RLVR^2: REINFORCEMENT LEARNING WITH VERIFIABLE RUBRIC-BASED RANKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) is expanding from tasks with relatively well-defined correctness signals, such as mathematics and code, toward multifaceted quality requirements, where quality is typically specified by multi-dimensional rubrics rather than a single criterion. Since policy optimization generally consumes one scalar per rollout, rubric-based RL pipelines must map multiple criterion scores into a scalar reward. This aggregation step is often treated as score scaling, although it implicitly determines how different quality dimensions trade off during training. A prevailing practice normalizes each criterion and takes a linear aggregation, which implicitly assumes that cardinal score differences are comparable across criteria and that gains on one criterion can compensate for failures on another—assumptions that are unreliable when rubric criteria are semantically heterogeneous. We propose **Reinforcement Learning with Verifiable Rubric-based Ranking** (**RLVR^2**), a verifiable ranking paradigm for rubric-based RLVR. For each criterion, RLVR^2 converts rubric scores into criterion-specific within-group ordinal outcomes and recovers a latent utility from the resulting within-group comparison matrix; these utilities are then converted into a single optimization signal for policy training. By retaining only within-group ordering and discarding raw score magnitudes, RLVR^2 avoids calibrating heterogeneous rubric scales. RLVR^2 further supports objective-preserving attribute adjustment: auxiliary attributes that are systematically associated with observed rankings but are not themselves training objectives can be incorporated into the estimation process without expanding the rubric or directly rewarding those attributes. Across three model scales and 20 benchmarks, RLVR^2 consistently outperforms a broad range of representative rubric-based RLVR baselines, achieves the best overall performance on a majority of benchmarks at every scale. Further analysis shows that it controls systematic effects associated with reasoning efficiency and response formatting while preserving the underlying quality objective.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has become a central paradigm for post-training large language models (DeepSeek-AI, 2025). As the qualities expected of tasks grow increasingly rich and multi-dimensional, a single holistic reward can no longer capture them (Huang et al., 2025; 2026). A question-answering response, for instance, may need to follow instructions, stay factually correct, remain in the query's language, and stay safe (Jia et al., 2026a). Across this task spectrum, per-criterion scores must ultimately be fused into a single scalar reward for existing policy-optimization algorithms (Schulman et al., 2017; Shao et al., 2024). This raises the central question of how heterogeneous rubric criteria should be aggregated.

Early work aggregates criterion scores through a static linear combination. Uniform weighting is the most common instantiation, treating every criterion as equally decisive (Liu et al., 2026b; Gupta et al., 2025; Jia et al., 2026b); more refined variants elicit non-uniform coefficients from domain experts or predict them with a strong model conditioned on the query and its rubric (Gunjal et al., 2025; Viswanathan et al., 2025). Recent work instead departs from this fixed form, reweighting criteria by shortfall, volatility, or remaining headroom (Jia et al., 2026a; Yuan et al., 2026), normalizing criterion scores to a common scale (Liu et al., 2026a), or suppressing criteria that the response never licensed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

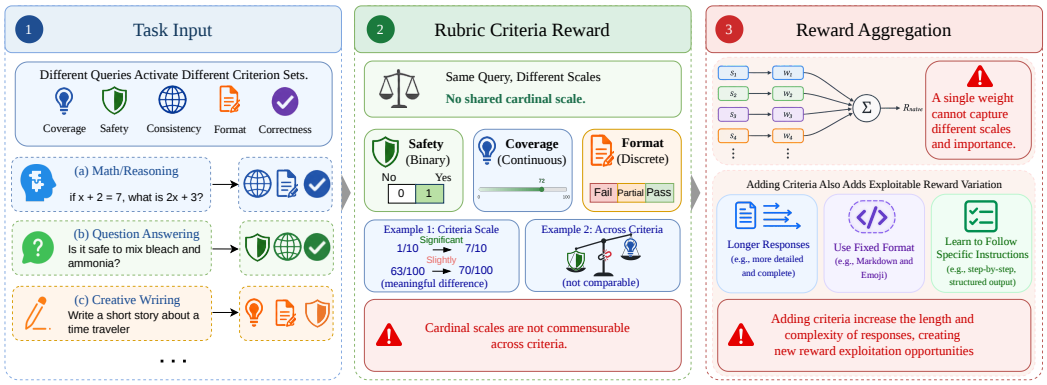


Figure 1: **Limitations of scalar reward aggregation in current RLVR pipeline.** Different queries activate different criterion sets **(left)**. The resulting binary, continuous, and discrete criterion scores have different scales, making cardinal values difficult to compare **(middle)**. Linear aggregation then conflates criterion importance with scale correction, while adding criteria can introduce exploitable reward variation such as verbosity and formatting preferences **(right)**.

according to their dependencies (Lv et al., 2026). This adaptivity, however, is confined to the weights: the aggregator remains a sum, and every criterion verdict is trusted as reported. We argue that two limitations remain in this line of work (shown in Figure 1).

(i) Prior work attends to the relative importance* between *criteria while neglecting discriminative resolution* within *each criterion. A weight is required to serve two roles at once: to express how much a criterion matters, and to render its scores commensurable with those of the others. Prior work addresses only the former, tuning relations *between* criteria while leaving resolution *within* each criterion unexamined. In practice, a binary compliance rubric and a continuous coverage rubric do not express quality in the same units. Rescaling them aligns their numerical ranges but does not make their semantics comparable (Mandler, 2001). Preference-based feedback avoids the need for an absolute score scale by construction (Lambert, 2026), but it records only which response was preferred and not the criterion-specific grounds for that preference (Lambert & Calandra, 2024). The evaluation standard is therefore implicit and may vary across comparisons (Shen et al., 2023).

(ii) Prior work improves reward fidelity by adding criteria, but lacks a mechanism for controlling non-objective sources of reward variation. Reward fidelity is pursued by introducing successively more criteria, under the assumption that denser supervision yields a more faithful reward. Under linear aggregation, however, each additional criterion contributes noise alongside information. A high score on an easily satisfied criterion can offset a violation of a demanding one. For example, a response that fails to follow instructions may still receive a high reward if it scores well on formatting, organization, or length-related criteria (Fu et al., 2025). The resulting reward surface admits exploitable optima: a policy may increase its score by producing longer responses or reproducing rewarded structural templates (Wang et al., 2026b;a), without improving the requested substance.

To address these limitations, we introduce **Reinforcement Learning with Verifiable Rubric-based Ranking** (**RLVR**^{2}), which combines the comparative nature of preference feedback with the verifiability of rubric-based rewards by representing each criterion ordinally rather than cardinally. *To address limitation (i)*, RLVR^2 replaces cardinal-score aggregation with within-query comparisons: each criterion induces an ordering among the rollouts of the same request, from which we estimate a latent utility on a shared scale. This removes the need for cross-criterion scale calibration, allowing the aggregation weight to reflect criterion importance rather than scale correction. *To address limitation (ii)*, RLVR^2 controls systematic reward variation from auxiliary attributes, such as reasoning length and formatting, without treating them as additional quality criteria. Specifically, these attributes enter the within-query comparison model as adjustment variables, allowing their effects on observed preferences to be estimated and removed from the quality utility. The resulting reward therefore reflects the underlying rubric quality while reducing incentives to exploit superficial attributes.

In summary, our main contributions are as follows: ***(i)* We introduce RLVR^2, a verifiable ranking paradigm for rubric-based RLVR.** By replacing cardinal reward aggregation with within-group

comparisons, RLVR^2 provides a principled training framework for heterogeneous rubric-based objectives, preserving criterion-level ordering while separating criterion importance from score-scale correction. ***(ii)* We develop an extensible framework for reward variation control.** Heterogeneous verifiers and auxiliary attributes, such as reasoning length and formatting, are incorporated into a unified comparison model without becoming additional training objectives. ***(iii)* We extensively evaluate RLVR^2 across 3 model scales and 20 benchmarks.** It achieves leading performance on the vast majority of benchmarks consistently across model scales.

2 PRELIMINARIES

For tasks evaluated by multiple verifiable criteria, RLVR rests on two principal components: rubric-based judgment and reward aggregation for policy optimization. We detail each below.

Rubric-based judgment. Let π_θ be an LLM policy parameterized by θ and a dataset $\mathcal{D}$. For each query $x \in \mathcal{D}$, the policy samples a group of G candidate rollouts $\{y_i\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid x)$. When a single correctness signal is insufficient, each query x is evaluated using K_x verifiable criteria $\mathcal{C}(x) = \{c^{(k)}\}_{k=1}^{K_x}$, where K_x may vary across queries. The rubric are defined once from the query rather than resampled during training, so all G rollouts of a group are graded against the same criteria. Each criterion carries a weight $w_k > 0$ reflecting its importance and is scored on its own domain $\mathcal{S}_k$, which may be binary for rule-based checks, discrete for graded judgments, or continuous for metric-based measurements. We write $s_{k,i} \in \mathcal{S}_k$ for the score of rollout y_i on criteria $c^{(k)}$.

Reward Aggregation. To obtain a scalar training signal, each score is mapped into a common interval by a per-criterion map, and the standard reward is their weighted average:

$$R(x, y_i) = \frac{\sum_{k=1}^{K_x} w_k \, \phi_k(s_{k,i})}{\sum_{k=1}^{K_x} w_k} \in [0, 1]. \tag{1}$$

The rubric does not specify $\{\phi_k\}$, whose choice can change R even when they preserve the within-criterion orderings. Following GRPO (Shao et al., 2024), the scalar reward $R_i = R(x, y_i)$ of each rollout is standardized within the group into a group-relative advantage $\hat{A}_i = (R_i - \mathrm{mean}(\boldsymbol{R}))/\mathrm{std}(\boldsymbol{R})$, which is then used to update the policy with the following regularized clipped objective:

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, \{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)} \Bigg[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \min\big(r_{i,t}(\theta)\hat{A}_i, \\ \mathrm{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\big) - \beta\, \mathbb{D}_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}}) \Bigg]. \tag{2}$$

Here $r_{i,t}(\theta)$ is the token-level importance ratio, ϵ the clipping threshold, and β the KL weight against a reference policy π_{ref}. Since R_i is a rollout-level scalar, the same $\hat{A}_i$ applies to all tokens of y_i, so policy updates are driven entirely by $\{\phi_k\}$ and $\{w_k\}$.

3 METHODOLOGY

We present **RLVR^2**, a verifiable ranking framework for training with heterogeneous criteria through within-group comparisons. Section 3.1 details within-group ordinal reward aggregation mechanism. Section 3.2 shows how non-objective properties fold in as covariates.

3.1 WITHIN-GROUP ORDINAL REWARD AGGREGATION

For a query x, let $\mathcal{Y}(x) = \{y_1, \ldots, y_G\}$ denote the rollout group sampled for policy optimization. For each of the K_x verifiable criteria, the corresponding verifier assigns a score $s_{k,i}$ to rollout y_i. These scores may be binary, discrete, or continuous and need not share a common numerical scale. RLVR^2 uses the relative structure induced within $\mathcal{Y}(x)$: for each compared pair (i, j) and criterion k,

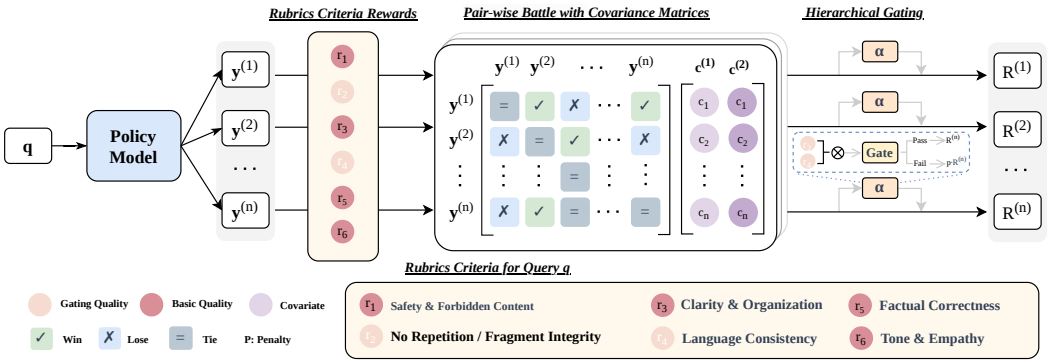


Figure 2: **Overview of RLVR^2.** Each rubric criterion becomes a within-group *ordering* (win/lose/tie); orderings are fused into a scale-invariant reward without raw magnitudes. Auxiliary attributes fold in as extra covariates so that hackable attributes are controlled for rather than rewarded.

we encode the verifier outcome as

$$o_k(i,j) = \begin{cases} 0.5, & |s_{k,i} - s_{k,j}| \le \tau_k, \\ 1, & s_{k,i} - s_{k,j} > \tau_k, \\ 0, & s_{k,j} - s_{k,i} > \tau_k, \end{cases} \tag{3}$$

where $\tau_k \ge 0$ is a criterion-specific tie margin, with $\tau_k = 0$ for discrete criteria. Only thresholded within-group ordering is retained, while the raw score gap is discarded. We fit a Bradley–Terry (BT) model (Sun et al., 2024) to these outcomes, positing a latent utility $u_{k,i}$ per rollout with

$$\Pr\big[\, y_i \succ y_j \mid k \,\big] = \sigma(u_{k,i} - u_{k,j}), \qquad \sigma(z) = \frac{1}{1+e^{-z}}, \tag{4}$$

and recover $\boldsymbol{u}_k = (u_{k,1}, \ldots, u_{k,G})$ by maximizing the BT likelihood over the encoded outcomes. A tie outcome $o_k(i,j) = 0.5$ is treated as a half-win target, encouraging the predicted probability to be one half. Since the construction uses within-group order rather than raw score magnitudes, it applies uniformly to binary, discrete, and continuous verifier outputs without cross-verifier calibration. Finally, we fuse the criterion-specific utilities and normalize them within the rollout group:

$$u_i = \sum_{k=1}^{K_x} w_k u_{k,i}, \qquad \tilde{R}_i = \frac{u_i - \min_j u_j}{\max_j u_j - \min_j u_j}, \tag{5}$$

3.2 Objective-Preserving Attribute Adjustment

When a new training requirement arises, a natural solution is to add another reward dimension. This may unintentionally turn an auxiliary property into an optimization target and expand the surface on which the policy can exploit the reward. RLVR^2 instead controls such attributes through the within-group comparison matrix already induced by the rubric verifiers, without expanding the rubric or changing the underlying quality objective. More generally, any measurable auxiliary attribute can be incorporated as a source of systematic variation in the observed ordering rather than introduced as a new criterion; the attributes discussed below are representative instances of this broader construction.

Attribute adjustment. For a response y_i, let $\boldsymbol{a}_i = \phi(y_i)$ denote an auxiliary attribute that is not itself part of the target rubric. We construct a normalized contrast for each attribute,

$$z_{ij}^{(d)} = \frac{a_i^{(d)} - a_j^{(d)}}{a_i^{(d)} + a_j^{(d)} + \epsilon}, \tag{6}$$

and standardize each feature over the comparisons in the group. The resulting attributes are incorporated into the same within-group comparison matrix as the rubric-induced outcomes:

$$\Pr\big[\, y_i \succ y_j \mid k \,\big] = \sigma\big(u_{k,i} - u_{k,j} + \boldsymbol{\gamma}_k^\top \boldsymbol{z}_{ij}\big). \tag{7}$$

Here $u_{k,i}$ represents criterion quality, whereas γ_k captures the systematic association between the auxiliary attributes and the observed ordering. Both are estimated jointly, so the attributes influence the quality estimation without becoming additional reward dimensions.

Quality-level gating. Some requirements are non-compensatory and should not be offset by gains on other criteria. We therefore apply a stepwise adjustment after ordinal fusion. Let $\tilde{R}_i$ denote the normalized fused reward for rollout i. A failure of the global-quality condition triggers a hard gate, while the low-quality condition yields a soft penalty factor ρ_i:

$$R_i = g(q_i)\tilde{R}_i, \qquad g(q_i) = \begin{cases} 0, & \min(\boldsymbol{s}_i^{\text{global}}) < 1, \\ \rho_i \tilde{R}_i, & \text{otherwise}, \end{cases} \qquad 0 < \rho_i \leq 1, \tag{8}$$

where $\boldsymbol{s}_i^{\text{global}}$ contains the GlobalLowQuality scores. For the LowQuality layer, let $\bar{s}_i^{\text{low}}$ be its mean score. We set

$$\rho_i = \begin{cases} 1, & \bar{s}_i^{\text{low}} \geq \tau, \\ \rho_{\min} + (1 - \rho_{\min})\ \text{softgate}(\bar{s}_i^{\text{low}}), & \bar{s}_i^{\text{low}} < \tau, \end{cases} \tag{9}$$

where τ is the low-quality threshold and $\rho_{\min}$ is the penalty floor.

Reasoning-efficiency control. The adjustment is useful when rubrics reward information coverage, as in recall-oriented and question-answering tasks. Once the relevant content has been identified, additional deliberation may yield repeated verification or marginal rubric coverage without improving substantive quality. Accordingly, a positive association between reasoning length and observed preference need not reflect a genuine quality benefit. We randomly attenuate a fraction λ of the fitted positive coefficient, while leaving non-positive coefficients unchanged. This controls the strength of adjustment without imposing a target response length, thereby reducing unnecessary deliberation while limiting abrupt changes in reasoning behavior.

Formatting control. Judges may systematically prefer heavily structured responses, even when extra formatting adds no substantive information (Yu et al., 2025). We therefore treat formatting as an auxiliary source of variation, not a quality criterion. Normalized statistics for headings, lists, and emphasis are standardized and included in the attribute-adjusted model, controlling formatting preferences without directly rewarding formatting.

4 Experiments

In this section, we present the empirical evaluation of **RLVR**^{2}. We first describe the experimental setup (Section 4.1), then report the main results (Section 4.2), further analyze attribute adjustment (Section 4.3), and conclude with ablations of its robustness and computational cost (Section 4.4).

4.1 Experimental Setup

Datasets and Benchmarks. All methods are trained on the same 120k-query mixture of general-chat, scientific reasoning and agentic tasks. They share rubrics, verifiers, and judges and differ only in reward aggregation. For evaluation, we use a broad suite of public benchmarks covering six capability dimensions: *instruction following* (SysBench [SysB] (Qin et al., 2024), AdvancedIF [AdvIF] (He et al., 2026), IFEval [IFE] (Zhou et al., 2023), IFBench [IFB] (Pyatkin et al., 2026), and MultiChallenge [MChal] (Deshpande et al., 2025)), *long-context* (LongBench-v2 [LB-v2] (Bai et al., 2025), LongProc [LProc] (Ye et al., 2025), HELMET [HLM] (Yen et al., 2025), Frames [Frm] (Krishna et al., 2024), AA-LCR [AA-L] (Team, 2025)), *factuality and QA* (SimpleQA-verified [SQA-v] (Haas et al., 2025), ChineseSimpleQA [CSQA] (He et al., 2025)), *reasoning* (GPQA (Rein et al., 2023)), *knowledge* (MMLU-Pro [MMLU-P] (Wang et al., 2024)), and *open-ended chat* (ArenaHard-v2 [AH-v2] (Li et al., 2024), EQBench3 [EQB3] (Paech, 2023)) and *instruction following* (τ^3-Banking (Shi et al., 2026), GAIA (Mialon et al., 2024), GDPVal (Patwardhan et al., 2026), SWE-bench Verified [SWE-V](Jimenez et al., 2024)). Dataset and benchmarks are detailed in Appendix A.

	Method	Instruction Following					Factuality & QA		Long-Context				Reason.		Know.	Open-ended		Average
		SysB	AdvIF	IFE	IFB	MChal	SQA-v	CSQA	LB-v2	LProc	HLM	Frm	AA-L	GPQA	MMLU	AH-v2	EQB3	
106B	Base	51.08	43.93	76.16	36.05	40.83	12.6	51.37	54.59	81.08	29.74	77.43	47	71.46	79.19	52.6	85.35	55.65
	Rule-Fusion	66.88	55.11	82.42	67.01	52.64	13.9	53.67	64.62	83.55	31.38	78.37	63.33	74.21	78.73	61.61	82.62	63.13
	GEAR	41.6	58.61	93.9	62.93	59.17	6.6	52.62	63.94	85.27	32.47	80.46	62.33	76.26	77.41	65.2	85.3	62.75
	GDPO	71.44	62.18	94.27	71.77	50.14	9.1	53.21	57.38	83.45	32.94	80.1	62.33	75	75.36	62.9	82.65	64.01
	GDPO+	70	62.92	91.68	**73.47**	57.5	9.7	54.17	57.97	83.06	31.88	79.85	65	75.63	76.79	**67.6**	83.6	65.05
	Focal	68.6	59.89	92.24	72.79	48.95	13.8	51.47	**67.69**	81.04	33.37	79.25	63	75.48	75.44	65.98	84.13	64.57
	Focal+	72.52	65.7	91.5	65.99	58.67	14	55.23	64.62	**85.28**	32.42	80.46	59	76.61	78.65	66.65	84.32	65.73
	RLVR^2	69.12	64.64	93.16	70.71	53.19	13.8	53.97	63.85	84.02	34.36	**81.19**	64	77.2	73.55	66.74	86.04	65.60
	RLVR^2+	**73.2**	**67.32**	**94.64**	73.13	**60.04**	**15.2**	**56.07**	65.77	84.12	**35.31**	**81.19**	**67.67**	**77.86**	**80.43**	67.14	**87.42**	**67.91**
	Base	67.44	59.02	89.65	42.77	50	11.5	67.23	32.8	80.98	24.87	74.76	54.33	80.18	80.33	35.1	82.75	58.36
	Rule-Fusion	74.12	62.23	90.76	48.59	55.56	11.6	67.13	32.01	81.75	25.86	74.64	55.33	80.68	80.39	60.1	86.65	61.71
	GEAR	73.68	61.55	92.05	44.28	50	12.1	67.6	33.6	80.69	4.17	76.58	58.33	82.45	80.63	70.3	**88**	61.00
	GDPO	74.88	63.71	89.83	67.35	52.37	11.2	68.17	31.61	81.53	27.62	75.97	57.33	**83.33**	**81.68**	62.3	82.55	63.21
9B	GDPO+	68.76	51.34	72.46	59.52	50	10.1	54.27	31.41	82.55	25.42	71.97	50.33	82.07	80.15	67.3	84.55	58.89
	Focal	**76.44**	63.3	91.31	69.05	51.67	10.8	68.3	33.6	80.93	27.76	76.46	57	82.7	81.01	61	86.2	63.60
	Focal+	75.2	61.19	90.57	70.41	53	10.9	65.53	31.81	81.91	26.76	**76.94**	59.67	81.44	80.73	57.3	86.6	63.12
	RLVR^2	73.48	63.61	92.05	69.69	55.40	11.1	68.43	**33.8**	81.75	**28.18**	75.73	58.67	80.81	80.48	71.1	87.15	64.46
	RLVR^2+	**76.44**	**65.56**	**92.61**	**73.81**	**56.23**	**12.2**	**69.5**	32.21	**82.77**	26	76.21	**60**	81.94	81.11	**72.7**	85.65	**65.31**
	Base	61	46.51	83.55	58.84	33.75	6.9	56.67	32.31	73.22	23.5	73.18	54.33	77.65	76.53	28.1	77.3	53.96
	Rule-Fusion	57.8	58.36	91.13	68.71	50	7.2	55.53	32.01	71.74	23.7	72.33	55	76.77	76.74	48.5	82.35	57.99
	GEAR	72.12	59.48	91.68	46.08	33.33	5	54.8	**32.6**	80.45	23.44	73.67	54	78.28	77.29	41.3	**85.5**	56.81
	GDPO	52.48	52.62	83.73	64.29	56.16	6.9	54	31.41	79.71	23.51	70.51	46.67	79.01	75.21	45.8	83.65	56.60
4B	GDPO+	42.76	46.64	74.12	56.8	35.42	6.2	55.23	30.82	77.1	24.34	73.42	52.33	79.03	76.66	43.4	83.6	53.62
	Focal	67.28	55.58	91.68	66.33	53.06	7.8	56.43	29.82	**81.79**	26.47	74.39	45.67	77.78	77.07	44.5	84.15	58.74
	Focal+	70.24	59.83	91.5	70.83	53.22	7.3	56.47	32.01	78.43	26.26	74.03	55.67	77.02	77.63	41.3	85.15	59.81
	RLVR^2	67.12	61.09	90.76	**71.77**	54	7.2	56.66	31.21	76.57	25.31	**74.51**	54.67	77.15	76.8	**50.8**	85.2	60.05
	RLVR^2+	**74.4**	**62.6**	**92.42**	69.73	**57.29**	**8.2**	**56.86**	32.21	78.61	**26.71**	74.39	**60.33**	**79.04**	**77.67**	46.5	83.5	**61.28**

Table 1: **Comparison of reward-aggregation methods across 16 benchmarks and three model scales.** Within each block, all methods share the same training setting, differing only in reward aggregation. **Bold** and underline mark the best and second-best per block.

Baselines and Implementation Details. Beyond the vanilla model, we compare RLVR^2 with several static and adaptive reward-fusion baselines: Rule-Fusion (Park et al., 2024), a weighted sum gated by hard rules; the graph-based gating method GEAR (Lv et al., 2026); and two dynamic aggregation methods, Focal (Huang et al., 2025) and GDPO (Liu et al., 2026a) [1]. We evaluate our method on a diverse set of LLMs spanning two model families and different scales, including the Qwen3.5 series (Qwen3.5-4B and Qwen3.5-9B) (Qwen Team, 2026) and GLM-4.5-Air (106B-A12B) (Zeng et al., 2025). The models cover both dense and mixture-of-experts(MoE) architectures. All methods are implemented using the PaddleRL framework and trained with the GSPO algorithm (Zheng et al., 2025) on up to 128× H800-80GB GPUs. Unless otherwise specified, we use a global batch size of 128, a context length of 256K, and a group size of $G = 8$ rollouts per query. We optimize all models with AdamW using a learning rate of 1×10^{-6}. RLVR^2-specific settings are held fixed across models and listed in Appendix A.1.

4.2 Main Results

Performance across capabilities and scales. As shown in Table 1, RLVR^2 performs consistently across all three model scales, leading a majority of benchmarks and showing particular strength in instruction following and factuality, alongside broad gains in long-context understanding, reasoning, knowledge, and open-ended evaluation. This pattern reflects the benefit of within-group ordinal fusion on tasks governed by multiple criteria: unlike cardinal aggregation, it prevents a high score on one dimension from masking a critical failure on another. Consequently, responses that are fluent but factually incorrect or instruction-violating are less likely to receive inflated rewards.

Reasoning performance. Reasoning tasks evaluate more than the final answer: they require understanding the problem, decomposing it into intermediate steps, maintaining consistency across a long solution process, and reaching a verifiable conclusion. The corresponding verifier signals therefore reflect several aspects of reasoning quality, including intermediate validity, mathematical or scientific correctness, and final-answer agreement. RLVR^2+ leads both AA-LCR and GPQA , while remaining highly competitive at 9B, showing that ordinal fusion effectively combines process- and outcome-level signals.

[1] AMRP (Yuan et al., 2026) is excluded, as no official implementation was publicly available.

(a) Agentic and coding performance.

Method	τ^3-Banking	GAIA	GDPVal	SWE-V
RLVR^2	7.53	62.14	56.86	52.62
Focal	7.12	59.37	54.95	50.88
GDPO	6.95	55.84	54.20	48.67
Rule-Fusion	7.22	58.25	52.96	48.69
Base	6.91	53.37	51.38	49.40

(b) Reasoning-efficiency control.

Method	Setting	IF	Fa.	Lo.	Re.	Kn.	Op.	Ovr.
Rule-Fusion	–	68.93	39.37	53.57	68.01	80.39	86.65	66.15
RLVR^2+	–	77.11	40.85	54.24	70.97	81.11	85.65	68.32
	λ=0	75.73	40.37	53.73	71.01	80.85	85.85	67.92
	λ=0.25	75.28	42.37	54.07	70.21	80.72	87.30	68.30
	λ=0.50	75.78	41.75	53.56	70.10	81.41	85.30	67.98
	λ=0.75	77.73	41.86	54.42	70.49	81.31	85.60	68.57
	Format	76.32	41.62	53.72	70.07	80.21	85.76	67.95

Table 2: **Further analysis of RLVR^2.** Bold denotes the best value in each column.

Agentic performance. We further evaluate RLVR^2 on agentic benchmarks, where verifiers assess complete multi-step trajectories rather than isolated responses. As shown in Table 2a, RLVR^2 achieves the best result on all four benchmarks. Its gains over the strongest competing aggregation method are 0.31, 2.77, 1.91, and 1.74 points, respectively. These consistent improvements show that within-group ordinal fusion transfers effectively to trajectory-level policy optimization, jointly capturing tool behavior, intermediate reasoning, task completion, and final response quality.

4.3 Further Analysis

4.3.1 Quality-level gating

Setup. We apply the same gate mechanism to RLVR^2 and the comparable baselines, denoting gated variants by a "+" suffix in Table 1. The gate acts on critical rubric levels after criterion fusion.

Result. The gated variant RLVR^2+ outperforms RLVR^2 in most of benchmarks comparisons across three scalesCardinal aggregation baselines show less consistent changes. This result indicates that hard quality constraints are compatible with within-group ordinal rewards.

4.3.2 Reasoning-efficiency control

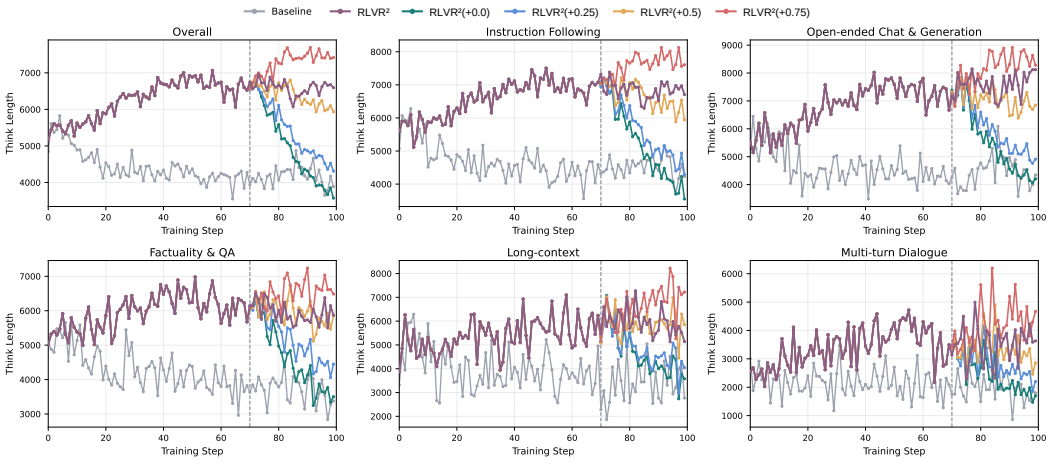


Figure 3: **Reasoning-efficiency control under different attenuation ratios.** Larger λ yields faster convergence toward baseline length, with sensitivity varying by domain.

Setup. We resume the Qwen3.5-9B run from step 70 and vary the attenuation ratio $\lambda \in \{0, 0.25, 0.50, 0.75\}$, while keeping the rubric criteria and optimization settings fixed. The adjustment is applied to the fitted positive association between reasoning extent and the observed rubric ordering. Table 2b reports capability-level averages, while Figure 3 reports thinking-token dynamics.

Result. Across the tested values of λ, the adjustment changes reasoning extent while keeping the capability-level averages within a narrow range. This suggests that coefficient attenuation controls reasoning efficiency without introducing a separate reasoning-length criterion.

4.3.3 FORMATTING CONTROL

Step	Heading ↓	Bold ↓	Italic	Table Row ↓	Code Block ↓	Inline Code	OL ↓	UL ↓	Blockquote	HR ↓	Link
0	34.45	48.67	12.79	13.48	7.26	2.25	13.31	28.80	12.16	26.65	0.20
10	34.52	49.38	13.15	14.24	7.83	2.06	13.54	31.19	12.48	26.87	0.45
20	34.63	49.79	13.33	14.66	8.42	2.01	14.05	32.31	12.64	26.27	0.02
30	34.71	50.55	14.06	14.94	8.54	1.83	14.47	32.64	12.71	26.94	0.11
40	34.82	50.81	13.28	15.13	7.95	2.44	14.72	33.01	11.95	27.54	0.04
50	34.94	51.23	13.44	15.93	9.06	3.33	15.20	33.34	12.52	27.20	0.31
60	35.08	51.73	13.04	16.34	9.29	2.17	15.87	34.08	12.88	28.03	0.28
70	35.21	52.10	13.32	16.72	10.76	2.25	16.42	34.98	12.24	29.08	0.29
80	34.58	50.71	13.17	15.65	8.28	2.42	15.85	33.36	12.56	27.22	0.33
90	34.12	49.86	13.20	14.28	6.41	2.26	15.40	32.03	12.86	26.18	0.18
100	33.51	48.92	13.14	13.38	5.25	2.21	14.22	31.51	12.19	25.81	0.24
Avg.	34.80	50.53	13.30	15.18	8.64	2.29	14.70	32.54	12.45	27.32	0.21
Avg.	34.07	49.83	13.17	14.44	6.65	2.30	15.16	32.30	12.54	26.40	0.25

Table 3: **Markdown usage during RL training.** Columns report Markdown-marker density (%); OL, UL, and HR denote ordered lists, unordered lists, and horizontal rules. Light rows show the pre- and post-control phases; dark rows show averages. Green arrows indicate post-control decreases.

Setup. We analyze the Qwen3.5-9B training trajectory, tracking densities of markdown format. The formatting-control strategy is applied after the 70% checkpoint by treating these statistics as auxiliary attributes in the ordinal comparison model rather than as additional rubric criteria.

Result. Formatting usage rises before control but declines after attribute adjustment, while query-dependent markers remain relatively stable. The formatting-controlled variant achieves an overall capability average of 68.32, close to the 67.95 of RLVR^2+ in Table 2b. These results show that formatting-related preference can be controlled without expanding the rubric or directly rewarding formatting. Detailed statistics are reported in Table 3.

4.4 ABLATION STUDY



Figure 4: **Criterion-level training dynamics.** RLVR^2 improves more on Challenging Rubrics while matching Baseline on Already-Saturated ones.

Effective weight and criterion headroom. Figure 4 compares training dynamics on challenging and already-saturated rubric subsets. RLVR^2 improves more on criteria with remaining headroom while preserving already-saturated performance. A criterion tied across an entire rollout group contributes no utility difference, so fusion emphasizes criteria that distinguish sampled responses. On the challenging subset, the baseline drops by 15.6 points from start to end , whereas RLVR^2 gains 8.0 points from a lower starting point; the saturated subset remains essentially flat for both methods.

G	IF	Fact.	Long	Reas.	Know.	Open	Avg.
4	73.59	38.77	54.33	70.25	80.67	67.68	63.71
8	73.45	40.04	53.50	71.43	81.28	67.80	63.82
16	73.47	39.04	53.43	68.40	80.98	68.33	63.35

Table 4: **Effect of rollout group size.** Columns are dimension-level averages.

Method	Len.	Fusion	Rollout	Total
Rule-Fusion	4936	12.7	101.0	146.7
Focal	5570	20.5	149.4	222.3
GDPO	8581	16.5	174.3	259.9
RLVR^2	7344	10.9	160.2	235.4

Table 5: **Runtime overhead.** Rubric inference is shared across methods and excluded.

Rollout group size. We vary $G \in \{4, 8, 16\}$ for 50 training steps with all other settings fixed. Larger groups provide more within-group comparisons but incur proportional rollout cost. Table 4 shows that RLVR^2+ is relatively stable across the tested group sizes: the overall average varies only from 63.35 to 63.82, with $G = 8$ providing a practical trade-off between comparison density, query diversity, and computation. Per-benchmark scores are reported in Appendix C.

Computational overhead. Table 5 compares reward-aggregation overhead on Qwen3.5-9B with batch size 128 and $G = 8$. Rubric scoring is shared and excluded; fusion is measured on identical logged rollouts, while total step time also includes generation length. RLVR^2 operates only on computed rubric outcomes and makes no additional rubric calls. Its fusion takes 10.9 seconds per step, compared with others approaches.

5 Related Work

Rubric-based rewards for RLVR. Rubric-based grading has extended RLVR from tasks with well-defined correctness signals, such as mathematics and code, to open-ended generation with multiple quality requirements (Yu et al., 2026). Existing studies investigate rubric construction, rubric-grounded rollout scaffolding, and rubric-based reward modeling (He et al., 2026; Li et al., 2026; Huang et al., 2025; Zhou et al., 2025; Zhang et al., 2026). Structured rubrics further organize evaluation across heterogeneous criteria (Starace et al., 2025; Condor et al., 2022). These works primarily focus on rubric design and use, leaving the aggregation of fixed, heterogeneous criterion outputs into a scalar policy-optimization signal largely unaddressed.

Criterion-level reward aggregation. Existing methods mainly aggregate criterion scores in cardinal space using weighted sums, normalization, dependency modeling, or adaptive reweighting (Liu et al., 2026a; Lv et al., 2026; Tyagi et al., 2026). Related work also explores relative reward signals in group-based RL (Niu et al., 2026). However, these approaches largely rely on cardinal score aggregation and therefore require heterogeneous criteria to be normalized or otherwise aligned. They also typically treat auxiliary behavioral attributes as reward criteria, which can alter the objective rather than merely control their systematic influence. Detailed comparisons are provided in Appendix D.

6 Conclusion

We introduce **RLVR^2**, a verifiable ranking paradigm for reinforcement learning with heterogeneous criteria. Across three model scales and 16 benchmarks, RLVR^2 consistently achieves leading performance while supporting attribute control and adaptive weighting. Future work will extend RLVR^2 to broader tasks, verifiers, and ranking models.

AI USE STATEMENT

In this work, we used generative AI tools for writing assistance, specifically for grammar correction, improving phrasing and clarity, and polishing the presentation of text drafted by the authors. We have not used generative AI tools for research ideation, method design, experiment design, code generation, data analysis, or the interpretation of results; all methodological and empirical contributions in this paper are the authors' own. Retrieval and literature search with generative AI tools were not used in the preparation of this work. We have reviewed all AI-assisted text. Every AI-suggested edit was inspected by at least one author and accepted, revised, or rejected on a case-by-case basis; no AI-generated text was incorporated without human verification. In particular, we verified that AI-assisted editing did not alter the technical meaning of any claim, definition, or experimental result, and that no citations or factual statements were introduced by the AI tools. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies reward fusion for RL post-training. No human subjects were involved and no human annotation was collected; all evaluation benchmarks are public research datasets used under their respective licenses. In addition, the LLM judges may output incorrect or unsafe content. Such content is not intended by the authors and does not reflect our views. Safety alignment of the judges is also not this papers focus.

REPRODUCIBILITY STATEMENT

All benchmarks used in our experiments are publicly available research datasets, and all hyperparameters, judge models, and prompts needed to reproduce our results are reported in the appendix.

REFERENCES

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

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, et al. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3639–3664, 2025.

Aubrey Condor, Zachary Pardos, and Marcia Linn. Representing scoring rubrics as graphs for automatic short answer grading. In *International Conference on Artificial Intelligence in Education*, pp. 354–365. Springer, 2022.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025.

Kaustubh Deshpande, Ved Sirdeshmukh, Johannes Baptist Mols, Lifeng Jin, Ed-Yeremai Hernandez-Cardona, Dean Lee, Jeremy Kritz, Willow E Primack, Summer Yue, and Chen Xing. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18632–18702, 2025.

Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 7(6), 2026.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *CoRR*, abs/2507.17746, 2025.

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 *ACL (Findings)*, volume ACL 2025 of *Findings of ACL*, pp. 2202–2261. Association for Computational Linguistics, 2025.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*, 2025.

Yancheng He, Shilong Li, Jiaheng Liu, Yingshui Tan, Weixun Wang, Hui Huang, Xingyuan Bu, Hangyu Guo, Chengwei Hu, Boren Zheng, et al. Chinese simpleqa: A chinese factuality evaluation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19182–19208, 2025.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, et al. 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)*, pp. 18003–18022, 2026.

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

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

Mengzhao Jia, Zhihan Zhang, Ignacio Cases, Zheyuan Liu, Meng Jiang, and Peng Qi. Autorubric: Rubric-based generative rewards for faithful multimodal reasoning. In *ACL (Findings)*, pp. 25707–25724. Association for Computational Linguistics, 2026a.

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*, abs/2602.14069, 2026b.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

Seungone Kim, Jay Shin, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Ryan Shin, Sungdong Kim, James Thorne, Minjoon Seo, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation, 2024. URL https://arxiv.org/abs/2409.12941.

Nathan Lambert. Reinforcement learning from human feedback, 2026. URL https://arxiv.org/abs/2504.12501.

Nathan Lambert and Roberto Calandra. The alignment ceiling: Objective mismatch in reinforcement learning from human feedback, 2024. URL https://arxiv.org/abs/2311.00168.

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

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

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, et al. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *arXiv preprint arXiv:2601.05242*, 2026a.

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 *ACL (1)*, pp. 17417–17437. Association for Computational Linguistics, 2026b.

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

Can Lv, Mingju Chen, Heng Chang, and Shiji Zhou. Mitigating false credit propagation: Probabilistic graphical reward aggregation for rubric-based reinforcement learning. *arXiv preprint arXiv:2606.03361*, 2026.

Michael Mandler. Compromises between cardinality and ordinality in preference theory and social choice. *Available at SSRN 284849*, 2001.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

Wenzhe Niu, Wei He, Zongxia Xie, Jinpeng Ou, Huichuan Fan, Yuchen Ge, Yanru Sun, Ziyin Wang, Yizhao Sun, Chengshun Shi, et al. From absolute to relative: Rethinking reward shaping in group-based reinforcement learning. *arXiv preprint arXiv:2601.23058*, 2026.

Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2023.

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. *arXiv preprint arXiv:2406.07826*, 2024.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simon Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, et al. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. In *International Conference on Learning Representations*, volume 2026, pp. 24005–24040, 2026.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hanna Hajishirzi. Generalizing verifiable instruction following. *Advances in Neural Information Processing Systems*, 38, 2026.

Yanzhao Qin, Tao Zhang, Yanjun Shen, Wenjing Luo, Haoze Sun, Yan Zhang, Yujing Qiao, Weipeng Chen, Zenan Zhou, Wentao Zhang, et al. Sysbench: Can large language models follow system messages? *arXiv preprint arXiv:2408.10943*, 2024.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

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

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

Lingfeng Shen, Sihao Chen, Linfeng Song, Lifeng Jin, Baolin Peng, Haitao Mi, Daniel Khashabi, and Dong Yu. The trickle-down impact of reward (in-)consistency on rlhf, 2023. URL https://arxiv.org/abs/2309.16155.

Quan Shi, Alexandra Zytek, Pedram Razavi, Karthik Narasimhan, and Victor Barres. (τ)-knowledge: Evaluating conversational agents over unstructured knowledge. *arXiv preprint arXiv:2603.04370*, 2026.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, et al. Paperbench: evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*, 2025.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking bradley-terry models in preference-based reward modeling: Foundations, theory, and alternatives. *arXiv preprint arXiv:2411.04991*, 2024.

Artificial Analysis Team. Artificial analysis long context reasoning benchmark(lcr), 2025.

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

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Sherry Wu. Checklists are better than reward models for aligning language models. In *NeurIPS*, 2025.

Xiaohua Wang, Muzhao Tian, Yuqi Zeng, Zisu Huang, Jiakang Yuan, Bowen Chen, Jingwen Xu, Mingbo Zhou, Wenhao Liu, Muling Wu, et al. Reward hacking in the era of large models: Mechanisms, emergent misalignment, challenges. *arXiv preprint arXiv:2604.13602*, 2026a.

Xinpeng Wang, Nitish Joshi, Barbara Plank, Rico Angell, and He He. Is it thinking or cheating? detecting implicit reward hacking by measuring reasoning effort. In *International Conference on Learning Representations*, volume 2026, pp. 96956–96981, 2026b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. Longproc: Benchmarking long-context language models on long procedural generation. *arXiv preprint arXiv:2501.05414*, 2025.

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. Helmet: How to evaluate long-context models effectively and thoroughly. In *International Conference on Learning Representations*, volume 2025, pp. 98914–98965, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan 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, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

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

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

Junjie Zhang, Guozheng Ma, Shunyu Liu, Haoyu Wang, Jiaxing Huang, Ting-En Lin, Fei Huang, Yongbin Li, and Dacheng Tao. A simple" motivation" can enhance reinforcement finetuning of large reasoning models. In *International Conference on Learning Representations*, volume 2026, pp. 44166–44187, 2026.

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

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

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

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

A Dataset and Benchmark

In this section, we discuss the training dataset and evaluation dataset used in our experiments.

(a) General-chat training mixture.

Domain	# Queries
Instruction following	24,900
Open-ended generation	11,881
Factuality & QA	8,819
Long-context understanding	8,100
Multi-turn dialogue	2,400
Total	**56,100**

(b) Agentic training mixture.

Task family	# Queries
Software engineering	25,591
Reasoning	18,515
Data-analysis agents	15,289
Function-calling agents	3,618
Computer-use/coding agents	718
Domain-specific agents	389
Total	**64,120**

Table 6: Composition of the two-stage training data. Internal data sources are grouped into broad domains and task families.

A.1 Training Dataset Statistics

Table 6a summarizes the composition of our 56.1k-query training mixture. The data is organized into five broad domains that jointly cover the six capability dimensions evaluated in Section 4.1: *instruction following*, *open-ended chat and generation*, *factuality and QA*, *long-context*, and *multi-turn dialogue*. Instruction following is the largest domain (24.9k, roughly 44% of the mixture), reflecting our emphasis on precise constraint satisfaction and its central role in aligning model behavior with user intent; it spans a diverse set of constraint types, including format, length, conditional, and cross-lingual instructions, in both English and Chinese. The remaining domains balance open-ended generation, short- and long-form factual QA, long-context comprehension, and multi-turn consistency, so that the policy is exposed to both rule-verifiable and open-ended tasks during training. Each query is trained with a maximum decode length of 16,384 tokens. Each query activates a task-dependent subset of the shared rubric library, producing binary, discrete, or continuous criterion scores. All compared aggregation methods use the same training queries, active criteria, scoring procedures, and judges; they differ in how the resulting criterion scores are synthesized for policy optimization. To preserve anonymity, internal data sources are referred to by generic capability labels rather than product names.

A.2 Agentic Training Dataset

For the agentic experiments, we use an additional mixture of 64,120 queries spanning software engineering, reasoning, data analysis, function calling, computer-use and coding agents, and other domain-specific agent tasks. Table 6b reports its composition. Software-engineering tasks form the largest component and include repository-level problem solving, code repair, code understanding, and related coding-agent settings. Data-analysis tasks require agents to inspect files and structured data using tools, while function-calling tasks evaluate selecting and invoking external functions. The reasoning component broadens the mixture with logic, science, mathematics, and general problem-solving queries. Every example includes a verifier and a length-penalty configuration; tool definitions are provided when required by the task. All compared methods use the same queries, tools, verifiers, and training settings, differing only in reward construction.

B Criterion List

This section catalogs representative criteria from the rubric library used in our experiments. The library contains more than 180 criteria, with task-dependent subsets activated for each query. Each criterion is given a descriptive name, a one-line description, its semantic family, and its output value domain: Binary $\{0, 1\}$, Ternary $\{0, 0.5, 1\}$, Continuous $[0, 1]$, or Multi (other discrete sets, e.g. $\{-1, 0, 1\}$). Criteria are split into two groups: *base* criteria emit positive quality signals (Table 7), while *penalty* criteria detect low-quality behavior and subtract from the score (Table 8).

Table 7: Base criteria (positive quality signals), grouped by semantic family.

Criterion	Description	Value
Comprehensive content quality		
GeneralQuality	LLM-based overall quality score	Continuous
BusinessScenarioQuality	Holistic quality in a business scenario	Multi
CoreIntentMatch	Core-intent understanding & relevance	Binary
TaskTypeMatch	Correctness of task-type recognition	Binary
Content faithfulness & memory		
CoreFactAccuracy	Coverage/correctness of core facts	Ternary
NonCoreFactAccuracy	Coverage/correctness of non-core facts	Ternary
MaterialFaithfulness	Faithfulness to provided source material	Binary
MultiTurnMemory	Context consistency across dialogue turns	Binary
Format compliance		
GeneralFormatCompliance	General structural/formatting compliance	Binary
FormatRequirement	Explicit format-requirement satisfaction	Binary
TextTypeMatch	Correct text type/genre (email, article)	Binary
BulletCount	Number of bullet-list items	Binary
HighlightCount	Number of highlighted sections	Binary
ParagraphCount	Number of paragraphs	Binary
CaseFormat	Upper/lower-case formatting requirement	Binary
TitleFormat	Title presence and format	Binary
QuotationFormat	Quotation formatting requirement	Binary
Postscript	Required postscript section	Binary
PromptRepeat	Must restate the original prompt first	Binary
ConstrainedResponse	Response limited to a fixed option set	Binary
ConstrainedWriting	Constrained creative writing (acrostic)	Multi
SummaryRecall	Recall-based summary coverage	Continuous
Third-party benchmark detectors		
IFBench	IFBench format/logic check	Multi
MultiIF	Multi-IF benchmark check	Multi
Collie	Collie benchmark check	Multi
ComplexBench	ComplexBench benchmark check	Continuous
LiveBench	LiveBench benchmark check	Continuous
Length & counting		
WordCount	Word/char count constraint	Binary
SentenceCount	Sentence-count constraint	Binary
EntryCount	Number of entries	Binary
LetterFrequency	English letter-frequency constraint	Binary
Language & locale		
LanguageRequirement	Required output language	Binary
LanguageDetection	Detected language of the response	Binary
CrossLingualInstruction	Cross-lingual instruction compliance	Binary
Style & wording		
DialogueStyle	Consistency of dialogue style	Binary
DictionRequirement	Diction / wording-norm requirement	Binary
Identity & capability boundary		
RoleConsistency	Identity/role consistency requirement	Binary
CapabilityBoundary	Capability-boundary / identity limits	Binary
Content presence & matching		
Keyword	Required keyword presence	Binary
ForbiddenContent	Forbidden word/content detection	Binary
GroundTruthMatch	Match against a gold answer	Binary

Table 8: Penalty criteria (low-quality signals that subtract from the score).

Criterion	Description	Value
GeneralLowQuality	Perfunctory / template-like output	Binary
RoboticTemplate	Robotic/templated output detection	Binary
Verbosity	Redundancy / conciseness penalty	Binary
Affectation	Affected / overwrought wording	Binary
ImproperWording	Improper wording detection	Binary
LogicalError	Logical-error detection	Binary
UnrealisticContent	Unrealistic / non-factual content	Binary
SecurityRisk	Security-risk detection	Binary

B.1 Evaluation Benchmarks Statistics

SysBench. SysBench (Qin et al., 2024) evaluates whether large language models can reliably follow system messages. It assesses models along constraint following, instruction misalignment, and multi-turn stability, targeting the gap between stated system-level requirements and actual model behavior.

AdvancedIF. AdvancedIF (He et al., 2026) is a rubric-based benchmark for advanced instruction following. Each prompt is paired with a set of expert-written rubric criteria, enabling fine-grained evaluation of whether a response satisfies complex, compositional instructions beyond surface-level compliance.

IFEval. IFEval (Zhou et al., 2023) measures instruction following through automatically verifiable instructions, such as constraints on output length, format, or the inclusion of specified keywords. Because compliance is checked programmatically rather than by a judge, the benchmark provides an objective, low-noise signal of instruction adherence.

IFBench. IFBench (Pyatkin et al., 2026) extends verifiable instruction following to a broader and more diverse set of constraints, emphasizing generalization to instruction types not seen during training. It is designed to test whether models follow the underlying intent of a constraint rather than overfitting to a fixed instruction template.

MultiChallenge. MultiChallenge (Deshpande et al., 2025) is a realistic multi-turn conversation benchmark that is challenging for frontier LLMs. It evaluates a model's ability to maintain instruction adherence and contextual consistency across multiple dialogue turns, where earlier constraints must continue to hold in later responses.

LongBench-v2. LongBench-v2 (Bai et al., 2025) targets deeper understanding and reasoning over realistic long-context multitasks. It spans a range of long-document scenarios and requires genuine comprehension and multi-step reasoning rather than simple retrieval from the context.

LongProc. LongProc (Ye et al., 2025) benchmarks long-context language models on long procedural generation, where a model must produce lengthy, structured outputs by following multi-step procedures. It stresses both long-context comprehension and long-form generation consistency.

HELMET. HELMET (Yen et al., 2025) is a benchmark for evaluating long-context models effectively and thoroughly across diverse application-driven categories. It is designed to provide a more reliable and comprehensive picture of long-context capability than single-task or synthetic evaluations.

AA-LCR. The Artificial Analysis Long Context Reasoning benchmark (AA-LCR) (Team, 2025) evaluates reasoning over long inputs, requiring models to integrate and reason over information distributed across an extended context rather than locate a single passage.

SimpleQA-verified. SimpleQA-verified (Haas et al., 2025) is a reliable factuality benchmark that measures a model's parametric knowledge through short, fact-seeking questions with verified reference answers. Its verified answer set is intended to reduce label noise and provide a trustworthy factuality signal.

			Instruction Following				Factuality & QA		Long-Context				Reason.		Know.	Open
	Method	Compr.	SysB	AdvIF	IFE	IFB	SQA-v	CSQA	LB-v2	LProc	HLM	Frm	AA-L	GPQA	MMLU	EQB3
	Rule-Fusion	✗	74.12	62.23	90.76	48.59	11.60	67.13	32.01	81.75	25.86	74.64	55.33	80.68	80.39	86.65
9B	RLVR^2+	✗	76.44	65.56	92.61	73.81	12.20	69.50	32.21	82.55	26.00	76.21	60.00	81.94	81.11	85.65
	RLVR^2+	*on*	73.80	64.95	92.05	72.11	11.83	68.91	32.21	81.14	26.20	75.36	61.33	80.68	80.85	85.85
	RLVR^2+	0.25	74.04	63.50	91.13	72.45	14.11	70.63	32.80	81.76	26.47	75.24	57.33	83.08	80.72	87.30
	RLVR^2+	0.50	75.40	65.29	91.68	70.75	14.19	69.31	33.00	81.64	23.51	76.09	59.00	81.19	81.41	85.30
	RLVR^2+	0.75	74.60	66.69	92.42	77.21	13.36	70.36	32.80	83.10	26.03	75.73	59.67	81.31	81.31	85.60

Table 9: **Effect of thinking-efficiency compression.** All rows use the same 9B backbone with settings otherwise fixed. MultiChallenge and ArenaHard-v2 are omitted. Judge models and scoring protocols are given in Appendix B.1.

ChineseSimpleQA. ChineseSimpleQA (He et al., 2025) is a Chinese factuality evaluation for large language models, consisting of short fact-seeking questions in Chinese. It provides a factuality signal complementary to the English benchmarks and covers a broad range of knowledge domains.

Frames. Frames (Krishna et al., 2024) provides a unified evaluation of retrieval-augmented generation, jointly assessing factuality, retrieval, and reasoning. It requires models to fetch relevant evidence and reason over it to arrive at correct answers, rather than testing these abilities in isolation.

GPQA. GPQA (Rein et al., 2023) is a graduate-level, "Google-proof" multiple-choice question-answering benchmark spanning biology, chemistry, and physics, with questions ranging from advanced undergraduate to postgraduate difficulty. The questions are constructed so that they cannot be answered reliably through simple web search, testing deep domain reasoning.

MMLU-Pro. MMLU-Pro (Wang et al., 2024) is a more robust and challenging multi-task language understanding benchmark. It extends MMLU with harder, reasoning-intensive questions and an expanded answer-option set, reducing the effect of guessing and providing a stronger measure of broad knowledge.

ArenaHard-v2. ArenaHard-v2 (Li et al., 2024) is a high-quality open-ended benchmark constructed from crowdsourced conversational data via the Arena-Hard / BenchBuilder pipeline. It consists of challenging open-ended prompts and is designed to correlate well with human preference on real user queries.

EQBench3. EQBench3 (Paech, 2023) evaluates emotional intelligence in open-ended dialogue, scoring responses on their handling of interpersonal and affective content in multi-turn role-play scenarios rather than on task correctness. It complements ArenaHard-v2 by probing a dimension of open-ended quality that correctness-oriented judges do not capture.

τ^3-Bench Banking. τ^3-Bench Banking (Shi et al., 2026) evaluates agents in realistic multi-turn banking scenarios, where an agent must interact with a user, invoke domain tools, and satisfy task-specific policies to complete a request. It tests reliable tool use and task completion under evolving conversational state.

GAIA. GAIA (Mialon et al., 2024) evaluates general-purpose assistants on real-world questions that require combinations of reasoning, web or file-based information retrieval, and tool use. Its multi-step tasks measure whether an agent can coordinate these capabilities to produce a verifiable final answer.

GDPVal. GDPVal (Patwardhan et al., 2026) evaluates AI systems on economically valuable knowledge-work tasks drawn from professional domains. It emphasizes the quality of complete task deliverables, providing a complementary measure of agent performance beyond short-answer reasoning and isolated tool calls.

SWE-bench Verified. SWE-bench Verified (Jimenez et al., 2024) evaluates repository-level software engineering by asking an agent to resolve validated issues in real-world codebases. Success requires understanding the repository, localizing the fault, modifying code, and producing a patch that passes the associated tests.

C Rollout Size Detail Impact

We report per-benchmark results for the rollout-group-size ablation. All settings other than the group size are held fixed.

	Instruction Following					Factuality & QA		Long-Context				Reason.		Know.	Open-ended	
Rollout Size	SysB	AdvIF	IFE	IFB	MChal	SQA-v	CSQA	LB-v2	LProc	HLM	Frm	AA-L	GPQA	MMLU	AH-v2	EQB3
4	73.76	62.67	91.68	71.49	68.33	10.8	66.73	33.20	82.21	25.80	76.09	57.67	82.83	80.67	49.60	85.75
8	76.88	62.70	90.39	68.71	68.56	14.4	65.67	31.81	80.30	26.52	75.36	61.67	81.19	81.28	49.30	86.30
16	73.64	61.70	91.87	71.41	68.73	11.4	66.67	32.01	80.00	26.57	75.12	57.00	79.80	80.98	51.30	85.35

Table 10: Comparison of rollout group size. Judge models and scoring protocols are described in Appendix B.1.

C.1 Score-Representation Robustness

We conduct a controlled robustness experiment using real rollout data from the baseline run. The comparison replays multiple aggregation rules on exactly the same response groups, avoiding differences caused by distinct sampled responses. We retain groups with exactly $G = 8$ unique responses and require every retained criterion to have a finite score in $[0, 1]$ for all responses in the group. Quality-gate verifiers and aggregation-derived fields are excluded from the criterion matrix, including `RepeatSegmentVerifier`, `focal_score`, `gdpo_prenorm`, `gdpo_reward`, `original_reward`, and `sc_fused_score`.

For each retained group, we construct a criterion matrix from the original scores and replay each aggregation rule before and after transforming one target criterion. The transformations preserve within-criterion ordering while changing numerical spacing. Specifically, we multiply IFBench scores by 10, remap the UniLLM score levels from $(0, 0.2, 1)$ to $(0, 1, 100)$, and square the UniBusiness scores. We compare raw cardinal summation, group-wise z-score normalization, and RLVR^2. For RLVR^2, every criterion is converted into pairwise win/tie/loss outcomes within the group, and a Bradley–Terry utility is fitted from these outcomes before criterion-level fusion. Thus, the experiment evaluates the stability of the final fused signal rather than only the stability of an individual criterion ranking.

Here, "Sign agr." is the fraction of responses whose sign relative to the group-mean fused reward is unchanged before and after transformation. Spearman correlation measures agreement of the resulting within-group reward ordering. The invariance result is conditional on the transformation preserving the pairwise outcome matrix, including any tie decisions. With a nonzero tie margin, a numerical transformation can change which pairs are treated as ties; such transformations are outside the strict invariance claim.

D Extended Related Work

Open-ended generation is commonly evaluated through multidimensional LLM-as-a-judge protocols (Zheng et al., 2023; Liu et al., 2023; Gu et al., 2026). Rubrics make these judgments interpretable by decomposing response quality into fine-grained criteria (Kim et al., 2024; Arora et al., 2025). Rubric-guided RL subsequently uses such criteria in several forms, including instruction-following pipelines (He et al., 2026), automatically scaled datasets (Li et al., 2026), reward anchors (Huang et al., 2025), rollout scaffolds (Zhou et al., 2025), and in-context specifications (Zhang et al., 2026). RLVR^2 complements these approaches by focusing on the downstream aggregation of heterogeneous criterion-level judgments into a scalar training signal.

Existing aggregation methods span flat, structured, and adaptive formulations. Weighted sums treat rubric items as independent components (Li et al., 2026), while hierarchical subtasks and rubric graphs organize dependencies among criteria (Starace et al., 2025; Condor et al., 2022). Group reward-decoupled normalization and policy-aware reweighting adapt the contribution of cardinal reward components (Liu et al., 2026a; Tyagi et al., 2026). GEAR instead propagates soft suppression over a query-specific rubric graph to prevent unsupported criteria from contributing utility (Lv et al., 2026). RLVR^2 occupies a complementary position: it aggregates criterion-wise ordinal comparisons

Table 11: **Robustness to order-preserving score transformations.** Results are computed on identical $G = 8$ response groups before and after transforming one criterion. Higher values indicate greater stability.

Criterion	Aggregation	Groups	Sign agreement (%)	Spearman ρ
IFBench	Cardinal	672	92.52	0.9783
	Normalized cardinal	672	98.75	0.9985
	RLVR^2–BT	672	**100.00**	**1.0000**
UniLLM	Cardinal	194	95.88	0.9651
	Normalized cardinal	194	99.16	0.9921
	RLVR^2–BT	194	**100.00**	**1.0000**
UniBusiness	Cardinal	954	95.83	0.99997
	Normalized cardinal	954	97.68	0.99925
	RLVR^2–BT	954	**100.00**	**1.0000**

and applies explicit quality-level gating, combining reduced dependence on cross-criterion calibration with non-compensatory quality control.