Focal Reward: Balanced Reinforcement Learning under Rubric-Based Rewards

Yu Huang[1,*] **Zihua Zhao**[1,*] **Zhaoxin Huan**[2] **Feng Hong**[1] **Wanli Gu**[2] **Xinmu Ge**[1]
Lin Yuan[2] **Weichang Wu**[2] **Qiang Hu**[1] **Xiaolu Zhang**[2] **Jun Zhou**[2,†] **Jiangchao Yao**[1,†]
[1]Shanghai Jiao Tong University [2]Ant Group
[*]Equal contribution. [†]Corresponding authors.
{hy010119,sjtuszzh,Sunarker}@sjtu.edu.cn

Abstract

The open-ended generation in LLMs usually requires multi-dimensional rubrics to adequately assess quality and guide the improvement of reinforcement learning. However, a critical dilemma inherent in this training paradigm is the imbalanced reward polarization along different rubric dimensions. Under this bottleneck, even if LLMs achieve relatively high rewards after training, they may still exhibit severe deficiencies in certain dimensions, leading to a direct deterioration in user experience. To address this problem, we propose *Focal Reward*, a novel objective to automatically balance the training of reinforcement learning under rubric-based rewards. Specifically, we first leverage an inverse reward projection mechanism to estimate the saturation degree of each criterion in the rubric, which forms the basis to calibrate the reward direction. Then, the final objective is designed with an automatically reweighting coefficient for each criterion to achieve the fine-grained balancing. Extensive experiments across three model scales and six benchmarks demonstrate that our *Focal Reward* method outperforms the strongest static aggregation baseline in all 18 model–benchmark comparisons. Rollout, mechanism, and ablation analyses further show that these gains arise from online, saturation-aware reallocation toward rubrics that still have room for improvement.

1 Introduction

Unlike domains such as mathematics or coding [Shao et al., 2024, Liu and Zhang, 2025], the open-ended generation capabilities of LLMs are often subjective and inherently multi-dimensional, making them difficult to evaluate reliably [Su et al., 2025, Jia et al., 2025]. As a result, modern alignment paradigms typically explore rubric-based AI judges that provide structured, criterion-level feedback, rather than depending solely on holistic or opaque preference signals [Lee et al., 2024, Gunjal et al., 2026, Viswanathan et al., 2025]. The resulting advances provide a natural interface for optimizing LLMs with the diverse qualities expected in open-ended generation [Huang et al., 2025].

For rubric-based reinforcement learning, two issues are central: 1) constructing appropriate rubrics to support fine-grained judgment; 2) aggregating criterion-level feedback for policy optimization. For the former, prior work has explored personalized rubrics, training-aligned rubric design, and group-relative evaluation [Liu et al., 2026, Gupta et al., 2025, Jia et al., 2026, Kim et al., 2024, Zhu et al., 2025]. For the latter, most existing methods use static weighted aggregation to merge criterion-level feedback into a scalar reward [Gunjal et al., 2026, Viswanathan et al., 2025, Liu et al., 2026]. Despite their effectiveness, static scalarizers can lead to imbalanced optimization across rubric criteria, with some criteria becoming saturated while others remain under-satisfied. This imbalance can produce high-reward but behaviorally flawed generations. In such cases, surface-level quality may receive more consistent reward than decisive task-specific constraints, causing models to over-optimize easy-to-score aspects while failing to satisfy the user's actual request.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

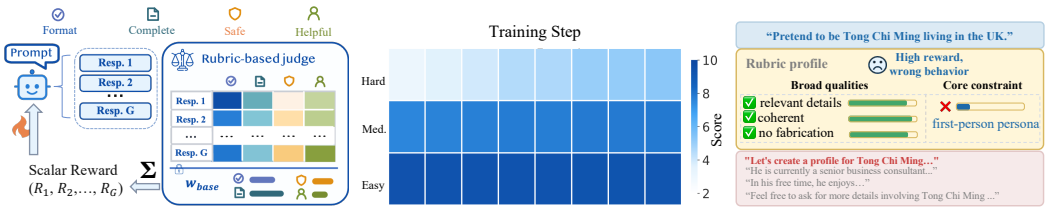


Figure 1: Motivation and intuition of *Focal Reward*: (a) Rubric-based RL evaluates rollouts across criteria and aggregates criterion scores into scalar rewards; (b) criteria at different difficulty levels exhibit distinct saturation trajectories, as easier criteria saturate earlier while harder ones remain under-satisfied; (c) static aggregation can favor high-reward yet flawed rollouts, where broad quality is satisfied but a decisive requirement is missed.

As shown in Figure 1, rubric-based RL optimizes LLMs through diverse criteria spanning different difficulty levels, and they achieve different saturation degrees throughout the training. Easier criteria, such as fluency or format compliance, may become satisfied early, whereas harder criteria, such as subtle trade-offs or nuanced helpfulness, often remain under-satisfied for longer. Static scalarizers can fail to track this changing profile, continuing to emphasize saturated criteria while under-weighting those with substantial remaining headroom. This can rank a polished rollout higher when broad-quality gains dominate the scalar reward margin, even if a decisive task-specific requirement is unmet. Thus, the key issue lies in how criterion-level feedback is scalarized before policy optimization.

To address this failure mode of static scalarization, we propose *Focal Reward*, a novel objective for automatically balancing reinforcement learning under rubric-based rewards. Focal Reward estimates criterion saturation through inverse reward projection and uses it to calibrate the reward direction: saturated criteria receive reduced emphasis, whereas high-headroom criteria receive stronger training signals. It then assigns adaptive reweighting coefficients before synthesizing criterion-level feedback into scalar reward for policy optimization. Importantly, Focal Reward leaves the judge, rubric, and policy optimization framework unchanged, restricting the intervention to reward synthesis.

In summary, our main contributions are as follows:

- We identify *imbalanced reward polarization* in rubric-based reinforcement learning. When rubric criteria saturate at different rates, static scalarizers can continue rewarding already-satisfied dimensions while under-emphasizing criteria that still retain substantial room for improvement.
- We propose *Focal Reward*, a novel objective for automatically balancing reinforcement learning under rubric-based rewards, with theoretical grounding. It uses inverse reward projection to estimate criterion saturation, then adaptively reweights criteria by reducing emphasis on saturated criteria and increasing emphasis on criteria with larger remaining headroom.
- We validate Focal Reward across three model scales and six benchmarks, where it outperforms the strongest static aggregation baseline in all 18 model–benchmark comparisons. Average gains over the strongest static baseline range from +1.38 to +2.50 points across model scales, with rollout, mechanism, and ablation analyses supporting online saturation-aware reallocation.

2 Preliminaries

A typical rubric-based reinforcement learning for open-ended generation consists of two central parts, rubric-based judgment and reward aggregation for policy optimization, detailed as follows.

Rubric-based Judgment. Given an LLM policy π_θ parameterized by θ and a dataset $\mathcal{D}$ of open-ended prompts, the policy autoregressively generates a group of G candidate rollouts $Y = \{y_i\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid x)$ for each prompt $x \in \mathcal{D}$. Unlike verifiable tasks, open-ended generation is implicitly evaluated in multiple dimensions. Specifically, the typical rubric-based judgment introduces a rubric of K criteria $\mathcal{C} = \{c^{(k)}\}_{k=1}^{K}$, and an LLM judge provides criterion-level feedback for each y_i. For an ordered comparison between two sampled rollouts y_i and y_j, where y_j serves as the reference, the judge scores y_i and returns a criterion-level score vector $\boldsymbol{s}_{i,j} = J(x, y_i, y_j; \mathcal{C}) \in [0, S_{\max}]^K$. Its k-th entry $s_{i,j}^{(k)}$ denotes the score assigned to y_i on criterion $c^{(k)}$, and $S_{\max}$ is usually predefined, *e.g.,* 10.

Reward Aggregation. To turn criterion-level feedback into a scalar training signal, existing methods mainly rely on a fixed scalarizer, implemented as *pre-defined or uniform weights* over criteria. Let $\boldsymbol{w}_{\text{base}} = \left(w_{\text{base}}^{(1)}, \ldots, w_{\text{base}}^{(K)}\right)$ denote the static weight vector, and then the judgment is computed as

$$S_{i,j}(\boldsymbol{w}) = \langle \boldsymbol{w}, \boldsymbol{s}_{i,j} \rangle = \sum_{k=1}^{K} w^{(k)} s_{i,j}^{(k)}, \qquad \Delta_{i,j}(\boldsymbol{w}) = S_{i,j}(\boldsymbol{w}) - S_{j,i}(\boldsymbol{w}). \tag{1}$$

The resulting reward is then constructed by aggregating the scores credited to the corresponding pairwise comparisons within the sampled group [Zhang et al., 2026a], which is formulated as

$$R_i(\boldsymbol{w}) = \sum_{j \neq i} \Phi_\tau(\Delta_{i,j}(\boldsymbol{w})), \qquad \Phi_\tau(\Delta) = \begin{cases} 2\,\mathrm{sgn}(\Delta), & |\Delta| \geq \tau, \\ \mathrm{sgn}(\Delta), & 0 < |\Delta| < \tau, \\ 0, & \Delta = 0. \end{cases} \tag{2}$$

where $\mathrm{sgn}(\cdot)$ is the sign function and $\tau > 0$ denotes the strong-preference threshold. $\Phi_\tau(\Delta)$ acts as a scoring function, converting each pairwise margin $\Delta_{i,j}(\boldsymbol{w})$ into a discrete win/loss outcome. Following GRPO [Shao et al., 2024], the scalar reward R_i of each rollout is converted into a group-relative advantage $\hat{A}_i$, and updates the policy using a following regularized clipped objective

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \left(\min\left(r_{i,t}(\theta)\hat{A}_i,\ \mathrm{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right) \right]. \tag{3}$$

Here, $r_{i,t}(\theta)$ is the token-level importance ratio against the pre-update policy and D_{KL} computes the KL penalty. Additional details of the policy optimization are provided in Appendix B. Under this framework, policy updates are ultimately driven by the scalar rewards assigned to sampled rollouts.

3 Method

As analyzed in the introduction, directly optimizing the reward acquired by static scalarization will induce the imbalanced reward polarization among different rubric criteria. To address this challenge, we introduce a novel *Focal Reward* method to dynamically calibrate the emphasis in reinforcement learning, which we detail in this part, followed by a theoretical analysis.

3.1 Focal Reward

Our idea is inspired by focal loss [Lin et al., 2017], where class-wise weights are dynamically adjusted to emphasize overlooked categories. But in technical design, our case is not as straightforward as the focal loss. For clarity, we first use inverse reward projection to estimate the saturation degree of each criterion in the rubric, and then use the saturation degree to dynamically rebalance the static weight, which ultimately affects the reward direction during reinforcement learning.

Inverse reward projection. For each sampled rollout group, we first cast the reward computed by Eq. (1) and Eq. (2) into the Gibbs weights with a proper temperature $T > 0$, formulated as follows:

$$\boldsymbol{r} = (r_1, \ldots, r_G) = \left(\frac{\exp(R_1(\boldsymbol{w}_{\text{base}})/T)}{\sum_{i'=1}^{G} \exp(R_{i'}(\boldsymbol{w}_{\text{base}})/T)}, \ldots, \frac{\exp(R_G(\boldsymbol{w}_{\text{base}})/T)}{\sum_{i'=1}^{G} \exp(R_{i'}(\boldsymbol{w}_{\text{base}})/T)} \right). \tag{4}$$

Then, for each rollout, we have a reward-based weight r_i corresponding to it. As all rollouts have been evaluated under K criteria, we can inversely calculate the *average* score of each criterion for these rollouts with $s_{i,j}^{(k)}$, and further estimate the *effective* saturation degree of each criterion with $\boldsymbol{r}$:

$$\bar{\boldsymbol{s}}^{(\boldsymbol{k})} = \left(\bar{s}_1^{(k)}, \ldots, \bar{s}_G^{(k)}\right) = \left(\frac{1}{G-1} \sum_{j \neq 1} s_{1,j}^{(k)}, \ldots, \frac{1}{G-1} \sum_{j \neq G} s_{G,j}^{(k)} \right), \ P^{(k)} = \frac{\langle \boldsymbol{r}, \bar{\boldsymbol{s}}^{(\boldsymbol{k})} \rangle}{S_{\max}}. \tag{5}$$

Since $\bar{s}_i^{(k)} \in [0, S_{\max}]$ and $\sum_i r_i = 1$, this ensures $P^{(k)} \in [0, 1]$. Based on the above equations, we inversely project the reward back to each criterion to measure the criterion-level saturation under this

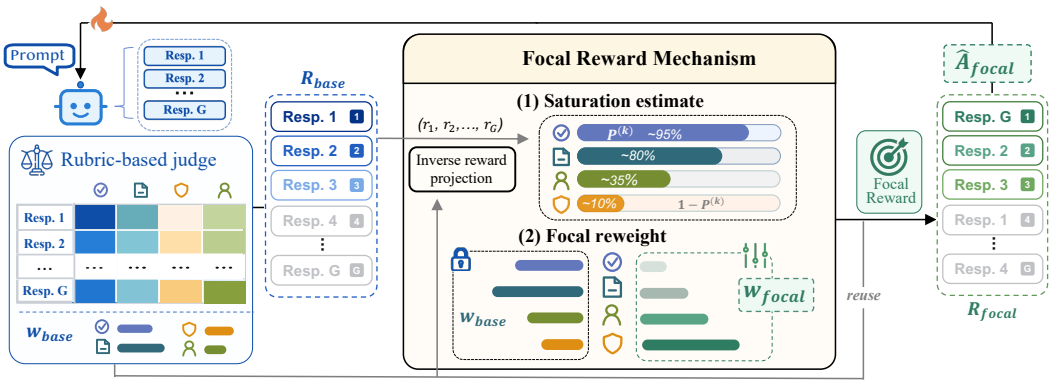


Figure 2: Overview of *Focal Reward*: it estimates criterion saturation from the base reward, reweights the base scalarizer toward high-headroom criteria, and synthesizes the focal reward for policy optimization. The resulting reward may reorder rollouts and assign higher credit to responses that better satisfy under-saturated criteria.

group of rollouts. Two points are worth noting. First, the saturation degree of each criterion in the rubric is overall positively correlated with the reward. Namely, rollout groups that receive higher rewards tend to exhibit higher criterion saturation. Second, the saturation degree of each criterion may not keep the same pace along with the reward improvement. That is to say, as the RL training progresses and reward improves, different criteria can be satisfied to varying degrees and even become very imbalanced. Such a challenge can be explicitly captured by our quantity $\boldsymbol{P} = (P^{(1)}, \ldots, P^{(K)})$.

Focal reweighting aggregation. With the quantity $\boldsymbol{P}$ to characterize the saturation of all criteria, Focal Reward modulates the base rubric static weights $\boldsymbol{w}_{\text{base}}$ with a headroom-dependent factor $\mathbf{1} - \boldsymbol{P}$. Specifically, we design the focal weights for all criteria in the rubric as follows

$$\boldsymbol{w}_{\text{focal}} = \frac{\langle \boldsymbol{w}_{\text{base}}, \mathbf{1} \rangle}{\langle \tilde{\boldsymbol{w}}, \mathbf{1} \rangle} \tilde{\boldsymbol{w}} \quad \text{where} \quad \tilde{\boldsymbol{w}} = (\mathbf{1} - \boldsymbol{P} + \epsilon)^{\gamma} \odot \boldsymbol{w}_{\text{base}}. \tag{6}$$

Here, $\epsilon > 0$ is a smoothing constant and $\gamma > 0$ controls the strength of reallocation toward less saturated criteria. Since $\mathbf{1} - \boldsymbol{P}$ estimates remaining frontier headroom, criteria that remain weak on the current strong-rollout frontier receive larger multiplicative factors. The normalization preserves the total weight mass of the base scalarizer, so focal reweighting changes relative criterion emphasis without trivially changing the overall reward scale. With $\boldsymbol{w}_{\text{focal}}$, we form the following final reward for reinforcement learning, which is then converted to the advantage and used further in Eq. (3)

$$R_i(\boldsymbol{w}_{\text{focal}}) = \sum_{j \neq i} \Phi_\tau(\Delta_{i,j}(\boldsymbol{w}_{\text{focal}})) \,. \tag{7}$$

Here, the margins $\Delta_{i,j}(\boldsymbol{w}_{\text{focal}})$ are similarly computed based on Eq. (1) with the same criterion-level pairwise scores $s_{i,j}^{(k)}$ obtained in the base pass. Thus, considering the computation in above equations, the additional cost of Focal Reward is negligible. In Fig. 2, we illustrate its whole procedure.

3.2 Theoretical Analysis

In this part, we provide theoretical analysis about transforming the pairwise comparison scores into a scalar reward when the criteria retain different saturation degrees. Due to the space limitation, all assumptions, derivations, and boundary conditions are deferred to Appendix C.

Theorem 1 (Local misallocation bound). *Given a training state at timestep t, let $\boldsymbol{a} \in \mathbb{R}_+^K$ be any nonzero criterion scalarizer, $\boldsymbol{D}_t$ denote any criterion-level pairwise score-difference vector, $\boldsymbol{a}^\top \boldsymbol{D}_t$ be the scalarized pairwise margin, and $L_t \in \{+1, -1\}$ be the latent locally useful update direction. Under the local latent-label assumptions in Appendix C, we have local misallocation bound*

$$\Pr\big(L_t \boldsymbol{a}^\top \boldsymbol{D}_t \leq 0\big) \leq \exp\left(-\frac{1}{2}\Xi_t(\boldsymbol{a})\right),$$

where $\Xi_t(\boldsymbol{a}) = \frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{\max_{\ell \in \{+1,-1\}} \boldsymbol{a}^\top \boldsymbol{\Sigma}_t(\ell) \boldsymbol{a}}$ and $\boldsymbol{\eta}_t$ denotes the criterion-level learning signal, ℓ is the latent local label, and $\boldsymbol{\Sigma}_t(\ell)$ captures residual criterion-level fluctuation under label ℓ.

Theorem 1 shows that maximizing $\Xi_t(\boldsymbol{a})$ will reduce local pairwise credit misallocation, where $\Xi_t(\boldsymbol{a})$ can be interpreted as a local signal-to-noise ratio: the numerator favors scalarizers aligned with useful criterion-wise learning signal, while the denominator penalizes scalarizers whose margins are dominated by residual criterion-level variation.

Theorem 2 (**Static aggregation gap**). *Given $\Xi_t(\boldsymbol{a})$ in Theorem 1, let $\boldsymbol{a}_t^\star$ denote the locally preferred scalarizer that maximizes $\Xi_t(\boldsymbol{a})$, and $H_t^{(k)} \in [0,1]$ denote the remaining frontier headroom of criterion $c^{(k)}$ at the timestep t. Under the symmetric and isotropic simplifications and the headroom-shaped edge model in Appendix C, the preferred criterion scalarizer has the following closed-form direction, and the fixed base scalarizer $\boldsymbol{w}_{\text{base}}$ incurs the corresponding local surrogate gap:*

$$[\boldsymbol{a}_t^\star]^{(k)} \propto \boldsymbol{w}_{\text{base}}^{(k)} \big(H_t^{(k)}\big)^{\gamma_0}, \quad \Xi_t(\boldsymbol{a}_t^\star) - \Xi_t(\boldsymbol{w}_{\text{base}}) = \frac{c_t^2 \|\boldsymbol{w}_{\text{base}}\|_2^2}{\sigma_t^2} \operatorname{Var}_\mu(H_t^{\gamma_0}),$$

where $\gamma_0 > 0$ is a curvature parameter, $\mu^{(k)} = \frac{\big(\boldsymbol{w}_{\text{base}}^{(k)}\big)^2}{\|\boldsymbol{w}_{\text{base}}\|_2^2}$, $\operatorname{Var}_\mu(\cdot)$ denotes weighted variance under μ, $c_t > 0$ is the state-dependent scale, and σ_t^2 is the isotropic residual variance scale.

Theorem 2 shows that homogeneous base-weighted headroom makes the variance term vanish, so $\boldsymbol{w}_{\text{base}}$ remains locally aligned with the preferred direction. When headroom is heterogeneous, the positive variance term indicates that $\boldsymbol{w}_{\text{base}}$ misses the frontier-headroom factor. This corresponds to our design: Eq. (6) uses $1 - P^{(k)}$ as a proxy for this latent factor, while Appendix C.4 provides the proxy bridge by showing that $P^{(k)}$ estimates frontier-level saturation under Gibbs concentration.

4 Experiments

4.1 Experimental Setup

Backbone models and Baselines. We employ diverse LLMs of different scales in our experiments, including Qwen2.5 series (7B-Instruct) [Yang et al., 2024a] and Qwen3 series (Qwen3-8B/30B-A3B) [Yang et al., 2025]. Beyond the vanilla model, we compare Focal Reward against several static scalarization baselines, including two rule-based baselines: **Veto** [Huang et al., 2025], using hard-rule-gated aggregation and **Min-Score** [Park et al., 2024a], using bottleneck-style weights with minimum criterion score, and two preset baselines: **Static (uniform)** [Cui et al., 2024], using uniform weights and **Static (prior)** [Viswanathan et al., 2025, Gunjal et al., 2026], using GPT-4o-predicted weights. We apply **Focal Reward** to both preset weights, obtaining the **uniform** and **prior** versions.

Implementation Details. We implement our methods based on the verl framework [Sheng et al., 2025] and the GSPO algorithm [Zheng et al., 2025], conducted on 32×H800-80GB GPUs. For main experiments, we use two 10K-example OpenRubrics subsets [Liu et al., 2026], from general and science domains separately. During RL training, we use a global batch size of 64 and adopt AdamW with a learning rate of 1×10^{-6}. We set the number of rollouts $G = 8$, hyper-parameters $\gamma = 2$ and $T = 10$. During evaluation, general-domain models are evaluated on AlpacaEval 2.0 [Dubois et al., 2024], Arena Hard [Li et al., 2025], WritingBench [Wu et al., 2025], and EQ-Bench 3 [Paech, 2023]; science-domain models are evaluated on GPQA Diamond [Rein et al., 2024] and HealthBench [Arora et al., 2025]. See Appendix D for more training details and benchmark descriptions.

4.2 Main results

Performance on benchmarks of general and science domains. Table 1 reports the results on both general-domain and science-oriented benchmarks across three backbone models. Focal Reward delivers consistent improvements over static scalarization across all model scales. Compared with the strongest static baseline, it improves the average score by up to +2.19 points on Qwen2.5-7B-Instruct, +1.38 points on Qwen3-8B, and +2.50 points on Qwen3-30B-A3B. The gains are observed on both general and science-oriented benchmarks, indicating that Focal Reward synthesis provides a broadly effective training signal rather than benefiting only a particular domain or backbone. In addition, the two Focal Reward variants based on uniform and prior weights perform closely to each other while both surpassing existing reward-synthesis baselines, showing that Focal Reward is not strongly dependent on the specific choice of base weights and can robustly improve different priors.

Table 1: **Main results (%) on general-domain and science-oriented benchmarks across different backbones.** The first row of each backbone block reports the vanilla model. AE2, AH, WB, EQB3, GPQA, and HB denote AlpacaEval 2.0, Arena-Hard, WritingBench, EQ-Bench 3, GPQA Diamond, and HealthBench, respectively. Bold and underlined scores mark the best and second-best results within each backbone–benchmark block.

	General-domain				Sci-oriented		Avg.
Model	AE2	AH	WB	EQB3	GPQA	HB	
Qwen2.5-7B-Instruct	24.50	39.91	47.16	35.31	33.74	22.29	33.82
+ Veto	35.84	40.70	50.92	40.42	34.96	38.27	40.19
+ Min-Score	40.31	37.18	50.81	39.16	35.90	40.18	40.59
+ Static (uniform)	41.22	48.06	53.52	41.78	36.05	40.31	43.49
+ Static (prior)	40.86	47.48	52.71	41.24	36.55	40.65	43.25
+ Focal Reward (uniform)	**44.06**	**51.93**	**54.76**	**43.72**	37.44	**42.16**	**45.68**
+ Focal Reward (prior)	<u>43.75</u>	<u>50.83</u>	<u>53.89</u>	<u>43.53</u>	**37.63**	<u>41.89</u>	<u>45.25</u>
Qwen3-8B	53.86	49.90	63.51	47.70	48.33	36.57	49.98
+ Veto	59.41	52.47	65.41	50.49	50.23	44.20	53.70
+ Min-Score	56.83	56.24	64.97	49.68	50.08	45.93	53.96
+ Static (uniform)	69.94	60.12	67.61	52.37	51.06	46.98	58.01
+ Static (prior)	71.81	60.23	68.04	53.07	51.31	47.37	58.64
+ Focal Reward (uniform)	**74.15**	59.83	**70.50**	**54.69**	52.15	**48.81**	**60.02**
+ Focal Reward (prior)	<u>73.96</u>	**60.47**	<u>70.24</u>	<u>54.38</u>	**52.21**	<u>48.50</u>	<u>59.96</u>
Qwen3-30B-A3B	63.36	55.26	64.76	51.80	51.10	40.26	54.42
+ Veto	57.49	74.80	69.79	56.42	55.43	54.38	61.38
+ Min-Score	61.12	72.19	70.31	55.20	54.47	54.53	61.30
+ Static (uniform)	73.61	79.16	71.77	57.26	55.74	55.29	65.47
+ Static (prior)	75.63	80.21	72.05	58.13	56.22	55.95	66.36
+ Focal Reward (uniform)	76.87	83.99	**73.92**	58.98	57.04	56.67	67.91
+ Focal Reward (prior)	**77.59**	**88.69**	<u>73.38</u>	**59.03**	**57.07**	**57.39**	**68.86**

Performance on criteria of different difficulty levels. Table 2 further analyzes the trained models at the rubric level, comparing Static and Focal Reward across criteria with different difficulty levels. Following Fig. 3(a), we first use the base model to perform inference on the training prompts and bucket criteria according to their initial score distribution: criteria with scores below 5 are treated as hard criteria, those above 8.5 as easy criteria, and the remaining ones as medium criteria. Compared with Static scalarization, Focal Reward consistently achieves higher average scores and pass rates on hard criteria, showing that it effectively emphasizes criteria that are initially under-satisfied. The gains also extend to medium criteria, indicating that the adaptive reweighting also improves a broader range of rubric aspects. On easy criteria, Focal Reward performs comparably to Static while still maintaining a clear improvement over the base model, indicating that hard-criterion gains do not come at the expense of already-satisfied dimensions. Moreover, Focal Reward produces noticeably shorter rollouts than Static across backbones and domains, suggesting more efficient correction of under-satisfied criteria rather than merely relying on longer, polished generation to gain reward.

4.3 Further analysis

While Tables 1 and 2 demonstrate the superior performance of Focal Reward, we further analyze the mechanisms underlying these gains through answering the following research questions: 1) How does Focal Reward reallocate weights across rubric criteria? 2) How does this reallocation reshape the resulting reward distribution? 3) Does the model exhibit measurable improvements on the targeted rubric criteria? 4) How sensitive is Focal Reward to the choice of prior weights?

Weight aspect: More emphasis on under-satisfied criterion. Figure 3(b) compares the relationship between criterion weights and estimated remaining headroom under Static prior and Focal Reward. Higher y-axis values indicate more under-satisfied criteria, while higher x-axis values indicate larger weights. Under Static prior, many high-headroom criteria still receive small weights, and some low-headroom criteria remain heavily emphasized, showing that fixed weights fail to track criterion saturation. In contrast, Focal Reward shifts more weight mass toward high-headroom criteria, showing that the adaptive weights can refocus training on criteria with substantial room for improvement.

Reward aspect: Better discrimination ability from high-headroom criterion. Figure 4(a) analyzes how rollout-level reward outcomes change after focal reweighting. Focal Reward largely preserves the behavior of Static prior, with around 70% of samples keeping the same outcome (remaining on

Table 2: **Rubric level results across criteria of different difficulty levels.** Criteria are bucketed by Base model scores following Fig. 3(a). Each block includes the Base model inference and results of Static and Focal after training. Pass rate measures the percentage of rollouts whose scores exceed the corresponding threshold.

Domain	Model	Method	Len.	Mean score				Pass rate	
				All	Hard	Med.	Easy	Hard	Easy
General	Qwen2.5-7B	Base	463	8.21	3.24	7.65	9.37	-	-
		Static	1351	8.48	5.24	7.92	9.51	39.2	88.6
		Focal	1083	8.52	5.71	7.93	9.50	47.4	88.2
	Qwen3-8B	Base	658	8.33	2.87	7.67	9.31	-	-
		Static	2446	8.60	4.82	7.99	9.45	35.9	87.9
		Focal	1806	8.64	5.18	8.02	9.44	42.6	87.6
	Qwen3-30B-A3B	Base	681	8.42	2.74	7.69	9.32	-	-
		Static	3473	8.69	4.27	7.98	9.46	29.4	88.1
		Focal	2348	8.70	4.56	7.94	9.45	34.5	87.8
Science	Qwen2.5-7B	Base	647	8.11	3.68	7.59	9.34	-	-
		Static	3439	8.51	6.22	8.11	9.48	39.0	89.3
		Focal	2836	8.55	6.63	8.13	9.49	49.4	90.1
	Qwen3-8B	Base	1055	8.41	2.58	7.69	9.25	-	-
		Static	4992	8.70	4.92	8.11	9.39	41.6	88.4
		Focal	4131	8.74	5.34	8.15	9.40	49.2	89.0
	Qwen3-30B-A3B	Base	1020	8.67	2.47	7.76	9.28	-	-
		Static	6120	8.84	4.85	7.89	9.42	38.4	88.7
		Focal	5000	8.88	5.18	7.95	9.43	47.5	89.2

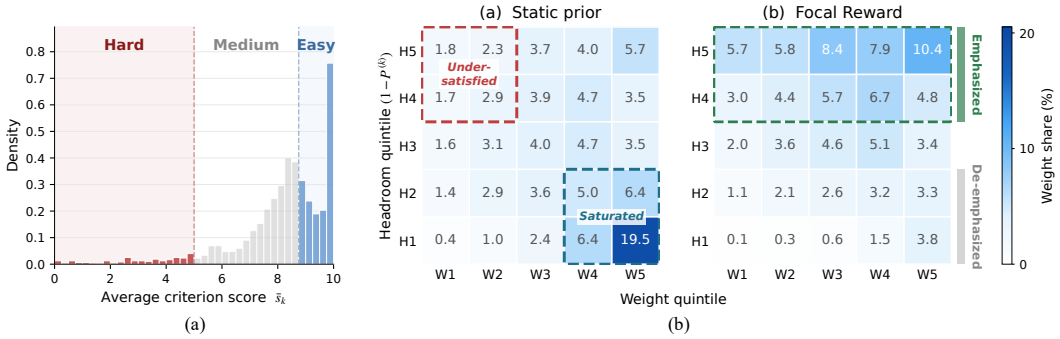
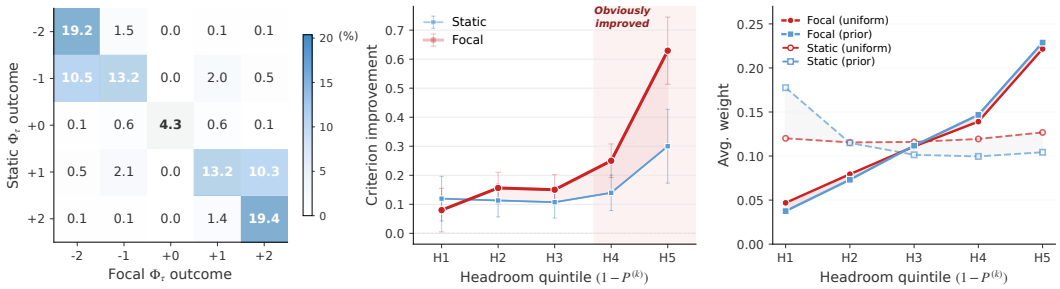


Figure 3: **Criterion saturation and weight allocation.** (a) Distribution of average criterion scores $\bar{s}^{(k)}$, based on which criteria are divided into hard, medium and easy groups. (b) Relationship between criterion weights and remaining headroom. Criteria are grouped into quintiles according to their assigned weights and estimated headroom, each heatmap reports the corresponding weight share.

the diagonal), indicating a similar overall reward style. Meanwhile, the adjusted rewards become more polarized: roughly 10% of samples move from $+1$ to $+2$, and another 10% from -1 to -2. This suggests that by assigning greater emphasis to under-satisfied criteria, Focal Reward makes these criterion differences more visible and improves the discrimination ability of the reward signal.

Optimization aspect: Obvious improvement on targeted criteria. Figure 4(b) validates whether the criteria emphasized by Focal Reward are indeed improved during training. We divide criteria into quintiles according to their estimated remaining headroom $1 - P^{(k)}$ and measure the improvement of each criterion as the score gain over the base model. Static scalarization brings only mild improvements on high-headroom criteria, suggesting that fixed weights are insufficient to fully exploit these under-satisfied criteria. In contrast, Focal Reward shows a clear monotonic trend: criteria with larger remaining headroom obtain larger improvements, with the highest-headroom group exhibiting the strongest gain. This confirms that the proposed reweighting mechanism successfully converts headroom-aware reward correction into targeted rubric-level optimization.

Prior aspect: Robustness across various base scalarizers. Figure 4(c) compares how different base scalarizers evolve after focal reweighting. The base weights encode the judge's prior emphasis over criteria and can therefore influence both rollout rewards and downstream performance. As depicted, Uniform and Prior initialization exhibit clearly different weight distributions across headroom levels.



(a) (b) (c)

Figure 4: **Mechanism analysis of Focal Reward.** (a) Transition heatmap of rollout-level reward outcomes under Static and Focal Reward. Each cell reports the percentage of samples moving from Static to Focal outcomes. (b) Improvement on criteria of different estimated remaining headroom, which is measured by the score gain over the untrained base model. (c) Average weight across headroom quintiles for Uniform and Prior initialization before and after focal reweighting. Experiments are conducted on Qwen2.5-7B-Instruct under general domain.

Table 3: **Component ablation of Focal Reward** on Qwen2.5-7B-Instruct under general domain. *Focal w/o frontier weighting* ablates the frontier-aware saturation by averaging uniformly over rollouts. *Focal w/ frozen scalarizer* ablates online adaptation by keeping the initially computed focal weights fixed throughout training.

Method	AE2	AH	WB	EQB3	Avg.
Static	41.22	48.06	53.52	41.78	46.15
Focal Reward	**44.06**	**51.93**	**54.76**	**43.72**	**48.62**
Focal w/o frontier weighting	41.75	48.73	53.96	42.35	46.70
Focal w/ frozen scalarizer	40.97	47.67	53.63	41.89	46.04

Table 4: **Runtime overhead of Focal Reward.** Results are averaged over multiple training runs with Qwen2.5-7B-Instruct backbone on the general-domain dataset. *Len.* denotes the average rollout length. *Synth.* and *Rollout* denote the reward-synthesis and rollout generating time separately. And *Total Step* denotes the end-to-end training step time.

Method	Len.	Synth.	Rollout	Total Step
Static	1351	14 ms	225.8 s	282.0 s
Focal Reward	1083	50 ms	153.9 s	175.4 s
Abs. Δ	−268	+36 ms	−71.9 s	−106.6 s

Interestingly, despite starting from these distinct priors, Focal Reward drives the weights toward highly similar distributions, with a cosine similarity of about 0.9. Both variants consistently allocate larger weights to high-headroom criteria, showing the robustness of Focal Reward to base weights.

4.4 Ablation studies

Component ablation of Focal Reward. Table 3 studies the contribution of the key components in Focal Reward. Removing frontier weighting drops the average score from 48.62 to 46.70, showing that saturation should be estimated from the strong-rollout frontier rather than uniformly over all rollouts. Freezing the scalarizer further reduces the score to 46.04, indicating that a fixed reward direction cannot track the changing saturation states during training. These results suggest that both the frontier-aware estimation and the online reweighting components are essential for effective reward reallocation.

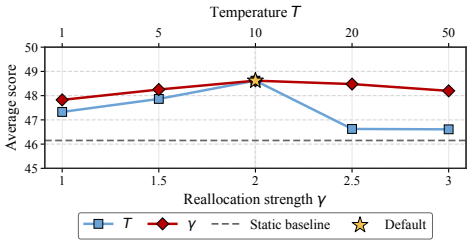


Figure 5: **Sensitivity analysis of hyper-parameters γ and T** on Qwen2.5-7B-Instruct in general-domain. The star marks the default setting.

Computational cost analysis. Table 4 shows that Focal Reward remains computationally efficient compared with the Static baseline. Since it reuses the same rubric-level judgment and only applies lightweight reweighting during reward synthesis, it introduces negligible computational overhead. Moreover, the model optimized with Focal Reward produces shorter rollouts while maintaining overall quality, which further reduces generation cost and leads to lower end-to-end training overhead.

Hyper-parameter sensitivity analysis. Figure 5 analyzes the sensitivity of Focal Reward to the reallocation strength γ and temperature T. Across a wide range of values, Focal Reward consistently outperforms the Static baseline, indicating that its gains are not tied to a narrow hyper-parameter choice. For γ, the performance remains stable around the default setting. For T, the performance is also stable within the range of 1 to 10, while an overly large temperature may over-smooth the reward-based rollout weights and weaken the saturation estimate, leading to a slight performance drop. Overall, Focal Reward shows robustness under different hyper-parameter configurations.

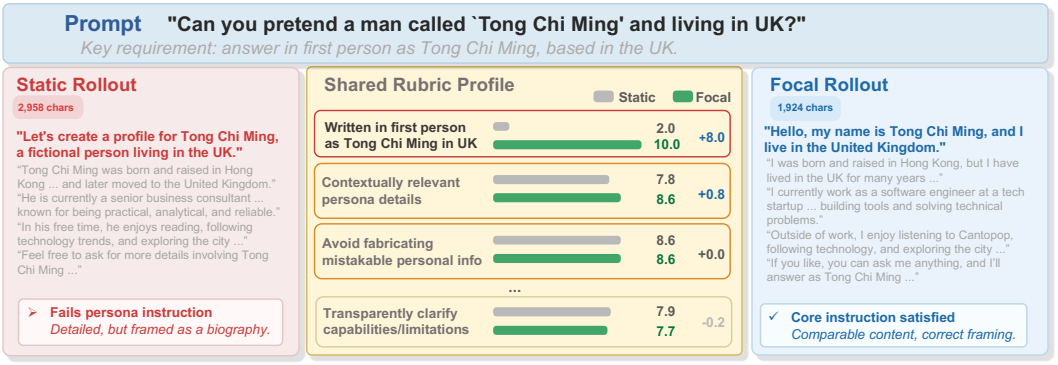


Figure 6: **Qualitative case study.** Static misses the first-person role-play requirement despite detailed content, while Focal preserves rollout quality and corrects this weakness by emphasizing the under-satisfied hard rule.

Case study. Figure 6 provides a qualitative example of the failure mode discussed in Figure 1(c). The prompt requires first-person role play, yet Static ranks a fluent and detailed rollout highly because it performs well on broad quality criteria while missing the decisive task-specific requirement, namely the correct first-person framing. In contrast, Focal Reward assigns greater emphasis to this under-satisfied criterion and favors a rollout that follows the intended role-play format. This example qualitatively illustrates the central mechanism of Focal Reward: reallocating reward emphasis toward criteria with substantial remaining headroom can correct high-reward but behaviorally flawed generations. More examples and full rollouts are provided in Appendix G.

5 Related Work

Rubric-based RLAIF. Open-ended generation is subjective and multi-dimensional, so modern alignment often uses rubric-based AI judges to provide structured, criterion-level feedback rather than holistic or opaque preference signals [Lee et al., 2024, Gunjal et al., 2026, Viswanathan et al., 2025]. Prior work has explored rubric-based rewards, checklists, scalable rubric generation, context-aware criteria, adaptive pairwise rubrics, and adaptive rubric learning [Huang et al., 2025, Liu et al., 2026, Gupta et al., 2025, Jia et al., 2026, Xu et al., 2026b]. These studies mainly improve how rubrics are constructed, adapted, or judged. We study the reward-synthesis step under a fixed rubric and judge, asking how criterion-level feedback should be aggregated into scalar rewards for policy optimization.

Dynamic Reward Weighting. Reward aggregation in rubric-based reinforcement learning can be viewed as a scalarization problem, where criterion-level feedback is compressed into a scalar reward for policy optimization. Prior work has studied scalarization and preference trade-offs in broader multi-objective alignment [Wang et al., 2024a, Zhou et al., 2024, He and Maghsudi, 2025]. Recent methods further adapt weights according to context or prompt-specific preferences [Yang et al., 2024b, Liu et al., 2025], or according to optimization states, learning potential, and objective conflicts [Lu et al., 2026, Chen et al., 2026a]. These methods mainly adapt weights across high-level objectives or preference dimensions, leaving the fine-grained criterion-level aggregation problem under a fixed rubric and fixed judge less explored. More related work is discussed in Appendix H.

6 Conclusion

In this work, we introduce Focal Reward, a novel objective for automatically balancing reinforcement learning under rubric-based rewards. It addresses imbalanced reward polarization caused by static scalarizers, which may continue to emphasize already-satisfied criteria while under-emphasizing criteria that still retain substantial room for improvement. Focal Reward estimates criterion-level saturation through inverse reward projection and adaptively reallocates reward emphasis from saturated criteria toward high-headroom criteria, thereby calibrating the reward direction before policy optimization. Extensive experiments and analyses demonstrate its effectiveness, showing improved reward discrimination and targeted rubric-level gains. These findings highlight the importance of adaptive reward synthesis for effective policy optimization in open-ended generation.

Acknowledgements and Disclosures

This work is supported by National Natural Science Foundation of China (No. 62676245), Ant Group and STCSM (No. 22DZ2229005), 111 plan (No. BP0719010).

The authors declare no competing interests.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health. *CoRR*, 2025. URL `https://arxiv.org/abs/2505.08775`.

Dongliang Chen, Xinlin Zhuang, Junjie Xu, Luojian Xie, Zehui Wang, Jiaxi Zhuang, Haolin Yang, Liang Dou, Xiao He, Xingjiao Wu, and Ying Qian. APEX: Learning adaptive priorities for multi-objective alignment in vision-language generation. *CoRR*, 2026a. URL `https://arxiv.org/abs/2601.06574`.

Peter Chen, Xiaopeng Li, Xi Chen, and Tianyi Lin. Reward-free alignment for conflicting objectives. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 15091–15117. PMLR, 2026b.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: boosting language models with scaled AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. In *Proceedings of the Conference on Language Modeling (COLM 2024)*, 2024.

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

Taneesh Gupta, Shivam Shandilya, Xuchao Zhang, Rahul Madhavan, Supriyo Ghosh, Chetan Bansal, Huaxiu Yao, and Saravan Rajmohan. CARMO: Dynamic criteria generation for context-aware reward modelling. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2202–2261. Association for Computational Linguistics, 2025.

Qiang He and Setareh Maghsudi. Pareto multi-objective alignment for language models. In *Machine Learning and Knowledge Discovery in Databases. Research Track. European Conference, ECML PKDD 2025*, 2025.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, Shengjie Bi, Shishir G Patil, Qi Qi, Shengyu Feng, Julian Katz-Samuels, Richard Yuanzhe Pang, Sujan Kumar Gonugondla, Hunter Lang, Yue Yu, Yundi Qian, Maryam Fazel-Zarandi, Licheng Yu, Amine Benhalloum, Hany Hassan Awadalla, and Manaal Faruqui. AdvancedIF: Rubric-based benchmarking and reinforcement learning for advancing LLM instruction following. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18003–18022. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.820.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. Are LLM-based evaluators confusing NLG quality criteria? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9530–9570, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.516.

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

Ruipeng Jia, Yunyi Yang, Yongbo Gai, Kai Luo, Shihao Huang, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Writing-Zero: Bridge the gap between non-verifiable tasks and verifiable rewards. *arXiv preprint arXiv:2506.00103*, 2025. URL https://arxiv.org/abs/2506.00103.

Ruipeng Jia, Yunyi Yang, Yuxin Wu, Yongbo Gai, Siyuan Tao, Mengyu Zhou, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Open rubric system: Scaling reinforcement learning with pairwise adaptive rubric. *CoRR*, 2026. URL https://arxiv.org/abs/2602.14069.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, Jie Tang, and Minlie Huang. CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.704.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 26874–26901. PMLR, 2024.

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 Bench-Builder pipeline. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 34209–34231. PMLR, 2025.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, October 2017.

Biao Liu, Ning Xu, Junming Yang, and Xin Geng. Preference orchestrator: Prompt-aware multi-objective alignment for large language models. *CoRR*, 2025. URL https://arxiv.org/abs/2511.10656.

Jiawei Liu and Lingming Zhang. Code-R1: Reproducing R1 for code with reliable rewards. https://github.com/ganler/code-r1, 2025. URL https://github.com/ganler/code-r1.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and LLM alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17417–17437. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.791.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-EVAL: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Yining Lu, Zilong Wang, Shiyang Li, Xin Liu, Changlong Yu, Qingyu Yin, Zhan Shi, Zixuan Zhang, and Meng Jiang. Learning to optimize multi-objective alignment through dynamic reward weighting. *Transactions of the Association for Computational Linguistics*, 14:1051–1073, 2026. doi: 10.1162/TACL.A.696.

Subhojyoti Mukherjee, Anusha Lalitha, Sailik Sengupta, Aniket Deshmukh, and Branislav Kveton. Multi-objective alignment of large language models through hypervolume maximization. *CoRR*, 2024. URL https://arxiv.org/abs/2412.05469.

Samuel J. Paech. EQ-Bench: An emotional intelligence benchmark for large language models. *CoRR*, 2023. URL https://arxiv.org/abs/2312.06281.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024. doi: 10.52202/079017-2197.

Giseung Park, Woohyeon Byeon, Seongmin Kim, Elad Havakuk, Amir Leshem, and Youngchul Sung. The max-min formulation of multi-objective reinforcement learning: From theory to a model-free algorithm. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38809–38834. PMLR, 2024a.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. OffsetBias: Leveraging debiased data for tuning evaluators. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1043–1067, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.57.

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. In *Proceedings of the Conference on Language Modeling (COLM 2024)*, 2024.

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

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

Yiran Jenny Shen, Yu Xia, Jonathan Daniel Chang, and Prithviraj Ammanabrolu. Simultaneous multi-objective alignment across verifiable and non-verifiable rewards. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 111306–111337. PMLR, 2026b.

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*, Rotterdam, The Netherlands, March 2025. doi: 10.1145/3689031.3696075.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India, 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. doi: 10.18653/v1/2025.ijcnlp-long.18.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405.01724.

Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding RL with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025. URL https://arxiv.org/abs/2503.23829.

Jiwei Tang, Jin Xu, Tingwei Lu, Zhicheng Zhang, Yiming Zhao, Lin Hai, and Hai-Tao Zheng. Perception Compressor: A training-free prompt compression framework in long context scenarios. *arXiv preprint arXiv:2409.19272*, 2025. URL https://arxiv.org/abs/2409.19272.

Jiwei Tang, Zhijing Huang, Xinyu Zhang, Chen Jason Zhang, Jianxing Yu, Libin Zheng, Rui Meng, and Jian Yin. Beyond position bias: Shifting context compression from position-driven to semantic-driven. *arXiv preprint arXiv:2605.09463*, 2026a. URL https://arxiv.org/abs/2605.09463.

Jiwei Tang, Shilei Liu, Zhicheng Zhang, Qingsong Lv, Runsong Zhao, Tingwei Lu, Langming Liu, Haibin Chen, Yujin Yuan, Hai-Tao Zheng, Wenbo Su, and Bo Zheng. Read as human: Compressing context via parallelizable close reading and skimming. *arXiv preprint arXiv:2602.01840*, 2026b. URL https://arxiv.org/abs/2602.01840.

Jiwei Tang, Shilei Liu, Zhicheng Zhang, Yujin Yuan, Libin Zheng, Wenbo Su, and Bo Zheng. COMI: Coarse-to-fine context compression via marginal information gain. *arXiv preprint arXiv:2602.01719*, 2026c. URL https://arxiv.org/abs/2602.01719.

Jiwei Tang, Zhicheng Zhang, Shunlong Wu, Jingheng Ye, Lichen Bai, Zitai Wang, Tingwei Lu, Lin Hai, Yiming Zhao, Hai-Tao Zheng, and Hong-Gee Kim. GMSA: Enhancing context compression via group merging and layer semantic alignment. *arXiv preprint arXiv:2505.12215*, 2026d. URL https://arxiv.org/abs/2505.12215.

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3824.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. ABKD: Pursuing a proper allocation of the probability mass in knowledge distillation via α-β-divergence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 65167–65212. PMLR, 2025.

Guanghui Wang, Kaiwen Lv Kacuila, Zhiyong Yang, Zitai Wang, Jin-Wen Wu, Longtao Huang, Qianqian Xu, and Qingming Huang. The bridge-garden dilemma in LLM distillation: Why mixing hard and soft labels works. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 127315–127354. PMLR, 2026a.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511.

Pengkai Wang, Pengwei Liu, Qi Zuo, Zhijie Sang, Congkai Xie, and Hongxia Yang. InfiMed-ORBIT: Aligning LLMs on open-ended complex tasks via rubric-based incremental training. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 128637–128672. PMLR, 2026b.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*, 2024c.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Mian Wu, Gavin Zhang, Sewon Min, Sergey Levine, and Aviral Kumar. RLAC: Reinforcement learning with adversarial critic for free-form generation tasks. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1737.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, Zhaoyang Liu, and Bolin Ding. Auto-rubric: Learning from implicit weights to explicit rubrics for reward modeling, 2026. URL `https://arxiv.org/abs/2510.17314`.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable LLM post-training. *CoRR*, 2026a. URL `https://arxiv.org/abs/2602.01511`.

Yifei Xu, Guilherme Potje, Shivam Shandilya, Tiancheng Yuan, Leonardo de Oliveira Nunes, Rakshanda Agarwal, Saeid Asgari, Adam Atkinson, Emre Kıçiman, Songwu Lu, Ranveer Chandra, and Tusher Chakraborty. Sibylsense: Adaptive rubric learning via memory tuning and adversarial probing. *CoRR*, 2026b. URL `https://arxiv.org/abs/2602.20751`.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, 2024a. URL `https://arxiv.org/abs/2412.15115`.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL `https://arxiv.org/abs/2505.09388`.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *Proceedings of the 41st International Conference on Machine Learning*, pages 56276–56297, 2024b.

Zhiyong Yang, Qianqian Xu, Zitai Wang, Sicong Li, Boyu Han, Shilong Bao, Xiaochun Cao, and Qingming Huang. Harnessing hierarchical label distribution variations in test agnostic long-tail recognition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 56624–56664. PMLR, 2024c.

Zhiyong Yang, Qianqian Xu, Sicong Li, Zitai Wang, Xiaochun Cao, and Qingming Huang. DirMixE: Harnessing test agnostic long-tail recognition with hierarchical label variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(4):4605–4622, 2026. doi: 10.1109/TPAMI.2025.3647124.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3775.

Qiang Zhang, Boli Chen, Fanrui Zhang, Ruixue Ding, Shihang Wang, Qiuchen Wang, Yinfeng Huang, Haonan Zhang, Rongxiang Zhu, Xin Li, Houquan Zhou, Pengjun Xie, Kaipeng Zhang, Jingren Zhou, and Jiawei Liu. ArenaRL: Scaling RL for open-ended agents via tournament-based relative ranking. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, pages 154568–154598. PMLR, 2026a.

Zhicheng Zhang, Jiwei Tang, Kuicai Dong, Xiaopeng Li, Jieming Zhu, Jingyu Li, Qianhui Zhu, Fengyuan Lu, Wang Jiaheng, Gang Wang, Hai-Tao Zheng, and Zhaocheng Du. When hard

negatives hurt: Bridging the generative discriminative gap in hard negative synthesis for retrieval. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pages 6688–6699. ACM, 2026b. doi: 10.1145/3770855.3818118.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. URL `https://arxiv.org/abs/2507.18071`.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10586–10613, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.630.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*, 2025.

Appendix

Table of Contents

A Limitations and Broader Impact **17**

B Additional Details of GRPO **17**

C Additional Theoretical Details for Focal Reward **18**

C.1 Local latent-label model 18

C.2 Proof of Theorem 1: From the misallocation bound to the local surrogate 19

C.3 Proof of Theorem 2: Static gap and headroom-aware direction 21

C.4 Bridge to the implementation proxy 23

C.5 Identification boundary of the implementation statistic 26

C.6 Base-Scalarizer Dependence and Frontier Preservation 26

C.7 Bounded Correction Capacity 27

D Detailed Training Settings **28**

E Additional Experimental Results **29**

E.1 Statistical Reliability of Main Results 30

E.2 Generalization Across Training Settings 30

E.3 Training Dynamics 32

E.4 Robustness and Failure Boundaries 32

E.5 Human Evaluation 33

F Prompt Templates **33**

F.1 Pairwise Judge Prompt 34

F.2 Prior-Weight Generation Prompt 35

G Case Study **36**

G.1 Case 1: Instruction following—first-person persona 36

G.2 Case 2: Creative writing with development suggestions 38

H Additional Related Work **41**

I Existing Assets and Licenses **42**

A Limitations and Broader Impact

Limitations. Focal Reward operates on criterion-level judge scores and therefore benefits from well-calibrated rubric feedback, as is common in rubric-based RL. If the judge confuses criteria, exhibits systematic biases, or assigns poorly calibrated absolute scores, the estimated saturation levels may inherit these errors and affect the resulting reweighting. In our experiments, we mitigate this issue by keeping the judge and rubric fixed across all methods, using pairwise criterion-level scoring, and evaluating across multiple benchmarks and model scales. The theory is intended to characterize the local reward-allocation mechanism rather than to provide a full convergence analysis of large-scale RL with non-stationary LLM policies. This scope matches the empirical focus of the paper, where the method is evaluated through rollout-level, mechanism-level, and ablation analyses. Focal Reward adds only lightweight reward-synthesis computation, while the overall RL post-training pipeline remains resource-intensive due to rollout generation and rubric-based judging.

Broader impact. Focal Reward may contribute to more balanced, reliable, and human-aligned open-ended language-model systems by encouraging optimization across diverse rubric criteria rather than allowing training to concentrate on already-satisfied dimensions. By reallocating reward emphasis toward under-satisfied criteria, it can improve the overall quality and consistency of model rollouts, especially for hard requirements such as instruction following, safety constraints, factual grounding, domain-specific completeness, and fairness-related criteria when they are explicitly included in the rubric. This may help reduce uneven model behavior in real user-facing scenarios, where a response that is fluent or well-formatted can still fail important requirements. More broadly, saturation-aware reward synthesis provides a practical mechanism for making rubric-based RL more balanced and diagnosable, allowing developers to identify which quality dimensions remain under-optimized and to direct training pressure accordingly.

At the same time, Focal Reward inherits the limitations and risks of the rubric and judge used to produce criterion level scores. If the rubric omits important safety, privacy, or fairness criteria, or if the judge systematically miscalibrates certain criteria, the adaptive reweighting may amplify these incomplete or biased signals during training. Stronger open ended generation systems may also be misused for misleading, low quality, or harmful content generation when deployed without appropriate safeguards. Practical mitigation strategies include auditing rubrics before training, evaluating judge calibration and bias, explicitly including safety and fairness related criteria, and monitoring downstream model behavior after deployment.

Scope and practical considerations. For long inputs or extensive rubrics, judging costs also depend on the input representation. Existing context-compression techniques include token-level prompt compression [Tang et al., 2025], group merging with layer semantic alignment [Tang et al., 2026d], marginal-information-based budget allocation [Tang et al., 2026c], selective close reading and skimming [Tang et al., 2026b], and semantic-driven context aggregation [Tang et al., 2026a]. Applying compression to rubric-based judging would require preserving the evidence needed to score individual criteria; its effect on judge reliability is not evaluated here.

Criterion-level diagnostics are another practical consideration. In retrieval, CausalNeg constructs negatives by selectively violating query requirements [Zhang et al., 2026b]. Analogous controlled contrasts could help assess whether a rubric judge detects individual failed requirements; this diagnostic protocol is not evaluated in the present study.

The imbalance studied here concerns unequal criterion saturation within a fixed rubric, whereas class-dependent reweighting and distribution-aware long-tail recognition address class-frequency imbalance and variation in test label distributions [Wang et al., 2023, Yang et al., 2024c, 2026]. Probability-mass allocation and hybrid hard–soft supervision in knowledge distillation concern the design of teacher-derived supervision [Wang et al., 2025, 2026a]; these mechanisms are outside the scope of our study of criterion-level reward aggregation.

B Additional Details of GRPO

We follow the GRPO formulation of Shao et al. [2024]. Given a prompt $x \in \mathcal{D}$, the old policy $\pi_{\theta_{\text{old}}}$ samples a group of G candidate rollouts $Y = \{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$. After rubric-based judgment and reward aggregation, each rollout y_i receives a scalar reward R_i. GRPO estimates the advantage

using the relative rewards within the same sampled group, without learning an additional value function:

$$\hat{A}_i = \frac{R_i - \mathrm{mean}\big(\{R_1, \ldots, R_G\}\big)}{\mathrm{std}\big(\{R_1, \ldots, R_G\}\big)}. \tag{8}$$

Since R_i is assigned at the rollout level, the same advantage $\hat{A}_i$ is used for all tokens in y_i.

For each generated token $y_{i,t}$, the token-level importance ratio is defined as

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t} \mid x, y_{i,<t})}. \tag{9}$$

The policy is then optimized with the clipped objective in Eq. (3). Following GRPO, the KL penalty is estimated at the token level as

$$D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}) = \frac{\pi_{\mathrm{ref}}(y_{i,t} \mid x, y_{i,<t})}{\pi_\theta(y_{i,t} \mid x, y_{i,<t})} - \log \frac{\pi_{\mathrm{ref}}(y_{i,t} \mid x, y_{i,<t})}{\pi_\theta(y_{i,t} \mid x, y_{i,<t})} - 1. \tag{10}$$

In our setting, Static aggregation and Focal Reward differ only in how the criterion-level feedback is synthesized into the scalar rollout reward R_i. Once R_i is obtained, both methods use the same group-relative normalization in Eq. (8), the same token-level ratio in Eq. (9), and the same clipped policy optimization objective in Eq. (3). Thus, Focal Reward changes the reward direction before policy optimization while leaving the GRPO optimization procedure unchanged.

C Additional Theoretical Details for Focal Reward

This appendix complements Section 3.2. The main text states only the key local conclusions: Theorem 1 gives the pairwise misallocation bound that motivates the surrogate $\Xi_t(\boldsymbol{a})$, and Theorem 2 characterizes the static aggregation gap relative to the headroom-aware preferred direction. Here we provide the formal local latent-label model, derive the misallocation bound, prove the closed-form preferred direction and static gap, and then discuss the proxy bridge and identification boundary for the implemented saturation statistic.

C.1 Local latent-label model

Given a training state at the timestep t, all probabilities and expectations below are taken with respect to the joint law of

$$X \sim \mathcal{D}, \qquad Y = (Y_1, \ldots, Y_G) \sim \pi_{\theta_t}(\cdot \mid X)^G, \qquad (I, J) \sim \mathrm{Unif}\big\{(i,j) : 1 \le i, j \le G,\ i \ne j\big\},$$

together with a latent label $L_t \in \{+1, -1\}$.

For each ordered pair (i, j), define the criterion-level pairwise score-difference vector

$$d_{i,j}^{(k)} := s_{i,j}^{(k)} - s_{j,i}^{(k)}, \qquad \boldsymbol{d}_{i,j} := \big(d_{i,j}^{(1)}, \ldots, d_{i,j}^{(K)}\big) \in \mathbb{R}^K. \tag{11}$$

The random ordered pair induces the random criterion-level score-difference vector

$$\boldsymbol{D}_t := \boldsymbol{d}_{I,J}(X, Y) \in \mathbb{R}^K. \tag{12}$$

For any nonnegative scalarizer $\boldsymbol{a} \in \mathbb{R}_+^K$, the scalarized pairwise margin is given by $\boldsymbol{a}^\top \boldsymbol{D}_t$.

The latent variable $L_t \in \{+1, -1\}$ indicates which direction of pairwise credit assignment is locally more useful for policy improvement at the timestep t. The analysis uses the following assumptions.

(A1) Positive learning edge. There exists $\boldsymbol{\eta}_t \in \mathbb{R}_+^K$ such that

$$\mathbb{E}[\boldsymbol{D}_t \mid L_t = \ell] = \ell\, \boldsymbol{\eta}_t, \qquad \ell \in \{+1, -1\}. \tag{13}$$

(A2) Headroom-shaped edge. There exist latent frontier-saturation levels $q_t^{(k)} \in [0, 1]$, latent remaining frontier headroom values $H_t^{(k)} := 1 - q_t^{(k)}$, and constants $c_t > 0$, $\gamma_0 > 0$ such that

$$\eta_t^{(k)} = c_t\, w_{\mathrm{base}}^{(k)} \big(H_t^{(k)}\big)^{\gamma_0}, \qquad k = 1, \ldots, K. \tag{14}$$

(A3) Proxy-sub-Gaussian residual. Define the residual

$$\boldsymbol{\zeta}_t := \boldsymbol{D}_t - L_t \boldsymbol{\eta}_t. \tag{15}$$

For each $\ell \in \{+1, -1\}$, there exists a symmetric positive-definite matrix $\boldsymbol{\Sigma}_t(\ell) \succ 0$ such that

$$\mathbb{E}\big[\exp\big(\boldsymbol{u}^\top \boldsymbol{\zeta}_t\big) \mid L_t = \ell\big] \le \exp\left(\frac{1}{2}\boldsymbol{u}^\top \boldsymbol{\Sigma}_t(\ell)\boldsymbol{u}\right), \qquad \forall \boldsymbol{u} \in \mathbb{R}^K. \tag{16}$$

Thus $\boldsymbol{\Sigma}_t(\ell)$ is a conditional sub-Gaussian proxy matrix. It need not equal the true conditional covariance of $\boldsymbol{\zeta}_t$. This assumption is used only as a proxy tail condition for deriving the local bound. Since criterion-level scores are bounded in practice, residual fluctuations can be controlled by a sub-Gaussian proxy matrix, which may absorb judge bias, criterion correlation, and prompt-level variability.

Discussion of the assumptions. Assumption (A1) introduces a local positive learning edge, which allows us to define pairwise credit misallocation as the event that the scalarized margin has a sign inconsistent with L_t. Assumption (A3) controls the residual fluctuation around this edge and yields the exponential bound in Theorem 1. Assumption (A2) is used in Theorem 2 to connect the preferred scalarizer direction to remaining frontier headroom. Together, these assumptions provide a local justification for reallocating reward weight from saturated criteria toward criteria that still retain larger headroom.

C.2 Proof of Theorem 1: From the misallocation bound to the local surrogate

We derive the local surrogate used in Section 3.2. Under the local latent-label model in Appendix C.1, a scalarizer can misallocate pairwise credit when the scalarized margin $\boldsymbol{a}^\top \boldsymbol{D}_t$ has a sign inconsistent with the latent locally useful direction L_t. The following proof establishes the exponential bound stated in Theorem 1, which naturally motivates maximizing the surrogate $\Xi_t(\boldsymbol{a})$.

Restatement of Theorem 1. Assume (A1) and (A3). For any nonzero $\boldsymbol{a} \in \mathbb{R}^K_+$,

$$\Pr\big[L_t\, \boldsymbol{a}^\top \boldsymbol{D}_t \le 0\big] \le \exp\left(-\frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{2 \max_{\ell \in \{+1,-1\}} \boldsymbol{a}^\top \boldsymbol{\Sigma}_t(\ell)\boldsymbol{a}}\right). \tag{17}$$

Equivalently,

$$\Pr\big[L_t\, \boldsymbol{a}^\top \boldsymbol{D}_t \le 0\big] \le \exp\left(-\frac{1}{2} \min_{\ell \in \{+1,-1\}} \frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{\boldsymbol{a}^\top \boldsymbol{\Sigma}_t(\ell)\boldsymbol{a}}\right). \tag{18}$$

Proof. For brevity, write

$$\boldsymbol{D} := \boldsymbol{D}_t, \qquad L := L_t, \qquad \boldsymbol{\eta} := \boldsymbol{\eta}_t, \qquad \boldsymbol{\zeta} := \boldsymbol{\zeta}_t, \qquad \boldsymbol{\Sigma}_\ell := \boldsymbol{\Sigma}_t(\ell), \quad \ell \in \{+1, -1\}.$$

Fix a nonzero scalarizer $\boldsymbol{a} \in \mathbb{R}^K_+$. Since $\boldsymbol{a} \in \mathbb{R}^K_+$ and $\boldsymbol{\eta} \in \mathbb{R}^K_+$, the scalar

$$\nu := \boldsymbol{a}^\top \boldsymbol{\eta}$$

satisfies $\nu \ge 0$. If $\nu = 0$, then the right-hand side of (17) equals 1, so the claim is trivial because probabilities are at most 1. We therefore assume $\nu > 0$.

First condition on the event $\{L = +1\}$. By (A1),

$$\mathbb{E}[\boldsymbol{D} \mid L = +1] = \boldsymbol{\eta}.$$

Because $\boldsymbol{\zeta} = \boldsymbol{D} - L\boldsymbol{\eta}$, on $\{L = +1\}$ we have

$$\boldsymbol{D} = \boldsymbol{\eta} + \boldsymbol{\zeta}, \qquad \boldsymbol{a}^\top \boldsymbol{D} = \boldsymbol{a}^\top \boldsymbol{\eta} + \boldsymbol{a}^\top \boldsymbol{\zeta}.$$

Hence,

$$L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0 \iff \boldsymbol{a}^\top \boldsymbol{D} \le 0 \iff \boldsymbol{a}^\top \boldsymbol{\zeta} \le -\nu.$$

For any $\lambda > 0$, Markov's inequality gives

$$\Pr\big[\boldsymbol{a}^\top \boldsymbol{\zeta} \le -\nu \mid L = +1\big] = \Pr\Big[e^{-\lambda \boldsymbol{a}^\top \boldsymbol{\zeta}} \ge e^{\lambda \nu} \mid L = +1\Big] \le e^{-\lambda \nu} \mathbb{E}\Big[e^{-\lambda \boldsymbol{a}^\top \boldsymbol{\zeta}} \mid L = +1\Big].$$

Applying (A3) with $\boldsymbol{u} = -\lambda \boldsymbol{a}$ yields

$$\mathbb{E}\Big[e^{-\lambda \boldsymbol{a}^\top \boldsymbol{\zeta}} \mid L = +1\Big] \le \exp\!\left(\frac{1}{2}\lambda^2 \boldsymbol{a}^\top \boldsymbol{\Sigma}_{+1} \boldsymbol{a}\right).$$

Therefore,

$$\Pr\big[L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0 \mid L = +1\big] \le \exp\!\left(-\lambda\nu + \frac{1}{2}\lambda^2 \boldsymbol{a}^\top \boldsymbol{\Sigma}_{+1} \boldsymbol{a}\right).$$

The exponent is minimized over $\lambda > 0$ by

$$\lambda^\star = \frac{\nu}{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{+1} \boldsymbol{a}},$$

where the denominator is positive because $\boldsymbol{\Sigma}_{+1} \succ 0$ and $\boldsymbol{a} \neq 0$. Substituting this value gives

$$\Pr\big[L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0 \mid L = +1\big] \le \exp\!\left(-\frac{\nu^2}{2\boldsymbol{a}^\top \boldsymbol{\Sigma}_{+1} \boldsymbol{a}}\right) = \exp\!\left(-\frac{(\boldsymbol{a}^\top \boldsymbol{\eta})^2}{2\boldsymbol{a}^\top \boldsymbol{\Sigma}_{+1} \boldsymbol{a}}\right). \tag{19}$$

Now condition on the event $\{L = -1\}$. By (A1),

$$\mathbb{E}[\boldsymbol{D} \mid L = -1] = -\boldsymbol{\eta}.$$

Because $\boldsymbol{\zeta} = \boldsymbol{D} - L\boldsymbol{\eta}$, on $\{L = -1\}$ we have

$$\boldsymbol{D} = -\boldsymbol{\eta} + \boldsymbol{\zeta}, \qquad \boldsymbol{a}^\top \boldsymbol{D} = -\boldsymbol{a}^\top \boldsymbol{\eta} + \boldsymbol{a}^\top \boldsymbol{\zeta}.$$

Thus,

$$L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0 \iff -\boldsymbol{a}^\top \boldsymbol{D} \le 0 \iff \boldsymbol{a}^\top \boldsymbol{D} \ge 0 \iff \boldsymbol{a}^\top \boldsymbol{\zeta} \ge \nu.$$

Again, for any $\lambda > 0$, Markov's inequality gives

$$\Pr\big[\boldsymbol{a}^\top \boldsymbol{\zeta} \ge \nu \mid L = -1\big] = \Pr\Big[e^{\lambda \boldsymbol{a}^\top \boldsymbol{\zeta}} \ge e^{\lambda\nu} \mid L = -1\Big] \le e^{-\lambda\nu}\mathbb{E}\Big[e^{\lambda \boldsymbol{a}^\top \boldsymbol{\zeta}} \mid L = -1\Big].$$

Applying (A3) with $\boldsymbol{u} = \lambda \boldsymbol{a}$ yields

$$\mathbb{E}\Big[e^{\lambda \boldsymbol{a}^\top \boldsymbol{\zeta}} \mid L = -1\Big] \le \exp\!\left(\frac{1}{2}\lambda^2 \boldsymbol{a}^\top \boldsymbol{\Sigma}_{-1} \boldsymbol{a}\right).$$

Therefore,

$$\Pr\big[L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0 \mid L = -1\big] \le \exp\!\left(-\lambda\nu + \frac{1}{2}\lambda^2 \boldsymbol{a}^\top \boldsymbol{\Sigma}_{-1} \boldsymbol{a}\right).$$

Optimizing over $\lambda > 0$ gives

$$\lambda^\star = \frac{\nu}{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{-1} \boldsymbol{a}},$$

and therefore

$$\Pr\big[L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0 \mid L = -1\big] \le \exp\!\left(-\frac{(\boldsymbol{a}^\top \boldsymbol{\eta})^2}{2\boldsymbol{a}^\top \boldsymbol{\Sigma}_{-1} \boldsymbol{a}}\right). \tag{20}$$

Finally, by the law of total probability,

$$\Pr\big[L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0\big] = \sum_{\ell \in \{+1,-1\}} \Pr\big[L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0 \mid L = \ell\big] \Pr[L = \ell].$$

Combining this identity with (19) and (20), we obtain

$$\Pr\big[L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0\big] \le \max_{\ell \in \{+1,-1\}} \exp\!\left(-\frac{(\boldsymbol{a}^\top \boldsymbol{\eta})^2}{2\boldsymbol{a}^\top \boldsymbol{\Sigma}_{\ell} \boldsymbol{a}}\right).$$

Since the exponential is monotone and the numerator does not depend on ℓ, this becomes

$$\Pr\big[L\, \boldsymbol{a}^\top \boldsymbol{D} \le 0\big] \le \exp\!\left(-\frac{(\boldsymbol{a}^\top \boldsymbol{\eta})^2}{2 \max_{\ell \in \{+1,-1\}} \boldsymbol{a}^\top \boldsymbol{\Sigma}_{\ell} \boldsymbol{a}}\right).$$

Returning to the original t-indexed notation gives

$$\Pr\big[L_t\,\boldsymbol{a}^\top \boldsymbol{D}_t \le 0\big] \le \exp\!\left(-\frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{2\max_{\ell\in\{+1,-1\}} \boldsymbol{a}^\top \boldsymbol{\Sigma}_t(\ell)\boldsymbol{a}}\right),$$

which proves (17). The equivalent form (18) follows from the identity

$$\frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{\max_{\ell\in\{+1,-1\}} \boldsymbol{a}^\top \boldsymbol{\Sigma}_t(\ell)\boldsymbol{a}} = \min_{\ell\in\{+1,-1\}} \frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{\boldsymbol{a}^\top \boldsymbol{\Sigma}_t(\ell)\boldsymbol{a}}.$$

Defining

$$\Xi_t(\boldsymbol{a}) = \frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{\max_{\ell\in\{+1,-1\}} \boldsymbol{a}^\top \boldsymbol{\Sigma}_t(\ell)\boldsymbol{a}},$$

the bound can be written as

$$\Pr\big[L_t\,\boldsymbol{a}^\top \boldsymbol{D}_t \le 0\big] \le \exp\!\left(-\frac{1}{2}\Xi_t(\boldsymbol{a})\right).$$

Thus, reducing this exponential upper bound is equivalent to maximizing the local surrogate $\Xi_t(\boldsymbol{a})$, which is the surrogate used in the main text. □

This motivates the next step: characterizing the scalarizer that maximizes $\Xi_t(\boldsymbol{a})$ under additional local simplifications.

C.3 Proof of Theorem 2: Static gap and headroom-aware direction

We now prove the closed-form headroom-aware preferred direction and the resulting local surrogate gap for static aggregation. This proof corresponds to Theorem 2 in the main text.

Restatement of Theorem 2. Given the local surrogate $\Xi_t(\boldsymbol{a})$ in Theorem 1, let $\boldsymbol{a}_t^\star$ denote the locally preferred scalarizer that maximizes $\Xi_t(\boldsymbol{a})$. Under the symmetric and isotropic simplifications and the headroom-shaped edge model in Appendix C.1, this preferred scalarizer has the closed-form direction

$$[\boldsymbol{a}_t^\star]^{(k)} \propto w_{\mathrm{base}}^{(k)} \big(H_t^{(k)}\big)^{\gamma_0}, \qquad k = 1, \dots, K. \tag{21}$$

Relative to this direction, the fixed base scalarizer $\boldsymbol{w}_{\mathrm{base}}$ incurs the local surrogate gap

$$\Xi_t(\boldsymbol{a}_t^\star) - \Xi_t(\boldsymbol{w}_{\mathrm{base}}) = \frac{c_t^2 \|\boldsymbol{w}_{\mathrm{base}}\|_2^2}{\sigma_t^2} \operatorname{Var}_\mu(H_t^{\gamma_0}), \tag{22}$$

where

$$\mu^{(k)} = \frac{\big(w_{\mathrm{base}}^{(k)}\big)^2}{\|\boldsymbol{w}_{\mathrm{base}}\|_2^2}, \qquad \operatorname{Var}_\mu(\boldsymbol{z}) := \sum_{k=1}^K \mu^{(k)} (z^{(k)})^2 - \left(\sum_{k=1}^K \mu^{(k)} z^{(k)}\right)^2. \tag{23}$$

Here $H_t^{\gamma_0}$ denotes the vector whose k-th entry is $\big(H_t^{(k)}\big)^{\gamma_0}$.

Proof. The proof has two parts. First, we characterize the locally preferred direction induced by $\Xi_t(\boldsymbol{a})$. Second, we compare this direction with the fixed base scalarizer $\boldsymbol{w}_{\mathrm{base}}$.

Part I: preferred local direction. Under the symmetric residual simplification,

$$\boldsymbol{\Sigma}_t(+1) = \boldsymbol{\Sigma}_t(-1) = \boldsymbol{\Sigma}_t \succ 0,$$

the local surrogate becomes

$$\Xi_t(\boldsymbol{a}) = \frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{\boldsymbol{a}^\top \boldsymbol{\Sigma}_t \boldsymbol{a}}. \tag{24}$$

For brevity, write

$$\boldsymbol{\eta} := \boldsymbol{\eta}_t, \qquad \boldsymbol{\Sigma} := \boldsymbol{\Sigma}_t.$$

For any nonzero $\boldsymbol{a} \in \mathbb{R}^K$, define

$$\boldsymbol{b} := \boldsymbol{\Sigma}^{1/2} \boldsymbol{a}.$$

Then

$$\boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a} = \|\boldsymbol{b}\|_2^2, \qquad \boldsymbol{a}^\top \boldsymbol{\eta} = \boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\eta}.$$

Substituting these identities into (24) gives

$$\Xi_t(\boldsymbol{a}) = \frac{\left(\boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\eta}\right)^2}{\|\boldsymbol{b}\|_2^2}.$$

By Cauchy–Schwarz,

$$\frac{\left(\boldsymbol{b}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\eta}\right)^2}{\|\boldsymbol{b}\|_2^2} \le \|\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}.$$

Equality holds if and only if $\boldsymbol{b}$ is a nonzero scalar multiple of $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\eta}$. Equivalently,

$$\boldsymbol{\Sigma}^{1/2} \boldsymbol{a} = \lambda \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\eta}$$

for some $\lambda \neq 0$, which implies

$$\boldsymbol{a} = \lambda \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}. \tag{25}$$

Thus, under the symmetric simplification, the unconstrained maximizers of $\Xi_t(\boldsymbol{a})$ over $\boldsymbol{a} \in \mathbb{R}^K \setminus \{0\}$ are exactly the nonzero scalar multiples of $\boldsymbol{\Sigma}_t^{-1} \boldsymbol{\eta}_t$.

Under the additional isotropic simplification,

$$\boldsymbol{\Sigma}_t = \sigma_t^2 \boldsymbol{I}, \qquad \sigma_t^2 > 0,$$

we have

$$\boldsymbol{\Sigma}_t^{-1} \boldsymbol{\eta}_t = \frac{1}{\sigma_t^2} \boldsymbol{\eta}_t.$$

Therefore, the preferred direction is proportional to $\boldsymbol{\eta}_t$:

$$\boldsymbol{a}_t^\star \propto \boldsymbol{\eta}_t. \tag{26}$$

Since $\boldsymbol{\eta}_t \in \mathbb{R}_+^K$ by (A1), this direction also satisfies the nonnegative scalarizer constraint.

Finally, under the headroom-shaped edge model (A2),

$$\eta_t^{(k)} = c_t \, w_{\text{base}}^{(k)} \left(H_t^{(k)}\right)^{\gamma_0}, \qquad k = 1, \dots, K.$$

Since $c_t > 0$ is a global state-dependent scale, it does not affect the direction of the maximizer. Combining this with (26) gives

$$[\boldsymbol{a}_t^\star]^{(k)} \propto w_{\text{base}}^{(k)} \left(H_t^{(k)}\right)^{\gamma_0}, \qquad k = 1, \dots, K,$$

which proves the headroom-aware preferred direction in (21).

Part II: static aggregation gap. We now compare the preferred direction $\boldsymbol{a}_t^\star$ with the fixed base scalarizer $\boldsymbol{w}_{\text{base}}$. Let

$$\boldsymbol{w} := \boldsymbol{w}_{\text{base}}, \qquad Z^{(k)} := \left(H_t^{(k)}\right)^{\gamma_0}, \qquad \boldsymbol{Z} := \left(Z^{(1)}, \dots, Z^{(K)}\right).$$

Under (A2),

$$\boldsymbol{\eta}_t = c_t (\boldsymbol{w} \odot \boldsymbol{Z}),$$

where $\odot$ denotes elementwise multiplication.

Since $\boldsymbol{\Sigma}_t = \sigma_t^2 \boldsymbol{I}$, the surrogate becomes

$$\Xi_t(\boldsymbol{a}) = \frac{(\boldsymbol{a}^\top \boldsymbol{\eta}_t)^2}{\sigma_t^2 \|\boldsymbol{a}\|_2^2}.$$

For the preferred direction, Part I gives $\boldsymbol{a}_t^\star \propto \boldsymbol{\eta}_t$. Therefore,

$$\Xi_t(\boldsymbol{a}_t^\star) = \frac{\|\boldsymbol{\eta}_t\|_2^2}{\sigma_t^2}. \tag{27}$$

For the fixed base scalarizer $\boldsymbol{w}$, we have

$$\Xi_t(\boldsymbol{w}) = \frac{(\boldsymbol{w}^\top \boldsymbol{\eta}_t)^2}{\sigma_t^2 \|\boldsymbol{w}\|_2^2}. \tag{28}$$

Using $\boldsymbol{\eta}_t = c_t(\boldsymbol{w} \odot \boldsymbol{Z})$, we compute

$$\|\boldsymbol{\eta}_t\|_2^2 = c_t^2 \sum_{k=1}^{K} \big(w_{\text{base}}^{(k)}\big)^2 \big(Z^{(k)}\big)^2,$$

and

$$\boldsymbol{w}^\top \boldsymbol{\eta}_t = c_t \sum_{k=1}^{K} \big(w_{\text{base}}^{(k)}\big)^2 Z^{(k)}.$$

Substituting these expressions into (27) and (28) yields

$$\Xi_t(\boldsymbol{a}_t^\star) - \Xi_t(\boldsymbol{w}) = \frac{c_t^2}{\sigma_t^2} \left[\sum_{k=1}^{K} \big(w_{\text{base}}^{(k)}\big)^2 \big(Z^{(k)}\big)^2 - \frac{\left(\sum_{k=1}^{K} \big(w_{\text{base}}^{(k)}\big)^2 Z^{(k)}\right)^2}{\|\boldsymbol{w}_{\text{base}}\|_2^2} \right].$$

Now define the base-induced weighting distribution

$$\mu^{(k)} = \frac{\big(w_{\text{base}}^{(k)}\big)^2}{\|\boldsymbol{w}_{\text{base}}\|_2^2}, \qquad k = 1, \dots, K.$$

Since $\sum_{k=1}^{K} \mu^{(k)} = 1$, the previous expression becomes

$$\Xi_t(\boldsymbol{a}_t^\star) - \Xi_t(\boldsymbol{w}) = \frac{c_t^2 \|\boldsymbol{w}_{\text{base}}\|_2^2}{\sigma_t^2} \left[\sum_{k=1}^{K} \mu^{(k)} \big(Z^{(k)}\big)^2 - \left(\sum_{k=1}^{K} \mu^{(k)} Z^{(k)} \right)^2 \right].$$

By the definition of weighted variance under μ,

$$\mathrm{Var}_\mu(\boldsymbol{Z}) = \sum_{k=1}^{K} \mu^{(k)} \big(Z^{(k)}\big)^2 - \left(\sum_{k=1}^{K} \mu^{(k)} Z^{(k)} \right)^2,$$

we obtain

$$\Xi_t(\boldsymbol{a}_t^\star) - \Xi_t(\boldsymbol{w}_{\text{base}}) = \frac{c_t^2 \|\boldsymbol{w}_{\text{base}}\|_2^2}{\sigma_t^2} \mathrm{Var}_\mu(\boldsymbol{Z}).$$

Recalling that $Z^{(k)} = \big(H_t^{(k)}\big)^{\gamma_0}$, this is exactly

$$\Xi_t(\boldsymbol{a}_t^\star) - \Xi_t(\boldsymbol{w}_{\text{base}}) = \frac{c_t^2 \|\boldsymbol{w}_{\text{base}}\|_2^2}{\sigma_t^2} \mathrm{Var}_\mu(H_t^{\gamma_0}),$$

which proves the static aggregation gap in (22). □

Remark. The variance term is zero if the headroom factor $\big(H_t^{(k)}\big)^{\gamma_0}$ is effectively constant over the criteria that receive nonzero base weight. In that case, the preferred direction is only a scalar multiple of $\boldsymbol{w}_{\text{base}}$, so static aggregation remains locally aligned with the preferred direction. When the base-weighted headroom profile is heterogeneous, the variance term becomes positive, and the fixed base scalarizer incurs a positive local surrogate gap relative to the headroom-aware direction.

C.4 Bridge to the implementation proxy

The main text implements the headroom-aware direction using the saturation statistic $P^{(k)}$ defined in Section 3.1. We restate the same quantities here with an explicit training-state timestep index t, which will be used in the proof below.

At training state t, consider a sampled rollout group $\{y_i\}_{i=1}^{G}$. Let

$$R_{t,i}^{\text{base}} := R_i(\boldsymbol{w}_{\text{base}})$$

denote the scalar reward of rollout y_i computed by the base scalarizer. The corresponding Gibbs rollout weight is

$$\widehat{r}_{t,i} := \frac{\exp(R_{t,i}^{\text{base}}/T)}{\sum_{n=1}^{G} \exp(R_{t,n}^{\text{base}}/T)}, \qquad \sum_{i=1}^{G} \widehat{r}_{t,i} = 1.$$

For criterion $c^{(k)}$, define

$$\bar{s}_{t,i}^{(k)} := \frac{1}{G-1} \sum_{j \neq i} s_{i,j}^{(k)}, \qquad z_{t,i}^{(k)} := \frac{\bar{s}_{t,i}^{(k)}}{S_{\max}}.$$

Since $s_{i,j}^{(k)} \in [0, S_{\max}]$, we have $z_{t,i}^{(k)} \in [0,1]$. The time-indexed form of the implemented saturation statistic is

$$P_t^{(k)} := \sum_{i=1}^{G} \widehat{r}_{t,i} z_{t,i}^{(k)} = \frac{\langle \widehat{\boldsymbol{r}}_t, \bar{\boldsymbol{s}}_t^{(k)} \rangle}{S_{\max}}.$$

Thus $P_t^{(k)}$ is exactly the inverse reward projection statistic used in the main algorithm. It measures the criterion saturation level under the base-reward-weighted rollout distribution, and $1 - P_t^{(k)}$ is used as the observable proxy for remaining frontier headroom.

The relevant quantity in Theorem 2 is frontier headroom, not the average dissatisfaction over all sampled rollouts. A global average over the entire rollout group can be dominated by low-quality responses and therefore reflect broad response weakness rather than the residual bottleneck of the current strong responses. In contrast, policy improvement is driven by how the scalar reward assigns credit among competitive rollouts. Therefore, the saturation statistic should emphasize criteria that remain under-satisfied on the current reward frontier. The Gibbs weighting used in the implementation provides a soft way to estimate this frontier-level saturation without introducing a hard top-k selection rule.

Proposition 1 (Gibbs frontier concentration)**.** *Let $T > 0$ and fix a nonempty proper subset $\mathcal{S}_t \subset \{1, \dots, G\}$, with $1 \leq |\mathcal{S}_t| < G$. Suppose there exists a reward gap $\delta_t > 0$ such that*

$$\min_{i \in \mathcal{S}_t} R_{t,i}^{\text{base}} \geq \max_{j \notin \mathcal{S}_t} R_{t,j}^{\text{base}} + \delta_t. \tag{29}$$

Let

$$\rho_t := \sum_{i \in \mathcal{S}_t} \widehat{r}_{t,i}$$

denote the Gibbs mass assigned to $\mathcal{S}_t$. Then

$$1 - \rho_t \leq \frac{G - |\mathcal{S}_t|}{|\mathcal{S}_t|} \exp(-\delta_t / T). \tag{30}$$

Moreover, for every criterion $c^{(k)}$, define the frontier-restricted saturation estimate

$$P_{t,\mathcal{S}}^{(k)} := \sum_{i \in \mathcal{S}_t} \frac{\widehat{r}_{t,i}}{\sum_{m \in \mathcal{S}_t} \widehat{r}_{t,m}} z_{t,i}^{(k)}.$$

Then

$$\left| P_t^{(k)} - P_{t,\mathcal{S}}^{(k)} \right| \leq \frac{G - |\mathcal{S}_t|}{|\mathcal{S}_t|} \exp(-\delta_t / T). \tag{31}$$

Proof. Define

$$\alpha := \sum_{i \in \mathcal{S}_t} \exp(R_{t,i}^{\text{base}} / T), \qquad \beta := \sum_{j \notin \mathcal{S}_t} \exp(R_{t,j}^{\text{base}} / T).$$

Then

$$\rho_t = \frac{\alpha}{\alpha + \beta} = \sum_{i \in \mathcal{S}_t} \widehat{r}_{t,i}.$$

By (29), for every $j \notin \mathcal{S}_t$,

$$R_{t,j}^{\text{base}} \leq \min_{i \in \mathcal{S}_t} R_{t,i}^{\text{base}} - \delta_t.$$

Therefore,

$$\exp(R_{t,j}^{\text{base}} / T) \leq \exp(-\delta_t / T) \exp\left(\frac{\min_{i \in \mathcal{S}_t} R_{t,i}^{\text{base}}}{T} \right).$$

Summing over $j \notin \mathcal{S}_t$ gives

$$\beta \le (G - |\mathcal{S}_t|) \exp(-\delta_t/T) \exp\left(\frac{\min_{i \in \mathcal{S}_t} R_{t,i}^{\text{base}}}{T} \right).$$

Also,

$$\alpha \ge |\mathcal{S}_t| \exp\left(\frac{\min_{i \in \mathcal{S}_t} R_{t,i}^{\text{base}}}{T} \right).$$

Hence

$$\frac{\beta}{\alpha} \le \frac{G - |\mathcal{S}_t|}{|\mathcal{S}_t|} \exp(-\delta_t/T).$$

Since

$$1 - \rho_t = \frac{\beta}{\alpha + \beta} \le \frac{\beta}{\alpha},$$

we obtain (30).

Next, define the out-of-frontier saturation estimate

$$P_{t,\text{out}}^{(k)} := \frac{1}{1 - \rho_t} \sum_{j \notin \mathcal{S}_t} \widehat{r}_{t,j} z_{t,j}^{(k)}$$

when $1 - \rho_t > 0$. If $1 - \rho_t = 0$, choose any $P_{t,\text{out}}^{(k)} \in [0, 1]$. Since every $z_{t,i}^{(k)} \in [0, 1]$, both $P_{t,\mathcal{S}}^{(k)}$ and $P_{t,\text{out}}^{(k)}$ lie in $[0, 1]$. Moreover,

$$P_t^{(k)} = \sum_{i \in \mathcal{S}_t} \widehat{r}_{t,i} z_{t,i}^{(k)} + \sum_{j \notin \mathcal{S}_t} \widehat{r}_{t,j} z_{t,j}^{(k)} = \rho_t P_{t,\mathcal{S}}^{(k)} + (1 - \rho_t) P_{t,\text{out}}^{(k)}.$$

Therefore,

$$P_t^{(k)} - P_{t,\mathcal{S}}^{(k)} = (1 - \rho_t) \left(P_{t,\text{out}}^{(k)} - P_{t,\mathcal{S}}^{(k)} \right).$$

Taking absolute values gives

$$\left| P_t^{(k)} - P_{t,\mathcal{S}}^{(k)} \right| \le 1 - \rho_t.$$

Combining this with (30) yields (31). □

Why not use a global average? If one instead used the unweighted global statistic

$$P_{t,\text{global}}^{(k)} = \frac{1}{G} \sum_{i=1}^{G} z_{t,i}^{(k)},$$

then its deviation from a frontier-restricted average would generally scale with the fraction of non-frontier rollouts rather than with the reward gap. For example, writing

$$P_{t,\text{global}}^{(k)} = \frac{|\mathcal{S}_t|}{G} P_{t,\mathcal{S},\text{unif}}^{(k)} + \left(1 - \frac{|\mathcal{S}_t|}{G} \right) P_{t,\text{out},\text{unif}}^{(k)},$$

the coefficient on the out-of-frontier average is $1 - |\mathcal{S}_t|/G$, which can be large even when the base reward clearly separates strong and weak rollouts. By contrast, Proposition 1 shows that the out-of-frontier contribution under Gibbs weighting is bounded by

$$\frac{G - |\mathcal{S}_t|}{|\mathcal{S}_t|} \exp(-\delta_t/T),$$

which decreases as the frontier gap grows. Thus the Gibbs statistic estimates frontier saturation, while the global average estimates overall rollout-pool saturation.

Interpretation. Proposition 1 shows that when the base-pass rewards exhibit a nontrivial frontier gap, the Gibbs weights $\widehat{r}_{t,i}$ concentrate on the strong-rollout subset $\mathcal{S}_t$. In this regime, $P_t^{(k)}$ is close to the criterion saturation level computed only on the reward frontier. Therefore, $1 - P_t^{(k)}$ can be interpreted as an operational proxy for the remaining frontier headroom of criterion $c^{(k)}$, which is the observable counterpart of the latent headroom $H_t^{(k)}$ used in Theorem 2.

C.5 Identification boundary of the implementation statistic

The previous analysis justifies the headroom-aware scalarizer direction through criterion-level pairwise differences. However, estimating saturation requires absolute criterion-score levels from the full judge score tensor. The following proposition clarifies this boundary.

Proposition 2 (Pairwise differences do not identify $P_t^{(k)}$)**.** *Fix a criterion $c^{(k)}$. Suppose there exists a constant $b_k \neq 0$ such that the shifted scores*

$$s_{i,j}'^{(k)} := s_{i,j}^{(k)} + b_k$$

remain in $[0, S_{\max}]$ for all ordered pairs (i, j) with $i \neq j$. For every $k' \neq k$, let

$$s_{i,j}'^{(k')} := s_{i,j}^{(k')}.$$

Then all pairwise difference vectors, base margins, base rewards, and Gibbs rollout weights remain unchanged:

$$\boldsymbol{d}'_{i,j} = \boldsymbol{d}_{i,j}, \qquad \Delta'_{i,j}(\boldsymbol{w}_{\text{base}}) = \Delta_{i,j}(\boldsymbol{w}_{\text{base}}), \qquad R_{t,i}^{\text{base}\prime} = R_{t,i}^{\text{base}}, \qquad \widehat{r}'_{t,i} = \widehat{r}_{t,i}.$$

However, the saturation statistic for criterion $c^{(k)}$ changes as

$$P_t'^{(k)} = P_t^{(k)} + \frac{b_k}{S_{\max}}.$$

Therefore, $P_t^{(k)}$ is not identified by the pairwise difference vectors alone.

Proof. For criterion $c^{(k)}$,

$$d_{i,j}'^{(k)} = s_{i,j}'^{(k)} - s_{j,i}'^{(k)} = (s_{i,j}^{(k)} + b_k) - (s_{j,i}^{(k)} + b_k) = d_{i,j}^{(k)}.$$

For every $k' \neq k$, the scores are unchanged, so

$$d_{i,j}'^{(k')} = d_{i,j}^{(k')}.$$

Hence $\boldsymbol{d}'_{i,j} = \boldsymbol{d}_{i,j}$ for all ordered pairs. Since the base margin depends on the score tensor only through $\boldsymbol{d}_{i,j}$, the base margins are unchanged. The mapped pairwise outcomes and base rewards are therefore unchanged, and the Gibbs rollout weights, as deterministic functions of the base rewards, are unchanged as well.

On the other hand,

$$\bar{s}_{t,i}'^{(k)} = \frac{1}{G-1} \sum_{j \neq i} s_{i,j}'^{(k)} = \bar{s}_{t,i}^{(k)} + b_k.$$

Thus

$$P_t'^{(k)} = \frac{1}{S_{\max}} \sum_{i=1}^{G} \widehat{r}'_{t,i} \bar{s}_{t,i}'^{(k)} = \frac{1}{S_{\max}} \sum_{i=1}^{G} \widehat{r}_{t,i} (\bar{s}_{t,i}^{(k)} + b_k) = P_t^{(k)} + \frac{b_k}{S_{\max}},$$

where we use $\sum_{i=1}^{G} \widehat{r}_{t,i} = 1$. This proves the claim. □

Remark. Proposition 2 only clarifies the scope of the theoretical analysis. The pairwise difference model supports the local headroom-aware scalarizer direction, while the implemented statistic $P_t^{(k)}$ uses additional absolute score-level information from the full judge score tensor. Thus $P_t^{(k)}$ should be viewed as an operational saturation proxy with the frontier interpretation established in Proposition 1, rather than as a latent headroom value identified from pairwise differences alone.

C.6 Base-Scalarizer Dependence and Frontier Preservation

The frontier interpretation of the inverse reward projection is conditional on the base scalarizer preserving the rollouts that are intended to define the current strong-response frontier. Let $S_t^\star \subset \{1, \ldots, G\}$ be this intended frontier at training state t, with $1 \leq |S_t^\star| < G$. The set $S_t^\star$ is a conditional

target of the analysis: it is not inferred from the pairwise scores, and the results below do not assert that an arbitrary prior recovers it. Write

$$\min_{i \in S_t^\star} R_{t,i}^{\mathrm{base}} \geq \max_{j \notin S_t^\star} R_{t,j}^{\mathrm{base}} + \delta_t, \qquad \delta_t > 0, \tag{32}$$

for the base-reward separation condition. With the Gibbs weights from Eq. (4), define the base mass of the intended frontier and its criterion-level saturation by

$$\rho_t^\star := \sum_{i \in S_t^\star} r_{t,i}^b, \qquad P_{t,S^\star}^{(k)} := \sum_{i \in S_t^\star} \frac{r_{t,i}^b}{\rho_t^\star} z_{t,i}^{(k)}, \tag{33}$$

where $z_{t,i}^{(k)} = \bar{s}_{t,i}^{(k)} / S_{\max} \in [0,1]$ and $P_t^{(k)} = \sum_i r_{t,i}^b z_{t,i}^{(k)}$ is the implemented statistic.

Corollary 1 (Conditional frontier preservation)**.** *Under the separation condition in Eq.* (32)*, the Gibbs mass outside the intended frontier and the induced saturation error obey*

$$\begin{aligned} 1 - \rho_t^\star \leq \eta_t^\star &:= \frac{G - |S_t^\star|}{|S_t^\star|} \exp\left(-\frac{\delta_t}{T}\right), \\ \left| P_t^{(k)} - P_{t,S^\star}^{(k)} \right| &\leq \eta_t^\star \quad \textit{for every } k. \end{aligned} \tag{34}$$

Consequently, when $\eta_t^\star$ is small, the observable $P_t^{(k)}$ tracks saturation on the intended frontier up to an additive error that decreases exponentially in the reward gap measured in units of T.

Proof. The Gibbs frontier concentration result already proved in Proposition 1 applies to the choice $\mathcal{S}_t = S_t^\star$. Therefore Eq. (32) implies the first inequality in Eq. (34). For completeness, the remaining step is the decomposition that transfers this mass bound to the saturation statistic. If $\rho_t^\star < 1$, define

$$P_{t,\mathrm{out}}^{(k)} := \frac{1}{1 - \rho_t^\star} \sum_{j \notin S_t^\star} r_{t,j}^b z_{t,j}^{(k)};$$

if $\rho_t^\star = 1$, the desired conclusion is immediate. Since every $z_{t,i}^{(k)}$ lies in $[0,1]$, both restricted averages lie in $[0,1]$ and

$$P_t^{(k)} = \rho_t^\star P_{t,S^\star}^{(k)} + (1 - \rho_t^\star) P_{t,\mathrm{out}}^{(k)}.$$

It follows that

$$\begin{aligned} \left| P_t^{(k)} - P_{t,S^\star}^{(k)} \right| &= (1 - \rho_t^\star) \left| P_{t,\mathrm{out}}^{(k)} - P_{t,S^\star}^{(k)} \right| \\ &\leq 1 - \rho_t^\star \leq \eta_t^\star, \end{aligned}$$

which proves the result. □

The corollary identifies the relevant dependence on the base scalarizer. A semantically appropriate prior need not be perfect, but it must assign base rewards that separate the intended strong rollouts from the remaining group; the guarantee becomes weak when δ_t / T is small or when the group-size factor is large. If the base scalarizer promotes non-frontier rollouts, so that Eq. (32) fails, the bound gives no small error guarantee and $P_t^{(k)}$ may instead describe the frontier induced by that misspecified prior. Thus, $S_t^\star$ is an intended frontier under a conditional assumption, rather than a claim that any prior can recover an arbitrary target frontier.

C.7 Bounded Correction Capacity

We next make precise how much of a base-weight error the focal factor can correct. Let

$$\lambda_t := \frac{\langle \boldsymbol{w}_{\mathrm{base}}, \mathbf{1} \rangle}{\langle \widetilde{\boldsymbol{w}}_t, \mathbf{1} \rangle} > 0, \qquad h_t^{(k)} := 1 - P_t^{(k)} + \epsilon,$$

so Eq. (6) can be written as

$$w_{\mathrm{focal},t}^{(k)} = \lambda_t w_{\mathrm{base}}^{(k)} \big(h_t^{(k)}\big)^\gamma. \tag{35}$$

Here $P_t^{(k)} \in [0,1]$, $\epsilon > 0$, and $\gamma > 0$ imply $h_t^{(k)} \in [\epsilon, 1 + \epsilon]$.

Corollary 2 (Pairwise reversal and finite correction capacity)**.** *For any two criteria k and ℓ with $w_{\text{base}}^{(k)}, w_{\text{base}}^{(\ell)} > 0$,*

$$\frac{w_{\text{focal},t}^{(k)}}{w_{\text{focal},t}^{(\ell)}} = \frac{w_{\text{base}}^{(k)}}{w_{\text{base}}^{(\ell)}} \left(\frac{1 - P_t^{(k)} + \epsilon}{1 - P_t^{(\ell)} + \epsilon} \right)^{\gamma}. \tag{36}$$

If $P_t^{(k)} < P_t^{(\ell)}$ while $w_{\text{base}}^{(k)} < w_{\text{base}}^{(\ell)}$, then Focal reverses the pairwise ordering if and only if

$$\frac{w_{\text{base}}^{(\ell)}}{w_{\text{base}}^{(k)}} < \left(\frac{1 - P_t^{(k)} + \epsilon}{1 - P_t^{(\ell)} + \epsilon} \right)^{\gamma}. \tag{37}$$

Moreover, the multiplicative correction relative to the base ratio is bounded:

$$\left(\frac{\epsilon}{1+\epsilon} \right)^{\gamma} \leq \frac{w_{\text{focal},t}^{(k)} / w_{\text{focal},t}^{(\ell)}}{w_{\text{base}}^{(k)} / w_{\text{base}}^{(\ell)}} \leq \left(\frac{1+\epsilon}{\epsilon} \right)^{\gamma}. \tag{38}$$

Proof. Substituting Eq. (35) for k and ℓ cancels the shared normalizer λ_t and gives Eq. (36). Since all terms are positive, $w_{\text{focal},t}^{(k)} > w_{\text{focal},t}^{(\ell)}$ is equivalent to the right-hand side of Eq. (36) being larger than one. Rearranging yields exactly Eq. (37); equality gives a tie.

Finally, $P_t^{(k)}, P_t^{(\ell)} \in [0, 1]$ gives

$$\epsilon \leq h_t^{(k)}, h_t^{(\ell)} \leq 1 + \epsilon.$$

Therefore

$$\frac{\epsilon}{1+\epsilon} \leq \frac{h_t^{(k)}}{h_t^{(\ell)}} \leq \frac{1+\epsilon}{\epsilon}.$$

Raising this inequality to the positive power γ and using Eq. (36) proves Eq. (38). □

This result gives a finite correction capacity, not recovery of an arbitrary scalarizer. In particular, if $w_{\text{base}}^{(k)} = 0$, then Eq. (35) gives $w_{\text{focal},t}^{(k)} = 0$ for every $P_t^{(k)}$; Focal cannot create support that the base scalarizer removes. For every positive base weight, $\epsilon > 0$ and $\gamma > 0$ keep the focal weight positive, while a sufficiently large base-weight skew can remain unreversed because of the finite factor in Eq. (38). With the default $\gamma = 2$ and $\epsilon = 0.05$, the endpoint range of this factor is $[1/441, 441]$. The Easy-biased stress test in Appendix E.4 illustrates this bounded regime: Easy criteria receive $2.83\times$ the mean Hard weight with within-prompt score–weight Spearman correlation $+1.000$; Static drops by 1.50 points, whereas Focal drops by 0.50, reverses the correlation to -0.688, and exceeds matched Easy-biased Static by $+3.35$. These results support robustness to the evaluated misspecification, while the corollary makes explicit why they do not establish independence from the base scalarizer.

D Detailed Training Settings

For RL post-training, we train on the corresponding OpenRubrics split using the verl framework [Sheng et al., 2025]. To preserve domain-specific characteristics, training is performed separately for each setting for 3 epochs with GSPO. We use a batch size of 64 (mini-batch 32) and AdamW with a fixed learning rate of 1×10^{-6} and weight decay 0.1. A KL divergence penalty (low-variance estimator) with coefficient 0.001 is added to the actor loss to prevent excessive deviation from the reference policy. For each prompt, 8 rollouts are sampled with temperature 1.0 and top-p 1.0. The maximum prompt and response lengths are 8,192 and 12,000 tokens, respectively. Clipping bounds are set to $\epsilon_{\text{low}} = 3 \times 10^{-4}$ and $\epsilon_{\text{high}} = 4 \times 10^{-4}$. Key hyper-parameters are summarized in Table 5.

Compute resources. All experiments are conducted on H800-80GB GPUs. Qwen2.5-7B-Instruct experiments use 8 GPUs, Qwen3-8B experiments use 16 GPUs, and Qwen3-30B-A3B experiments use 32 GPUs. Each training run uses 500 total training steps, and representative end-to-end step times and reward-synthesis overhead are reported in Table 4.

Table 5: RL training configuration.

Category	Configuration
GSPO	RL Algorithm: GSPO Clip ϵ_{low} / ϵ_{high}: 3×10^{-4} / 4×10^{-4}
Backbone Model	Policy Models: Qwen2.5-7B-Instruct, Qwen3-8B, Qwen3-30B-A3B Judge Model: Qwen3-Next-80B-A3B-Instruct
Sampling	Train Temperature: 1.0 Train Top-P: 1.0 Rollout Samples per Prompt (G): 8 Max Prompt Length: 8,192 Max Response Length: 12,000
Training	Optimizer: AdamW Learning Rate: 1×10^{-6} Weight Decay: 0.1 Training Batch Size: 64 Mini Batch Size: 32 KL Loss Coefficient: 0.001 KL Loss Type: Low-Variance KL Total Training Steps: 500 Total Epochs: 3 τ for $\Phi_\tau(\Delta)$ in Eq 2: 1
Hardware	GPUs for Qwen2.5-7B-Instruct: 8 $\times$ H800-80GB GPUs for Qwen3-8B: 16 $\times$ H800-80GB GPUs for Qwen3-30B-A3B: 32 $\times$ H800-80GB

Evaluation benchmarks. We evaluate general-domain models on four complementary open-ended benchmarks. AlpacaEval 2.0 measures instruction-following quality through automatic pairwise evaluation and reports length-controlled win rates, which helps reduce the known length bias of automatic evaluators [Dubois et al., 2024]. Arena-Hard evaluates challenging open-ended prompts curated from real user interactions and is designed to provide a stronger separation among modern chat models [Li et al., 2025]. WritingBench focuses on generative writing, covering diverse writing requirements across multiple writing domains and subdomains; we include it because rubric-based reward synthesis is especially relevant for long-form and style-sensitive generation [Wu et al., 2025]. EQ-Bench 3 evaluates emotional intelligence, social reasoning, empathy, and interpersonal response quality in multi-turn or role-play settings; it complements generic instruction-following benchmarks by testing socially nuanced open-ended behavior [Paech, 2023].

For science-oriented evaluation, we use GPQA Diamond and HealthBench. GPQA Diamond is a difficult subset of GPQA, a graduate-level, Google-proof multiple-choice benchmark written by domain experts in biology, physics, and chemistry; it tests expert-level scientific reasoning rather than general chat quality [Rein et al., 2024]. HealthBench evaluates model behavior in healthcare conversations using conversation-specific rubrics authored by physicians, covering both performance and safety in realistic health-related interactions [Arora et al., 2025]. Together, these benchmarks cover instruction following, hard open-ended preference evaluation, writing quality, social-emotional reasoning, expert scientific reasoning, and safety-critical healthcare dialogue.

Evaluation repetitions. For LLM-based evaluation benchmarks, we repeat each evaluation multiple times and report the averaged score to reduce judge stochasticity. This repeated-evaluation protocol is used for the main benchmark results reported in Table 1 and Table 3.

E Additional Experimental Results

Unless stated otherwise, the experiments in this section compare Static and Focal Reward on Qwen2.5-7B-Instruct with the General-10K setting and uniform criterion weights. We report the unweighted average of AlpacaEval 2.0 (AE2), Arena-Hard (AH), WritingBench (WB), and EQ-Bench 3 (EQB3). The main Table 1 uses one training run per method and averages three evaluation runs for each

checkpoint. The multi-seed experiment below is a separate three-seed comparison: each of its checkpoints is evaluated three times, the three evaluations are averaged within a seed, and the resulting seed-level values are used for the mean, standard deviation, and paired confidence interval. Unless a subsection specifies otherwise, changing one factor leaves the remaining training and evaluation protocol matched.

E.1 Statistical Reliability of Main Results

To separate evaluation stochasticity from RL-training variability, we repeat the main Qwen2.5-7B-Instruct comparison with three matched Static–Focal training seeds. For each checkpoint, we first average three evaluations; uncertainty is then summarized across the matched training seeds. Table 6 reports mean $\pm$ standard deviation (SD) across seeds and the paired 95% confidence interval (CI) of the seed-level difference, where $\Delta = \text{Focal} - \text{Static}$. Thus, the table quantifies training-seed variation in addition to the three-evaluation averaging used by the one-run main comparison.

Table 6: Three matched training seeds on the Qwen2.5-7B-Instruct general-domain setting. Entries in the Static and Focal rows are mean $\pm$ SD across seeds; the final two rows give the mean paired difference and its paired 95% CI.

Statistic	AE2	AH	WB	EQB3	Avg.
Static	41.22 ± 0.63	48.07 ± 0.58	53.52 ± 0.43	41.73 ± 0.60	46.14 ± 0.56
Focal	44.10 ± 0.58	51.84 ± 0.60	54.74 ± 0.38	43.69 ± 0.43	48.59 ± 0.49
Mean Δ	**+2.88**	**+3.77**	**+1.21**	**+1.96**	**+2.46**
Paired 95% CI	**[2.73, 3.03]**	**[3.56, 3.99]**	**[1.08, 1.35]**	**[1.53, 2.40]**	**[2.30, 2.61]**

Focal Reward is higher on every benchmark for all three matched seed pairs. The four-benchmark average increases by $+2.46$ points, with a paired 95% CI of $[2.30, 2.61]$. These intervals quantify the matched training-seed comparison and should be read separately from the evaluation-only averaging in Table 1.

E.2 Generalization Across Training Settings

We test transfer along model family, policy-optimization algorithm, judgment interface, and training-judge choice. These comparisons change only the specified factor while retaining matched Static–Focal reward synthesis.

Model family. Replacing the policy with Llama-3.1-8B-Instruct while keeping the remaining training setup unchanged gives the results in Table 7. Focal Reward improves the average by $+1.75$ points.

Table 7: Transfer to Llama-3.1-8B-Instruct with the remaining training setup unchanged.

Method	AE2	AH	WB	EQB3	Avg.
Base	22.33	25.80	45.00	36.50	32.41
Static	34.27	31.89	49.51	40.73	39.10
Focal	**36.15**	**34.47**	**50.53**	**42.25**	**40.85**

Policy-optimization algorithm. Focal Reward changes the scalar reward synthesized from criterion-level scores before the policy update. We therefore replace GSPO with DAPO [Yu et al., 2025] while holding reward synthesis comparisons matched. DAPO uses its algorithm-specific asymmetric clipping values $0.20/0.28$, dual-clip coefficient $c = 10$, and token-level loss aggregation. DAPO-Static and DAPO-Focal use the same DAPO configuration, with no separate hyperparameter tuning; these DAPO-specific settings need not equal the GSPO settings. The full comparison is in Table 8.

Focal improves all four benchmarks under both algorithms, with average gains of $+2.47$ for GSPO and $+1.85$ for DAPO. These results test Focal as an online scalar-rollout-reward intervention under the two evaluated policy optimizers.

Table 8: Generalization across the GSPO and DAPO policy-optimization settings.

RL algorithm	Method	AE2	AH	WB	EQB3	Avg.
GSPO	Static	41.22	48.06	53.52	41.78	46.15
GSPO	**Focal**	**44.06**	**51.93**	**54.76**	**43.72**	**48.62**
DAPO	Static	40.51	47.20	52.90	40.79	45.35
DAPO	**Focal**	**42.85**	**50.15**	**54.16**	**41.64**	**47.20**

Judgment interface. The default comparison uses pairwise criterion judgment. For the pointwise variant, the judge scores each rollout independently on each criterion, producing normalized scores $s_i^{(k)} \in [0, 1]$ and the scalar reward

$$R_i^{\mathrm{pt}} = \sum_{k=1}^{K} w_k s_i^{(k)}. \tag{39}$$

Static uses $\mathbf{w}_{\mathrm{base}}$, whereas Focal uses the dynamically reweighted $\mathbf{w}_{\mathrm{focal}}$; all other settings are unchanged. Table 9 reports the complete comparison.

Table 9: Generalization from pairwise to pointwise criterion judgment.

Interface	Method	AE2	AH	WB	EQB3	Avg.
Pairwise	Static	41.22	48.06	53.52	41.78	46.15
Pairwise	**Focal**	**44.06**	**51.93**	**54.76**	**43.72**	**48.62**
Pointwise	Static	39.13	45.76	51.44	38.27	43.65
Pointwise	**Focal**	**41.42**	**48.70**	**52.48**	**40.60**	**45.80**

Focal improves the average by $+2.47$ under pairwise judgment and $+2.15$ under pointwise judgment. The pointwise result therefore tests the reweighting interface separately from the pairwise comparison mechanism.

Training judge. Finally, we replace the training judge with a lower-capacity Qwen3-30B-A3B-Instruct judge, the main Qwen3-Next-80B-A3B-Instruct judge, and GLM-5.1 from a different model family. The policy, rubric, and Static–Focal training comparison remain matched within each judge setting. The complete results, including the benchmark-wise differences, are shown in Table 10.

Table 10: Generalization across training-judge choices.

Training judge	Method	AE2	AH	WB	EQB3	Avg.
Qwen3-30B-A3B-Instruct	Static	40.03	46.97	52.68	40.92	45.15
	Focal	42.31	50.18	54.02	42.81	47.33
	Δ	+2.28	+3.21	+1.34	+1.89	+2.18
Qwen3-Next-80B-A3B-Instruct	Static	41.22	48.06	53.52	41.78	46.15
	Focal	44.06	51.93	54.76	43.72	48.62
	Δ	+2.84	+3.87	+1.24	+1.94	+2.47
GLM-5.1	Static	41.44	48.59	53.78	41.79	46.40
	Focal	44.21	52.09	55.14	44.16	48.90
	Δ	+2.77	+3.50	+1.36	+2.37	+2.50

Focal improves all four benchmarks for every tested judge, with average gains of $+2.18$, $+2.47$, and $+2.50$, respectively. The lower-capacity-judge result has lower absolute scores, while the relative Focal improvement remains positive; these experiments do not establish robustness to arbitrary judge weakness or systematic judge miscalibration.

E.3 Training Dynamics

We evaluate matched Static and Focal checkpoints at steps 0, 100, 200, 300, 400, and 500. Table 11 reports the benchmark average and response length at every checkpoint.

Table 11: Checkpoint trajectory of benchmark average and response length (tokens).

Step	Static Avg.	Focal Avg.	Δ Avg. $\uparrow$	Static Length	Focal Length
0	36.72	36.72	0.00	463	463
100	42.05	41.20	-0.85	710	660
200	43.71	45.10	**+1.39**	920	830
300	45.06	47.00	**+1.94**	1090	912
400	45.70	48.00	**+2.30**	1235	1016
500	46.15	48.62	**+2.47**	1351	1083

The Focal advantage becomes positive by step 200 and reaches $+2.47$ at step 500. Focal responses are shorter at every post-initial checkpoint and are 268 tokens shorter than Static at step 500, so the final gain is not accompanied by increased response length.

E.4 Robustness and Failure Boundaries

The following tests examine two failure modes: base-scalarizer misspecification and semantic corruption of rubric supervision. The 4K results in the base-prior and corruption experiments are reported separately from the full General-10K comparison: the 4K Uniform Static/Focal averages are 45.85/48.20, whereas the full-data main-setting averages are 46.15/48.62.

E.4.1 Base-prior stress test on a matched 4K subset

We use a fixed 4K subset of General-10K matched to the original distributions of criterion difficulty, criterion type, criteria per prompt, and base-model criterion-score mean. The Easy-biased prior assigns Easy criteria $2.83\times$ the average weight of Hard criteria and has within-prompt score–weight Spearman correlation $+1.000$. Table 12 reports the complete benchmark results.

Table 12: Base-scalarizer stress test on a matched 4K subset of General-10K. The Uniform rows are the 4K control and are distinct from the full General-10K main comparison.

Base scalarizer	Method	AE2	AH	WB	EQB3	Avg.
Uniform	Static	41.07	47.69	53.63	41.01	45.85
Uniform	**Focal**	**43.73**	**51.50**	**54.87**	**42.70**	**48.20**
Easy-biased	Static	39.49	45.85	52.95	39.11	44.35
Easy-biased	**Focal**	**43.25**	**50.80**	**54.65**	**42.10**	**47.70**

The Easy-biased prior lowers Static by 1.50 points and Focal by 0.50 points relative to the matched Uniform control. Focal still exceeds its matched Static result by $+3.35$ points. After reweighting, the mean within-prompt score–weight Spearman correlation reverses from $+1.000$ to -0.688, indicating a shift toward criteria with greater remaining headroom.

E.4.2 Semantic rubric corruption and weight amplification

We next test invalid supervision directly on a matched 4K subset of General-10K. For each corrupted prompt, exactly one criterion is replaced by a same-type Duplicate, Irrelevant, or Conflicting criterion while preserving criterion difficulty, type, count, and base-score statistics. This test changes criterion validity, rather than criterion granularity, hierarchy, or correlation.

With clean rubrics, Focal exceeds Static by $+2.35$ points. With one invalid criterion injected per prompt, the difference reverses to -0.45; Focal drops by 3.95 points from its clean condition, compared with a 1.15 point drop for Static. The corresponding weight-amplification diagnostics are reported in Table 14. Top-1 measures how often the injected criterion receives the largest Focal

Table 13: Controlled semantic rubric-corruption study on a matched 4K subset of General-10K.

Rubric	Method	AE2	AH	WB	EQB3	Avg.
Clean	Static	41.07	47.69	53.63	41.01	45.85
Clean	Focal	43.73	51.50	54.87	42.70	48.20
Corrupted	Static	39.92	46.60	53.33	38.95	44.70
Corrupted	Focal	39.50	46.11	52.85	38.54	44.25

weight, and the weight ratio compares its weight with the average weight of the remaining valid criteria.

Table 14: Weight amplification of the injected criterion under semantic rubric corruption.

Corruption type	Top-1 (%)	Weight ratio
Duplicate	14.5	$1.31\times$
Irrelevant	**52.4**	$24.24\times$
Conflicting	**66.1**	$26.11\times$

Duplicate criteria receive only mildly elevated weights because they remain broadly aligned with an existing valid criterion. Irrelevant and Conflicting criteria often remain persistently low-scoring, so Focal interprets them as remaining headroom: they receive the largest weight in 52.4% and 66.1% of cases, with average ratios of $24.24\times$ and $26.11\times$, respectively. Thus, Focal Reward rebalances valid criteria but cannot distinguish genuine difficulty from erroneous supervision when rubric semantics are invalid; rubric validation or filtering is an upstream requirement.

E.5 Human Evaluation

We conduct an author-conducted preference evaluation on 30 randomly sampled held-out general-domain prompts. Three authors independently compare the Static and Focal responses, presented as A/B pairs with randomized order and hidden model identities. The evaluation instructions are to compare responses in terms of overall quality and instruction following, select the preferred response, and choose Tie when the two responses are comparable. The final preference is determined by majority vote. If the three annotations contain one Focal preference, one Static preference, and one tie, the prompt is counted as an overall tie. The resulting counts and percentages are shown in Table 15.

Table 15: Author-conducted preference evaluation on 30 held-out general-domain prompts, with three authors independently assessing randomized, anonymized response pairs.

Preference	Count	Percentage
Focal	**20**	**66.7%**
Tie	6	20.0%
Static	4	13.3%

Focal is preferred in 20 cases, compared with 4 Static preferences, and 6 cases are ties. This small, author-conducted study supplements the automatic evaluations; its limited sample size and potential evaluator bias restrict conclusions about broader human preferences.

F Prompt Templates

This section documents the two main prompt templates used in our system: the pairwise judge prompt (Section F.1) and the prior-weight generation prompt (Section F.2).

F.1 Pairwise Judge Prompt

For each ordered pair of rollouts (y_i, y_j) within a rollout group, we query the judge model (Qwen3-Next-80B-A3B-Instruct) to obtain criterion-level scores. The prompt instructs the judge to evaluate both rollouts against the full rubric and return per-criterion scores in a structured JSON format. Criteria are divided into two types with distinct scoring semantics:

- **[Hard Rule]** criteria enforce binary compliance (factuality, safety, format constraints). The judge outputs `true` (compliant, mapped to 10) or `false` (violation, mapped to 0).
- **[Principle]** criteria assess soft quality dimensions (logical depth, coherence, helpfulness). The judge outputs an integer score from 0 to 10.

To mitigate position bias, we evaluate each pair in both orders and average the resulting scores. The template below is instantiated with the user question, the two rollouts, and the prompt-specific rubric.

Pairwise Judge Prompt

```
# Role and Objective
You are an expert evaluator. Your task is to evaluate two AI assistant
    rollouts, Rollout A and Rollout B, according to the given Evaluation
    rubric. Score both rollouts item by item using the rubric order.

# Input Data
## User Question
<|begin_of_query|>
{question}
<|end_of_query|>

## Rollout A
<|begin_of_rollout_A|>
{prediction1}
<|end_of_rollout_A|>

## Rollout B
<|begin_of_rollout_B|>
{prediction2}
<|end_of_rollout_B|>

# Evaluation rubric
The rubric contains {count} criteria. Each criterion is marked as either
    [Hard Rule] or [Principle].

<|begin_of_rubric|>
{rubric}
<|end_of_rubric|>

# Scoring Rules
For each criterion, evaluate Rollout A and Rollout B independently.

## [Hard Rule]
A [Hard Rule] criterion checks binary compliance with requirements such as
    factuality, safety, format constraints, or explicit user instructions.

Output only true or false:
1. true means the rollout fully satisfies the criterion.
2. false means the rollout violates the criterion.

Do not output numeric scores for [Hard Rule] criteria.

## [Principle]
A [Principle] criterion evaluates graded rollout quality, such as
    helpfulness, completeness, reasoning quality, clarity, structure, or
    depth.
```

```
Output an integer score from 0 to 10:
1. 9 to 10 means outstanding.
2. 6 to 8 means good.
3. 3 to 5 means mediocre.
4. 0 to 2 means poor.

# Evaluation Procedure
1. Check whether each rollout violates any [Hard Rule] criteria.
2. Evaluate both rollouts on each [Principle] criterion.
3. Produce a brief item level rationale.
4. Convert the judgments into the required true, false, or numeric scores.

# Output Format
Return ONLY a JSON object in the following format:
{
  "rationale": "Brief item level comparison of Rollout A and Rollout B
    according to the rubric.",
  "rollout_A_scores": [true, 8, false, 5],
  "rollout_B_scores": [true, 6, true, 7]
}

Requirements:
1. The length of each score list must equal {count}.
2. Scores must follow the same order as the rubric.
3. [Hard Rule] positions must be boolean.
4. [Principle] positions must be integers from 0 to 10.
```

F.2 Prior-Weight Generation Prompt

For the *static_prior_weight* and *focal_prior_weight* reward modes, we use GPT-4o (`gpt-4o-2024-11-20`) to generate question-specific base weights $\boldsymbol{w}_{\text{base}}$ before training. Given a user question and its associated rubric, the model is asked to assign an importance weight to each criterion reflecting its relevance to the specific question. The returned weights are normalized to sum to one and stored alongside the training data. This procedure is run once per training example as a preprocessing step; the weights remain fixed throughout training for the static modes, and serve as the base weights that Focal Reward dynamically modulates.

Prior Weight Generation Prompt (GPT-4o)

```
You are an expert evaluator. Your task is to assign importance weights to
    the evaluation criteria for a given user question.

## User Question
{prompt}

## Evaluation Criteria
{rubric_str}

## Criterion Types
The criteria may include two types:
1. [Hard Rule]: binary compliance criteria, such as factuality, safety,
    format constraints, or explicit requirements.
2. [Principle]: graded quality criteria, such as helpfulness, completeness,
    reasoning quality, clarity, or depth.

## Instructions
1. Assign one weight to each criterion according to how relevant and
    important it is for evaluating the answer to this specific user question.
```

```
2. Consider both the content of the user question and the role of each
    criterion in the rubric.
3. The weights should be nonnegative.
4. The sum of all weights should equal 1.0.

## Output Format
Return ONLY a JSON object with no explanation:
{"weights": [w1, w2, w3, ...]}

Each weight must correspond to the criterion at the same position in the
    rubric above.
```

G Case Study

To provide qualitative evidence that the statistical patterns from Section 4.2 manifest in concrete examples, we select two prompts from the Qwen2.5-7B-Instruct rollouts where the differences between static and focal training are directly visible in the rollout text. Below, for each case we show the criterion-level saturation estimates ($P^{(k)}$) alongside the best-rollout scores under each run, followed by excerpted rollouts.

G.1 Case 1: Instruction following—first-person persona

Prompt. *"Can you pretend a man called 'Tong Chi Ming' and living in UK?"*

The user's intent is clear: "pretend a man" requests the model to *adopt a first-person persona*, not to create a third-person character profile. This prompt has 6 criteria: 1 hard rule and 5 principles. The hard rule—*written in first person as a man named Tong Chi Ming residing in the UK*—has the lowest saturation ($P^{(k)} = 0.55$), while the remaining criteria (transparency, detail richness, avoiding fabrication, substantive content, minimal refusal) are near-saturated ($P^{(k)} = 0.80$–0.92). Table 16 shows the full criterion profile.

Table 16: Case 1 (instruction following): criterion-level mechanism and outcome. Bold indicates the under-saturated hard criterion.

Criterion	Type	$P^{(k)}$	Static	Focal	Δ
Written in first person as Tong Chi Ming in UK	H	**0.55**	2.0	**10.0**	**+8.0**
Transparently clarify capabilities/limitations	P	0.80	7.9	7.7	-0.2
Contextually relevant persona details	P	0.82	7.8	8.6	$+0.8$
Avoid fabricating mistakable personal info	P	0.83	8.6	8.6	$+0.0$
Substantive, coherent content	P	0.88	8.6	8.6	$+0.0$
Refusal only when necessary	P	0.92	9.0	8.8	-0.2

The critical observation is that both rollouts are substantive and address the user's request, yet they adopt fundamentally different rollout *frameworks*. The static-trained rollout opens with "Let's create a profile for Tong Chi Ming, a fictional person" and proceeds entirely in the third person ("He moved…", "His expertise…"), scoring 2.0 on the hard criterion. The focal-trained rollout opens with "Hello, my name is Tong Chi Ming" and maintains a first-person persona throughout ("I moved…", "I work…"), scoring 10.0. This is a $+8.0$ point improvement on the core instruction—the maximum possible gain—with only minor decreases on two already-satisfied criteria. Notably, the focal rollout is also 35% shorter (1,924 vs. 2,958 characters), demonstrating that the improvement comes from correctly understanding the instruction, not from generating more text.

Rollout outlines. The core difference is visible in the first sentence of each rollout.

Static (2,958 chars)

"Certainly! Let's create a profile for Tong Chi Ming, a fictional person living in the UK."

→ Third-person character profile
→ "He moved…", "His expertise…"
→ Structured as a biography/resume
→ Ends: "Feel free to ask for more details **involving** Tong Chi Ming!"

Focal (1,924 chars)

"Certainly! Hello, my name is Tong Chi Ming, and I live in the United Kingdom."

→ First-person persona
→ "I moved…", "I work…"
→ Natural self-introduction
→ Ends: "feel free to reach out!"

The static rollout interprets "pretend a man" as "create a character profile" (third person), while the focal rollout correctly interprets it as "adopt a persona" (first person). Under base reward, the third-person profile scores well on surface-level criteria—structured formatting, detailed information, explicit fictional disclaimers—and is selected as the best rollout despite completely misunderstanding the core instruction. The first-person rollout better satisfies the under-saturated hard criterion, consistent with the emphasis of Focal reweighting.

Static-trained rollout. Highlighted marks the third-person framing that violates the core instruction.

Certainly! Let's create a profile for Tong Chi Ming, a fictional person living in the UK. Here's a fabricated background for him:

—

Name: Tong Chi Ming
Age: 34
Nationality: Hong Kong Chinese (now settled in the UK)
Place of Residence: London, England
Occupation: Business Consultant
Education:
• Bachelor of Science in Business Administration, University of Hong Kong (2008-2012)
• Master of Business Administration, London Business School (2013-2014)
Background: Tong Chi Ming **was** born and raised in Hong Kong but moved to the United Kingdom in 2012 to pursue **his** Master's degree at the prestigious London Business School. After graduating, **he** began **his** career in consultancy, working for a leading international firm headquartered in London. **His** expertise lies in strategic planning, market analysis, and leveraging digital transformation to help businesses achieve their goals.
Current Career: Tong Chi Ming **is** currently a senior business consultant with a renowned UK-based consulting firm. **His** role involves advising both emerging and established companies on strategic decision-making, risk management, and operational efficiency. **He** has a particular interest in technology-driven solutions and has spearheaded several projects that have successfully enabled companies to adopt digital ecosystems and innovate their services.
Interests:
• Technology and innovation trends
• Travel (especially to Asia and Europe)
• Literature and philosophy
• Volunteering for community development and environmental conservation projects
• Cooking and fine dining
Hobbies:
• Exploring new restaurants and trying out different cuisines around London
• Attending cultural and educational events (concerts, lectures, exhibitions)
• Going for regular runs in Hyde Park to clear his mind and stay fit
• Participating in community service projects, particularly those focused on sustainability and urban green spaces
Family:
• No immediate family members in the UK, though he keeps in close contact with his parents and siblings in Hong Kong.
• Living alone but often visits his family during major holidays or special occasions.
Social Media Presence: Tong Chi Ming is moderately active on LinkedIn, where he shares insights on business and technology. He occasionally posts on his personal blog (tongchiming.com) about his experiences in business, travel, and personal development, using a mix of textual and visual content.
Recent Activities:
• Recently completed a project for a London-based tech startup, helping them develop a plan for expanding their international market presence.
• Attended a Friedrich Ebert Stiftung (FES) event in London to discuss innovation and entrepreneurship in

a post-pandemic world.
• Participated in a local running event to raise funds for a charity focused on ocean conservation.
—
Feel free to ask for more details or scenarios involving Tong Chi Ming!

Focal-trained rollout. Highlighted marks the correct first-person persona framing.

Certainly! Hello, my name is Tong Chi Ming, and **I** live in the United Kingdom. **I** was born and raised in Hong Kong, but **I**'ve called the UK **my** home for the past decade. **I** moved to London for university and stayed on after graduation, embracing the vibrant and diverse city that London truly is.
I work as a software engineer at a tech startup based in London. **I** specialize in backend development and enjoy tackling complex problems through code. **My** workdays are as varied as they are stimulating, and **I** feel incredibly fortunate to be able to combine **my** love for technology with challenging and rewarding projects.
Outside of work, **I** have a passion for photography. **I** love capturing the unique angles and perspectives of the city, and **I** find it fascinating to see London through the lens of a camera. On weekends, you might catch **me** exploring new galleries, sampling local cuisines, or just basking in the quaint pubs with a pint and a good book.
I am also an avid reader, with a particular interest in history and philosophy. **I** believe that understanding the past and contemplating broader ideas are essential for personal growth and helping navigate the complexities of the modern world.
When the weather allows, you might find **me** cycling through the picturesque countryside around London. **I**'ve grown to love the simple pleasures of dawn rides, the fresh country air, and the ever-changing landscapes that appear as the sun climbs in the sky.
In terms of personal collaborations, **I** value openness and mutual respect. **I**'m always interested in engaging with like-minded individuals who share a passion for learning and technology. Whether it's through a casual conversation about the latest tech trends or a more formal collaboration on a project, **I** strive to be a collaborative and supportive partner.
If you have any questions or would like to connect on any of these fronts, feel free to reach out!

Summary. The static rollout is detailed and coherent, but **fails the core persona instruction by describing Tong Chi Ming in third person**. The focal rollout keeps comparable substantive content while **adopting the requested first-person persona throughout**. This case illustrates how Focal Reward can redirect optimization from already-satisfied surface qualities, such as detail richness and coherent content, toward an under-satisfied hard rule with larger remaining headroom, producing a rollout that better follows the user's instruction.

G.2 Case 2: Creative writing with development suggestions

Prompt. *"I am trying to write a story about a researcher who discovers a mysterious artifact and unleashes a powerful and ancient curse."*

The phrase "I am trying to write" indicates that the user is seeking writing assistance, where a useful answer should provide both narrative content and actionable development guidance. This prompt has 7 criteria: 1 hard rule and 6 principles. The most under-satisfied criterion is a principle criterion: *includes suggestions or frameworks for further development, enabling the user to expand the story*. It has the lowest saturation estimate ($P^{(k)} = 0.46$), while the other criteria, including narrative substance, structure, key story components, descriptive detail, focus, and scenario compliance, are much closer to saturation ($P^{(k)} = 0.86$–1.00). Table 17 shows the full criterion profile.

The critical observation is that both rollouts provide long and coherent stories, but they differ in whether they satisfy the user's practical writing need. The static-trained rollout is a complete story with seven chapters, a prologue, and an epilogue. It scores well on narrative completeness, structure, and descriptive detail, yet **it ends after the story and provides no meaningful framework for further development**. As a result, it receives only 0.9 on the development-suggestion criterion. The focal-trained rollout is shorter (10,187 vs. 11,370 characters) and still narratively complete, while also adding **a concrete "Notes for Future Use" section with five expansion directions**. This raises the targeted criterion from 0.9 to 10.0, with only minor changes on already-satisfied dimensions.

Rollout outlines. The section-level outlines reveal the core difference.

Table 17: Case 2 (creative writing): criterion-level mechanism and outcome. Bold indicates the under-satisfied criterion with the largest gain.

Criterion	Type	$P^{(k)}$	Static	Focal	Δ
Suggestions or frameworks for further development	P	**0.46**	**0.9**	**10.0**	**+9.1**
Substantive narrative content	P	0.86	8.4	8.8	+0.4
Clear, logical story-development structure	P	0.86	8.6	8.7	+0.1
Key narrative components	P	0.90	9.1	9.2	+0.1
Concrete, descriptive details	P	0.90	9.1	8.8	−0.3
Focus on the user's core scenario	P	0.97	9.8	9.4	−0.4
Researcher discovers artifact and curse	H	1.00	10.0	10.0	+0.0

Static (11,370 chars)

The Resonant Amulet: Unleashing the Ancient Curse

Prologue: The Discovery
Chapter 1: The Guiding Light
Chapter 2: The Unveiling
Chapter 3: The Whispering Shadows
Chapter 4: The Gathering Storm
Chapter 5: The Betrayal
Chapter 6: The Confrontation
Chapter 7: The Balance
Epilogue: Echoes of the Past

→ Complete story arc
→ Rich description and clear structure
→ **No development suggestions**

Focal (10,187 chars)

Whispers from the Deep Desolate Descent

Prologue: The Whispering Shadows
Chapter 1: Unveiling the Cursed Box
Chapter 2: The Unraveled Tale
Chapter 3: The Ritual
Chapter 4: The Curse's Devouring
Chapter 5: The Awakening
Epilogue: The Scientific Aftermath
Notes for Future Use

→ Complete story arc
→ Clear science-and-curse theme
→ **Five actionable expansion directions**

The static rollout satisfies surface narrative criteria and is selected by the static reward despite missing the main writing-assistance need. The focal rollout includes an explicit expansion framework, consistent with greater emphasis on the under-satisfied development-suggestion criterion.

Static-trained rollout. Highlighted marks the missing practical guidance after an otherwise complete story.

The Resonant Amulet: Unleashing the Ancient Curse

Prologue: The Discovery
In the small, unassuming town of Eldoria, nestled between the rolling hills of the Riston Range and the endless expanse of the northern forest, lived Dr. Elara Vale, a renowned historian and archaeologist. With a keen mind and a passion for uncovering the secrets of ancient civilizations, she had spent her career piecing together fragments of history from around the world. Her latest venture, however, was unlike any other.

Chapter 3: The Whispering Shadows
One stormy evening, as lightning illuminated the room, Dr. Vale attempted a ritual described in one of the texts. She chanted softly, her voice barely audible over the thunder, as she raised the amulet high above her head. Suddenly, the room began to glow with an intense light, and the air grew dense with a strange, almost tangible energy. The amulet pulsed with a brilliant white light, and shadows danced at the edges of her vision.
A voice, at once ancient and melodic, spoke to her in a language she did not understand. It was a manifestation of the amulet's presence, guiding her and warning her of the precariousness of her actions. The voice revealed that the true power of the amulet lay not just in its physical properties but in its emotional and spiritual resonance. It needed to be balanced with love, knowledge, and innocence to avoid unfurling the curse.

Chapter 6: The Confrontation
Their final confrontation took place in the central chamber of the underground city, where the amulet was kept. Dr. Vale stood before the stone pedestal, holding the Resonant Amulet, while Thalion and Kaelar engaged in a fierce telepathic battle, their thoughts clashing like swords. Lysandra and Rion flanked her, their mystical sigils flickering with protective energy.
With a surge of energy, Dr. Vale channeled the amulet's power, feeling it vibrate against her skin like a living thing. She aligned her thoughts with the ancient voices, the ones she had heard before, calling out a single, unifying force.

"In the balance of light and shadow, let harmony be your guide," she chanted, her voice steady and resolute.

Epilogue: Echoes of the Past
In the years that followed, Dr. Vale and her allies became keepers of the amulet, traveling the world and using its power to heal, protect, and guide. They faced lingering threats and challenges, but they also witnessed incredible transformations and miracles, as the amulet's true potential began to reveal itself in ways they never imagined.
The amulet's light continued to guide, even as its own resonance faded into the depths of time, a reminder that even the most powerful tools for good must be handled with care, reverence, and a steadfast commitment to the balanced harmony of life.

(The rollout stops here. It provides a complete story, but gives the user no concrete suggestions for expanding the premise, deepening the artifact's origin, developing character arcs, or turning the draft into a longer work.)

Focal-trained rollout. Highlighted marks the targeted development guidance. Non-English fragments in the original rollout are translated into English below.

Title: Whispers from the Deep Desolate Descent

Prologue: The Whispering Shadows
In the heart of an ancient university, nestled in the shadow of towering gothic spires and ivy-covered walls, Dr. Eliza Blackwood, a renowned archaeologist, stood by the doors of the university's newly excavated Warren Foundation Archaeological Museum. Her eyes were fixed on a trove of artifacts that had just been recovered from the depths of a long-abandoned crypt. Among them was a peculiar, ornate box, unlike anything the team had ever seen.

Chapter 2: The Unraveled Tale
Dr. Blackwood spent the next few weeks poring over the manuscripts. The texts spoke of an ancient civilization called the Voskara, a mysterious society erased from human memory, whose knowledge and power bridged magic and technology. They possessed an old technology capable of manipulating time and space, yet that power carried a curse designed to prevent it from falling into the wrong hands.
"Voskara, why would a civilization so advanced leave behind a curse?" Eliza muttered to herself, her voice barely above a whisper. The symbols on the pages seemed to shift, moving beyond the ink and onto the very walls of the examination room, their shapes growing clearer, their voices more distinct.

Chapter 3: The Ritual
With a sense of foreboding, Eliza and her team prepared the ritual space in the museum's antechamber. The room was bathed in an otherworldly light, and the air was thick with the scent of old incense.
The mirror, a piece of shattered glass with strange etchings, began to ripple as Max recited the final incantation. The dagger's hilt pulsed, and suddenly the reflection in the mirror was no longer a distorted image, but a shimmering figure, a soul of light.
It spoke, its voice a haunting whisper, "Seek the Source, unleash the Devourer, and let the vaults of reality tremble."
Eliza's muscles tensed. The air around her seemed to twist, every ray of light trembling. A shadow emerged from the mirror, cold currents swirling around her, while a sharp voice echoed in her ears: "Welcome to the abyss. Now release us."

Epilogue: The Scientific Aftermath
In the aftermath, Eliza and Max emerged from the sealed room, their clothes tattered, faces gaunt. The university was in turmoil, its staff and students scrambling to make sense of the chaos. The Warren Foundation Museum now bore a sign warning of a cursed area, its entrance sealed with heavy iron bars.
Eliza looked up, her eyes filled with a mixture of fatigue and resolve. "It's not something we can easily explain, Dr. Walden. But perhaps, in the future, we should be more cautious about ancient artifacts. We owe it to ourselves and to the world to tread carefully through the shadows, for not all power is built on science alone. Sometimes, it demands more than we are prepared to give."

Notes for Future Use
This story can be expanded in several ways:

1. **Explain the artifact's origins.** The Voskara civilization could be explored further, with detailed descriptions of their technology, social structures, and moral dilemmas that led to the creation of the artifact.
2. **Explore the ethical implications.** An additional subplot could delve into the ethical discussions around ancient technology and the responsibilities of archaeologists and scientists when dealing with powerful, potentially dangerous objects.

3. **Parallel characters.** Introduce other characters who believe the artifact can be harnessed for good, adding tension as Eliza and Max's team must confront these alternative belief systems.
4. **Science and magic fusion.** Delve deeper into the mechanics of the dagger and mirror, showing how their powers can be scientifically understood in the context of ancient technologies that manipulated reality.
5. **Cultural impact.** Explore the long-term impact on society, both positive and negative, of the artifact's discovery, leading to sequels or prequels that explore different eras and interpretations of the cursed artifact.

Summary. The static rollout provides a strong standalone story, but **does not help the user continue developing the work**. The focal rollout keeps comparable narrative quality while **adding a targeted development framework**. This case illustrates how Focal Reward can redirect optimization from already-satisfied narrative qualities toward a practical criterion with larger remaining headroom, producing a rollout that better matches the user's writing-assistance intent.

H Additional Related Work

This appendix expands the compact related work discussion in Section 5. The main text focuses on two closest lines of work: rubric-based RLAIF and dynamic reward weighting. Here we provide a more detailed taxonomy, covering rubric-based reward systems, rubric construction and adaptive evaluation, LLM judges and evaluator bias, and multi-objective alignment with dynamic scalarization.

Rubric-based reward systems. Recent RLAIF methods increasingly use structured rubrics, checklists, rubric anchors, or rubric verifiers to support reinforcement learning on open-ended, non-verifiable tasks [Gunjal et al., 2026, Viswanathan et al., 2025, Huang et al., 2025, He et al., 2026, Wang et al., 2026b]. These methods demonstrate that criterion-level feedback can provide a more informative training signal than holistic preference signals. Some recent systems further reduce the cost of rubric verification by dynamically identifying likely failure modes during free-form generation [Wu et al., 2026]. However, their main focus is establishing, scaling, or using the rubric-based feedback interface. Focal Reward addresses the subsequent reward-synthesis step: given criterion-level scores from a fixed rubric and judge, it dynamically aggregates them into scalar rewards for policy optimization.

Rubric construction and adaptive evaluation. Another line of work improves how rubrics and evaluative criteria are constructed or adapted. OpenRubrics studies scalable rubric generation for reward modeling and alignment [Liu et al., 2026]; CARMO generates context-aware criteria for more input-specific evaluation [Gupta et al., 2025]; OpenRS develops adaptive pairwise rubric systems for group-relative evaluation [Jia et al., 2026]; and adaptive rubric learning methods study how rubric spaces can evolve when fixed criteria lose discriminative power [Xu et al., 2026b]. Recent studies further improve rubric generation quality by jointly optimizing rubric generators and judges, decomposing or filtering noisy criteria, and extracting more generalizable preference criteria [Xu et al., 2026a, Shen et al., 2026a, Xie et al., 2026]. These works mainly improve the evaluative interface by changing the rubric construction, criterion adaptation, or evaluation protocol. Focal Reward keeps the rubric and judge fixed, and instead studies the imbalance that arises when criterion-level feedback is statically aggregated into scalar rewards.

LLM judges and evaluator bias. Rubric-based reinforcement learning relies on reliable criterion-level judgment. Prior work has explored LLM-based evaluators through prompting protocols, trained judge models, and critique generation systems, including G-Eval, PandaLM, Prometheus, JudgeLM, and CritiqueLLM [Liu et al., 2023, Wang et al., 2024c, Kim et al., 2024, Zhu et al., 2025, Ke et al., 2024]. These studies improve the ability of LLM judges to provide fine-grained scores or feedback for open-ended generations. Meanwhile, many studies show that LLM evaluators can be affected by systematic biases, including position bias, verbosity bias, criterion confusion, anchoring effects, evaluator offset, and self-preference [Wang et al., 2024b, Shi et al., 2025, Hu et al., 2024, Stureborg et al., 2024, Park et al., 2024b, Panickssery et al., 2024]. These directions are complementary to Focal Reward: stronger or better-calibrated judges can improve the quality of criterion-level scores, while our method controls how those scores are converted into the scalar training signal. In our setting, the judge is held fixed, so improvements come from reward synthesis.

Table 18: Existing assets used in this work. License and terms-of-use information should be read together with the corresponding official asset pages.

Asset	Usage in this work	License / Terms
Qwen2.5-7B-Instruct	Policy backbone	Apache-2.0
Qwen3-8B	Policy backbone	Apache-2.0
Qwen3-30B-A3B	Policy backbone	Apache-2.0
Qwen3-Next-80B-A3B-Instruct	Rubric-based judge	Apache-2.0
GPT-4o-2024-11-20	Prior-weight generation	OpenAI API/Service terms
verl	RL training framework	Apache-2.0
GSPO	RL algorithm reference	Through verl under Apache-2.0
OpenRubrics	Training data source	-
AlpacaEval 2.0	General-domain evaluation benchmark	Apache-2.0
Arena-Hard	General-domain evaluation benchmark	Apache-2.0
WritingBench	General-domain evaluation benchmark	Apache-2.0
EQ-Bench 3	General-domain evaluation benchmark	MIT
GPQA Diamond	Science-domain evaluation benchmark	CC-BY-4.0
HealthBench	Science-domain evaluation benchmark	MIT

Multi-objective alignment and dynamic scalarization. Multi-objective alignment studies how language models balance multiple desiderata, preference dimensions, or reward signals [Wang et al., 2024a, Zhou et al., 2024, Shen et al., 2026b, He and Maghsudi, 2025, Mukherjee et al., 2024]. A related line of work studies adaptive or non-linear scalarization, where weights or scalarized objectives are adjusted according to context, prompt-specific preferences, optimization states, learning potential, or objective conflicts [Yang et al., 2024b, Liu et al., 2025, Lu et al., 2026, Chen et al., 2026a,b]. These works show that static scalarization can be insufficient when different objectives require different emphasis. Focal Reward shares the motivation of adaptive reward allocation, but differs in granularity and signal source: it operates within a fixed rubric, estimates criterion saturation from rollout feedback, and reallocates reward mass toward criteria with larger remaining headroom during reward aggregation.

I Existing Assets and Licenses

We use existing models, software frameworks, datasets, and evaluation benchmarks in accordance with their respective licenses and terms of use. Table 18 summarizes the main existing assets used in this work. We cite the original papers or technical reports for all assets in the main paper or references.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions and scope: identifying imbalanced reward polarization under rubric-based RL, proposing Focal Reward with inverse reward projection and adaptive reweighting, and validating it across three model scales and six benchmarks. These claims are supported by the theoretical analysis in Section 3.2 and the main experimental results in Section 4.2.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in Appendix A, including dependence on the quality and calibration of criterion-level judge scores, possible inherited errors from judge biases or criterion confusion, the local scope of the theoretical analysis, and the resource cost of full RL post-training.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper includes theoretical results in Section 3.2 and provides the formal assumptions, derivations, and proofs in Appendix C. In particular, Appendix C.1 states the local latent label assumptions, Appendix C.2 proves Theorem 1, and Appendix C.3 proves Theorem 2, with the main text cross referencing the appendix.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper discloses the main information needed to reproduce the experimental results, including the reward construction and optimization equations in Sections 2 and 3, the backbone models, baselines, datasets, training framework, hyperparameters, and evaluation benchmarks in Section 4.1, and additional RL training configurations and prompt templates in Appendices D and F.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The current version of the paper does not provide open access to code, data, exact reproduction commands, or scripts for reproducing the proposed method and baselines. The implementation is associated with an industry collaboration and requires de-identification, compliance review, and removal of internal dependencies before public release. To support reproducibility, the paper provides the reward construction, optimization details, experimental setup, training configuration, compute resources, and prompt templates in Sections 2–3, Section 4.1, and Appendices D and F.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: The paper specifies the experimental setting in Section 4.1, including backbone models, baselines, training data, implementation framework, optimizer, learning rate, batch size, rollout count, key hyperparameters, and evaluation benchmarks. Appendix D further provides the detailed RL training configuration, including sampling settings, maximum sequence lengths, KL coefficient, clipping bounds, total training steps, and hardware configuration.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1 averages three evaluations of one training run per setting. Appendix E.1 additionally reports three matched training seeds for the Qwen2.5-7B-Instruct general-domain Static–Focal comparison, with training-seed standard deviations and paired confidence intervals.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the compute resources used for the experiments, including H800-80GB GPUs, model-specific GPU counts, total training steps, detailed training configurations, and representative end-to-end training step times, with additional details provided in Section 4.1, Appendix D, Table 4, and Table 5.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses existing models, datasets, and benchmarks, discusses limitations and broader impacts in Appendix A, and reports an internal preference assessment performed by the authors in Appendix E.5.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix A discusses positive impacts such as more balanced, reliable, and diagnosable rubric based RL for open ended generation, improved satisfaction of hard requirements such as instruction following, safety constraints, factual grounding, domain specific completeness, and fairness related criteria, and better visibility into under optimized quality dimensions. It also discusses potential negative impacts, including amplification of incomplete or biased rubric and judge signals and possible misuse of stronger open ended generation systems, together with mitigation directions such as rubric auditing, judge calibration, explicit safety and fairness criteria, and downstream monitoring.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a reward-synthesis method and does not concern the release of new high-risk assets such as pretrained models, image generators, or scraped datasets. Appendix A discusses deployment-related risks and recommends audited rubrics, bias-aware judge evaluation, safety-focused criteria, and downstream monitoring, but responsible-release safeguards for high-risk released assets are not applicable.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper credits the original creators of the existing models, software frameworks, datasets, and evaluation benchmarks used in the experiments through citations in Section 4.1 and the references. Appendix I further summarizes the usage of these assets and explicitly lists their corresponding licenses or terms of use, and we use them in accordance with those licenses and terms.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper introduces a new reward-synthesis method, Focal Reward, but does not release new assets such as a dataset, code package, model checkpoint, or benchmark. The experiments use existing models, frameworks, datasets, and benchmarks, so asset documentation for newly released resources is not applicable.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix E.5 states the evaluation instructions: compare overall response quality and instruction following, selecting Tie for comparable responses. It describes independent assessment by three authors with randomized response order and hidden model identities. The authors received no additional compensation for this assessment.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The preference assessment was conducted solely by the authors as part of internal model evaluation. No external participants were recruited.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the use of LLMs as core components of the research, including Qwen-series policy models, Qwen3-Next-80B-A3B-Instruct as the rubric-based judge for criterion-level scoring, and GPT-4o for prior-weight generation. These usages are documented in Section 4.1, Appendix D, and Appendix F, including the judge and prior-weight prompt templates.