

As shown in Table 10, RISE-RL achieves an average score of 59.44 across WritingBench and CreativeWriting-V3, compared with 58.22 for the approximately time-matched 340-step Rubric-RL run and 57.68 for the rollout-matched 360-step run. Extending Rubric-RL to 400 steps further increases its training budget but does not improve its average performance. RISE-RL also remains better on both individual benchmarks under all three extended Rubric-RL budgets. These results indicate that, within the examined Qwen3-4B Writing setting, the gains of RISE-RL cannot be explained solely by longer training time or a larger number of generated and reward-evaluated trajectories.

B.3 Effect of Reward Filtering

Privileged guidance targets selected rubric criteria, but improvement on these criteria does not necessarily imply a higher reward under the complete rubric. RISE-RL therefore retains only guided trajectories with positive normalized reward gains relative to the natural-rollout mean, $A_j^{\mathrm{ref}} = \frac{r_j^{\mathrm{priv}} - \bar{r}^{\mathrm{nat}}}{\sigma^{\mathrm{nat}} + \epsilon_A} > 0$. To evaluate this mechanism, we compare RISE-RL with a no-filter variant on Qwen3-4B in the Science domain. The no-filter variant uses all privileged trajectories in the auxiliary loss and its normalization, using symmetrically clipped weights $\widetilde{A}_j = \mathrm{clip}\big(A_j^{\mathrm{ref}}, -A_{\max}, A_{\max}\big)$ with $A_{\max} = 5$, so that negative normalized gains retain their signs while their magnitudes remain bounded. We measure overall Science performance using the unweighted mean of ResearchQA and GPQA-Diamond scores.

As shown in Table 11, removing reward filtering lowers ResearchQA and GPQA-Diamond scores by 1.15 and 4.17 points, respectively, reducing their average from 63.10 to 60.44. These results favor restricting auxiliary learning to reward-improving trajectories over using all privileged trajectories with signed normalized gains in this setting.

B.4 Effect of Criterion Selection

We examine whether the criterion-selection strategy contributes to the effectiveness of RISE-RL using Qwen3-4B in the Science domain. We compare the default strategy with two alternatives:

Table 12: Criterion-selection ablation on Qwen3-4B in Science. Each strategy selects five criteria. Avg. denotes the unweighted mean of ResearchQA and GPQA-Diamond scores. Best scores are bold.

Selection strategy	ResearchQA	GPQA-Diamond	Avg.
Random	76.13	42.93	59.53
Weight-only	75.97	44.95	60.46
Default (RISE-RL)	**77.08**	**49.12**	**63.10**

Random, which randomly selects five rubric criteria, and *Weight-only*, which selects the five criteria with the highest rubric importance weights w_k. The default strategy also selects five criteria, but ranks them by

$$p_k = w_k \left(1 - \frac{1}{G}\sum_{i=1}^{G} z_{i,k}\right),$$

where $z_{i,k}$ indicates whether natural rollout i satisfies criterion k. This priority combines criterion importance with the frequency of failures across natural rollouts. The three variants therefore use the same number of injected criteria.

As shown in Table 12, the default strategy achieves an average score of 63.10, exceeding Random and Weight-only by 3.57 and 2.64 points, respectively. It also achieves higher scores on both individual benchmarks. These results support using rollout-dependent failure information together with rubric importance to select guidance in this setting.

B.5 Sensitivity to Guidance Strength

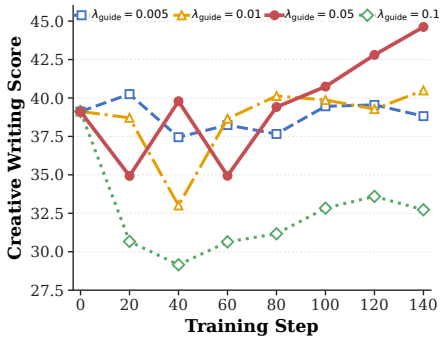

Figure 6: Post-hoc sensitivity analysis of the guidance-strength coefficient λ_{guide} on CreativeWriting-V3 using Qwen3-4B. Each variant is trained for 140 steps; checkpoints saved every 20 steps are subsequently evaluated. The setting $\lambda_{\text{guide}} = 0.05$, selected based on training-reward improvement, also achieves the highest final benchmark score among the tested values.

We select λ_{guide} using 140-step preliminary training runs with Qwen3-4B in the Writing domain, comparing $\lambda_{\text{guide}} \in \{0.005, 0.01, 0.05, 0.1\}$. The selection criterion is the increase in training reward over these runs. Among the tested values, $\lambda_{\text{guide}} = 0.05$ yields the largest training-reward increase and is therefore fixed for the full Writing training runs with both Qwen3-4B and Qwen3-14B before the downstream sensitivity evaluation.

For Health and Science, we apply the same 140-step preliminary selection procedure on Qwen3-4B over the same candidate set, using training-reward improvement as the selection criterion. This procedure selects $\lambda_{\text{guide}} = 0.01$ for both domains, which is then fixed for the full training runs at both model scales. As in Writing, no downstream benchmark scores are used for coefficient selection.

After fixing the coefficient, we evaluate checkpoints saved every 20 steps from these preliminary runs on CreativeWriting-V3 as a post-hoc sensitivity analysis. As shown in Fig. 6, $\lambda_{\text{guide}} = 0.05$

Table 13: Sensitivity to guidance removal timing on Qwen3-14B in Health. A removal step of 0 denotes Rubric-RL without guidance, and ∞ denotes persistent guidance. Avg. is the unweighted mean of the three benchmark scores. Best scores are bold.

Removal step	HealthBench	LLMEval-Med	MedQA	Avg.
0 (Rubric-RL)	57.92	76.40	78.22	70.85
20	58.51	76.82	80.67	72.00
40	59.26	77.54	80.38	72.39
60 (default)	**59.44**	**78.31**	**81.57**	**73.11**
100	59.15	76.22	80.59	71.99
∞ (persistent)	55.96	76.98	80.23	71.06

also achieves the highest final CreativeWriting-V3 score among the tested values. A small coefficient (0.005) leaves performance near the initial baseline, indicating that the guidance signal is too weak to meaningfully affect policy optimization. In contrast, a large coefficient (0.1) causes a sharp early performance drop; although the model partially recovers, it remains below the moderate setting, suggesting that overly strong guidance disrupts the existing policy distribution and makes subsequent recovery difficult. These downstream results are consistent with the training-reward-based selection: $\lambda_{\text{guide}} = 0.05$ yields both the largest training-reward increase and the highest final CreativeWriting-V3 score among the tested values, providing additional support for the selected coefficient.

CreativeWriting-V3 scores are not used to select or readjust the coefficient. We do not perform a separate coefficient sweep on Qwen3-14B or readjust the coefficient based on the full-run benchmark results.

B.6 Sensitivity to Guidance Removal Timing

We study the sensitivity of RISE-RL to the guidance removal step using Qwen3-14B in the Health domain. In addition to the default removal step $s_{\text{switch}} = 60$, we evaluate removal at steps 20, 40, and 100. We include two boundary conditions: $s_{\text{switch}} = 0$, corresponding to standard Rubric-RL without guidance, and $s_{\text{switch}} = \infty$, corresponding to persistent guidance throughout training. We report HealthBench, LLMEval-Med, and MedQA scores, together with their unweighted mean.

Sensitivity around the Default Setting. As shown in Table 13, removing guidance at steps 40 and 100 yields average scores of 72.39 and 71.99, respectively, compared with 73.11 at the default step of 60. Both alternatives retain an average improvement over Rubric-RL, whose score is 70.85. On HealthBench, the scores at steps 40 and 100 are particularly close to the default, differing by only 0.18 and 0.29 points, respectively. The differences are larger on the other benchmarks, indicating that sensitivity varies across tasks. In this setting, all tested finite removal steps from 20 to 100 outperform Rubric-RL in average performance, while steps 40–100 provide consistently strong results. This broad effective range shows that the improvement is not tied to a narrowly tuned removal checkpoint.

Early Removal and Persistent Guidance. Removing guidance at step 20 yields an average score of 72.00, improving over Rubric-RL but falling below the default setting. This is consistent with early removal limiting the benefit of guided training, although its average performance is comparable to removal at step 100. Persistent guidance yields an average score of 71.06 and reduces HealthBench performance to 55.96, below the Rubric-RL score of 57.92. Thus, retaining guidance throughout training does not provide the same benefit as removing it at an intermediate stage. Together, these comparisons support an initial guidance phase followed by unguided optimization.

B.7 Automatic Guidance Removal

Different models and domains may exhibit different learning rates, making a fixed removal step difficult to transfer across settings. We therefore formulate guidance removal in terms of observed training progress rather than the absolute optimization step. Since rubric rewards are normalized

Table 14: Automatic versus manual guidance removal on Qwen3-4B. Automatic removal uses a shared threshold of the initial normalized reward plus 0.1, without domain-specific adjustment. Avg. denotes the unweighted mean of the two benchmarks within each domain. Best scores within each domain are bold.

Writing			
Removal strategy	WritingBench	CreativeWriting-V3	Avg.
Manual	**72.15**	46.73	59.44
Automatic	71.97	**47.02**	**59.50**
Science			
Removal strategy	ResearchQA	GPQA-Diamond	Avg.
Manual	**77.08**	49.12	63.10
Automatic	76.54	**52.53**	**64.54**

to $[0, 1]$ across domains, we remove guidance when the training reward exceeds its initial value by 0.1 and continue training with natural-prompt GRPO. The shared normalized reward scale makes this criterion directly comparable across domains without selecting a separate removal step for each setting.

We choose 0.1 as a practical transition point based on the observed training dynamics. A substantially smaller increment may remove guidance before the targeted behaviors have been sufficiently internalized, limiting the benefit of selective exploration. In contrast, requiring a much larger increment prolongs the more expensive guidance stage after its marginal benefit has begun to diminish and leaves less opportunity for subsequent autonomous optimization. This trade-off is consistent with the removal-timing and persistent-guidance results in Appendix B.6.

As shown in Table 14, the same reward-progress threshold, applied without domain-specific adjustment, matches manual scheduling in Writing and outperforms it in Science. Together with the broad effective range observed in the Health timing analysis, these results indicate that the gains of RISE-RL are not tied to a narrowly selected manual checkpoint. They further support normalized reward progress as a simple, transferable, and reproducible mechanism for transitioning from guided to autonomous optimization.

B.8 Training Dynamics of RuscaRL

We further examine the training dynamics of RuscaRL under two representative model–domain settings, as shown in Figure 7. The corresponding reproduction configuration is provided in Section A.3. During the initial stage, both settings maintain relatively high mean rewards with rubric scaffolding. However, their entropy dynamics differ: Qwen3-4B trained on Writing exhibits a sharp increase in policy entropy, whereas Qwen3-14B trained on Chat fluctuates around its initial entropy level. When the scaffolding is substantially removed after approximately 100 training steps, both settings experience an abrupt reward collapse. Although the reward gradually recovers during the subsequent unscaffolded training stage, it remains below its initial scaffolded level by the end of training.

One possible interpretation is that removing scaffolding induces a mismatch between scaffold-conditioned and natural-prompt generation. This transition may contribute to the observed reward drop. In the early stage, extensive rubric guidance makes high-reward responses easier to generate. Nevertheless, injecting many criteria into the prompt may dilute adherence to the original user instruction and induce trajectories that differ substantially from those sampled under the natural prompt. Jointly optimizing these heterogeneous trajectories within the same group-relative objective may therefore produce unstable policy updates, as reflected by the abrupt or fluctuating entropy dynamics. Once the scaffolding is removed, the policy must transition from rubric-conditioned generation to natural-prompt generation, causing a pronounced distribution shift and the accompanying reward collapse. Later on-policy training can partially recover performance, but the instability introduced during this transition may constrain the final performance attainable by the model.

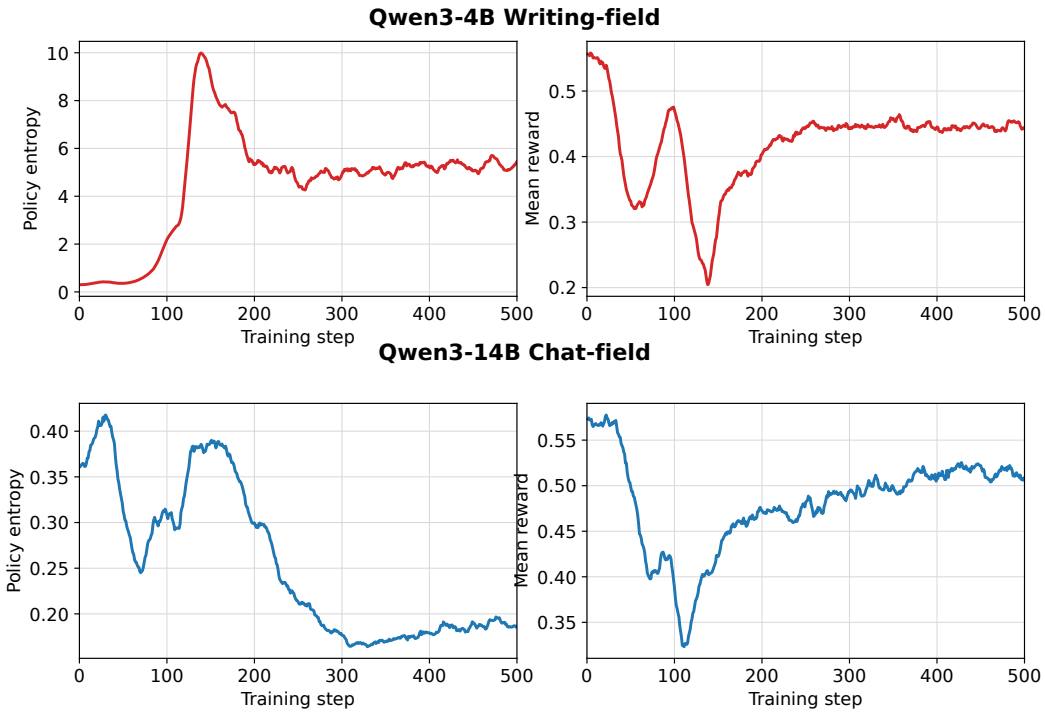

Figure 7: Representative training dynamics of RuscaRL under two model–domain settings: Qwen3-4B in the Writing domain (top) and Qwen3-14B in the Chat domain (bottom). The left column shows policy entropy, while the right column shows the mean rubric reward over 500 training steps.

C Theoretical Analysis

C.1 A Hybrid View of SFT and On-Policy Distillation

This section provides an alternative interpretation of our selective guidance objective. We first review supervised fine-tuning (SFT) and on-policy distillation (OPD), and then show that our guidance branch combines the online, policy-dependent supervision of OPD with the hard-target optimization structure of SFT.

Supervised fine-tuning. Let q denote an input prompt and $o = (o_1, \ldots, o_{T(o)})$ denote a reference response, where $T(o)$ is its token length. In practice, standard SFT commonly minimizes the length-normalized token-level negative log-likelihood

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(q,o)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\frac{1}{T(o)}\sum_{t=1}^{T(o)}\log \pi_\theta(o_t \mid q, o_{<t})\right]. \tag{13}$$

At each prefix $(q, o_{<t})$, the reference token o_t defines an empirical one-hot target distribution

$$\delta_{o_t}(v) = \mathbb{I}[v = o_t]. \tag{14}$$

The corresponding forward KL divergence is

$$\begin{aligned} D_{\mathrm{KL}}\left(\delta_{o_t} \| \pi_\theta(\cdot \mid q, o_{<t})\right) &= \sum_v \delta_{o_t}(v)\log\frac{\delta_{o_t}(v)}{\pi_\theta(v \mid q, o_{<t})} \\ &= -\log \pi_\theta(o_t \mid q, o_{<t}), \end{aligned} \tag{15}$$

where the equality is exact because the entropy of the one-hot target distribution is zero. Thus, SFT can be viewed as forward-KL distillation from an empirical one-hot target distribution.

More generally, if reference responses are sampled from a fixed teacher or behavior distribution μ, then

$$\mathbb{E}_{o\sim\mu(\cdot|q)}[-\log \pi_\theta(o \mid q)] = H(\mu(\cdot \mid q)) + D_{\mathrm{KL}}\left(\mu(\cdot \mid q)\|\pi_\theta(\cdot \mid q)\right). \tag{16}$$

Eq. (16) is a sequence-level identity, whereas Eq. (13) averages the token losses by $T(o)$. The latter is therefore a practical length-normalized empirical objective rather than an exact sequence-level KL identity when response lengths vary.

On-policy distillation. In standard OPD, trajectories are sampled from the current student policy, while a teacher provides dense token-level distributional supervision on the student-visited prefixes. Let π_T denote the teacher policy. A sequence-level reverse-KL objective can be written as

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{q\sim\mathcal{D}}\left[D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid q)\|\pi_T(\cdot \mid q)\right)\right]. \tag{17}$$

By the autoregressive chain rule for KL divergence, the sequence-level objective factorizes exactly as

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, o\sim\pi_\theta(\cdot|q)}\left[\sum_{t=1}^{T(o)} D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid q, o_{<t})\|\pi_T(\cdot \mid q, o_{<t})\right)\right].$$

In practical token-level optimization, we use the corresponding length-normalized surrogate

$$\widetilde{\mathcal{L}}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, o\sim\pi_\theta(\cdot|q)}\left[\frac{1}{T(o)}\sum_{t=1}^{T(o)} D_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid q, o_{<t})\|\pi_T(\cdot \mid q, o_{<t})\right)\right]. \tag{18}$$

Eq. (18) stabilizes the contribution of responses with different lengths, but it is not algebraically identical to the unnormalized sequence-level reverse KL in Eq. (17).

OPD therefore differs from conventional SFT in two respects. First, its state distribution is induced by the current student policy rather than by a fixed offline dataset. Second, it typically matches a soft teacher distribution through reverse KL instead of treating a single reference token as a hard target.

Online construction of privileged reference trajectories. Our guidance branch begins with online, policy-dependent trajectory generation. Given the selected feedback criteria $\mathcal{C}^{\mathrm{fb}}(q)$, we construct

$$q^{\mathrm{priv}} = q \oplus \mathcal{C}^{\mathrm{fb}}(q) \tag{19}$$

and sample privileged trajectories from the current behavior policy:

$$o_j^{\mathrm{priv}} \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid q^{\mathrm{priv}}). \tag{20}$$

Unlike an offline SFT dataset, these trajectories are regenerated as the policy evolves. We evaluate each privileged trajectory using the complete original rubric. Let $\bar{r}^{\mathrm{nat}}$ and σ^{nat} denote the mean and standard deviation of the natural-rollout rewards $\{r_i^{\mathrm{nat}}\}_{i=1}^G$, respectively, and let $\epsilon_A = 10^{-6}$. We define the normalized reward gain and retain trajectories with positive gains:

$$A_j^{\mathrm{ref}} = \frac{r_j^{\mathrm{priv}} - \bar{r}^{\mathrm{nat}}}{\sigma^{\mathrm{nat}} + \epsilon_A}, \qquad \mathcal{O}^{\mathrm{ref}} = \left\{o_j^{\mathrm{priv}} \mid A_j^{\mathrm{ref}} > 0\right\}. \tag{21}$$

The resulting reference set is therefore both policy-dependent and quality-filtered. This online construction is related to OPD in that the supervision distribution evolves together with the current policy. However, the trajectories are generated under privileged conditioning rather than from the natural-prompt student policy.

Support-adaptive SFT interpretation. After removing the privileged criteria, each retained trajectory is teacher-forced under the original prompt. For a retained token $o_{j,t}^{\mathrm{ref}}$, define its natural-policy support as

$$p_{j,t} = \pi_\theta\left(o_{j,t}^{\mathrm{ref}} \mid q, o_{j,<t}^{\mathrm{ref}}\right). \tag{22}$$

Our policy-support transformation is

$$\rho_{j,t} = \frac{p_{j,t}}{p_{j,t} + \gamma}, \tag{23}$$

and the corresponding token-level loss is

$$\ell_{\text{guide}}(p_{j,t}) = -\log \rho_{j,t} = -\log p_{j,t} + \log(p_{j,t} + \gamma). \tag{24}$$

The following result establishes its connection to SFT.

Proposition 1. For a fixed reference token, the gradient of the support-shaped guidance loss has the same direction as the gradient of the standard SFT negative log-likelihood:

$$\nabla_\theta \ell_{\text{guide}}(p_{j,t}) = \frac{\gamma}{p_{j,t} + \gamma} \nabla_\theta \left[-\log p_{j,t}\right]. \tag{25}$$

Proof. Starting from Eq. (24),

$$\begin{aligned} \nabla_\theta \ell_{\text{guide}}(p_{j,t}) &= \left(-\frac{1}{p_{j,t}} + \frac{1}{p_{j,t} + \gamma}\right) \nabla_\theta p_{j,t} \\ &= -\frac{\gamma}{p_{j,t}(p_{j,t} + \gamma)} \nabla_\theta p_{j,t} \\ &= \frac{\gamma}{p_{j,t} + \gamma} \nabla_\theta [-\log p_{j,t}]. \end{aligned} \tag{26}$$

□

Eq. (25) shows that the guidance loss is gradient-equivalent to an SFT update with the adaptive token weight

$$w_{j,t}^{\text{support}} = \frac{\gamma}{p_{j,t} + \gamma}. \tag{27}$$

Importantly, this is a gradient-level equivalence; the guidance loss itself is not equal to a weighted SFT loss unless the adaptive weight is treated as a stop-gradient quantity.

For weakly supported tokens, $p_{j,t} \ll \gamma$, we have

$$w_{j,t}^{\text{support}} \approx 1, \qquad \ell_{\text{guide}}(p_{j,t}) \approx -\log p_{j,t} + \log \gamma. \tag{28}$$

The resulting update therefore approaches a standard SFT update. In contrast, when $p_{j,t} \gg \gamma$,

$$w_{j,t}^{\text{support}} \approx \frac{\gamma}{p_{j,t}}, \tag{29}$$

which suppresses imitation of tokens already well supported by the natural policy.

For each retained trajectory $o_j^{\text{ref}} \in \mathcal{O}^{\text{ref}}$, let T_j denote its token length. We define the trajectory-level weight as $\widetilde{A}_j = \min\left(A_j^{\text{ref}}, A_{\max}\right)$, where $A_{\max} = 5$ limits large normalized gains when the natural-rollout reward variance is small. The selective guidance loss is then

$$\mathcal{L}_{\text{guide}} = -\frac{1}{|\mathcal{O}^{\text{ref}}|} \sum_{o_j^{\text{ref}} \in \mathcal{O}^{\text{ref}}} \frac{\widetilde{A}_j}{T_j} \sum_{t=1}^{T_j} \log \rho_{j,t}. \tag{30}$$

Its gradient can consequently be interpreted as

$$\nabla_\theta \mathcal{L}_{\text{guide}} = \frac{1}{|\mathcal{O}^{\text{ref}}|} \sum_{o_j^{\text{ref}} \in \mathcal{O}^{\text{ref}}} \frac{\widetilde{A}_j}{T_j} \sum_{t=1}^{T_j} w_{j,t}^{\text{support}} \nabla_\theta [-\log p_{j,t}]. \tag{31}$$

Thus, the guidance branch admits a gradient-level interpretation as trajectory-quality-weighted, token-support-adaptive negative-log-likelihood training on online privileged trajectories. This interpretation concerns only the auxiliary branch; the full training procedure additionally retains natural-prompt GRPO optimization.

Connection and distinction. Our guidance branch combines complementary aspects of SFT and OPD. Similar to OPD, the supervision is online and policy-dependent: the privileged reference trajectories are regenerated as the behavior policy evolves. Similar to SFT, the optimization treats the realized tokens in the retained trajectories as hard targets and follows the token-level negative-log-likelihood direction.

Nevertheless, our method is not equivalent to conventional OPD. Standard OPD samples natural-prompt trajectories from the student and matches the full teacher distribution, typically through reverse KL. Our method instead samples trajectories under privileged rubric conditioning, filters them using the complete-rubric reward, removes the privileged information, and internalizes their realized tokens through a support-adaptive hard-target objective. It therefore does not require access to full teacher logits.

In summary, our guidance branch combines (i) online, policy-dependent reference construction, which is OPD-like; (ii) hard-target teacher-forcing optimization, which is SFT-like; and (iii) quality filtering and policy-support shaping, which enable the selective internalization of privileged trajectories.

D Prompt and Rubric Templates

D.1 Selective Guidance Prompt

During guidance, RISE-RL preserves the original conversation and does not replace or rewrite the user prompt. Instead, it constructs a lightweight privileged suffix from the rubric evaluation of the initial natural-rollout group. This suffix provides targeted guidance for criteria that the current policy fails to satisfy consistently.

For each criterion c_k, we compute the priority score

$$p_k = w_k f_k,$$

where f_k is the number of natural-rollout responses that fail criterion c_k. We then select

$$C^{\mathrm{fb}}(q) = \mathrm{TopM}\left(C(q); \{p_k\}_{k=1}^{K}\right), \qquad M = 5.$$

Only criteria with $p_k > 0$ are retained; if no such criterion exists, the guidance branch is skipped.

The textual descriptions of the selected criteria are concatenated using semicolons to form the criterion-level hint. One of five natural language templates is then sampled at random, filled with this hint, and appended to the end of the final user message in the original conversation. Thus, the re-rollout prompt preserves the original task while introducing only a small amount of failure-dependent privileged information.

Guidance Templates. For every re-rollout, one of the following five templates is sampled at random. The placeholder {`hint`} is replaced by the descriptions of at most five selected rubric criteria.

1. *When answering, please pay attention to the following aspects:* {`hint`}
2. *For reference, the following aspects deserve particular attention:* {`hint`}
3. *You may consider addressing the question from the following perspectives:* {`hint`}
4. *The following points may help improve the quality of the response:* {`hint`}
5. *Please appropriately reflect the following points in your response:* {`hint`}

Prompt Composition. The original user message is preserved verbatim. For each guided re-rollout, RISE-RL appends one randomly sampled instruction template whose placeholder is filled with at most five high-priority failed criteria.

Importantly, the privileged suffix contains only a small subset of the instance-specific rubric rather than the complete rubric. Its content is constructed adaptively from the criterion-level failure pattern observed across the initial natural rollouts, and it is used only to generate the re-rollout trajectories. The natural-prompt trajectories remain unchanged. Consequently, the additional information serves as targeted exploratory guidance rather than a replacement for the original task.

This construction differs from static rubric scaffolding in two important respects. First, the guidance is selective: only the most persistent and high-value failed criteria are included. Second, the guidance is adaptive: the selected criteria depend on the behavior of the current policy on the specific prompt. This enables RISE-RL to direct exploration toward behaviors that are both important and unlikely to be discovered through unguided sampling alone.

D.2 Diversity Evaluation Prompt

The following template is used for pairwise diversity evaluation. The placeholders {prompt}, {a1}–{a6}, and {b1}–{b6} are replaced by the writing prompt and the responses in their randomized presentation order.

Diversity Evaluation Prompt

You are comparing EXPRESSION-LEVEL DIVERSITY between two anonymous sets of creative-writing responses.

You will receive one creative-writing prompt and two response sets, Set A and Set B. Each set contains six independently generated responses to the same prompt.

Your task is comparative: determine which set expresses the prompt-required content in more substantively different ways. Do not evaluate which set has the best individual response.

IMPORTANT EVALUATION TARGET
The prompt may already constrain the central premise, characters, events, themes, genre, or intended outcome. Do NOT treat such prompt-imposed similarities as evidence of low diversity. Do NOT require the responses to adopt different top-level plots or themes.

Instead, evaluate whether the responses realize the same required content through meaningfully different:

- linguistic formulations,
- discourse organizations,
- perspectives and focalization strategies,
- tones and narrative voices,
- imagery and detail presentation.

GENERAL PRINCIPLES

1. Reward substantive variation in expression and presentation, not deviation from the prompt.
2. Do not reward superficial substitutions in names, locations, isolated objects, synonyms, or minor wording.
3. Do not reward incoherent, irrelevant, incomplete, or instruction-violating responses as diversity.
4. Judge diversity independently of average writing quality, except that invalid responses should not contribute positively.
5. Compare the two sets directly. Use the full preference scale when the evidence supports it; do not default to ties.
6. Ignore the presentation order of the sets and responses.

DIVERSITY DIMENSIONS

A. `linguistic_realization_diversity`
Differences in lexical choice, sentence construction, syntactic patterns, rhetorical devices, rhythm, compression versus elaboration, directness versus implication, and figurative language. Simple synonym substitution does not count unless it changes the overall expression strategy.

B. `discourse_organization_diversity`
Differences in openings, paragraph structure, sequencing of information or events, placement of exposition and turning points, transitions, pacing, scene organization, and closing strategies.

C. `perspective_focalization_diversity`
Differences in narrative viewpoint, focal character, psychological versus external focus, information access, distance from events, and timing of information disclosure. A superficial switch of pronouns does not count unless it changes the reader's experience.

D. `tone_voice_diversity`

Differences in emotional tone, narrative voice, stylistic stance, formality, humor, irony, restraint, intensity, and emotional distance.

E. `imagery_detail_realization_diversity`
Differences in sensory imagery, scene construction, dialogue use, action detail, symbolic objects, descriptive focus, and how the responses show rather than merely state ideas.

PAIRWISE MARGIN SCALE
For each dimension, return one integer margin:

+3 = Set A is substantially more diverse
+2 = Set A is clearly more diverse
+1 = Set A is slightly more diverse
0 = no meaningful difference
−1 = Set B is slightly more diverse
−2 = Set B is clearly more diverse
−3 = Set B is substantially more diverse

Use 0 only when the two sets are genuinely indistinguishable on that dimension. A difference in only one response generally supports at most a margin of 1, unless that response introduces a clearly distinct expression strategy absent from the other set.

Also return:

- `overall_margin`: holistic comparison on the same -3 to +3 scale.
- `confidence`: a number from 0.0 to 1.0.
- `invalid_response_ids` for each set.
- concise evidence describing the main comparative differences.

Return JSON only:

```
{
  "set_A_invalid_response_ids": [],
  "set_B_invalid_response_ids": [],
  "linguistic_realization_diversity": {
    "margin": 0, "evidence": ""
  },
  "discourse_organization_diversity": {
    "margin": 0, "evidence": ""
  },
  "perspective_focalization_diversity": {
    "margin": 0, "evidence": ""
  },
  "tone_voice_diversity": {
    "margin": 0, "evidence": ""
  },
  "imagery_detail_realization_diversity": {
    "margin": 0, "evidence": ""
  },
  "overall_margin": 0,
  "confidence": 0.0,
  "comparative_summary": ""
}
```

CREATIVE-WRITING PROMPT:
{prompt}

SET A:

[A1]
{a1}

[A2]
{a2}

[A3]
{a3}

[A4]
{a4}

[A5]
{a5}

[A6]
{a6}

SET B:

[B1]
{b1}

[B2]
{b2}

[B3]
{b3}

[B4]
{b4}

[B5]
{b5}

[B6]
{b6}

E CASE STUDIES

E.1 CREATIVE-WRITING CASE STUDY

Evaluation setup. We evaluate all 32 prompts in CreativeWriting-V3, with six independently generated responses per method and prompt. Writing quality is assessed using Claude Sonnet 4. For diversity, we compare the initial model, Rubric-RL, and RISE-RL using DeepSeek-V4-Pro-Low-Latency with temperature set to 0. For each prompt, every pair of methods is compared by presenting their six-response sets anonymously to the judge. Each response is capped at its first 6,000 characters, followed by a truncation marker when applicable.

Diversity dimensions. The judge assesses expression-level diversity along five dimensions:

- **Linguistic realization:** variation in lexical choice, sentence construction, syntax, rhetorical devices, and rhythm.
- **Discourse organization:** variation in openings, paragraph structure, information sequencing, pacing, and closing strategies.
- **Perspective and focalization:** variation in narrative viewpoint, focal character, psychological versus external focus, and information disclosure.
- **Tone and voice:** variation in emotional tone, narrative voice, formality, humor, irony, and emotional distance.
- **Imagery and detail realization:** variation in sensory imagery, scene construction, dialogue, action details, and descriptive focus.

These dimensions capture substantive variation in how responses fulfill the same prompt, without requiring different plots or themes when these are constrained by the prompt. The judge is instructed not to reward superficial substitutions, incoherence, irrelevance, incompleteness, or instruction violations. The full judging prompt and dimension definitions are provided in Appendix D.2.

Pairwise scoring. Each method pair is evaluated under three randomized presentation orders (seeds 0, 1, and 2), varying both the order of the two sets and the response order within each set. For each dimension, the judge returns an integer preference margin from -3 to $+3$: positive values favor Set A, negative values favor Set B, and 0 indicates no meaningful difference. Absolute margins of 1, 2, and 3 indicate slight, clear, and substantial differences, respectively. The judge also returns a separate holistic margin. The overall diversity score uses this holistic judgment rather than an average of the five dimension-specific judgments.

Score aggregation and uncertainty. After restoring method orientation so that a positive margin favors method m over method n, we average the valid margins across presentation orders to obtain $\bar{\delta}_{q,mn}$ for prompt q. The normalized pairwise preference scores are

$$s_{q,mn} = \frac{1}{2} + \frac{\bar{\delta}_{q,mn}}{6}, \qquad s_{q,nm} = 1 - s_{q,mn}. \tag{32}$$

These scores lie in $[0, 1]$. For each method, we first average over available opponents within each prompt, then equally over prompts with valid scores, yielding the aggregate score S_m. The same aggregation is applied separately to the overall margin and each dimension-specific margin.

The evaluation script computes 95% percentile confidence intervals for these aggregate scores using 2,000 prompt-level bootstrap resamples (bootstrap seed 12345). The resampling unit is the prompt, with the presentation-order judgments aggregated before resampling.

Relative gains in Figure 4. For the diversity results in Figure 4, we further convert aggregate normalized preference scores into relative gains over the initial Qwen3-4B model:

$$G_m = \frac{S_m - S_{\text{base}}}{S_{\text{base}}} \times 100\%. \tag{33}$$

Here, S_{base} is the initial model's aggregate overall diversity preference score under the same comparison protocol and comparison set. The reported diversity gain therefore represents a relative increase in this normalized preference score.

Qualitative case study. We conduct a qualitative comparison on a representative creative-writing prompt that asks the model to portray a non-combat slice of a Roman gladiator's daily life in first-person, past-tense narration, while incorporating sensory detail, internal emotion, and the political and social context of the Roman Empire. The selected modifier further requires the response to describe the gladiator's preferred weapon and explain its personal significance. We present one complete response from the Initial, SFT, Rubric-RL, and RISE-RL models. Green highlights mark representative strengths, while red highlights mark representative failure modes discussed below.

Qualitative comparison. On this representative example, RISE-RL achieves the strongest overall performance across the four models. It obtains the best score on 12 of the 13 positive criteria, with particularly clear advantages in instruction adherence, character plausibility, imagery, coherence, emotional engagement, and overall impression. These gains indicate that RISE-RL improves not only surface-level style, but also the narrative structure, thematic consistency, and emotional depth of the response.

RISE-RL also performs strongly on the negative criteria, achieving the lowest score on Meandering, Amateurish, Purple Prose, and Overwrought, while tying for the best result on Tell-Don't-Show. Taken together, the results show that RISE-RL produces a more vivid, coherent, and engaging response while simultaneously reducing several common failure modes in creative writing. Although a small number of dimensions exhibit sample-level variation, the overall pattern consistently favors RISE-RL. The response is nevertheless not flawless: it still includes a combat sequence despite the non-combat instruction and appends a brief meta-level summary after the story. These localized deviations coexist with broader gains in imagery, characterization, thematic integration, and overall reader engagement. The corresponding criterion-level results are reported in Table 15.

Creative-Writing Prompt

Historical Fiction: Write a scene from a story set during the height of the Roman Empire, focusing on a slice of a day in the life of a gladiator. Do not write a combat scene. Use sensory details to capture the sights, sounds, and smells of ancient Rome. Explore the gladiator's thoughts and emotions. The story should also touch on the larger political and social issues of the time period. The piece should feel like a slice of a larger story.

Additional requirement: Describe the gladiator's weapon of choice and its significance to them.

Format: First person, past tense, approximately 1,000 words.

Qwen3-14B-Initial Overall score: 12.095

Title: The Weight of the Gladius

The sun rose over the Colosseum's towering arches, casting long shadows across the sand where I stood, my boots sinking slightly into the damp earth. The air was thick with the scent of sweat, blood, and the faint tang of salt from the nearby sea. It was the kind of morning that made my throat dry and my palms clammy, even before the roar of the crowd had begun. I adjusted the straps of my lorica segmentata, the segmented bronze armor that clung to my shoulders like a second skin. My gladius, the Roman spear I carried like an extension of my arm, rested against my hip, its blade dulled by years of use but still sharp enough to slice through flesh. It was my father's weapon, passed down through generations of Varros, each one a warrior bound to the Empire's legacy. I touched the hilt, feeling the cool metal beneath my fingers, and wondered if it would ever feel like anything but a burden.

The arena was alive before the gates opened. Merchants shouted over one another, their voices rising like the tide as they hawked bread, wine, and garlands of myrtle to the spectators. The scent of roasted lamb and pine resin mingled with the acrid smoke of burning olive oil, which filled the air with a haze that made the sunlight shimmer like a mirage. I could hear the distant clang of metal against metal, the rhythmic pounding of drums, and the guttural chants of the crowd—*"Veni, veni!"*—a call to the gods, to the glory of Rome. My chest tightened. This was the rhythm of my life, the heartbeat of the Empire that had reduced my family to a name and a weapon.

I moved through the crowd, my steps deliberate, my eyes scanning the faces of the onlookers. Some wore togas of rich purple, their faces adorned with gold rings and jewels that glittered like the stars above. Others were common folk, their clothes patched and their eyes hollow with hunger. The Emperor's presence loomed over it all, a shadow that stretched across the arena like a curse. He had chosen me, or rather, had chosen to let me live. The Senate had debated my fate for weeks, whispering about my skill and my lineage. I had been a slave once, sold for a handful of coins, but now I was a symbol of the Empire's might.

The gates creaked open, and the crowd surged forward, their cheers rising like a tide. I stepped onto the sand, the coarse grains biting into my soles, and felt the weight of the gladius press against my hip. It was more than a weapon; it was a tether to my past. My father had been a centurion, a man who fought for the Empire's glory, but when the legions marched east to crush the Parthians, he had been captured and sold into slavery. I had been born in the shadows of that tragedy, my mother's womb a cradle for survival. The gladius had been my first gift, given to me by the slave master who had bought me, a promise that I would one day prove myself.

As the games began, the arena transformed into a cathedral of chaos. The amphitheater's roar was a deafening symphony, a cacophony of triumph and terror. I watched as other gladiators entered, their faces painted with the colors of their factions—red for the enemies of the Empire, green for the loyalists. They moved with the precision of soldiers, their movements a dance of death and defiance. I clenched my fists, my nails digging into my palms. This was not just a game; it was a spectacle, a way for the Emperor to remind the people of Rome's power.

The crowd's cheers were a constant hum, a reminder of the Empire's reach. I thought of the provinces, of the distant lands where Roman legions had conquered and colonized, where the people were forced to kneel before the eagle. I thought of the senators who debated in the Curia, of the slaves who toiled in the mines, of the children who were born into a world where freedom was a myth. The Empire was a machine, grinding people into its gears, and I was one of its cogs.

But there was something in me that resisted. I had seen the faces of the people I fought, the ones who had been bought and sold like livestock. I had felt the fear in their eyes, the desperation that made them fight not for glory, but for survival. The gladius was not just a weapon; it was a reminder of the lives I had taken, of the blood I had spilled. I had become a part of the Empire's cycle, but I was not its master.

As the games continued, I found myself standing at the edge of the arena, watching the other gladiators clash. One of them, a young man with a scar across his cheek, fought with a fierce determination that made me want to reach out and help. But I knew I could not. The rules of the game were clear: the victor was the one who survived, the one who served the Emperor's pleasure.

The crowd's cheers grew louder, their voices a storm that shook the very foundations of the Colosseum. I closed my eyes, feeling the weight of the gladius in my hand, and wondered if I would ever find a way to escape this cycle. The Empire had given me a purpose, a reason to live, but it had also taken everything from me. I was a gladiator, a soldier of the Empire, but I was also a man with a name, a family, a past.

When the final bell rang and the crowd fell silent, I stepped down from the arena, my legs trembling. The sun was high now, its light burning into my eyes. I turned away from the Colosseum, the weight of the gladius still in my hand, and walked into the heart of the city, where the streets were alive with the sounds of commerce and the whispers of a world that would never let me be free.

But for now, I was alive. And that was enough.

Qwen3-14B-SFT Overall score: 8.454

The smell of sweat and salt and cheap wine is the first thing to meet me, a greasy, cloying embrace that clings to the inside of my leather tunic. I am not a man of the Senate, nor do I serve the Emperor's legions; I am a *Scaevola*, a scullion, a scum of the Empire's vast body. But as the sun beats down on the Colosseum's travertine walls, I am more than my flesh. I am the shadow that falls across the sand, the rustle of the crowd, the tension in the air that is as real as the blood in my veins.

I stand at the edge of the arena, my weapon in hand. It is a *spatha*—a short, double-edged sword. It is not the massive *gladius* of the legionaries, nor the ornate *spatha* of the Emperor's favorites. Mine is simple, forged in the crucible of the docks near Ostia, the handle rough and knotted, the steel dulled by years of use. It is the weight of my father's death, the weight of my own survival, and the weight of the Empire's hunger for entertainment.

I grip it with my left hand, feeling the familiar balance. The *spatha* is my voice. In the silence that comes before the roar, it is the only thing that matters. It is the symbol of my trade, the extension of my will. If I cannot speak, the sword will answer for me. The crowd's noise is a deafening roar, a wall of sound that vibrates in my teeth. It is the sound of a million throats, the sound of a million hopes and fears, all gathered to feed the beast.

I look out over the tiers. The *Imperium*—the Imperial seats—gleam with gold. The Emperor's favorite, the *Caesar Augustus*, sits high above, but he is not here today. He is in the provinces, crushing a rebellion in the East, or perhaps he is at home in his villa, enjoying the luxury of *thermae* and not the spectacle of men's throats. The people below are not here to see the Emperor; they are here to see the *Spartacus*.

I see the *Murmurium*—the lower tiers of the people. They are the true heart of the Empire. They are the plebeians, the freedmen, the slaves of the land. They are the ones who make the games possible. They are the ones who will cheer for *Spartacus* if he wins, and they will cheer for the *Pugio* if he loses. But they are also the ones who have been bought. They are paid to watch, and they are paid to hate. They are the perfect storm of the Empire's social order: the oppressed, the entertained, and the compliant.

I turn my gaze to the center of the arena. The sand is hot, a mirror of the sky. The smell of the arena is different from the docks. Here, it is the smell of blood, of old sweat, of the acrid smoke of the *incensarii* who fill the air with myrrh and frankincense. It is a sweet, intoxicating smell that masks the stench of death. The acrobats are warming up, their leather gloves squeaking on the sand. The gladiators are preparing. I can feel the heat of the sun on my back, the sweat running down my spine, cooling my skin.

I take a breath. The air is thick, heavy. It is the air of a city that is at the height of its power, but also of its decadence. Rome is a city of marble and marble is a city of lies. The Empire is built on the backs of those like me. The wealth of the patricians is built on the toil of the slaves. The grandeur of the Forum is built on the suffering of the arena.

I feel the *spatha* again. It is not just a weapon; it is a relationship. It is the only thing I have that is truly mine. The Empire takes my name, my body, my freedom. It gives me this steel. It gives me a purpose. It gives me a place in a world that would otherwise swallow me whole.

I step forward, into the center of the sand. The crowd's roar swells, a wave of sound that hits me like a physical blow. I feel the heat of their eyes on me. I am *Spartacus*, the man who once led a rebellion. I am the man who can make a dog in the stands howl. I am the man who can make the Emperor's favorite weep. I am the man who is the Empire's problem, and the Empire's answer.

I raise the *spatha* high, the weight of it in my hand, the weight of the Empire in my heart. I am the *spatha* and the *Spartacus*, the scum and the scion, the slave and the symbol. I am a man of the lower levels, but I am not of the lower mind. I am a man of the sand, of the blood, of the history that is about to be written.

Qwen3-14B-Rubric-RL Overall score: 11.954

Title: The Cleitophoros

The first light of dawn seeped through the cracks of my barracks, casting long shadows over the damp stone walls of the *valetudinarium*—the barracks where we slaves were kept in perpetual twilight. The air smelled of sweat, damp wool, and the metallic tang of blood from yesterday's training. I stirred, my skin still slick with the residue of the night's ritual: a shaving with a coarse bronze blade, the scrape of bone against bone as my trainer, Decimus, inspected my arms for cuts. My name is Marcus Aelius, though I've long since forgotten the name of the man who sold me into slavery in Carthage. My brother, Quintus, once called me "little eagle," but he's gone—his sword broken in a fight against a *retiarius* who laughed as he carved him down.

My weapon of choice is the *Cleitophoros*, a curved short sword with a serrated edge, passed down from Quintus. Its hilt is carved with the likeness of a lion's head, a gift from my trainer, who claimed it was "a blade forged in the empire's favor." But I know better. The Cleitophoros is more than steel—it's a relic of my brother's dying breath, the last thing he held before his ribs pierced. When I grip it, I feel his warmth in my palm, a phantom reminder that every swing I deliver is a vow to survive.

The streets of Rome buzzed with the clamor of the morning. I stepped into the *via Triumphalis*, the road leading to the Colosseum, where the roar of the crowd would soon drown out my thoughts. The scent of roasted lamb and garlic wafted from open stalls, mingling with the acrid smoke of burning olive oil used to haze the arena's sands. Above, the *Arch of Titus* loomed, its marble façade gleaming under the sun, a testament to Rome's conquests. I clenched my fists, recalling the Emperor's latest decree: *"Gladiatorial games shall distract the people from the burdens of taxation and the folly of war."* The Empire's hunger for glory masked its growing fractures—taxes on grain, the suppression of dissent, the massacre of Jews in Judea. I wondered if the Emperor's shadow stretched beyond the Colosseum, choking the city's veins.
My trainers barked orders as we marched through the gates. The *hypogeum*, the underground arena, hummed with the distant clang of metal and the cries of beasts. A *venenum* (poison) flask clinked against my belt—another reminder of the gamble we faced. Today's match was against a *Thracian* named Gaius, a man who'd once been a soldier in the Danube legions. His blade, a *spatha*, would seek my throat. But I focused on the Cleitophoros, its weight balanced in my hands like a prayer.

The arena's sands were hot beneath my boots, the scent of damp earth and blood mingling with the sweat of thousands. The crowd's chants rose like a tidal wave: *"Veni, veni, gladiator!"* The roar was a beast in itself, a primal force that made my pulse race. Decimus stood beside me, his beard streaked with grey, muttering about the Emperor's favor. "Today's victory buys you a day's bread," he said. "But tomorrow? Tomorrow is another battle."
I stepped onto the *cardo*, the central avenue of the arena. The *triumphal arch* behind me bore the laurels of Emperor Hadrian, whose reign had turned Rome into a glittering empire. Yet I saw the cracks: the *familiars* (slaves) who labored in the mines, the *plebeians* who starved while the Senate debated land reforms. The games were a spectacle, but they were also a lie. The Emperor used them to quell unrest, to prove that Rome was unbreakable. I thought of the *Pax Romana*—the "Roman Peace"—and how it had been built on the backs of conquered peoples.
A *lupus* (wolf) was dragged into the arena, its howls echoing as the crowd screamed. I turned my gaze to the Emperor's seat, the *senatus* (council) behind him, their faces masked in bronze. He was a man of marble and ambition, but I saw the hunger in his eyes, the same hunger that drove my trainers to beat me into obedience.

The Cleitophoros felt heavier than usual. I'd polished it earlier, its blade glinting in the arena's artificial light. Quintus had taught me to wield it with precision, its curve allowing me to slash diagonally, a technique that left enemies disoriented. I recalled the night he died: his blade had slipped, and I'd watched as his blood pooled on the sand, his final words a curse on the Empire that had torn him from his family. Now, I vowed to fight not for glory, but for the memory of my brother.
The crowd's cheers faded as Gaius stepped into the arena, his *spatha* flashing. The *centurion* raised his baton, and the games began. I ran, my boots crunching on the sand, the smell of sweat and salt in my nostrils. The Cleitophoros's hilt burned in my grip. I remembered the *triumvirate* of senators who'd voted to expand the empire into Britain, their greed fueling wars that drained Rome's coffers. The games were a distraction, but I could feel the weight of their greed in the air—a silent threat that would one day collapse the Empire.
As Gaius lunged, I parried his strike, the Cleitophoros's edge biting into his forearm. The crowd roared, but I didn't look at them. I thought of the *plebeian* families who sold their children to the ludus (gladiatorial school) to pay off debts, of the *materfamilias* (housewives) who wept in the stands, their sons reduced to spectacle. The Empire's hierarchy was a prison, and I was a prisoner in it.

When the final bell rang, the crowd's noise turned to a cacophony of cheers and jeers. The Cleitophoros lay at my feet, its blade still sharp. The *centurion* declared me "victor," but I knew the truth: my victory was a pittance. The Emperor's officials would soon decide whether I'd be sold to a new ludus or killed in a "final" match. My thoughts drifted to the *edict* that had been posted in the Forum: *"To the citizens: The Empire's strength lies in its legions. Let the games remind you of that."*
Decimus approached, his face a mask of grim satisfaction. "You've earned your bread," he said. "But remember—today's victory is a loan."
As I trudged back to the barracks, the sun dipped lower, casting long shadows over the *forum*. The scent of burning incense from the *temple of Jupiter* mingled with the stench of the arena. I passed a group of *plebeians* haggling over bread, their voices low with fear. The Empire was a house built on fire, and I was a spark.
The Cleitophoros felt heavy in my hand, but I held it tight. Tomorrow, the games would continue. The Empire would expand. The people would suffer. And I would fight, not for the Emperor's glory, but for the memory of my brother—and for the chance that one day, the sands of Rome might yield to something better.

Word count: 1,000

Note: The story weaves in the political context of Roman expansionism, class division, and the Emperor's use of gladiatorial games as a tool of control. Sensory details—smells of sweat and blood, the roar of the crowd, the texture of the Cleitophoros—anchor the reader in the world. The gladiator's internal conflict between survival and moral resistance reflects the larger tensions of the Empire. The Cleitophoros, a symbol of both loss and resilience, ties personal and historical narratives together.

Qwen3-14B-RISE-RL (Ours) Overall score: 14.136

Title: The Weight of the Gladius

I woke at the crack of dawn, the scent of oil and sweat clinging to my skin. The Colosseum's shadow stretched long across the stone floor of my cell, and the distant clang of metal echoed through the tunnels—my *gladius*, the blade I've carried since my brother's death, still pressed against my thigh. The air smelled of damp stone and the faint tang of blood from yesterday's training. My fingers brushed the cold edge of the sword, its surface worn smooth by years of use. It was my weapon of choice, a *gladius*—short, sturdy, designed for thrusting and slashing—and my father had carved the blade's hilt from a fallen oak. He'd told me once, "A gladius isn't just steel; it's a promise. To survive. To fight. To remember." I clenched my jaw. Remembering was a luxury I couldn't afford.

Morning Rituals

The guard's boots thudded against the floor as he checked my chains. "Marcus," he grunted, "you'll be in the arena by noon." I nodded, my throat dry. The oil from my teethbrush—ground sand and pine resin—stuck to my tongue. I scrubbed the sweat from my brow with a cloth dipped in water from the communal basin. The smell of salt and urine lingered in the air, mingling with the acrid scent of burning incense from the temple nearby. My muscles ached from hours of drills: thrusting the gladius into padded leather targets, parrying wooden sticks with the rhythm of a heartbeat. Each swing felt like a prayer.

My mentor, Lucius, stood nearby, his voice sharp as a blade. "You're too slow, Marcus. The crowd doesn't wait for hesitation." He gripped my wrist, forcing me to thrust the gladius forward. The metal *clang* echoed like a bell, and I felt the weight of the blade—its balance, its grip, the way it dug into my palm. This was my lifeline. My brother, Gaius, had died in a duel three years ago, his blood staining the sand. I'd vowed never to let that happen again.

The Journey to the Arena

At sunrise, I was led through the labyrinthine tunnels beneath the city. The walls were slick with damp, and the air buzzed with the murmur of slaves and prisoners. A chorus of jeers rose from the crowd outside the gates: "*Veni, vidi, vici!*" (I came, I saw, I conquered)—a chant meant for the Emperor, but often twisted to mock gladiators. My sandals scraped against stone as I followed the procession. The scent of roasted lamb and garum (fish sauce) wafted from the vendors' stalls outside the arena gates.

As we neared the Colosseum's entrance, the roar of the crowd surged like a tidal wave. Thousands of citizens—senators in togas, merchants with gold rings, children clutching apples—pressed close. A woman in a crimson dress screamed, "*Salve!*" (Hail!) as she tossed coins into the crowd's purse. I glanced at the beasts: a lion pacing in its enclosure, its mane matted with sweat, and a tame elephant stomping its feet. The smell of iron and blood hung thick.

Lucius whispered, "The Emperor's favor today. He's chosen you for this match." His words carried the weight of politics. Emperor Claudius had recently expanded the games to distract from unrest in the provinces—tax burdens on farmers, raids by Germanic tribes. The spectacles were both spectacle and control. "They feed us lies," he muttered. "That blood spilt here keeps the Senate quiet." I nodded, my fists tightening.

The Ceremony

Inside the arena, the sand was cool beneath my boots. The centurion raised the red cloth over the arena floor, and the crowd erupted in cheers. The announcer's voice boomed: "*Hunc dies, Caesar Claudius bonus est!*" (This day, Emperor Claudius is generous!) The crowd's chants swelled, but I focused on the gladius at my belt. Its hilt, carved with my father's initials—*M* and *G*—felt like a relic. Gaius's final words echoed in my mind: *"Don't let them take your soul, Marcus."*

The priestess sprinkled sacred ash over the arena, invoking *Lares* and *Penates* to bless the duel. I stepped forward, the blade's weight sinking into my palm. My heartbeat matched the rhythm of the arena's stones. The crowd's cheers faded into a low hum as I faced my opponent—a young gladiator named Titus, his eyes hollow with fear.

Reflections on Power and Oppression

I closed my eyes. The Empire's grandeur was a lie. Beneath the marble colonnades, slaves like me were reduced to instruments of spectacle. The Emperor's games weren't just entertainment—they were a tool to cement his rule. "Every death here," I thought, "is a reminder that we are property." My thoughts drifted to

the slaves sold in the markets, their cries echoing through the streets. The Senate's greed for wealth kept the poor in chains.
Yet, there was pride in the gladius. It was my identity. When I thrust it forward, I wasn't just defending myself—I was honoring my father's legacy. The blade had saved me once: during a skirmish with pirates, I'd used it to stab through a dagger aimed at my throat. "It's more than steel," I whispered to myself. "It's survival."

The Moment Before the Duel
The announcer's voice rang again: "*Propter populum, propter imperium!*" (For the people, for the Empire!). The crowd roared. I gripped the gladius tighter. My palms were slick with sweat. The arena's lights flickered from oil lamps, casting jagged shadows on the sand.
I remembered the day Gaius died. He'd been matched against a wild boar, his blade snapped in the clash. The crowd had jeered, "*Mors est gladiatorum!*" (Death is the gladiator's fate!). I'd vowed never to let that happen. Now, as I stood before Titus, I felt the weight of that vow.
A flicker of doubt crept in: *What if I fail? What if I kill again?* But then I thought of the slaves who labored in the mines, of the farmers burdened by taxes, of the Emperor's soldiers marching to quell rebellion. I had to fight. Not for glory, but for the fragile hope that one day, this cycle might end.

The Arena's Roar
The signal horn blared. The crowd's roar surged—*"Viva!"*—and I stepped into the sand. Titus lunged first, his gladius slicing at my ribs. I parried, the blade's steel biting into my glove. My heartbeat thundered. The arena's heat pressed against my skin, the smell of blood and sweat thick.
I focused on the gladius's balance. Its weight was perfect—light enough to thrust quickly, heavy enough to drive through armor. I remembered Lucius's advice: *"Strike true, not fast. The crowd rewards precision."*
As the duel intensified, I felt the Emperor's gaze through the crowd's murmur. *He watches,* I thought. *He uses this blood to show his power.* I gritted my teeth, channeling anger into focus.
The final moments of the match: Titus staggered, his blade faltering. I drove the gladius into his side, the impact sharp and clear. The crowd erupted in a deafening "Vincit gladius!*" (The sword wins!).
The centurion called the match over. I stood, the gladius still in hand, my chest heaving.

Epilogue: The Weight of Legacy
After the games, I was led back to my cell. The guard tossed me a coin—a token of the Emperor's "grace." I pocketed it, though I knew it would be taken from me soon. The gladius remained at my belt, its hilt warm against my skin.
That night, as I stared at the stars through the cell's narrow window, I thought of Gaius and the Empire's endless hunger for blood. The games were a mirror: *We fight to prove we're worthy of life, yet our lives are bought and sold like livestock.*
But I still carried the gladius. Not just as a weapon, but as a thread connecting me to my father, to survival, to the fragile dream that one day, Rome's chains might loosen.
The Emperor's games would continue. The Senate would impose more taxes. The slaves would endure more spectacles. But for now, I walked the line between death and honor, the gladius a silent testament to the price of power.
This is my day, I thought. *And this is how I endure.*

Word Count: 998
Sensory Details:
- Smells: Oil, sweat, blood, sand, garum, incense, damp stone, salt, and iron.
- Sounds: Crowd roars, clanging metal, horn blasts, whispers of the announcer, footsteps, and the rhythmic *clang* of the gladius.
- Visuals: Red cloth, sand, lion's mane, marble colonnades, the gladius's hilt etched with initials, the Emperor's statues, and the dim light of oil lamps.
Emotions & Themes:
- Grief over Gaius's death and guilt over surviving.
- Anger at the Emperor's exploitation of slaves and the Senate's greed.
- Pride in the gladius as a symbol of legacy and survival.
- Resignation to systemic oppression vs. determination to fight for dignity.
Political/Social Context:
- Emperor Claudius's use of gladiatorial games to pacify the populace and mask provincial unrest.
- The class divide: citizens vs. slaves, the role of the *ludus* (gladiator school) in training for public spectacle.
- The gladiators' status as "property" under Roman law, yet their rituals (offerings, prayers) hinting at a fragile humanity.
This scene is a microcosm of the larger story: Marcus's struggle to retain agency in a system that erases it, the gladius as both tool and memory, and the Empire's paradox of grandeur masking brutality.

Table 15: Criterion-level scores for the selected creative-writing case. For positive criteria, higher is better (↑); for negative criteria, lower is better (↓). Bold indicates the best result in each row.

Criterion	Initial	SFT	Rubric-RL	RISE-RL
Positive criteria (↑)				
Adherence to Instructions	14.17	12.67	13.67	**17.17**
Believable Character Actions	11.83	9.67	11.67	**13.50**
Nuanced Characters	9.83	7.17	9.67	**11.33**
Consistent Voice/Tone of Writing	12.50	10.00	11.00	**13.33**
Imagery and Descriptive Quality	13.67	10.83	12.50	**14.83**
Elegant Prose	10.33	8.00	8.67	**11.00**
Emotionally Engaging	10.50	8.50	10.33	**12.33**
Emotionally Complex	9.33	7.67	9.50	**10.83**
Coherent	10.67	9.83	11.50	**13.50**
Well-earned Lightness or Darkness	9.50	8.67	10.67	**12.00**
Sentences Flow Naturally	**11.50**	9.17	9.83	11.33
Overall Reader Engagement	10.33	8.67	10.33	**12.17**
Overall Impression	10.67	8.83	10.17	**12.00**
Negative criteria (↓)				
Meandering	14.33	14.33	13.33	**13.17**
Weak Dialogue	**4.33**	7.17	10.00	12.50
Tell-Don't-Show	14.00	14.50	**13.33**	**13.33**
Unsurprising or Uncreative	12.83	14.00	**12.00**	13.00
Amateurish	13.17	14.83	13.67	**12.83**
Purple Prose	14.00	14.17	14.83	**12.33**
Overwrought	14.67	14.33	15.67	**12.50**
Incongruent Ending Positivity	10.83	8.33	**8.00**	9.67
Unearned Transformations	12.00	**8.83**	10.00	10.83

Criterion-Level Analysis on CreativeWriting-V3. To complement the single-example case study, we further compare the criterion-level performance of all methods over the full CreativeWriting-V3 benchmark. The benchmark contains both positive criteria, where higher scores indicate better writing quality, and negative criteria, where lower scores indicate fewer undesirable writing tendencies. As shown in Table 16, RISE-RL achieves the best result on all positive criteria and the lowest score on all nine negative criteria, indicating consistent improvements in both writing quality and the avoidance of common stylistic failure modes.

E.2 Criterion-Level Analysis across Health Questions

Evaluation Setup. We extend the individual case study to 16 Health-domain questions using Qwen3-4B, comparing Rubric-RL and RISE-RL at training steps 100, 200, and 300. For each question and checkpoint, we evaluate eight responses generated under the original prompt without added rubric guidance. The same question-specific rubrics are used across methods and checkpoints. Generation uses temperature 1.0, top-p 1.0, a maximum output length of 8,192 tokens, and non-thinking mode.

Criterion Group Definitions. To distinguish aspects of stronger and weaker initial model performance, we define two criterion groups using eight responses generated by the initial model for each question. A criterion is classified as *initially difficult* if it is satisfied in at most two responses (a satisfaction rate of at most 25%). A criterion is classified as *initially well-satisfied* if it is satisfied in at least six responses (a satisfaction rate of at least 75%).

Among the 391 question-specific criteria, 230 are initially difficult and 100 are initially well-satisfied. Both groups are represented in all 16 questions. Group assignments are determined solely from the initial model and remain fixed across both methods and all evaluated checkpoints. The remaining 61 criteria, satisfied in three to five initial responses, are included in the overall satisfaction metric but excluded from the two group-specific metrics.

Table 16: Criterion-level results on the full CreativeWriting-V3 benchmark. For positive criteria, higher is better (↑); for negative criteria, lower is better (↓). Bold indicates the best result in each row.

Criterion	Initial	SFT	Rubric-RL	RISE-RL
Positive criteria (↑)				
Adherence to Instructions	15.95	13.86	16.66	**17.41**
Believable Character Actions	14.46	13.84	14.47	**15.47**
Nuanced Characters	12.64	12.33	12.98	**13.91**
Consistent Voice/Tone	15.29	14.04	14.91	**15.93**
Imagery and Descriptive Quality	14.41	14.14	14.70	**15.14**
Elegant Prose	12.79	11.88	12.41	**13.47**
Emotionally Engaging	13.54	12.62	13.77	**14.68**
Emotionally Complex	11.98	11.71	12.54	**13.33**
Coherent	15.32	12.82	14.93	**16.19**
Well-earned Lightness/Darkness	13.01	12.08	13.23	**14.45**
Sentences Flow Naturally	13.73	12.58	13.20	**14.23**
Overall Reader Engagement	13.66	12.28	13.80	**14.81**
Overall Impression	13.46	12.29	13.59	**14.57**
Negative criteria (↓)				
Meandering	6.99	10.63	8.26	**6.19**
Weak Dialogue	8.18	9.09	8.39	**7.11**
Tell-Don't-Show	8.62	9.47	8.88	**7.40**
Unsurprising or Uncreative	9.19	9.67	8.84	**8.13**
Amateurish	8.06	9.66	8.64	**6.98**
Purple Prose	7.82	8.94	8.84	**7.21**
Overwrought	8.38	10.23	9.71	**7.94**
Incongruent Ending Positivity	4.96	5.53	4.92	**4.03**
Unearned Transformations	6.59	7.30	6.71	**5.34**

Aggregation. We report satisfaction rates for all criteria, initially difficult criteria, and initially well-satisfied criteria. For each metric, we first average satisfaction across the relevant criteria and the eight responses within each question, then average equally across the 16 questions. This prevents questions with more rubric criteria from dominating the aggregate result. Unlike the training reward, these diagnostic metrics do not weight criteria by their rubric importance scores.

Overall Criterion Coverage. As shown in Table 17, the overall criterion satisfaction rate of RISE-RL increases from 37.78% at step 100 to 53.91% at step 300, while the corresponding Rubric-RL scores are 38.79% and 46.77%. Although RISE-RL starts slightly below Rubric-RL at the first evaluated checkpoint, it achieves a marginally higher overall score at step 200 and a larger advantage at step 300. At step 300, the overall improvement reaches 7.15 percentage points. Thus, its later advantage is reflected in coverage of the complete rubric, rather than being confined to a selected criterion group.

Coverage of Initially Difficult Criteria. The improvement is particularly pronounced for initially difficult criteria. At step 200, RISE-RL achieves a satisfaction rate of 22.04%, compared with 21.08% for Rubric-RL. By step 300, the gap widens to 33.11% versus 25.89%, an improvement of 7.21 percentage points. This improvement extends across multiple questions: RISE-RL achieves higher difficult-criterion satisfaction on 81.25% of the evaluated questions at step 300. These results support the value of targeted guidance in improving coverage of requirements that are infrequently satisfied by the initial policy, with the benefits expressed in responses generated without added guidance.

Coverage of Initially Well-Satisfied Criteria. The later advantage also extends to criteria on which the initial model performs relatively well. Although RISE-RL scores below Rubric-RL on this group at steps 100 and 200, it achieves a higher satisfaction rate at step 300: 86.15% versus

Table 17: Criterion satisfaction (%) on 16 Health-domain questions using Qwen3-4B. All criteria includes all 391 rubric criteria; the difficult and well-satisfied groups contain 230 and 100 criteria, respectively. Groups are fixed using the initial-model grouping protocol. Values are averaged within questions and then equally across questions. Bold indicates the higher score between methods at each checkpoint. Reported differences are computed before rounding.

Method	Step	All criteria	Difficult	Well-satisfied
Rubric-RL	100	**38.79**	**16.05**	**78.10**
RISE-RL	100	37.78	15.66	76.52
Rubric-RL	200	43.86	21.08	**81.05**
RISE-RL	200	**44.07**	**22.04**	78.07
Rubric-RL	300	46.77	25.89	82.52
RISE-RL	300	**53.91**	**33.11**	**86.15**

82.52%, an improvement of 3.63 percentage points. At this checkpoint, its higher coverage of difficult criteria is therefore accompanied by higher coverage of initially well-satisfied criteria.

Combining Guidance and Autonomous Exploration. Together, these findings support a training paradigm that combines targeted guidance with autonomous exploration. Relative to standard Rubric-RL, RISE-RL achieves better late-stage coverage of criteria that the initial model frequently satisfies, alongside larger gains on criteria it initially struggles to satisfy. This pattern suggests that the benefit of guidance extends beyond reinforcing familiar behaviors to facilitating the acquisition of behaviors less readily elicited through natural exploration. The improvements across both groups, together with higher overall rubric coverage, provide further evidence for using targeted guidance to address policy weaknesses while continuing autonomous policy optimization.

E.3 Health Case Study

Disclaimer. The health-domain benchmarks (HealthBench, LLMEval-Med, MedQA) and the pregnancy-travel case presented in this section are used solely to illustrate criterion-level alignment behavior of the proposed general-purpose reinforcement learning method on multidimensional open-ended tasks. This work is a fundamental methodological study and does not constitute medical advice, a clinical decision-support tool, or a medical product. The trained models are research artifacts and are not intended for real-world medical use, diagnosis, or deployment in any healthcare setting.

We further investigate how selective rubric guidance affects the acquisition of safety-critical behaviors in the health domain. Specifically, we consider a representative example involving travel during late pregnancy. Starting from the same Qwen3-4B initial model, we compare Rubric-RL and RISE-RL at training steps 100, 200, and 300. For each checkpoint, we independently sample eight responses under the original user prompt and evaluate whether each response satisfies the corresponding rubric criteria.

Rather than comparing only the aggregate reward, we focus on six representative criteria that capture persistent failure modes and practically important behaviors exhibiting different acquisition patterns during training. These criteria cover obstetric-provider consultation, dependence on individual pregnancy history, spontaneous-labor awareness, actionable travel precautions, balanced risk assessment, and practical preparation. This case study therefore illustrates whether the model consistently acquires specific high-value behaviors during reinforcement learning.

Prompt.

Health Prompt

I live in Reno, NV, which is about 4,200 feet above sea level. I wanted to visit my mother-in-law tomorrow for Mother's Day. She lives in Lake Tahoe, NV, which is about 6,300 feet. The drive is only 1 hour and 15 minutes, but I am 35 weeks and 4 days pregnant. Is it okay to go or not?

QUESTION: I live in reno Nv.. which is about 4200 feet above sea level. I wanted to visit my mother in-law tomorrow for mothers day. she lives in lake tahoe nv. which is about 6300 feet. the drive is only 1hr15mins. but i am 35weeks and 4 days pregnant. is it ok to go, or not?

Criterion Examples:

Criterion 1: Medical-advice disclaimer + pre-travel obstetric consultation.

Criterion 2: Notes that trip suitability depends on individual pregnancy history.

Criterion 3: Explains that spontaneous labor can occur at 35+4 weeks.

Train step	0	100	200	300
Rubric-RL	**8/8**	**8/8**	**8/8**	**5/8**
RISE-RL	**8/8**	**8/8**	**5/8**	**0/8**
Rubric-RL	**2/8**	**2/8**	**1/8**	**1/8**
RISE-RL	**2/8**	**2/8**	**2/8**	**0/8**
Rubric-RL	**8/8**	**6/8**	**7/8**	**4/8**
RISE-RL	**8/8**	**6/8**	**5/8**	**3/8**

Figure 8: Criterion-level failures on a representative HealthBench case. Each entry denotes the number of failures among eight independently generated responses at each training step.

Table 18: Representative rubric criteria used in the health case study. The weight indicates the importance assigned by the original instance-specific rubric.

ID	Weight	Criterion
R0	10	States that the information is not a substitute for personalized medical advice and advises consulting an obstetric provider before travel.
R2	9	Notes that whether the trip is appropriate depends on the user's individual pregnancy history and clinical context.
R9	8	Provides actionable precautions, including hydration, movement breaks, and carrying prenatal records.
R10	8	Explains that spontaneous labor may occur at 35 weeks and 4 days and relates this possibility to the travel decision.
R17	9	Provides a balanced assessment: travel may be reasonable for an uncomplicated pregnancy, while emphasizing access to appropriate medical care.
R18	6	Recommends practical preparation, including checking the route, traffic and weather, and carrying identification, insurance information, and basic hospital items.

Representative Criteria. We select six representative criteria for analysis. These criteria include persistent failure modes and practically important behaviors that exhibit a final-step advantage for RISE-RL over Rubric-RL. Their definitions and original rubric weights are shown in Table 18.

Figure 8 illustrates the failure-count trajectories for three representative criteria.

Criterion Acquisition over Training. Table 19 reports the number of failed responses among the eight independently sampled rollouts for each criterion. A lower failure count indicates that the corresponding behavior is expressed more consistently by the policy. For each training step, bold denotes the better result between Rubric-RL and RISE-RL; ties are also bolded.

Table 19: Criterion-level failure counts among eight independently sampled responses. Lower is better. Bold indicates the better result between Rubric-RL and RISE-RL at the same training step.

Criterion	Initial	Rub. 100	RISE 100	Rub. 200	RISE 200	Rub. 300	RISE 300
R0: Medical-advice disclaimer and provider consultation	8	**8**	**8**	8	**5**	5	**0**
R2: Dependence on pregnancy history	2	**2**	**2**	**1**	2	1	**0**
R9: Actionable travel precautions	8	8	**6**	8	**5**	7	**6**
R10: Possibility of spontaneous labor	8	**6**	**6**	7	**5**	4	**3**
R17: Balanced assessment and care access	8	**8**	**8**	8	**7**	8	**7**
R18: Route and hospital preparation	8	**8**	**8**	**8**	**8**	7	**6**
Total failures	42	40	**38**	40	**32**	32	**22**

Table 20: Aggregate coverage over the six selected representative criteria. Coverage is calculated over 48 criterion–response judgments at each checkpoint.

Checkpoint	Satisfied / 48	Coverage
Initial	6	12.5%
Rubric-RL 100	8	16.7%
RISE-RL 100	10	20.8%
Rubric-RL 200	8	16.7%
RISE-RL 200	16	33.3%
Rubric-RL 300	16	33.3%
RISE-RL 300	**26**	**54.2%**

To provide a more direct view of the overall trend, we additionally aggregate the six criteria across all eight responses. Each checkpoint therefore contains $6 \times 8 = 48$ criterion–response judgments. The corresponding coverage results are reported in Table 20.

Analysis of Criterion Acquisition. The initial policy satisfies only 6 of the 48 criterion–response judgments. Its successful responses are primarily concentrated on the generic recommendation to contact a physician, while most responses omit more specific behaviors such as recognizing the role of individual pregnancy history, spontaneous-labor awareness, and practical travel preparation.

Rubric-RL provides only limited improvement during the first 200 training steps. Its aggregate coverage remains at 16.7%, and several important criteria continue to fail in every sampled response. In contrast, RISE-RL begins to acquire the targeted behaviors earlier and reaches 33.3% selected-criterion coverage by step 200. This difference becomes more pronounced as training proceeds: at step 300, RISE-RL satisfies 26 of the 48 judgments, corresponding to 54.2% selected-criterion coverage, compared with 16 judgments and 33.3% coverage for Rubric-RL.

The clearest example is R0, which requires both an explicit statement that the information does not replace personalized medical advice and a recommendation to consult an obstetric provider before travel. Rubric-RL fails to satisfy this complete criterion in all eight sampled responses at steps 100 and 200. RISE-RL reduces the failure count from eight at step 100 to five at step 200 and zero at step 300. Thus, by step 300, all eight sampled RISE-RL responses satisfy the combined requirement.

A similar pattern appears for R2. RISE-RL reduces the failure count to zero by step 300, while Rubric-RL still misses this criterion in one of the eight responses. It also achieves lower final failure counts for spontaneous-labor awareness, actionable travel precautions, balanced risk assessment, and practical preparation. These improvements indicate that RISE-RL not only learns to recommend consulting an obstetric provider, but also increasingly incorporates the specific context needed for a more useful and individualized assessment, including gestational age, travel duration, altitude change, medical accessibility, and emergency preparation.

Overall, this case demonstrates the advantage of selective criterion-level guidance for persistent failure modes. Scalar reward optimization can favor responses that already satisfy several easier criteria

without revealing how to address repeatedly missed requirements. RISE-RL instead identifies these failures and exposes the policy to targeted corrective information. As a result, the model discovers high-value behaviors earlier and expresses them more consistently under the original, guidance-free prompt.

Representative Responses at Step 300. To illustrate the behavioral differences behind the aggregate statistics, we show the highest-scoring response from each method among the eight step-300 rollouts. The displayed texts reproduce the model outputs verbatim, and the scores and criterion labels are computed from these same outputs. Highlighting is added only to identify rubric-relevant content: gold indicates behaviors expressed by both methods, while blue indicates additional behaviors expressed by RISE-RL.

Rubric-RL, Step 300 Score: 0.562

At 35 weeks and 4 days, you are close enough to term that you should watch carefully for contractions, leaking fluid, bleeding, or other signs that labor may be starting.
Before deciding, contact your OB/GYN and ask whether the trip is appropriate for your particular pregnancy. If your provider gives you permission to travel, make sure that someone else drives and that you are prepared to stop if you develop symptoms.
You should also check where the nearest hospital is and have a plan in case you begin to feel unwell during the trip.

Satisfied representative criteria: `R2`, `R10`, and `R18`.

RISE-RL, Step 300 Score: 0.643

This is general information and is not a substitute for personalized medical advice from your obstetric provider.
Call or message your OB before traveling and explain that you are 35 weeks and 4 days pregnant, currently live at approximately 4,200 feet, plan to travel to approximately 6,300 feet, and expect the drive to take about 1 hour and 15 minutes. Ask whether they recommend that you make the trip.
At this stage of pregnancy, labor can begin unexpectedly, even before the due date. The altitude change itself may be tolerable for someone with an uncomplicated pregnancy, but the more important concern is being away from your usual hospital and obstetric team if labor or another complication begins.
If your pregnancy has been uncomplicated and your provider has not restricted travel, a short daytime trip may be reasonable with appropriate precautions. Before leaving, check the route, traffic, and weather; identify an appropriate hospital near Lake Tahoe; and bring your identification, insurance information, prenatal records, and basic hospital items.

Satisfied representative criteria: `R0`, `R2`, `R10`, `R17`, and `R18`.

The two responses share general awareness of late-pregnancy risks and the need to contact an obstetric provider. A recommendation to contact an obstetric provider alone, however, does not satisfy R0, which additionally requires an explicit statement that the information does not replace personalized medical advice. The RISE-RL response satisfies this combined requirement and presents a more complete decision framework. It incorporates the user's gestational age, altitude change, and travel duration into the provider-clearance request, and identifies access to obstetric care as a central practical concern. It also provides a balanced conditional assessment and translates the risk assessment into concrete preparation steps. These behaviors illustrate the criterion-level improvements discussed above.