RAR^2: RELIABILITY-AWARE RUBRICS AS REWARDS FOR REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

LLMs are increasingly deployed across open-ended tasks, where high-quality outputs must satisfy multiple, task-specific criteria. Rubric-based evaluation makes these criteria explicit and is used to assess complex outputs in specialized, high-stakes domains. The Rubrics as Rewards (RaR) framework extends this approach to reinforcement learning by aggregating rubric-level verdicts into structured reward signals. We characterize rubric signals along two dimensions: discrimination across policy-generated responses and stability across repeated evaluations. The two dimensions divide rubric signals into four quadrants, with stable and discriminative signals providing the most reliable guidance for policy learning. However, standard RaR implicitly assumes noise-free rubric verdicts and therefore cannot distinguish genuine differences between responses from those induced by unstable judgments. To address this limitation, we propose *Reliability-Aware Rubrics as Rewards* (RAR^2). RAR^2 measures stability through within-response variance and discrimination through between-response variance. A Beta–Binomial random-effects model then assesses whether each rubric's apparent discrimination is supported by stable judgments, and a reliability gate filters out rubric signals that lack such support. Experiments show that RAR^2 achieves the best performance across models of different sizes and diverse tasks. Additional analyses demonstrate that RAR^2 delivers consistent performance gains across judge reasoning-effort settings and improves agreement with expert assessments across judge models.

1 INTRODUCTION

As large language models (LLMs) are increasingly deployed in open-ended applications, evaluating response quality across multiple dimensions has become essential (Liang et al., 2023; Laskar et al., 2024). Reliable evaluation is particularly important in medicine, law, and finance, where domain-specific standards govern response quality and errors may have serious consequences (Arora et al., 2025; Akyürek et al., 2026). Reducing such multidimensional quality to a single holistic verdict obscures which requirements are met and leaves the judgment vulnerable to known biases and inconsistencies (Zheng et al., 2023; Saito et al., 2023; Yagubyan, 2026). Rubric-based evaluation instead decomposes quality into explicit criteria, yielding criterion-level judgments that are more transparent and actionable (Hashemi et al., 2024; Shen et al., 2026). Prompt-specific rubrics have been used to assess medical and professional reasoning (Arora et al., 2025; Akyürek et al., 2026). The Rubrics as Rewards framework (RaR; Gunjal et al., 2025) turns their criterion-level verdicts into structured, interpretable reward signals for on-policy reinforcement learning.

However, the effectiveness of rubric scores as reinforcement-learning rewards can be characterized along two complementary dimensions: judgment stability and response discrimination. Judgment stability quantifies the consistency of evaluations of the same response, whereas response discrimination captures the extent to which a rubric distinguishes among responses sampled from the current policy. Figure 1 highlights the importance of considering these dimensions jointly. Although response discrimination is essential for a rubric to yield an informative learning signal, it is insufficient unless the underlying judgments are also stable. For unstable rubrics, apparent discrimination may reflect judge noise rather than genuine response-dependent variation, thereby distorting credit assignment and impairing policy optimization. This two-dimensional characterization motivates our central research question:

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.

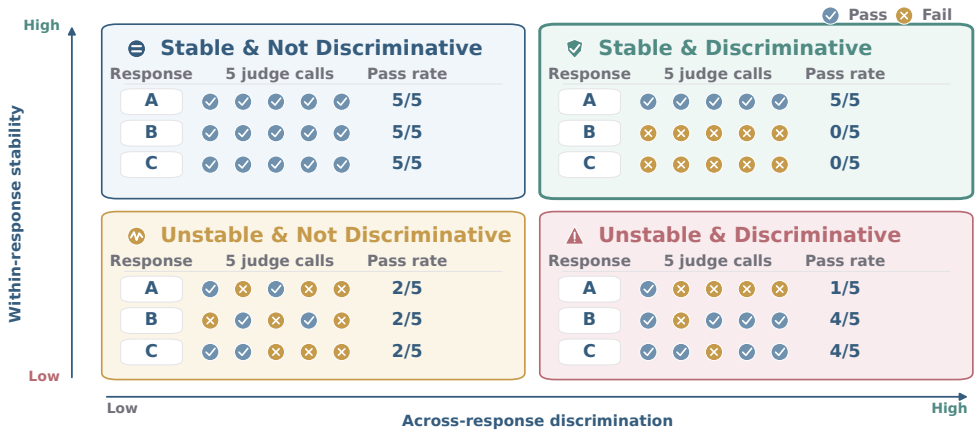

Figure 1: **A useful rubric reward needs both stability and discrimination.** Each row fixes one response and repeats the same rubric judgment five times. Across-response differences provide discrimination; within-response agreement establishes stability. Only the upper-right quadrant supports reliable credit assignment across responses. RAR^2 targets this stable, discriminative signal while filtering apparent contrast caused by inconsistent judgments.

> *How can we provide stable and discriminative rubric-based reward signals for reinforcement learning?*

To address this challenge, we introduce *Reliability-Aware Rubrics as Rewards* (RAR^2), a novel method for selecting rubric signals that are both stable and discriminative. Leveraging multiple evaluations under a fixed judge protocol, RAR^2 assesses judgment stability from within-response variance and response discrimination from between-response variance. Because direct variance estimates are imprecise for small response groups, a Beta–Binomial random-effects model estimates the posterior reliability of each rubric's response contrast. RAR^2 retains rubric signals whose discrimination is supported by stable judgments and filters unreliable signals from the training reward. The gate preserves the original weights of retained rubrics and defaults to retention when statistical evidence is insufficient. Moreover, RAR^2 exploits the conditional independence of repeated judgments to query the judge in parallel.

RAR^2 delivers consistent gains across multiple medical and professional reasoning benchmarks, substantially outperforming a suite of strong baselines. Further analyses demonstrate that these improvements persist across policy scales and judge reasoning-effort settings, while the resulting reliability-aware signals exhibit stronger agreement with expert assessments across judge models.

Our contributions are:

1. We propose *Reliability-Aware Rubrics as Rewards* (RAR^2), a novel method that addresses unreliable rubric-based rewards by retaining learning signals that are both stable and discriminative.
2. We characterize stability through within-response variance and discrimination through between-response variance, using a Beta–Binomial random-effects model to estimate their contributions.
3. We demonstrate that RAR^2 consistently outperforms strong baselines across multiple benchmarks and improves agreement between rubric-derived learning signals and expert assessments across judge models.

2 Preliminaries

To establish the formal basis for RAR^2, we first formalize rubric-based reward construction for reinforcement learning and introduce the associated notation. We then decompose rubric-level variation into stable differences across responses and within-response judgment noise, providing the basis for our definition of rubric reliability.

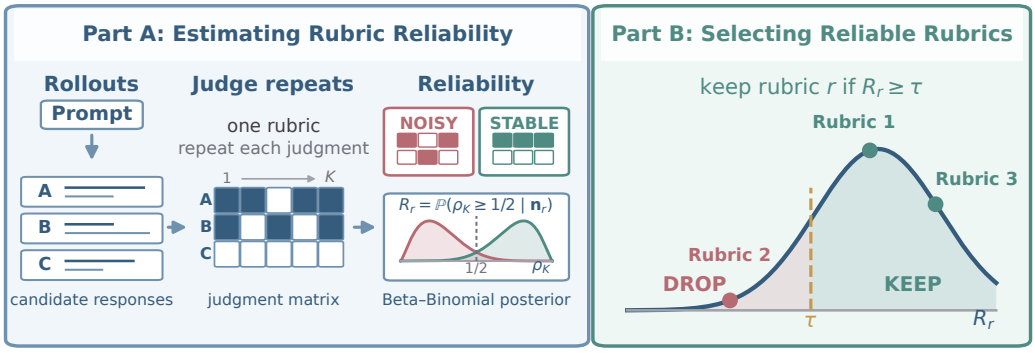

Figure 2: **Overview of RAR^2. Part A: Estimating Rubric Reliability.** Multiple evaluations determine whether each rubric produces stable response contrast or noisy judge decisions. **Part B: Selecting Reliable Rubrics.** Rubrics are positioned by R_r, the posterior probability that $\rho_K \geq \frac{1}{2}$; rubrics with $R_r < \tau$ are dropped, while the rest are retained for reward construction.

2.1 Rubrics as Rewards

Fix a prompt x and let $\mathcal{R} = \{r_1, \ldots, r_M\}$ denote its set of M rubrics. Each rubric $r \in \mathcal{R}$ has an importance weight w_r. For a response y_g, let $c_r(x, y_g)$ denote its score under rubric r. Rubrics as Rewards (RaR; Gunjal et al., 2025) aggregates these rubric-level scores as

$$S_g^{\mathrm{RaR}} := \frac{1}{W} \sum_{r \in \mathcal{R}} w_r c_r(x, y_g), \qquad W := \sum_{r \in \mathcal{R}} \max(w_r, 0) > 0. \tag{1}$$

When all weights are nonnegative, $W = \sum_{r \in \mathcal{R}} w_r$. If negative weights are present, they contribute to the weighted numerator but are excluded from the normalization denominator. This normalization makes rewards comparable across prompts with different rubric counts or weights.

2.2 Multiple evaluations and variance decomposition

To estimate judgment stability, RAR^2 evaluates G responses $y_1, \ldots, y_G$ under each rubric K times using a fixed judge protocol. We use g for responses, k for evaluations, and r for rubrics. For a fixed rubric r, let $v_{gk} \in \{0, 1\}$ denote the k-th verdict for response y_g, and collect these verdicts in

$$V_r := [v_{gk}]_{g=1,\ldots,G;\, k=1,\ldots,K} \in \{0,1\}^{G \times K}, \qquad \bar{v}_{g,r} := \frac{1}{K} \sum_{k=1}^{K} v_{gk}. \tag{2}$$

When r is clear from context, we write $\bar{v}_g$ for $\bar{v}_{g,r}$.

For the fixed rubric, let $q_g \in [0, 1]$ denote the latent judge pass probability for response y_g. It is the probability that the fixed judge protocol returns a positive verdict for that response. Conditional on q_g, the repeated verdicts $v_{g1}, \ldots, v_{gK}$ are independent Bernoulli draws. Across responses, the law of total variance gives

$$\mathrm{Var}(v_{gk}) = \underbrace{\mathrm{Var}(q_g)}_{\text{stable response contrast}} + \underbrace{\mathbb{E}[q_g(1 - q_g)]}_{\text{within-response judge noise}}. \tag{3}$$

With $K = 1$, the observed variance contains only this sum, so a large rollout-level standard deviation is not evidence of reproducible contrast. Multiple judgments hold the response fixed and expose judge instability. Comparing their means $\bar{v}_g$ then reveals stable differences across responses. Estimating this variance decomposition directly is unstable for the small response groups used during training. Appendix A provides the derivation.

3 Reliability-Aware Rubrics as Rewards

Reliability-Aware Rubrics as Rewards (RAR^2) addresses this challenge by estimating whether a rubric's observed contrast reflects stable differences among responses rather than variation in the judge's verdicts. RAR^2 proceeds in two stages: (i) estimate each rubric's reliability from multiple evaluations and (ii) select reliable rubrics to construct the reliability-aware training reward.

3.1 Estimating Rubric Reliability

Fix a rubric r and use the entries v_{gk} of its verdict matrix V_r. Summarize the K verdicts for response y_g by its positive-verdict count

$$n_g := \sum_{k=1}^{K} v_{gk}, \qquad \mathbf{n} := (n_1, \ldots, n_G). \tag{4}$$

Thus one rubric is represented by a length-G count vector rather than a three-index verdict tensor.

For $G \geq 2$ and $K \geq 2$, RAR^2 fits each rubric independently with a Beta–Binomial random-effects model (Skellam, 1948; Prentice, 1986):

$$\begin{aligned} q_g \mid \mu, \phi &\overset{\text{iid}}{\sim} \text{Beta}(\mu\kappa, (1-\mu)\kappa), \qquad \kappa = \phi^{-1} - 1, \\ n_g \mid \mu, \phi &\overset{\text{ind}}{\sim} \text{BetaBinomial}(K, \mu\kappa, (1-\mu)\kappa). \end{aligned} \tag{5}$$

Here μ is the rubric's mean judge pass probability across responses, while ϕ is the intraclass correlation between repeated verdicts for the same response. Under this parameterization, the concentration κ is fully specified by ϕ rather than estimated independently. Every rubric has its own (μ, ϕ) pair; its label remains suppressed within this independent fit.

For the fixed ensemble size K, define rubric reliability as the fraction of variance in $\bar{v}_g$ attributable to stable differences between responses:

$$\rho_K := \frac{\text{Var}(q_g)}{\text{Var}(\bar{v}_g)} = \frac{K\phi}{1 + (K-1)\phi}. \tag{6}$$

This is the Spearman–Brown reliability of an average of K exchangeable measurements when the single-measure intraclass correlation is ϕ (Shrout & Fleiss, 1979; Spearman, 1910).

Variance decomposition and interpretation. The model separates stable response variation $\sigma_B^2 := \text{Var}(q_g) = \mu(1-\mu)\phi$ from within-response judge noise $\sigma_W^2 := \mathbb{E}[q_g(1-q_g)] = \mu(1-\mu)(1-\phi)$. Averaging K conditionally independent verdicts leaves σ_B^2 unchanged and divides σ_W^2 by K, giving $\text{Var}(\bar{v}_g) = \sigma_B^2 + \frac{\sigma_W^2}{K} = \mu(1-\mu)\left(\phi + \frac{1-\phi}{K}\right)$. Hence, under Eq. 5, ρ_K in Eq. 6 is the fraction of mean-verdict variance explained by stable response differences; Appendix A gives the full derivation.

Identifiability from multiple evaluations. When $K = 1$, the count n_g is binary and its marginal likelihood is

$$p(n_g \mid \mu, \phi) = \mu^{n_g}(1-\mu)^{1-n_g}. \tag{7}$$

Equation 7 contains no information about ϕ. Rubric reliability therefore cannot be estimated from one judgment per response. When $K \geq 2$, the count variance is

$$\text{Var}(n_g \mid \mu, \phi) = K\mu(1-\mu)\left[1 + (K-1)\phi\right]. \tag{8}$$

Larger values of ϕ induce greater overdispersion in the response-level counts, reflecting stronger agreement among judgments of the same response. Accordingly, ϕ is identifiable from the count distribution when $K \geq 2$.

Effect of multiple evaluations. For fixed $0 < \phi < 1$, ρ_K increases monotonically with K and converges to one as $K \to \infty$. More judgments therefore reduce the contribution of judge noise to the mean verdict but increase evaluation cost. In practice, K should balance this cost against the reliability required for a stable learning signal.

We place independent priors $\mu \sim \text{Beta}(1, 1)$ and $\phi \sim \text{Beta}(1, 4)$. Combining these priors with the Beta–Binomial likelihood yields the posterior

$$p(\mu, \phi \mid \mathbf{n}) \propto p(\mu)p(\phi) \prod_{g=1}^{G} p(n_g \mid \mu, \phi). \tag{9}$$

Algorithm 1 Reliability-Aware Rubrics as Rewards

Require: prompt x, policy π_θ, rubric set $\mathcal{R} = \{r_1, \dots, r_M\}$ with weights w_r, group size G, judgments per response K, threshold τ

sample responses $y_1, \dots, y_G \sim \pi_\theta(\cdot \mid x)$
collect binary verdicts $v_{gk,r}$ for every response, judgment, and rubric
for all $r \in \mathcal{R}$ **do**
    $n_{g,r} \leftarrow \sum_{k=1}^{K} v_{gk,r}$ for $g = 1, \dots, G$
    $\mathbf{n}_r \leftarrow (n_{1,r}, \dots, n_{G,r})$
    $R_r \leftarrow \mathbb{P}(\rho_K \geq \frac{1}{2} \mid \mathbf{n}_r)$ under the Beta–Binomial model
    if evidence is insufficient or domain policy protects r **then**
        $m_r \leftarrow 1$
    else
        $m_r \leftarrow \mathbb{I}[R_r \geq \tau]$
$W_m \leftarrow \sum_{r \in \mathcal{R}} m_r \max(w_r, 0)$
if $W_m = 0$ **then**
    $m_r \leftarrow 1$ for every $r \in \mathcal{R}$
    $W_m \leftarrow \sum_{r \in \mathcal{R}} \max(w_r, 0)$
for $g = 1, \dots, G$ **do**
    $S_g \leftarrow \frac{1}{K W_m} \sum_{r \in \mathcal{R}} w_r m_r n_{g,r}$
$\bar{S} \leftarrow \frac{1}{G} \sum_{g=1}^{G} S_g$
update π_θ from the relative credits $\{S_g - \bar{S}\}_{g=1}^{G}$
Evaluation: skip gating by setting $m_r = 1$ for all $r \in \mathcal{R}$.

The prior on ϕ places greater mass on small values, so high reliability requires clear evidence from multiple judgments of stable variation across responses. For each rubric r, the resulting posterior is summarized by the gate statistic

$$R_r := \mathbb{P}(\rho_K \geq \tfrac{1}{2} \mid \mathbf{n}). \tag{10}$$

R_r is the posterior probability that stable between-response differences account for at least half of the variance in mean verdicts under the fixed judge protocol.

We assume the same number of valid judgments for every response. Algorithm 2 in Appendix B describes how these judgments are collected through parallel judge calls.

3.2 Selecting Reliable Rubrics

Let $\tau \in (0, 1)$ be the required posterior confidence and let $m_r \in \{0, 1\}$ indicate whether rubric r is retained. RAR^2 uses

$$m_r = \begin{cases} 1, & \text{if a fail-open safeguard applies,} \\ \mathbb{I}[R_r \geq \tau], & \text{otherwise.} \end{cases} \tag{11}$$

Fail-open safeguards retain a rubric when evidence is insufficient or when the rubric is protected for normative reasons. The threshold τ controls the tradeoff between retaining noisy rubrics and discarding reliable rubrics.

Reliability-aware reward construction. For illustration, consider a group-relative update such as GRPO (Shao et al., 2024), which assigns credit by comparing rewards within the G-response prompt group. For each rubric r, define its group-mean verdict as

$$\bar{v}_{\cdot,r} := \frac{1}{G} \sum_{j=1}^{G} \bar{v}_{j,r}. \tag{12}$$

Let $W_m := \sum_{r \in \mathcal{R}} m_r \max(w_r, 0)$ denote the retained positive weight mass. If $W_m = 0$, RAR^2 fails open by setting $m_r = 1$ for every $r \in \mathcal{R}$. The reliability-gated score and the centered contrast used for relative credit are

$$S_g = \frac{1}{W_m} \sum_{r \in \mathcal{R}} w_r m_r \bar{v}_{g,r}, \qquad S_g - \bar{S} = \frac{1}{W_m} \sum_{r \in \mathcal{R}} w_r m_r (\bar{v}_{g,r} - \bar{v}_{\cdot,r}). \tag{13}$$

Table 1: **Main results on held-out test sets (%).** Entries are mean ± standard deviation over three runs under high judge reasoning effort. RaR averages K judgments per rubric and explicitly aggregates all rubrics. Boldface marks the best mean for each policy model.

Method	HealthBench	PRBench		$OneMillion-Bench	
		Non-Hard	Hard	Global	CN
Policy model: Qwen3-4B					
RaR	46.23 ± 0.6106	39.56 ± 0.3300	18.67 ± 0.3510	23.57 ± 0.8217	25.06 ± 0.7901
ARL	47.17 ± 0.6956	40.73 ± 0.3020	19.22 ± 0.2398	23.88 ± 0.6864	25.49 ± 0.7759
POW3R	47.66 ± 0.5923	41.88 ± 0.1977	19.88 ± 0.2280	24.62 ± 0.5130	26.44 ± 0.7387
RAR^2	$\mathbf{49.71 \pm 0.2825}$	$\mathbf{43.01 \pm 0.2277}$	$\mathbf{20.76 \pm 0.1257}$	$\mathbf{26.11 \pm 0.2884}$	$\mathbf{28.96 \pm 0.6582}$
Policy model: Qwen3-30B-A3B					
RaR	69.74 ± 0.6389	47.46 ± 0.7360	38.70 ± 0.8109	31.04 ± 0.7825	32.30 ± 0.8856
ARL	70.33 ± 0.9875	48.73 ± 0.7192	39.24 ± 0.7717	31.31 ± 0.8888	33.18 ± 0.8331
POW3R	71.96 ± 0.6937	50.26 ± 0.6466	39.71 ± 0.5876	32.28 ± 0.5658	33.83 ± 0.7410
RAR^2	$\mathbf{74.34 \pm 0.5791}$	$\mathbf{52.23 \pm 0.3295}$	$\mathbf{41.67 \pm 0.4072}$	$\mathbf{33.83 \pm 0.6779}$	$\mathbf{35.47 \pm 0.5774}$

Here $\bar{S} = \frac{1}{G}\sum_{j=1}^{G} S_j$. The same rubric mask is applied to every response, removing unreliable rubric signals while retaining all responses. Training uses the gated score, with the retained rubric contributions normalized by W_m. Evaluation restores the full rubric set by setting $m_r = 1$ for every $r \in \mathcal{R}$.

4 Experimental Setup

Benchmarks. HealthBench provides conversation-specific physician rubrics (Arora et al., 2025). We construct a category-balanced split of 4,000 training and 1,000 test conversations. PRBench comprises 1,100 expert-authored tasks in finance and law (Akyürek et al., 2026). We split them evenly into 550 training and 550 test tasks and report results separately for the Non-Hard and Hard subsets, each pooled across the two domains. $OneMillion-Bench contains 400 expert-curated tasks across five professional domains (Yang et al., 2026b). We use balanced training and test splits of 200 tasks each, with equal coverage of the five domains and the Global and CN subsets.

Models. We evaluate Qwen3-4B-Instruct-2507 (Qwen3-4B) and Qwen3-30B-A3B-Instruct-2507 (Qwen3-30B-A3B) policies (Team, 2025; Qwen Team, 2025b;a) with a gpt-oss-120b rubric judge (OpenAI, 2025). All main experiments use high judge reasoning effort. We additionally evaluate Qwen3-4B on HealthBench and PRBench at low, medium, and high judge reasoning effort to measure sensitivity to this setting. We compare RAR^2 with three baselines. RaR (Gunjal et al., 2025) explicitly aggregates rubric verdicts without a reliability gate, ARL (Lan, 2026) optimizes one semantic rubric class at a time, and POW3R (Tyagi et al., 2026) emphasizes rubrics that exhibit contrast across rollouts. All methods use $G = 16$ rollouts per prompt. Appendix B details the experimental setup, and Table 2 lists the complete optimization settings. Appendix C.1 analyzes the choice of τ.

5 Results

Main results. As shown in Table 1, RAR^2 consistently achieves the best performance on HealthBench, PRBench, and $OneMillion-Bench, outperforming all three baselines. Relative to RaR, RAR^2 improves Qwen3-4B and Qwen3-30B-A3B by 3.09% and 3.66% on average, respectively. Notably, Qwen3-4B improves by 3.90% on the $OneMillion-Bench CN subset. The larger Qwen3-30B-A3B policy benefits more from RAR^2. On HealthBench, for example, its improvement over RaR is 4.60%, compared with 3.48% for Qwen3-4B. These results are consistent with our central hypothesis that retaining stable, discriminative rubrics concentrates optimization on reliable response-level contrasts, thereby improving the fidelity and discriminative power of the learning signal. RAR^2 also has the

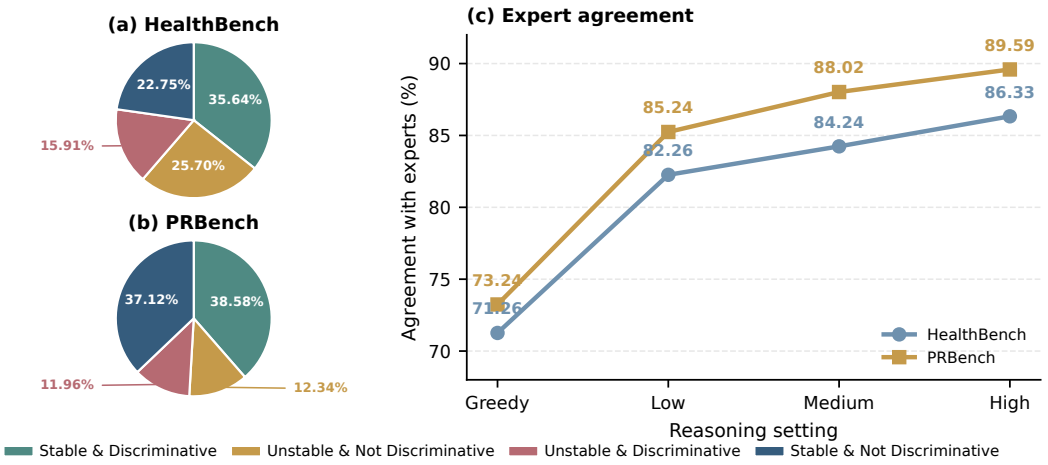

Figure 3: **Rubric dynamics and expert agreement across reasoning settings. (a)** Distribution of rubric types during Qwen3-4B training on HealthBench, evaluated by the judge model at high reasoning effort. **(b)** Distribution of rubric types during training on PRBench under the same setting. Rubric types are defined by judgment stability and within-group response discrimination. **(c)** Agreement (%) between the judge model and expert annotations under greedy decoding and low, medium, and high reasoning effort.

smallest run-to-run standard deviation in eight of the ten comparisons. Detailed results stratified by domain and difficulty level are presented in Appendix D.

Rubric outcomes during training. Figures 3(a) and 3(b) show that all four combinations of stability and discrimination occur at non-negligible rates during training, underscoring that these two properties capture distinct aspects of rubric quality. Unstable rubrics are prevalent on both benchmarks and account for more than 40% of the learning signals on HealthBench. Evaluator inconsistency is therefore not a marginal failure mode and, without explicit control, can exert substantial spurious optimization pressure. RAR^2 provides such explicit control by filtering out unstable rubric signals, thereby grounding policy updates in response-level differences that are reproducible across repeated judgments. Consistent with this mechanism, its gains over the strongest baseline are larger on HealthBench, where unstable rubrics are more prevalent than on PRBench.

Reasoning effort and expert agreement. Figure 3(c) compares the judge model's judgments with expert annotations across reasoning-effort settings. On both benchmarks, agreement increases with configured reasoning effort. Greedy decoding can improve stability by removing sampling randomness, although it does not guarantee identical outputs across runs. However, it lowers expert agreement by more than 15% relative to high effort on both benchmarks, so the judge model makes substantially more incorrect judgments. These results suggest a trade-off between expert alignment and judgment stability. RAR^2 addresses this trade-off by retaining high-effort judgments for their stronger observed expert alignment while explicitly identifying and filtering unstable rubric signals.

Performance across reasoning efforts. Figure 4 shows that RAR^2 consistently outperforms all baselines under every reasoning-effort setting of the judge model. The gains are most pronounced at low effort, where the relative improvement over the strongest baseline, POW3R, reaches 6.73% on HealthBench and 12.22% on PRBench. This may be partly attributed to the larger share of unstable and non-discriminative signals produced by the low-effort judge. By filtering them out, RAR^2 focuses policy updates on stable and discriminative rubric signals and thereby substantially improves training.

Number of judgments. We examine how the number of training judgments K affects final performance, as illustrated in Figure 5. As shown in Eq. 7, a single judgment is uninformative about ϕ, so RAR^2 reduces to RaR at $K = 1$. Once $K \geq 2$ renders ϕ identifiable, performance improves substantially but saturates by $K = 5$. Increasing K from 5 to 7 incurs 40% more judge calls yet changes performance by at most 0.54%, even degrading it in three of the four settings. As indicated

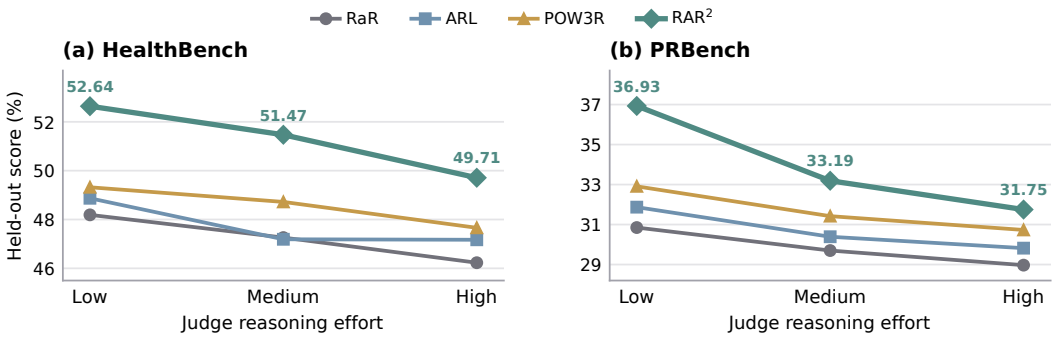

Figure 4: **Method performance across the judge model's reasoning efforts.** Held-out Qwen3-4B scores (%) at low, medium, and high judge reasoning effort on **(a)** HealthBench and **(b)** PRBench.

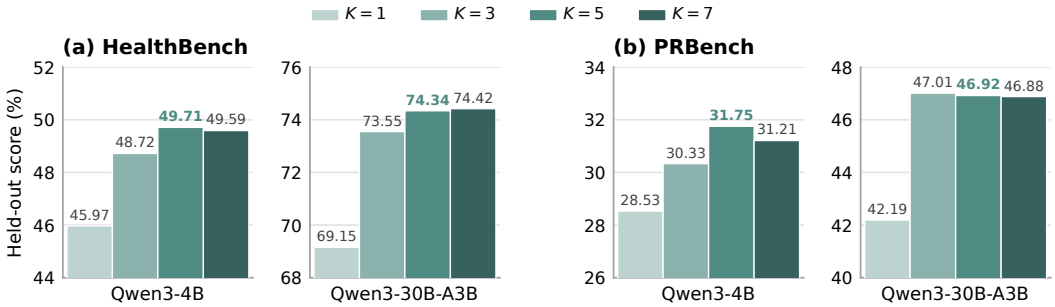

Figure 5: **Effect of the number of training judgments** K**.** Model performance (%) of Qwen3-4B and Qwen3-30B-A3B trained with RAR^2 on **(a)** HealthBench and **(b)** PRBench.

by the Spearman–Brown formula in Eq. 6, the marginal reliability gain decays as $O(K^{-2})$, whereas judge cost grows linearly. We therefore adopt $K = 5$ to balance performance and judge cost.

Training efficiency. Figure 6(a) compares the wall-clock training time of different methods under the fixed 1,000-step protocol. Relative to RaR, RAR^2 incurs only a 2.9% training-time overhead on PRBench and a 1.6% overhead on HealthBench. Its runtime is comparable to those of ARL and POW3R, within 1% of each on both benchmarks, and is even slightly shorter than that of ARL on HealthBench. This modest overhead stems from the mutual independence of judge calls, which enables their concurrent execution. Consequently, RAR^2 can estimate rubric reliability without a substantial increase in wall-clock time.

Cross-judge expert agreement. Figures 6(b) and 6(c) compare the agreement between expert assessments and the learning signals produced by various judge models during training. RAR^2 improves agreement for every judge–benchmark pair, with gains over RaR ranging from 13.74% to 23.69% on PRBench and from 10.96% to 26.10% on HealthBench. Hy4 preview exhibits the largest improvement, with an agreement gain of 26.10% on HealthBench. Agreement with RAR^2 remains above 95% across all eight settings, showing consistent gains across judge models.

6 Related Work

Rubric-based evaluation. Rubric-based LLM evaluation decomposes open-ended quality into explicit, task-specific criteria, making judgments more interpretable and actionable (Kim et al., 2024a; Hashemi et al., 2024). Expert-authored rubrics provide domain-specific evaluation objectives in medicine and professional reasoning (Arora et al., 2025; Akyürek et al., 2026). Related frameworks operationalize criteria through generated evaluation steps, instance-relevant skills, self-adaptive question-level rubrics, and open evaluator models (Liu et al., 2023; Ye et al., 2024; Fan et al., 2024; Kim et al., 2024b). RRD uses recursive decomposition and filtering to expand rubric coverage and remove misaligned or redundant criteria (Shen et al., 2026). Autorubric consolidates heterogeneous

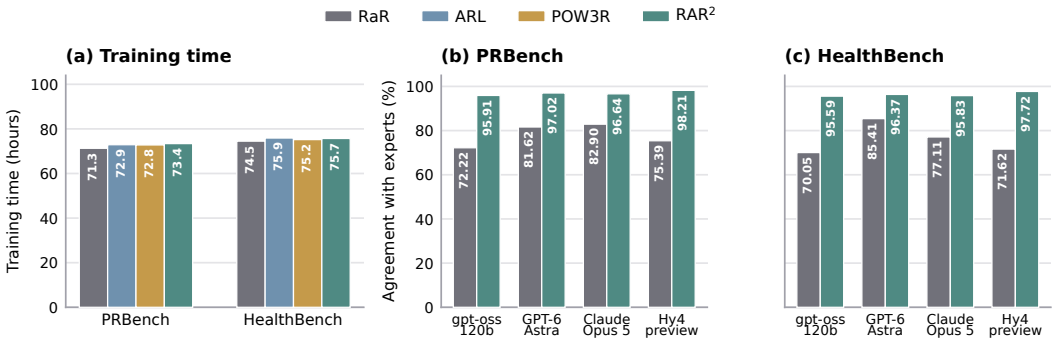

Figure 6: **Efficiency and cross-judge agreement with expert assessments. (a)** Training time for 1,000 steps with batch size 16. **(b)** and **(c)** Agreement of RaR and RAR^2 with expert assessments under four rubric judges. For each benchmark, both methods and all judges are evaluated on the same 100 samples drawn during training. Expert assessments are provided by three human experts.

rubric types, judge ensembles, and psychometric diagnostics within a unified evaluation framework (Rao & Callison-Burch, 2026). Rubrics as Rewards (RaR) further converts instance-specific rubric verdicts into structured signals for on-policy reinforcement learning (Gunjal et al., 2025).

Reliability and uncertainty of LLM feedback. LLM judgments can exhibit substantial run-to-run variability (Schroeder & Wood-Doughty, 2024; Stureborg et al., 2024b; Lee et al., 2025; Yagubyan, 2026). Complementary work quantifies aggregate ranking confidence, constructs conformal prediction intervals, or combines verdicts from heterogeneous judges (Stureborg et al., 2024a; Sheng et al., 2025; Verga et al., 2024). Classical intraclass-correlation models quantify stable between-target variation relative to measurement error, while the Spearman–Brown formula predicts the reliability gain from averaging multiple measurements (Shrout & Fleiss, 1979; Spearman, 1910). Related RL methods address other forms of uncertainty. UARM learns calibrated sample-level reward uncertainty (Pan et al., 2026), whereas EBPO shrinks uncertain prompt-level GRPO baselines (Han et al., 2026). RAR^2 instead targets rubric-level measurement reliability for current-policy responses.

Decomposed and adaptive rewards for RL. Fine-grained RLHF (Wu et al., 2023) and LLM-graded rule-based safety rewards (Mu et al., 2024) both provide decomposed training signals. Multi-objective alignment trades off diverse rewards through reward-specific model-weight interpolation (Ramé et al., 2023) or objective-weighted DPO (Zhou et al., 2024). Reward-model overoptimization has been characterized in synthetic proxy/gold-model settings (Gao et al., 2023), and conservative reward-model ensembles mitigate it in related settings (Coste et al., 2024). Existing methods adapt policy optimization according to observed rubric outcomes. POW3R emphasizes rubrics exhibiting rollout contrast (Tyagi et al., 2026), Focal Reward prioritizes unsaturated criteria (Huang et al., 2026), and ARL optimizes semantic rubric classes sequentially (Lan, 2026). Rubric Dropout instead randomizes group-shared criteria to reduce proxy exploitation (Yang et al., 2026a). These approaches differ in how they select or weight rubrics, yet all of them overlook judgment stability and thus cannot distinguish genuine response differences from contrast induced by judge noise.

7 Conclusion

This work studies rubric-based rewards for reinforcement learning and identifies judgment stability and response discrimination as two complementary requirements. Because rubric contrasts from a single evaluation can conflate genuine response differences with judge inconsistency, we introduce RAR^2, which uses multiple rubric-level verdicts and a Beta–Binomial posterior to assess whether an observed contrast is reproducible before assigning relative credit. By retaining stable and discriminative rubrics for training, RAR^2 grounds policy updates in reproducible response differences rather than judge noise. RAR^2 consistently outperforms strong baselines, demonstrating that explicitly controlling reward reliability improves both the fidelity and effectiveness of rubric-guided learning. More broadly, our findings suggest that when rubric judgments serve as learning signals, their stability and discrimination should be considered jointly.

Reproducibility statement

Appendix B describes the experimental details and hyperparameter settings. We will provide a link to the corresponding code repository in the public version.

Ethics statement

Our experiments use three publicly released benchmarks. HealthBench (Arora et al., 2025) is distributed under the MIT License, PRBench (Akyürek et al., 2026) under CC BY 4.0, and $OneMillion-Bench (Yang et al., 2026b) under the Apache License 2.0. Our use of these benchmarks, including re-splitting them into training and test sets, complies with their licenses. The human experts who provided the expert assessments were recruited through legitimate channels and compensated at an hourly rate commensurate with the local average wage for data annotation. This work introduces no private or personally identifiable data.

AI use statement

Generative AI tools were used to polish the language of this manuscript and to write code. The authors take full responsibility for the content of this paper, including its methods and experimental results.

References

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph IV, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego A Mares Buendia, Pavit Singh, Michael Liu, Subodh Chawla, Peter Cline, Lucy Ogaz, Ernesto Gabriel Hernández Montoya, Zihao Wang, Pavi Bhatter, Marcos Ayestaran, Bing Liu, and Yunzhong He. Prbench: Large-scale expert rubrics for evaluating high-stakes professional reasoning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 42297–42325. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1958. URL https://doi.org/10.18653/v1/2026.acl-long.1958.

Anthropic. Introducing Claude Opus 5. https://www.anthropic.com/research/claude-opus-5, July 2026.

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*, abs/2505.08775, 2025. doi: 10.48550/ARXIV.2505.08775. URL https://doi.org/10.48550/arXiv.2505.08775.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=dcjtMYkpXx.

Zhiyuan Fan, Weinong Wang, Xing Wu, and Debing Zhang. Sedareval: Automated evaluation using self-adaptive rubrics. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pp. 16916–16930. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.984. URL https://doi.org/10.18653/v1/2024.findings-emnlp.984.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10835–10866. PMLR, 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

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. doi: 10.48550/ARXIV.2507.17746. URL https://doi.org/10.48550/arXiv.2507.17746.

Kevin Han, Yuhang Zhou, Mingze Gao, Gedi Zhou, Serena Li, Abhishek Kumar, Xiangjun Fan, Weiwei Li, and Lizhu Zhang. EBPO: empirical bayes shrinkage for stabilizing group-relative policy optimization. *CoRR*, abs/2602.05165, 2026. doi: 10.48550/ARXIV.2602.05165. URL https://doi.org/10.48550/arXiv.2602.05165.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 13806–13834. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.745. URL https://doi.org/10.18653/v1/2024.acl-long.745.

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. doi: 10.48550/ARXIV.2605.26579. URL https://doi.org/10.48550/arXiv.2605.26579.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024a. URL https://openreview.net/forum?id=8euJaTveKw.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 4334–4353. Association for Computational Linguistics, 2024b. doi: 10.18653/V1/2024.EMNLP-MAIN.248. URL https://doi.org/10.18653/v1/2024.emnlp-main.248.

Guangchen Lan. Alternating reinforcement learning with contextual rubric rewards. *CoRR*, abs/2603.15646, 2026. doi: 10.48550/ARXIV.2603.15646. URL https://doi.org/10.48550/arXiv.2603.15646.

Md. Tahmid Rahman Laskar, Sawsan Alqahtani, M. Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee-Wei Tan, Md. Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 13785–13816. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.764. URL https://doi.org/10.18653/v1/2024.emnlp-main.764.

Noah Lee, Jiwoo Hong, and James Thorne. Evaluating the consistency of LLM evaluators. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 10650–10659. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.710/.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu

Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Trans. Mach. Learn. Res.*, 2023, 2023. URL https://openreview.net/forum?id=iO4LZibEqW.

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 Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.153. URL https://doi.org/10.18653/v1/2023.emnlp-main.153.

Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c4e380fb74dec9da9c7212e834657aa9-Abstract-Conference.html.

OpenAI. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925, 2025. doi: 10.48550/ARXIV.2508.10925. URL https://doi.org/10.48550/arXiv.2508.10925.

OpenAI. GPT-6 Astra: A new generation of intelligence. https://openai.com/index/gpt-6-astra/, September 2026.

Licheng Pan, Haocheng Yang, Haoxuan Li, Yichen Sun, Yunsheng Lu, Shijian Wang, Lei Shen, Yuan Lu, Zhixuan Chu, and Hao Wang. Uncertainty-aware reward modeling for stable RLHF. *CoRR*, abs/2606.19818, 2026. doi: 10.48550/ARXIV.2606.19818. URL https://doi.org/10.48550/arXiv.2606.19818.

Ross L. Prentice. Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, 81:321–327, 1986. URL https://api.semanticscholar.org/CorpusID:121104169.

Qwen Team. Qwen3-30B-A3B-Instruct-2507 model card. https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507, 2025a.

Qwen Team. Qwen3-4B-Instruct-2507 model card. https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507, 2025b.

Alexandre Ramé, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/e12a3b98b67e8395f639fde4c2b03168-Abstract-Conference.html.

Delip Rao and Christopher Callison-Burch. Autorubric: A unifying framework for rubric-based llm evaluation on non-verifiable tasks. 2026. URL https://api.semanticscholar.org/CorpusID:287199954.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *CoRR*, abs/2310.10076, 2023. doi: 10.48550/ARXIV.2310.10076. URL https://doi.org/10.48550/arXiv.2310.10076.

Kayla Schroeder and Zach Wood-Doughty. Can you trust LLM judgments? reliability of llm-as-a-judge. *CoRR*, abs/2412.12509, 2024. doi: 10.48550/ARXIV.2412.12509. URL https://doi.org/10.48550/arXiv.2412.12509.

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. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.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*, abs/2602.05125, 2026. doi: 10.48550/ARXIV.2602.05125. URL https://doi.org/10.48550/arXiv.2602.05125.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. Analyzing uncertainty of llm-as-a-judge: Interval evaluations with conformal prediction. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 11286–11328. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.569. URL https://doi.org/10.18653/v1/2025.emnlp-main.569.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86 2:420–8, 1979. URL https://api.semanticscholar.org/CorpusID:13168820.

J. Gordon Skellam. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the royal statistical society series b-methodological*, 10:257–261, 1948. URL https://api.semanticscholar.org/CorpusID:123856354.

Charles Edward Spearman. Correlation calculated from faulty data. *British Journal of Psychology*, 3: 271–295, 1910. URL https://api.semanticscholar.org/CorpusID:145312981.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Characterizing the confidence of large language model-based automatic evaluation metrics. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 2: Short Papers, St. Julian's, Malta, March 17-22, 2024*, pp. 76–89. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.EACL-SHORT.9. URL https://doi.org/10.18653/v1/2024.eacl-short.9.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators. *CoRR*, abs/2405.01724, 2024b. doi: 10.48550/ARXIV.2405.01724. URL https://doi.org/10.48550/arXiv.2405.01724.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Tencent Hy Team. Hy4 preview model card. https://github.com/Tencent-Hunyuan/Hy4-preview, 2026.

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. *CoRR*, abs/2605.20164, 2026. doi: 10.48550/ARXIV.2605.20164. URL https://doi.org/10.48550/arXiv.2605.20164.

Pat Verga, Sebastian Hofstätter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *CoRR*, abs/2404.18796, 2024. doi: 10.48550/ARXIV.2404.18796. URL https://doi.org/10.48550/arXiv.2404.18796.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/b8c90b65739ae8417e61eadb521f63d5-Abstract-Conference.html.

Abel Yagubyan. The coin flip judge? reliability and bias in llm-as-a-judge evaluation. *CoRR*, abs/2606.13685, 2026. doi: 10.48550/ARXIV.2606.13685. URL https://doi.org/10.48550/arXiv.2606.13685.

Minglai Yang, Xinyu Guo, Utkarsh Tyagi, Mian Zhang, Razvan Dumitru, Sunjie Hou, Yunzhong He, Daniel Yue Zhang, and Ying Liu. Rubric dropout: A simple way to mitigate reward hacking in rubric-as-reward RL. *CoRR*, abs/2608.11669, 2026a. doi: 10.48550/ARXIV.2608.11669. URL https://doi.org/10.48550/arXiv.2608.11669.

Qianyu Yang, Yang Liu, Jiaqi Li, Jun Bai, Hao Chen, Kaiyuan Chen, Tiliang Duan, Jiayun Dong, Xiaobo Hu, Zixia Jia, Yang Liu, Tao Peng, Yixin Ren, Ran Tian, Zaiyuan Wang, Yanglihong Xiao, Gang Yao, Lingyue Yin, Ge Zhang, Chun Zhang, Jianpeng Jiao, Zilong Zheng, and Yuan Gong. $onemillion-bench: How far are language agents from human experts? *CoRR*, abs/2603.07980, 2026b. doi: 10.48550/ARXIV.2603.07980. URL https://doi.org/10.48550/arXiv.2603.07980.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=CYmF38ysDa.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

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 (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 10586–10613. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.630. URL https://doi.org/10.18653/v1/2024.findings-acl.630.

A Statistical Derivations

A.1 Derivation of Eq. 6

For $q_g \sim \mathrm{Beta}(\alpha, \beta)$ with $\alpha = \mu\kappa$, $\beta = (1-\mu)\kappa$, and $\kappa = \phi^{-1} - 1$,

$$\mathbb{E}[q_g] = \mu, \qquad \mathrm{Var}(q_g) = \frac{\mu(1-\mu)}{\kappa+1} = \mu(1-\mu)\phi. \tag{14}$$

Using $\mathbb{E}[q_g(1-q_g)] = \mathbb{E}[q_g] - \mathbb{E}[q_g^2]$ and $\mathbb{E}[q_g^2] = \mathrm{Var}(q_g) + \mu^2$,

$$\mathbb{E}[q_g(1-q_g)] = \mu(1-\mu) - \mu(1-\mu)\phi = \mu(1-\mu)(1-\phi). \tag{15}$$

Conditional on q_g,

$$\mathbb{E}[\bar{v}_g \mid q_g] = q_g, \qquad \mathrm{Var}(\bar{v}_g \mid q_g) = \frac{q_g(1-q_g)}{K}. \tag{16}$$

The law of total variance therefore gives

$$\mathrm{Var}(\bar{v}_g) = \mathrm{Var}(q_g) + \frac{1}{K}\mathbb{E}[q_g(1-q_g)] \tag{17}$$

$$= \mu(1-\mu)\left[\phi + \frac{1-\phi}{K}\right]. \tag{18}$$

Dividing the latent between-response component by the total yields

$$\frac{\phi}{\phi + (1-\phi)/K} = \frac{K\phi}{1 + (K-1)\phi} = \rho_K. \tag{19}$$

For two distinct judgments $k \neq k'$ of the same response, conditional independence and the law of total covariance give $\mathrm{Cov}(v_{gk}, v_{gk'}) = \mathrm{Var}(q_g) = \mu(1-\mu)\phi$. Each verdict is marginally $\mathrm{Bernoulli}(\mu)$, so $\mathrm{Var}(v_{gk}) = \mu(1-\mu)$ and $\mathrm{Corr}(v_{gk}, v_{gk'}) = \phi$. Thus ϕ is the single-verdict intraclass correlation, and ρ_K is the Spearman–Brown reliability of the mean of K exchangeable verdicts.

A.2 Moment decomposition

Fix one rubric, suppress its label, and write $\bar{v}_\cdot := \frac{1}{G}\sum_{g=1}^{G} \bar{v}_g$. For $G \geq 2$, $K \geq 2$, and complete equal-K rows, the diagnostic moment estimator is

$$\widehat{\sigma}_W^2 = \frac{1}{G}\sum_{g=1}^{G} \frac{n_g(K-n_g)}{K(K-1)}, \tag{20}$$

$$s_{\bar{v}}^2 = \frac{1}{G-1}\sum_{g=1}^{G} (\bar{v}_g - \bar{v}_\cdot)^2, \tag{21}$$

$$\widehat{\sigma}_B^2 = \max\left\{ s_{\bar{v}}^2 - \widehat{\sigma}_W^2/K,\, 0 \right\}, \tag{22}$$

$$\widehat{\rho}_{\mathrm{mom}} = \frac{\widehat{\sigma}_B^2}{\widehat{\sigma}_B^2 + \widehat{\sigma}_W^2/K}. \tag{23}$$

Here $\widehat{\sigma}_W^2$ is the average within-response judge variance, $s_{\bar{v}}^2$ is the observed variance across response means, $\widehat{\sigma}_B^2$ estimates latent between-response variance, and $\widehat{\rho}_{\mathrm{mom}}$ is the corresponding estimate of ρ_K. Because $n_g \mid q_g \sim \mathrm{Binomial}(K, q_g)$, we have $\mathbb{E}[n_g(K-n_g) \mid q_g] = K(K-1)\, q_g(1-q_g)$, so $\widehat{\sigma}_W^2$ is unbiased for $\sigma_W^2 = \mathbb{E}[q_g(1-q_g)]$. By Eq. 17, the i.i.d. response means $\bar{v}_g$ have variance $\sigma_B^2 + \sigma_W^2/K$ with $\sigma_B^2 = \mathrm{Var}(q_g)$. Before truncation, $s_{\bar{v}}^2 - \widehat{\sigma}_W^2/K$ is therefore unbiased for σ_B^2. Whenever $s_{\bar{v}}^2 > 0$, the estimator reduces to $\widehat{\rho}_{\mathrm{mom}} = \max\{1 - \widehat{\sigma}_W^2/(K s_{\bar{v}}^2),\, 0\}$, which is the one-way $\mathrm{ICC}(1, K)$ of Shrout & Fleiss (1979) truncated at zero.

When the judge returns the same verdict for every response and every repetition, both $\widehat{\sigma}_W^2$ and $s_{\bar{v}}^2$ are zero, so the defining ratio is undefined and we set $\widehat{\rho}_{\mathrm{mom}} = 0$. With small G and low ϕ, sampling noise often pushes $s_{\bar{v}}^2 - \widehat{\sigma}_W^2/K$ below zero, and truncation then reports zero reliability even for rubrics with genuine between-response differences. For these reasons, the training gate of RAR^2 relies on the posterior probability R_r in Eq. 10 and the fail-open safeguards in Eq. 11, while $\widehat{\rho}_{\mathrm{mom}}$ serves only as a diagnostic.

B Implementation Details

B.1 Experimental configuration

Hardware and judges. We train Qwen3-4B on 8 NVIDIA H20 GPUs and Qwen3-30B-A3B on 32 H20 GPUs. The rubric judge gpt-oss-120b is served on a separate judge server with 16 H20 GPUs. The cross-judge analysis in Figure 6 additionally uses GPT-6 Astra (OpenAI, 2026), Claude Opus 5 (Anthropic, 2026), and Hy4 preview (Tencent Hy Team, 2026) as judges.

Training reward. Under RAR^2, each training response y_g receives the reliability-gated score S_g in Eq. 13. The training reward further includes a response-length penalty and a format reward with weight 0.1. Advantages follow GRPO (Shao et al., 2024), which centers and standardizes the rewards within each group of G responses. Table 2 lists the remaining hyperparameters.

Table 2: **Training and evaluation settings.** The gate threshold and the protected weight apply only to RAR^2.

Setting	HealthBench	PRBench	$OneMillion-Bench
Train / test prompts	4,000 / 1,000	550 / 550	200 / 200
Learning rate	1×10^{-6}	5×10^{-6}	5×10^{-6}
Warmup ratio	0	0.1	0.1
Training epochs	8	12	12
Prompt batch size	16		
Responses per prompt G	16		
Judgments per response K, train / test	5 / 3		
Rollout temperature	1.0		
Gate threshold τ	0.55		
Protected weight	$\lvert w_r\rvert \geq 8$		

B.2 Evaluation metric

We report the Expert Score on all benchmarks, following the rubric scoring rule of HealthBench (Arora et al., 2025) and PRBench (Akyürek et al., 2026). For the j-th test prompt x_j, $j = 1, \ldots, N$, let y_j be the policy response and let $c_r(x_j, y_j)$ be the verdict on rubric r averaged over the test judgments. The prompt score s_j is S^{RaR} in Eq. 1 computed on the full rubric set, that is, the weighted sum of rubric verdicts divided by the total positive weight W. The benchmark score, reported in percent, is

$$\mathrm{Score} = \max\Big\{0,\ \frac{1}{N}\sum_{j=1}^{N} s_j\Big\}. \tag{24}$$

The length penalty and the format reward are excluded from evaluation.

B.3 Distributed implementation

Row aggregation. During training, the G responses to a prompt are scored by several reward workers in parallel. Each response yields one row of K verdicts per rubric, and the rows of a group arrive asynchronously in different processes. Algorithm 2 shows how these rows are combined into a single rubric mask per prompt group. A coordinator stores rows under a stable prompt key and a round counter. A per-prompt asynchronous lock makes reading the round, appending a row, and finalizing the mask atomic, while judge calls remain outside the lock. The coordinator waits for all G rows up to a fixed time limit. Once the limit is reached, it stops waiting and computes the mask from the rows that have arrived. If fewer than three rows are available, the gate is skipped and every rubric is retained. The committed mask is cached, so late rows are scored with the same mask as the rest of their group.

Algorithm 2 Distributed reliability-gated rubric reward

Require: prompt key u, prompt x, rubric set $\mathcal{R} = \{r_1, \ldots, r_M\}$ with weights w_r, expected group size G, judgments per response K, threshold τ

sample responses $y_1, \ldots, y_G$
for all responses y_g in parallel **do**
    request K rubric-verdict vectors from the fixed judge protocol
    submit valid vectors under prompt key u and its current round
coordinator waits for all G rows or until the time limit
if fewer than three rows are available **then**
    return $m_r = 1$ for every $r \in \mathcal{R}$
else
    for all $r \in \mathcal{R}$ **do**
        form $\mathbf{n}$ for rubric r from the complete rows
        infer R_r
        if rubric r is protected or has insufficient evidence **then**
            $m_r \leftarrow 1$
        else
            $m_r \leftarrow \mathbb{I}[R_r \geq \tau]$
if $\sum_{r \in \mathcal{R}} m_r \max(w_r, 0) = 0$ **then**
    $m_r \leftarrow 1$ for every $r \in \mathcal{R}$
cache the shared rubric mask for late rows and advance the round
score each response with the same $(m_r)_{r \in \mathcal{R}}$
center and standardize scores within the group; update the policy

Posterior evaluation. Only complete rows, in which all K judgments are valid, enter the count vector $\mathbf{n}$ in Eq. 4. The posterior in Eq. 9 is evaluated on a 120×120 grid of equally spaced values of (μ, ϕ) in $[0.005, 0.995]^2$. Beta–Binomial log-probabilities are precomputed for every count and grid cell, so each rubric requires only a histogram of its counts, a weighted sum of cached terms, and a log-sum-exp normalization. The gate statistic R_r is the posterior mass of the grid cells with $\rho_K \geq \frac{1}{2}$, which is equivalent to $\phi \geq 1/(K+1)$.

C Further Analysis

C.1 Effect of the gate threshold τ

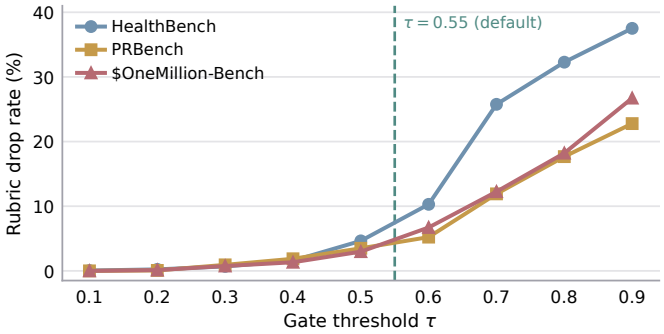

Figure 7: **Rubric drop rate under different gate thresholds τ.** The drop rate is the fraction of rubrics removed by the reliability gate during training. The dashed line marks the default $\tau = 0.55$.

Figure 7 compares the rubric drop rate during training under different thresholds τ, where the drop rate is the fraction of rubrics removed by the reliability gate in Eq. 11. For small τ, the gate retains almost all rubrics and RAR^2 effectively reduces to RaR. At $\tau = 0.1$, the drop rate is at most 0.05% on every benchmark. The drop rate rises sharply on all three benchmarks once τ reaches 0.7 and continues to increase for larger τ. Such aggressive filtering removes a substantial share of the learning signal available for training. We therefore set $\tau = 0.55$, below the point at which the drop rate rises sharply.

D Detailed Results on PRBench and $OneMillion-Bench

Table 3 reports PRBench results for both policy models across domains and difficulty levels. Tables 4 and 5 report $OneMillion-Bench results across subsets and domains for Qwen3-4B and Qwen3-30B-A3B, respectively.

Table 3: **Performance comparison on PRBench by domain and difficulty.** Boldface marks the best mean for each slice.

Slice	RaR	ARL	POW3R	RAR^2
Policy model: Qwen3-4B				
Finance Non-Hard	37.10 ± 0.3120	38.19 ± 0.3492	39.58 ± 0.2720	$\mathbf{40.33} \pm 0.2633$
Finance Hard	24.07 ± 0.4588	24.78 ± 0.3379	25.81 ± 0.2561	$\mathbf{26.16} \pm 0.2024$
Legal Non-Hard	42.52 ± 0.6220	43.77 ± 0.5157	44.65 ± 0.2873	$\mathbf{46.22} \pm 0.3887$
Legal Hard	12.19 ± 0.5414	12.55 ± 0.3376	12.76 ± 0.3965	$\mathbf{14.28} \pm 0.1322$
Policy model: Qwen3-30B-A3B				
Finance Non-Hard	48.15 ± 0.9361	49.71 ± 1.0592	51.20 ± 0.9027	$\mathbf{53.38} \pm 0.5051$
Finance Hard	38.96 ± 1.1789	39.02 ± 1.0860	39.28 ± 0.8673	$\mathbf{41.19} \pm 0.6233$
Legal Non-Hard	46.63 ± 1.1661	47.55 ± 0.9423	49.13 ± 0.9221	$\mathbf{50.86} \pm 0.3978$
Legal Hard	38.39 ± 1.0869	39.51 ± 1.0882	40.23 ± 0.7669	$\mathbf{42.24} \pm 0.4932$

Table 4: **Performance comparison of Qwen3-4B on $OneMillion-Bench subsets.** Each entry reports mean score (%) $\pm$ run-to-run variance over three runs. Boldface marks the best mean for each slice.

Domain	RaR	ARL	POW3R	RAR^2
Global subset				
Finance	21.10 ± 0.6391	22.03 ± 0.7902	22.92 ± 0.5329	$\mathbf{25.52} \pm 0.2362$
Law	22.46 ± 0.9863	22.41 ± 0.7126	24.07 ± 0.3298	$\mathbf{25.79} \pm 0.5683$
Healthcare	25.28 ± 0.7565	25.22 ± 0.9203	26.38 ± 0.8929	$\mathbf{27.82} \pm 0.1661$
Natural Sci.	23.59 ± 0.8546	23.63 ± 0.4685	22.94 ± 0.3827	$\mathbf{24.28} \pm 0.2816$
Industry	25.42 ± 0.8723	26.10 ± 0.5406	26.79 ± 0.4269	$\mathbf{27.14} \pm 0.1899$
CN subset				
Finance	24.49 ± 0.8674	24.72 ± 0.6989	25.35 ± 0.7587	$\mathbf{27.79} \pm 0.7855$
Law	27.28 ± 0.7854	28.16 ± 0.8848	29.61 ± 0.7859	$\mathbf{30.61} \pm 0.5897$
Healthcare	24.38 ± 0.8617	$\mathbf{24.85} \pm 0.6321$	23.85 ± 0.6571	24.76 ± 0.5983
Natural Sci.	21.84 ± 0.5224	22.03 ± 0.9051	23.63 ± 0.8762	$\mathbf{32.41} \pm 0.7304$
Industry	27.31 ± 0.9137	27.69 ± 0.7587	$\mathbf{29.77} \pm 0.6159$	29.25 ± 0.5871

Table 5: **Performance comparison of Qwen3-30B-A3B on $OneMillion-Bench subsets.** Each entry reports mean score (%) $\pm$ run-to-run variance over three runs. Boldface marks the best mean for each slice.

Domain	RaR	ARL	POW3R	RAR^2
Global subset				
Finance	28.36 ± 0.7467	28.51 ± 0.4135	29.39 ± 0.5005	$\mathbf{31.84} \pm 0.5496$
Law	29.87 ± 0.8361	29.69 ± 0.8315	30.33 ± 0.6224	$\mathbf{32.19} \pm 0.6884$
Healthcare	35.76 ± 0.5328	36.53 ± 1.0089	36.72 ± 0.6878	$\mathbf{37.26} \pm 0.7581$
Natural Sci.	29.17 ± 0.8433	29.83 ± 1.2671	31.52 ± 0.2348	$\mathbf{33.97} \pm 0.5731$
Industry	32.03 ± 0.9443	31.97 ± 0.9233	33.45 ± 0.7837	$\mathbf{33.89} \pm 0.8206$
CN subset				
Finance	30.02 ± 0.8522	31.24 ± 1.0085	31.97 ± 0.8942	$\mathbf{33.43} \pm 0.9182$
Law	32.79 ± 0.7875	32.55 ± 0.8269	33.16 ± 0.8152	$\mathbf{34.97} \pm 0.5285$
Healthcare	36.25 ± 0.6943	37.34 ± 0.8571	38.01 ± 0.7732	$\mathbf{39.44} \pm 0.3948$
Natural Sci.	29.03 ± 1.0420	29.98 ± 0.6882	31.20 ± 0.6942	$\mathbf{33.87} \pm 0.4339$
Industry	33.42 ± 1.0520	34.81 ± 0.7846	34.79 ± 0.5283	$\mathbf{35.64} \pm 0.6116$