RUBRIC REWARDS FROM ITEM RESPONSE THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Many language tasks have no single answer that can be checked automatically. Rubrics provide criteria for judging responses to these tasks. For reinforcement learning, the resulting verdicts must be combined into a scalar reward. A common approach sums the points assigned to satisfied criteria. Distinct verdict patterns can thus receive the same reward, and the fixed points encode how much each criterion should count, not how strongly its verdict distinguishes the current rollouts. Beyond this aggregation problem, judging the full rubric needs more judge requests as the criterion count grows. To address these limitations, Rubric Response Theory (RRT) measures quality and selects criteria when rubric criteria are monotone indicators of a shared target. Rather than adding assigned points, RRT uses a two parameter item response model that treats the verdict pattern as evidence about scalar quality specific to the rubric. Under this model, its likelihood score maximizes the local signal-to-noise ratio for quality. Its Response Parameter Network (RPN) reads the prompt and criterion text to predict criterion difficulty and discrimination. As the policy distribution changes during training, RRT uses online expectation maximization to update the RPN from current rollout verdicts. With Qwen3.5-4B as the policy, RRT's macro criterion score across Medical, Science, Rubrics as Rewards Science, and RubricBench is 1.7 points above that of group relative policy optimization (GRPO). On hard and very hard criteria in Medical and Science, RRT gains 2.8 to 5.6 points over GRPO. At half the criterion budget, adaptive Fisher selection with a frozen RPN keeps the macro criterion score across four datasets within 0.1 points of GRPO with full judging. These results show RRT can reduce judge requests while remaining competitive with GRPO. Code

1 INTRODUCTION

Reinforcement learning for language models has a well-defined reward when success is automatically verifiable, but many language tasks have no single correct answer. Such tasks use large language models as judges (Zheng et al., 2023) and rubrics specific to each prompt (Gunjal et al., 2026; Viswanathan et al., 2025). A rubric decomposes an evaluation target into natural language criteria such as required content, reasoning steps, output constraints, and errors to avoid. For policy training, the resulting binary criterion verdicts are mapped to a scalar reward for each rollout. Beyond this aggregation problem, judging the full rubric requires more judge requests as criteria are added.

To produce this scalar reward, the baseline takes a weighted average of satisfied criteria using their assigned points (Gunjal et al., 2026; Arora et al., 2025). This additive form can assign the same total to distinct verdict patterns and therefore gives them the same rubric reward in group relative policy optimization (GRPO) (Guo et al., 2025). The assigned points encode how much each criterion should count in the rubric, not how strongly its verdict distinguishes the current rollouts. The contribution of each verdict also does not depend on the rollout's other verdicts.

This work introduces Rubric Response Theory (RRT), which adapts item response theory (IRT) to on-policy reward inference and criterion selection for rubrics specific to each prompt. For rubrics whose criteria are monotone indicators of one shared target, RRT uses the verdicts to estimate rollout quality instead of adding assigned points. Like questions in a test, criteria can differ in difficulty and in how strongly they distinguish responses of different quality. RRT represents each rubric criterion with difficulty and discrimination parameters from IRT (Chen et al., 2025). Difficulty locates a criterion on the quality scale, while discrimination controls how sharply its pass probability changes with quality. The Response Parameter Network (RPN) predicts these parameters from the prompt

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.

and criterion text. Using these parameters, RRT finds the quality that best explains the full verdict pattern and uses it as the GRPO reward. Rubric points specify how much a criterion should count, while RRT measures how much its verdict reveals about shared quality. This distinction lets RRT distinguish rollouts that receive the same reward based on rubric points but whose verdict patterns provide different evidence about quality. Because the rollout distribution changes with the policy, RRT uses online hard expectation maximization (EM) to update the RPN from current rollout verdicts. It uses the same parameters to select criteria under a criterion budget.

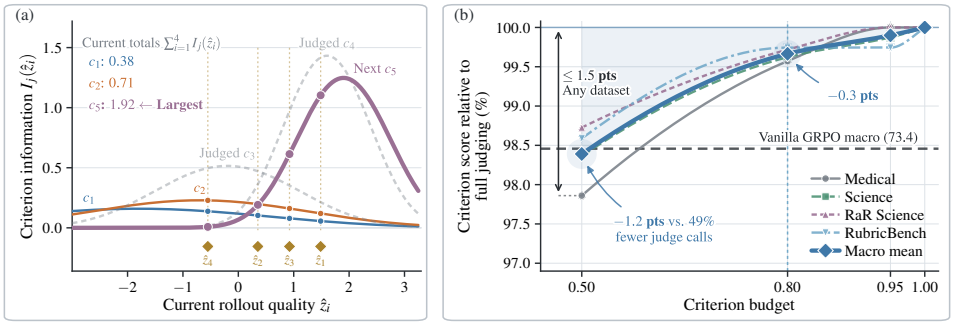

Figure 1: Criterion selection in RRT using Fisher information. (a) Criterion information for five criteria and four rollouts after judging two criteria. Gold diamonds show inferred rollout qualities. Dots show information from unjudged criteria at these qualities, dashed gray curves show judged criteria, and c_5 is selected next. (b) Criterion scores after RRT with adaptive Fisher selection, normalized to full judging for each dataset. Shading shows the macro gap from full judging.

The contributions are the following.

- RRT formulates rubric aggregation as Bayesian inference, using the posterior mode of quality as a scalar GRPO reward to distinguish rollouts with equal rubric point totals. It combines the likelihood of full verdict patterns with a quality prior. The RPN predicts criterion difficulty and discrimination from prompt and criterion text for unseen rubrics.
- RRT provides an online EM procedure to calibrate the RPN from current rollout verdicts as the policy changes. A theoretical analysis establishes that, in RRT's item response model, the rubric likelihood score maximizes the local signal-to-noise ratio (SNR) for small changes in rollout quality among all scalar functions of the verdicts, reaching the Fisher information bound.
- RRT extends adaptive Fisher selection to groups of rollouts, reducing judge requests under a criterion budget. It ranks unjudged criteria by total Fisher information at inferred rollout qualities and updates those qualities after each selected criterion is judged.

Key Findings: Across Medical, Science, Rubrics as Rewards (RaR) Science, and RubricBench, RRT preserves the gains in macro criterion score from Vanilla GRPO across three base policies. It exceeds Vanilla GRPO by 1.7 points on Qwen3.5-4B. On rollouts from trained policies, adaptive Fisher selection reaches 95.0% mean Pearson correlation with GRPO advantages from full rubric judging while leaving 21.0% of criteria unjudged. At half the criterion budget, RRT with adaptive Fisher selection keeps its macro criterion score across all four datasets within 0.1 points of Vanilla GRPO with full judging (Figure 1).

2 Related Work

Rubric and checklist supervision provides training signals specific to each prompt. Gunjal et al. (2026) generate rubrics from reference answers and train with GRPO on a weighted sum of criterion verdicts or one rating from a judge that reads the full rubric. They call fixed weights brittle and leave learned weighting to future work. Viswanathan et al. (2025) generate checklists from failure modes in candidate responses and combine judge and program verifier scores with generated importance weights. Both fix criterion weights in advance or leave aggregation to the judge, and judge the full rubric for every response. RRT complements rubric generation. It learns from current rollouts how much each verdict reveals about quality, to compute the reward and to choose which criteria to judge.

Recent work adapts rubric aggregation to current rollouts. POW3R (Tyagi et al., 2026) rescales human weights by contrast across rollouts, and DIVA (Cook et al., 2026) weights soft criterion scores

by their variance across responses. Both keep a weighted sum with one weight per criterion for all rollouts and judge every criterion. RRT replaces the weighted sum. It learns each criterion's difficulty and how well it separates good from poor responses, then finds the overall quality that best explains a rollout's full verdict pattern. Two rollouts with equal point totals can thus receive different rewards. This knowledge also lets RRT choose which criteria to judge and reduce judge requests.

IRT and learned rubric measurement serve calibration, assessment, data selection, and efficient evaluation. Hashemi et al. (2024) train a network to calibrate a judge's rubric answers to human ratings. Uto (2021) uses a Rasch model with item and rater effects to assess examinees. Lalor et al. (2019) fit IRT to many neural models' responses and filter training data by difficulty. Adaptive testing selects questions from a calibrated pool to measure ability with fewer questions (Weiss, 1982). Truong et al. (2025) predict question difficulty from text for adaptive testing of language models. Each measures a fixed examinee or model. RRT applies this measurement to policy training, where each rollout's measured quality becomes its reward. From the prompt and criterion text, RRT predicts each criterion's difficulty and how well it separates responses, so it works on unseen rubrics. Training verdicts then refine these estimates, so they stay current as the policy changes. The same estimates pick the most informative criteria for the current rollouts, so RRT judges fewer criteria per prompt.

3 METHOD

3.1 IRT FOR RUBRIC CRITERIA

For a prompt q, let rollouts $i = 1, \ldots, N$ be judged against rubric criteria $c_1, \ldots, c_K$. Each verdict is encoded as $G_{ij} \in \{0, 1\}$, where 1 denotes satisfaction of a positive criterion or avoidance of a pitfall (Appendix F.2). Let $w_j > 0$ be the available points for criterion j. The reward based on rubric points is $R_i^{\text{base}} = \sum_j w_j \, G_{ij} / \sum_j w_j \in [0, 1]$. RRT treats each rubric criterion as an item whose verdict provides evidence about scalar quality z_i for the rubric. The model assumes that every criterion's pass probability increases strictly with z_i. For a fixed rubric and criterion parameters, the expected reward based on rubric points therefore increases strictly with quality (Theorem 8 in Appendix A.5). The model also assumes local independence, so verdicts are conditionally independent given quality and criterion parameters (Chen et al., 2025).

Criterion j has difficulty b_j and discrimination a_j. The RPN ψ predicts these parameters from the prompt q and criterion text c_j, so $(a_j, b_j) := \psi(q, c_j)$ with $a_j > 0$. For rollout quality z_i, define $u_{ij} := a_j(z_i - b_j)$ and $P_{ij} := F(u_{ij})$. The response function $F : \mathbb{R} \to (0, 1)$ is differentiable and strictly increasing, with $F(0) = 0.5$. Thus b_j is the quality where P_{ij} crosses 0.5 with slope $a_j F'(0)$. This item response model has two parameters per criterion and lets discrimination vary (Birnbaum, 1968). With a logistic response function and $a_j = 1$ for every criterion, it reduces to the Rasch model with one parameter per criterion (Rasch, 1966). Appendix A.1 states the assumptions on F.

3.2 BAYESIAN REWARD INFERENCE AND ONLINE EM

RRT infers quality with the criterion parameters held fixed. Write $c = (c_1, \ldots, c_K)$ for the rubric and $G_i = (G_{i1}, \ldots, G_{iK})$ for rollout i's verdict vector. Since $G_{ij} \mid z_i, q, c_j, \psi \sim \text{Bernoulli}(P_{ij})$, local independence gives

$$p_\psi(G_i \mid z_i, q, c) = \prod_{j=1}^{K} p_\psi(G_{ij} \mid z_i, q, c_j) = \prod_{j=1}^{K} P_{ij}^{G_{ij}} (1 - P_{ij})^{1 - G_{ij}}. \tag{1}$$

RRT places a fixed Gaussian prior with mean zero and variance σ_z^2 on quality. If $\ell_{ij}(z)$ is the log of criterion j's Bernoulli factor, Bayes' rule gives

$$p_\psi(z_i \mid G_i, q, c) \overset{\text{Bayes}}{\propto} p(z_i) \prod_{j=1}^{K} P_{ij}^{G_{ij}} (1 - P_{ij})^{1 - G_{ij}}, \quad \ell_i(z) := \sum_{j=1}^{K} \ell_{ij}(z) - z^2/(2\sigma_z^2). \tag{2}$$

RRT uses the standard Gaussian CDF, $F = \Phi$, and the maximum a posteriori (MAP) reward $R_i = \hat{z}_i = \arg\max_z \ell_i(z)$ (Mislevy, 1986). The Gaussian CDF lets criterion difficulty affect reward ordering, while the logistic CDF orders rollouts by $\sum_j a_j G_{ij}$ (Theorems 4 and 5 in Appendix A.1). Figure 2 shows how the full verdict pattern determines this reward.

For fixed criterion parameters, the log posterior is strictly concave and has a unique maximizer. Satisfying an additional criterion increases the MAP reward when the other verdicts and criterion

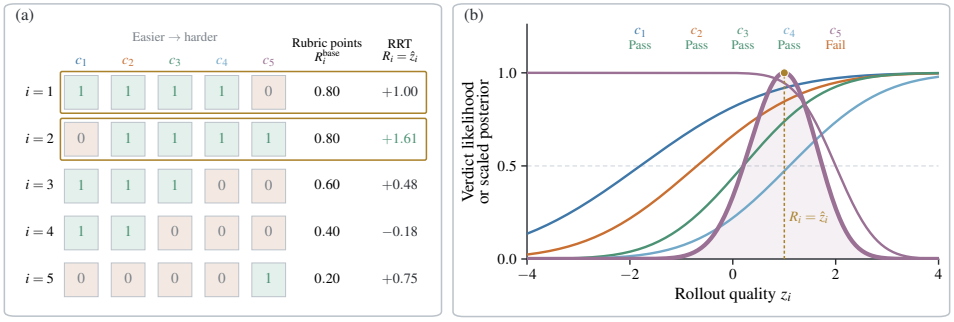

Figure 2: RRT reward inference for an illustrative rubric. (a) Verdict vectors for equally weighted criteria ordered by difficulty, with both rewards shown. (b) E-step for rollout $i = 1$. The thick curve combines the verdict likelihoods with the Gaussian prior, and its mode is the RRT reward.

parameters are fixed (Theorem 6 and Corollary 1 in Appendix A.2). RRT finds this maximizer by bisection using the log posterior gradient from Theorem 3 in Appendix A.1. RRT uses $R_i = \hat{z}_i$ as the GRPO reward (Guo et al., 2025), while Vanilla GRPO uses R_i^{base}, the normalized points score used in evaluation. Appendix B.2 specifies the GRPO update.

As the policy changes during joint training, RRT calibrates ψ from new rollout verdicts using one online EM sweep per policy step (Bach et al., 2015). The E-step reuses Eq. 2 to compute a posterior mode for each rollout. With the posterior modes held fixed, the M-step loss for one prompt group is

$$\mathcal{L}(\psi) = -\frac{1}{NK}\sum_{i=1}^{N}\sum_{j=1}^{K}\ell_{ij}(\hat{z}_i;\psi) + \frac{\lambda_a}{K}\sum_{j=1}^{K}(\log a_j)^2. \tag{3}$$

The regularizer pulls a_j toward one when $\lambda_a > 0$. Appendix B.1 gives the EM objective, algorithms, and implementation details.

3.3 Fisher Information for Aggregation and Selection

The item response model used for reward inference also quantifies each criterion's Fisher information about quality. Let ϕ be the standard Gaussian density and define the criterion likelihood score as $\mathcal{S}_{ij}(z_i) := \partial_{z_i}\ell_{ij}(z_i)$.

Theorem 1 (Fisher information of a rubric criterion)**.** *Under the item response model with $F = \Phi$, the criterion likelihood score has conditional mean zero. Its Fisher information about rollout quality is*

$$I_j(z_i) = \mathbb{E}\left[\mathcal{S}_{ij}(z_i)^2 \mid z_i\right] = a_j^2\frac{\phi(u_{ij})^2}{P_{ij}(1-P_{ij})} = -\mathbb{E}\left[\frac{\partial^2\ell_{ij}(z_i)}{\partial z_i^2}\,\middle|\, z_i\right]. \tag{4}$$

Appendix A.3 gives the proof. Summing criterion information gives a local bound for rubric aggregation. Write $\mathcal{S}_i(z_i) := \sum_j \mathcal{S}_{ij}(z_i)$ for the rubric likelihood score and $I(z_i) := \sum_j I_j(z_i)$ for the rubric information.

Theorem 2 (Local information bound for rubric aggregation)**.** *Under $F = \Phi$, fix a quality level z_i and the criterion parameters, and assume conditionally independent verdicts. For a statistic $T(G_i)$ with $0 < \mathrm{Var}(T \mid z_i) < \infty$, define its local SNR for quality as*

$$\mathrm{SNR}_T(z_i) := \left(\partial_{z_i}\mathbb{E}[T \mid z_i]\right)^2 / \mathrm{Var}(T \mid z_i). \tag{5}$$

Then $\mathrm{SNR}_T(z_i) \le I(z_i)$, with equality if and only if $T - \mathbb{E}[T \mid z_i] = \gamma\,\mathcal{S}_i(z_i)$ almost surely for some $\gamma \neq 0$. Up to a nonzero affine map, the rubric likelihood score is therefore the unique scalar verdict signal that achieves this bound. For the reward based on rubric points, equality holds if and only if $w_j \propto a_j\phi(u_{ij})/[P_{ij}(1-P_{ij})]$ for every j. Once two criteria have different criterion parameters, no fixed vector of rubric points achieves the bound at every quality level.

Appendix A.5 gives the proof. The MAP reward combines the rubric likelihood score and prior through realized posterior curvature. A Fisher scoring surrogate uses the expected local precision at a common quality level (Theorem 9 and Eq. 15 in Appendix A.5). Appendix A.4 derives the statistic with minimum variance subject to unit local response to quality and the expected local precision.

Fisher information also guides criterion selection under a budget. It measures expected information before judging, while the likelihood score measures evidence from an observed verdict (Figure 3 in Appendix A.3). Following classical adaptive testing (Weiss, 1982), RRT uses the criterion information from Theorem 1 to rank unjudged criteria by $\sum_i I_j(\hat{z}_i)$ at the qualities currently inferred for the rollout group. The qualities are updated after each selected criterion is judged.

4 EXPERIMENTS

The experiments are designed to address the three following research questions (RQs).

RQ1: How does RRT compare with the baselines in verdict prediction, reward variation, and ordering stability within each prompt?

RQ2: How does RRT affect criterion score, normalized points score, and training behavior across datasets and policies?

RQ3: What tradeoff does selection based on Fisher information provide between judge usage, reward fidelity, and evaluation scores?

4.1 EXPERIMENTAL SETUP

The default base policy is Qwen3.5-4B (Qwen Team, 2026), and the datasets are Medical, Science, Rubrics as Rewards (RaR) Science, and RubricBench. Transfer comparisons also use Qwen3.5-2B and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Appendix F.1 gives data sources, splits, and roles. Each condition has one training run, and each policy step uses 32 prompts, eight rollouts per prompt, and one proximal policy optimization (PPO) epoch. Each policy comparison keeps the data, base checkpoint, sampler, judge, and GRPO implementation fixed. GPT-5.5 (OpenAI, 2026) with reasoning disabled gives one binary verdict per rollout and criterion. Using these verdicts, Vanilla GRPO trains on the reward R_i^{base} based on rubric points. The main RRT condition applies one stochastic partial M-step per policy step.

Criterion score is the percentage of rubric criteria satisfied. Normalized points score is the weighted fraction of available points. Scores, ROC-AUC values, correlations, rates, and shares are reported as percentages, and differences between them are percentage points. Macro means are unweighted. Policy scores average three evaluation sampling seeds. Reported confidence intervals are 95% bootstrap intervals over prompt groups, paired when two conditions are compared on the same groups. They quantify variation across prompt groups with the trained policies fixed. Appendices F.2 and F.3 give the policy input and judge prompts, full configuration, and checkpoint selection.

4.2 VERDICT PREDICTION AND REWARD STABILITY

Leave-one-criterion-out prediction infers rollout quality from the other $K-1$ verdicts. For their verdicts $G_{i,-j}$ and texts c_{-j}, the prediction is $\widehat{P}_{ij}^{(-j)} = \int F(a_j(z-b_j))\, p_\psi(z \mid G_{i,-j}, q, c_{-j})\, dz$. Under this protocol, four methods predict G_{ij} from the same rollouts. RRT uses a frozen RPN. Its matched $a_j = 1$ baseline adjusts b_j to preserve marginal pass rate. The RPN baseline without rollout evidence uses $\mathbb{E}_{z\sim\mathcal{N}(0,\sigma_z^2)}[F(a_j(z-b_j))]$. The other baseline averages the remaining $K-1$ verdicts. The evaluation reports ROC-AUC within each criterion and pooled ROC-AUC. Results are averaged over Qwen3.5-2B and Llama-3.1-8B-Instruct. The macro mean covers the four datasets.

RRT gains 10.1 points in pooled macro ROC-AUC over the mean of other verdicts and also exceeds the RPN without rollout evidence on every dataset (Table 1). All three methods using rollout evidence reach 65.8% macro ROC-AUC within each criterion.

The RPN configuration ablations infer quality from all verdicts and vary the response function, parameter count, text conditioning, embedder, and policy. The adopted model improves macro ROC-AUC by 0.7 points over the model with $a_j = 1$ (Table 5 in Appendix C.1). The RPN representation analysis also compares predicted and empirical criterion difficulty, with Spearman correlations of 38.0% to 47.9% (Table 6 in Appendix C.2). To compare reward orderings, both response functions use $a_j = 1$ and the same criterion difficulties. The Gaussian CDF separates 83.3% of rollout pairs with equal pass counts on Medical and 39.3% on Science, while the logistic CDF leaves all such pairs tied (Table 7 in Appendix C.3).

Table 1: Leave-one-criterion-out verdict prediction by ROC-AUC. Results within each criterion use the RPN without rollout evidence as the reference, and pooled results use the mean of the other verdicts. Blue marks RRT, and green parentheses give gains from the reference.

Prediction method	Science	RaR Science	Medical	RubricBench	Macro mean
Rollout evidence: ROC-AUC within each criterion					
RPN without rollout evidence	50.0	50.0	50.0	50.0	50.0
Mean of other verdicts	69.6	67.2	63.8	62.5	**65.8**
RRT, $a_j = 1$	69.7	67.0	63.9	**62.8**	**65.8**
RRT	**70.0** (20.0)	**67.4** (17.4)	**64.0** (14.0)	61.8 (11.8)	**65.8** (15.8)
Joint rollout and criterion ranking: pooled ROC-AUC					
Mean of other verdicts	69.4	70.6	67.0	70.5	69.3
RPN without rollout evidence	74.6	76.5	76.7	68.8	74.1
RRT, $a_j = 1$	79.2	81.2	80.1	**77.2**	79.4
RRT	**79.7** (10.3)	**81.4** (10.8)	**80.3** (13.3)	76.6 (6.1)	**79.5** (10.1)

The calibration analysis compares online and frozen RPNs in predicting empirical criterion pass rates as the policy changes. Online updates increase macro Pearson correlation by 1.7 points on the next policy step, with gains on both datasets (Table 10 in Appendix C.6). Two criterion trajectories selected post hoc end with lower predicted difficulty and higher empirical pass rates (Figure 6). The RubricBench RPN is fitted with hard (posterior mode) and soft (full posterior) E-steps at two discrimination regularizer weights. At both weights, the hard E-step improves ROC-AUC by 0.6 to 1.3 points and reduces criterion loss by 0.011 to 0.017 (Table 11 in Appendix C.7).

Reward variation within prompts is compared between RRT and rubric points across policies and rollout group sizes. Relative to rubric points, RRT has 1.2 to 2.2 times the variance share within prompts and 2% to 58% fewer tied pairs in all 12 dataset and policy cells. With groups drawn from 48 rollouts per prompt, the tied pair share falls by 2.2 to 2.3 points on Medical and 1.5 to 1.6 on Science across tested group sizes from 2 to 32, including the training size $n = 8$ (Figure 5 and Table 8 in Appendix C.4). In the companion analysis, 52.3% to 80.5% of criteria receive the same verdict across all rollouts in a group (Table 9 in Appendix C.5).

Variation in judge verdicts is measured with three independent verdicts for each prompt, response, and criterion. Mean disagreement with the consensus verdict is 1.27% across all 12 dataset and policy cells (Tables 12 and 13 in Appendix D.1). Stable nonzero ordering share is the fraction of rollout pairs that remain separated and retain their order. On repeated verdicts, RRT with marginal calibration increases its macro mean by 5.4 points over rubric points. Under corruption of the consensus verdicts, mean gains over rubric points are 0.99 to 2.00 points with estimated a_j, exceeding gains with $a_j = 1$ at every tested level in both corruption channels. At the strongest tested level, 0.10, its mean order flip rate is lower than rubric points whether corruption targets split verdicts or all verdicts (Tables 14 and 15 in Appendix D.2).

The comparison of noise structures measures cosine similarity to advantages from clean criterion scores under controlled corruption. At level 0.20, RRT with marginal calibration gains 1.3 to 5.1 points in mean similarity over criterion score across four channels that corrupt criterion subsets. Criterion score leads by 0.8 to 1.2 points under lenient noise tied to response length and symmetric noise on all criteria (Table 16 in Appendix D.3). The companion comparison varies corruption level on a fixed 30% criterion subset and includes POW3R and DIVA. At each tested nonzero level, RRT with marginal calibration has the highest similarity on every dataset, with gains of 3.0 to 7.0 points over criterion score at 0.20 (Table 17 in Appendix D.3).

Variation across rollout samples is measured with pass rates from disjoint blocks of eight rollouts. Reliability, measured by Pearson correlation between blocks, is 93.6% to 96.0% over all criteria and 70.9% to 71.9% over uncertain criteria, whose empirical pass rates over 48 rollouts lie between 10% and 90% (Table 18 in Appendix D.4). For criterion sampling, orderings from random rubric halves are compared using Kendall's τ, with RRT parameters from the online RPN. RRT raises rank correlation over rubric points by 15.1 points on RaR Science and 0.5 to 3.6 points on the other datasets (Table 19 in Appendix D.5). Across quintiles of rank correlation from rubric points, RRT's gain falls from 14 points in the lowest quintile to 1 point in the highest (Table 20 in the same appendix).

To check local independence, the analysis measures dependence between criterion verdicts after accounting for fitted quality. Fitted quality explains 72.8% to 85.3% of pairwise mutual information

(Table 21 in Appendix D.6). For pairs with the most similar criterion text, the share with residual redundancy exceeds the bootstrap null rate by 3.4 to 11.9 points. For the least similar text, it stays within 0.7 points of the null rate.

4.3 Policy Training Results

This experiment compares policy performance across rubric rewards and criterion parameter settings. Baseline rewards include POW3R (Tyagi et al., 2026) and DIVA (Cook et al., 2026) (Appendix F.4). Empirical pass rate conditions use $a_j = 1$ and $b_j = 1 - 2\bar{G}_j$, using pass rates from batch or cached rollouts (Appendix D.4). RRT + frozen RPN and RRT + online RPN predict both a_j and b_j from text.

Table 2: Criterion and normalized points scores. Bold marks the highest score in each column. RRT variants use parameters from empirical pass rates or the RPN, with the same reward and E-step. RubricBench uses one point per criterion, so both scores are equal.

	Medical		Science		RaR Science		RubricBench	Macro mean	
Condition	Criterion score	Normalized points score	Criterion score	Normalized points score	Criterion score	Normalized points score	Criterion score	Criterion score	Normalized points score
Base policy	64.1	63.3	76.5	70.9	68.7	71.0	74.7	71.0	70.0
Vanilla GRPO	67.7 (3.6)	66.8 (3.5)	78.3 (1.8)	77.8 (6.9)	70.6 (1.9)	73.7 (2.7)	77.0 (2.3)	73.4 (2.4)	73.8 (3.8)
POW3R	70.6 (6.5)	69.6 (6.3)	79.0 (2.5)	78.7 (7.8)	70.1 (1.4)	73.2 (2.2)	77.4 (2.7)	74.3 (3.3)	74.7 (4.7)
DIVA	64.9 (0.8)	64.4 (1.1)	78.8 (2.3)	78.3 (7.4)	70.6 (1.9)	73.8 (2.8)	75.8 (1.1)	72.5 (1.5)	73.1 (3.1)
Same RRT reward $R_i = \hat{z}_i$ and E-step across criterion parameter settings									
RRT +									
↪ batch pass rate	68.0 (3.9)	67.0 (3.7)	78.3 (1.8)	77.8 (6.9)	70.8 (2.1)	73.5 (2.5)	77.3 (2.6)	73.6 (2.6)	73.9 (3.9)
↪ cached pass rate	67.8 (3.7)	66.8 (3.5)	78.3 (1.8)	77.9 (7.0)	70.0 (1.3)	73.0 (2.0)	77.4 (2.7)	73.4 (2.4)	73.8 (3.8)
↪ frozen RPN	70.1 (6.0)	**69.8** (6.5)	79.5 (3.0)	79.2 (8.3)	70.7 (2.0)	73.8 (2.8)	77.9 (3.2)	74.6 (3.6)	75.2 (5.2)
↪ online RPN	**70.9** (6.8)	**69.8** (6.5)	**79.8** (3.3)	**79.4** (8.5)	**70.9** (2.2)	**74.0** (3.0)	**78.7** (4.0)	**75.1** (4.1)	**75.5** (5.5)

RRT + online RPN is highest or tied in every column of Table 2, exceeding Vanilla GRPO by 1.7 and POW3R by 0.8 points on both macro metrics. The frozen RPN is 1.0 to 1.2 macro criterion points above empirical pass rate conditions, and the online RPN scores 0.5 points above the frozen RPN.

RRT's macro criterion score is 0.2 points below Vanilla GRPO on Qwen3.5-2B and 0.1 points above it on Llama-3.1-8B-Instruct. It scores higher on RubricBench with both policies, as on Qwen3.5-4B (Table 22 in Appendix E.1). Difficulty bands are defined by empirical criterion difficulty under the base policy. RRT + online RPN exceeds Vanilla GRPO by 2.8 to 5.6 points in seven of eight bands on Medical and Science, including every Medium, Hard, and Very hard band (Figure 7 in Appendix E.2). Policies trained on Medical and Science are also evaluated on HealthBench and ResearchQA, respectively. RRT gains 0.1 to 0.7 macro criterion points over Vanilla GRPO (Table 23 in Appendix E.3). The policy comparison also measures response length. RRT has lower median response length in all 12 dataset and policy combinations, with the mean of dataset medians 10.6% to 47.9% below Vanilla GRPO across policies (Figure 8 in Appendix E.4).

4.4 Criterion Selection Based on Fisher Information

The comparison finds the smallest criterion budget reaching 95.0% reward fidelity, the mean Pearson correlation between GRPO advantages from partial and full judging. All K criteria give $\mathbf{A}^{(K)}$ and m selected criteria give $\mathbf{A}^{(m)}$. Methods share a frozen RPN and verdict matrix per dataset. The selection methods are random, discrimination (a_j^2), static Fisher ($NI_j(0)$), and adaptive Fisher selection.

Table 3: Unjudged criteria at the smallest budget reaching 95.0% correlation with GRPO advantages from the full rubric on rollouts from trained policies. Parentheses give gains over random selection.

Selection method	Medical	Science	RaR Science	RubricBench	Macro mean
Random	10.7%	15.2%	5.1%	8.9%	10.0%
Discrimination (a_j^2)	20.0% (9.3)	18.0% (2.7)	5.1% (0.0)	8.9% (0.0)	13.0% (3.0)
Static Fisher	20.0% (9.3)	20.5% (5.3)	18.3% (13.2)	18.6% (9.6)	19.3% (9.3)
Adaptive Fisher	20.8% (10.1)	22.9% (7.6)	18.3% (13.2)	21.9% (13.0)	21.0% (11.0)

On rollouts from trained policies, adaptive Fisher leaves 11.0 points more criteria unjudged than random selection and 1.7 more than static Fisher in the macro mean (Table 3). Discrimination alone

matches random selection on RaR Science and RubricBench, where static Fisher leaves 13.2 and 9.6 points more criteria unjudged, respectively. On rollouts from base policies, macro gains over random selection are 10.6 points for static Fisher and 9.5 for adaptive Fisher (Table 24 in Appendix E.5).

To measure how these budgets affect policy training, this experiment compares criterion score and normalized points scores under full and partial judging. All RRT conditions share the frozen RPN. Random selection uses a criterion budget of 0.50 for Vanilla GRPO and RRT. Adaptive Fisher budgets are 0.50, 0.80, and 0.95, with full judging for reference.

Table 4: Scores under full and partial judging. Budget is the fraction of criteria judged. Parentheses give differences from Vanilla GRPO with full judging. Only criterion score is shown for RubricBench.

	Medical		Science		RaR Science		RubricBench	Macro mean	
Condition	Criterion score	Normalized points score	Criterion score	Normalized points score	Criterion score	Normalized points score	Criterion score	Criterion score	Normalized points score
Vanilla GRPO	67.7	66.8	78.3	77.8	70.6	73.7	77.0	73.4	73.8
RRT, full judging (1.00)	70.1 (2.4)	69.8 (3.0)	79.5 (1.2)	79.2 (1.4)	70.7 (0.1)	73.8 (0.1)	77.9 (0.9)	74.6 (1.2)	75.2 (1.4)
Random selection (0.50)									
↪ Vanilla GRPO	65.9 (1.8)	64.0 (2.8)	74.7 (3.6)	74.8 (3.0)	67.5 (3.1)	70.3 (3.4)	76.3 (0.7)	71.1 (2.3)	71.4 (2.5)
↪ RRT	67.8 (0.1)	67.0 (0.2)	77.2 (1.1)	77.0 (0.8)	70.3 (0.3)	72.2 (1.5)	76.5 (0.5)	73.0 (0.5)	73.2 (0.7)
RRT, adaptive Fisher									
↪ 0.95	70.1 (2.4)	69.8 (3.0)	79.4 (1.1)	79.1 (1.3)	70.7 (0.1)	73.8 (0.1)	77.7 (0.7)	74.5 (1.1)	75.1 (1.2)
↪ 0.80	69.8 (2.1)	69.4 (2.6)	79.2 (0.9)	78.8 (1.0)	70.5 (0.1)	73.5 (0.2)	77.7 (0.7)	74.3 (0.9)	74.9 (1.0)
↪ 0.50	68.6 (0.9)	68.2 (1.4)	78.2 (0.1)	77.8 (0.0)	69.8 (0.8)	72.8 (0.9)	76.8 (0.2)	73.4 (0.1)	73.9 (0.1)

At criterion budget 0.50, adaptive Fisher selection keeps both RRT macro scores within 0.1 points of Vanilla GRPO with full judging (Table 4). Relative to each method's full judging score, random selection at this budget lowers macro criterion scores by 2.3 points for Vanilla GRPO and 1.6 for RRT. Adaptive Fisher selection at this budget reduces RRT's judge requests by 49.0% on Medical and Science relative to full judging. Separate API measurements yield parseable verdicts for all 11,712 completed requests (Tables 29 and 30 in Appendix G.1).

The cost analysis measures generation, judging, and total time per policy step. Judging is the longest measured stage in every tested condition. Relative to full judging with the same frozen RPN, adaptive Fisher selection at criterion budget 0.50 reduces median judging time by 49.1% to 49.6% and total step time by 23.5% to 28.7% on Medical and Science. Generation time stays within 1.0% of full judging (Table 32 in Appendix G.2). Added computation for RRT + online RPN is 0.123% of a Vanilla GRPO Medical step (Table 31). One RPN warm start per dataset serves all RPN variants. Embedding and fitting from cached verdicts cost 0.89 to 7.64 GPU hours across datasets and embedder sizes (Table 33 in Appendix G.3). The trained policy adds zero parameters or inference components at deployment (Table 34 in Appendix G.4).

5 Discussion and Conclusion

RRT gives GRPO access to differences between verdict patterns that point totals discard (RQ1, Proposition 1 in Appendix A.5). Fewer ties preserve these distinctions across policies and rollout group sizes (Appendix C.4). The rubric likelihood score maximizes local SNR for quality under the item response model (Theorem 2). This supports likelihood aggregation because differences in criterion parameters make the optimal relative weights vary with quality. MAP rewards also depend on posterior curvature from realized verdicts (Theorem 9 in Appendix A.5). Unanimous criteria, common in rollout groups (Appendix C.5), can change MAP reward gaps through shared likelihood factors. RRT can thus use the full rubric, even unanimous criteria, to shape GRPO advantages.

Policy score gains with an RPN support estimating criterion parameters from text rather than empirical pass rates alone (Table 2). Criterion parameters complement rollout evidence in pooled verdict prediction (Table 1). Less reliable empirical pass rates on uncertain criteria further support combining text and rollout evidence (Appendix D.4). Text is informative about criterion difficulty, as predicted and empirical difficulty correlate positively (Table 6 in Appendix C.2). Estimating discrimination improves mean reward stability over fixed discrimination under corruption with marginal calibration (Appendix D.2). Theorems 4 and 5 in Appendix A.1 motivate the Gaussian CDF: difficulty can

change reward ordering, while logistic ordering depends only on totals weighted by discrimination. At $a_j = 1$ and matched difficulties, the Gaussian CDF separates observed rollout pairs with equal pass counts (Table 7 in Appendix C.3). As the policy changes, online EM reuses GRPO verdicts to update the RPN. These updates improve correlation between predicted pass probabilities and empirical criterion pass rates at the next policy step (Table 10 in Appendix C.6). The hard E-step has lower criterion loss than the tested soft variants (Table 11 in Appendix C.7). These results support an RPN warm start followed by online calibration from current rollout verdicts.

With criterion parameters fitted per rubric by marginal calibration, RRT preserves reward distinctions under judge variation. Its mean stable nonzero ordering share exceeds that of rubric points under repeated judging and at every tested corruption level. It combines fewer ties with better order preservation under repeated judging and has a lower mean order flip rate at the strongest tested corruption level (Tables 14 and 15 in Appendix D.2). RRT with marginal calibration better preserves advantages from clean criterion scores when errors concentrate on particular criteria, while symmetric errors on all criteria and lenient errors tied to response length favor criterion score (Table 16 in Appendix D.3). With criterion parameters from the online RPN, RRT is less sensitive to criterion sampling than rubric points on all tested datasets. The largest gain occurs in the lowest quintile of rank correlation under rubric points (Tables 19 and 20 in Appendix D.5). Dependence diagnostics support the item response model by showing that fitted quality explains most pairwise dependence, with residual redundancy concentrated among criteria with similar text (Table 21 in Appendix D.6). Modeling this dependence could address redundancy and known bias in unidimensional IRT models (Yen, 1984). Multiple quality targets could also accommodate rubrics with explicit tradeoffs.

RRT + online RPN improves criterion satisfaction over Vanilla GRPO, POW3R, and DIVA in the primary Qwen3.5-4B comparison (RQ2, Table 2). It exceeds Vanilla GRPO on criteria the base policy often misses (Figure 7 in Appendix E.2). Gains on other benchmarks support generalization beyond the training rubrics (Table 23 in Appendix E.3). Across the tested policies, RRT produces shorter responses. On Qwen3.5-2B and Llama-3.1-8B-Instruct, its macro criterion scores remain close to Vanilla GRPO's (Appendices E.1 and E.4).

Fisher information gives a common basis for reward inference and criterion selection (RQ3, Theorem 1). RRT connects the information principle used in adaptive testing and model evaluation (Weiss, 1982; Truong et al., 2025) to criterion selection during policy training with rubrics. Difficulty locates each criterion's informative quality range, and discrimination controls its information peak (Theorem 7 in Appendix A.3). Fisher selection's savings over random selection at matched reward fidelity show that these parameters can reduce judging while approximating GRPO advantages (Table 3, Appendix E.5). Discrimination alone matches random selection on RaR Science and RubricBench, while static Fisher selection reduces judging on both (Table 3). This supports selection using difficulty as well as discrimination. Mean savings are greatest with static Fisher selection on rollouts from base policies and adaptive Fisher selection on those from trained policies (Tables 3 and 24). This shift supports adapting judge allocation as the policy changes.

At half the criterion budget, RRT's macro criterion score falls less than Vanilla GRPO's under random selection. Adaptive Fisher selection keeps RRT's macro scores close to Vanilla GRPO with full judging at about half the judge requests (Table 4, Appendix G.1). Larger criterion budgets recover more of RRT's gains from full judging (Table 4). Since judging is the longest measured stage of a policy step, selecting fewer criteria shortens steps with nearly unchanged generation time (Table 32 in Appendix G.2). Online RPN updates add little computation (Table 31). One RPN warm start per dataset serves all RPN variants (Appendix G.3). These costs remain in training: RRT adds no policy parameters or inference components at deployment (Appendix G.4).

Judging cost and time are a major challenge in reinforcement learning with rubrics. For rubrics whose criteria are monotone indicators of a shared target, RRT addresses this challenge with a common model in which difficulty and discrimination determine the evidence in each verdict and the expected information of each unjudged criterion. RRT thus distinguishes verdict patterns with the same point total and selects informative criteria. This reduces judge usage and shortens policy steps while retaining the gains from Vanilla GRPO. Beyond these savings, RRT improves criterion satisfaction, including on difficult criteria, and stabilizes reward orderings under repeated judging with marginal calibration. The deployed policy has no added parameters or inference components.

Reproducibility Statement

An anonymous reference implementation and instructions for preparing datasets, judging rollouts, fitting the RPN, computing RRT rewards, and evaluating policies are available at Code. The experimental configuration, training parameter choices, computing resources, and checkpoint selection criteria are detailed in Appendix F.3. The data sources, processing steps, splits, and roles in the experiments are documented in Appendix F.1. Policy and judge prompts are provided in Appendix F.2. Appendix A states the model assumptions and gives complete mathematical derivations and proofs.

AI Use Statement

Generative AI tools assisted with the theoretical model, mathematical claims and proofs, hypotheses and experiment design, implementing the method, data preparation, and interpreting results. They were not used to generate datasets. Translation and qualitative analysis are not applicable. They also assisted with code, text drafting and editing, and literature search. The authors made all decisions, reviewed and modified all assisted work, checked every proof step, tested all code, verified all numbers and references, and take responsibility for the final content, including text, claims, or artifacts produced with the aid of generative AI.

Ethics Statement

The experiments use public research datasets and responses generated by language models. No new data were collected from human participants. Rubrics and automated judge verdicts can reflect biases in their source data and evaluation criteria. Appendix F.1 documents the data sources and their roles, and Appendix D.1 reports agreement across repeated judging.

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.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *arXiv preprint arXiv:2511.07025*, 2025.

Stephen Bach, Bert Huang, Jordan Boyd-Graber, and Lise Getoor. Paired-dual learning for fast training of latent variable hinge-loss mrfs. In *International Conference on Machine Learning*, pp. 381–390. PMLR, 2015.

Mark Bagnoli and Ted Bergstrom. Log-concave probability and its applications. *Economic theory*, 26(2):445–469, 2005.

Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*, 1968.

Yunxiao Chen, Xiaoou Li, Jingchen Liu, and Zhiliang Ying. Item response theory—a statistical framework for educational and psychological measurement. *Statistical Science*, 40(2):167–194, 2025.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Jonathan Cook, Tim Rocktäschel, Jakob Nicolaus Foerster, Dennis Aumiller, and Alex Wang. Check your work: Structured checklist feedback for improving large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16649–16688, 2026.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations*, volume 2026, pp. 127924–127945, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

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 *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 13806–13834, 2024.

John P Lalor, Hao Wu, and Hong Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4249–4259, 2019.

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.

David Magis. A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, 37(4):304–315, 2013.

Robert J Mislevy. Bayes modal estimation in item response models. *Psychometrika*, 51(2):177–195, 1986.

Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23, 1950.

OpenAI. GPT-5.5 system card. Technical report, OpenAI, 2026. URL `https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf`.

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

Georg Rasch. An item analysis which takes individual differences into account. *British journal of mathematical and statistical psychology*, 19(1):49–57, 1966.

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

Sang T. Truong, Yuheng Tu, Percy Liang, Bo Li, and Sanmi Koyejo. Reliable and efficient amortized model-based evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60238–60265. PMLR, 13–19 Jul 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.

Masaki Uto. A multidimensional generalized many-facet rasch model for rubric-based performance assessment. *Behaviormetrika*, 48(2):425–457, 2021.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(86):2579–2605, 2008.

Vijay Viswanathan, Yanchao Sun, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=RPRqKhjrr6.

David J Weiss. Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement*, 6(4):473–492, 1982.

Wendy M Yen. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2):125–145, 1984.

Li S Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. ResearchQA: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *Transactions of the Association for Computational Linguistics*, 14:1365–1389, 2026.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

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.

Junyi Zhou, Qiyuan Zhang, Yufei Wang, Fuyuan Lyu, Yidong Ming, Can Xu, Qingfeng Sun, Kai Zheng, Peng Kang, Xue Liu, and Chen Ma. RubricBench: Aligning model-generated rubrics with human standards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31179–31200. Association for Computational Linguistics, 2026.

A Theoretical Results and Proofs

A.1 Response Function Assumptions and Properties

The response function $F : \mathbb{R} \to (0, 1)$ of Section 3.1 maps the signed margin $a_j(z_i - b_j)$ to a probability. It must satisfy two conditions.

- $F(t) \to 0$ as $t \to -\infty$, $F(t) \to 1$ as $t \to +\infty$, and F is strictly increasing. Increasing a rollout's quality for the rubric never lowers its chance of satisfying a criterion.
- $F(0) = 0.5$, so $z_i = b_j \iff P_{ij} = 0.5$. Thus b_j is the criterion difficulty.

The logistic CDF σ and the Gaussian CDF Φ both satisfy them. RRT adopts Φ.

Fix a rollout i and use the signed margin u_{ij}, with $a_j > 0$ and prior variance $\sigma_z^2 > 0$, and assume that F is differentiable. Define the response weight function

$$s_F(u) := \frac{F'(u)}{F(u)(1 - F(u))} = \frac{d}{du} \operatorname{logit} F(u). \tag{6}$$

Theorem 3 (Log posterior gradient for general F)**.** *The log posterior objective ℓ_i for rollout i in Eq. 2 has gradient*

$$\ell_i'(z_i) = \sum_j a_j \, s_F(u_{ij}) \big(G_{ij} - F(u_{ij})\big) - \frac{z_i}{\sigma_z^2}. \tag{7}$$

Proof. Differentiating the log likelihood ℓ_{ij} of criterion j with respect to its margin u_{ij} gives

$$\frac{\partial \ell_{ij}}{\partial u_{ij}} = F'(u_{ij}) \left[\frac{G_{ij}}{F(u_{ij})} - \frac{1 - G_{ij}}{1 - F(u_{ij})} \right] = s_F(u_{ij}) \big(G_{ij} - F(u_{ij})\big).$$

Multiplying by $\partial u_{ij}/\partial z_i = a_j$, summing over j, and adding the prior derivative $-z_i/\sigma_z^2$ gives Eq. 7. □

Theorem 4 (Difficulty dependence of the verdict coefficient)**.** *Let F be twice continuously differentiable. The coefficient of G_{ij} in Eq. 7 is $a_j s_F(u_{ij})$, and*

$$\frac{\partial}{\partial b_j}\big[a_j s_F(u_{ij})\big] = -a_j^2\, s_F'(u_{ij}). \tag{8}$$

This coefficient is independent of b_j for all margins if and only if s_F is constant. For a centered F, this holds if and only if $F(u) = \sigma(\gamma_F u)$ for some $\gamma_F > 0$. In particular, for $F = \sigma$, the reward order depends only on $\sum_j a_j G_{ij}$ and not on b_j.

Proof. Theorem 3 gives the coefficient $a_j s_F(u_{ij})$, and $\partial u_{ij}/\partial b_j = -a_j$ yields Eq. 8, which vanishes for every margin if and only if $s_F' \equiv 0$. Since $s_F = \frac{d}{du}\operatorname{logit} F$, a constant $s_F \equiv \gamma_F$ integrates to $\operatorname{logit} F(u) = \gamma_F u + C$, and $F(0) = 1/2$ forces $C = 0$, so $F = \sigma(\gamma_F u)$. Strict increase requires $\gamma_F > 0$. The converse follows by differentiation.

For $F = \sigma$, $\sigma' = \sigma(1-\sigma)$ gives $s_\sigma \equiv 1$, so $\ell_i'(z_i) = \sum_j a_j G_{ij} - g(z_i)$ with

$$g(z_i) := \sum_j a_j \sigma(u_{ij}) + \frac{z_i}{\sigma_z^2}.$$

The map g does not involve the verdicts. Each $a_j\sigma(u_{ij})$ is nondecreasing in z_i and the term z_i/σ_z^2 is strictly increasing with range $\mathbb{R}$, so g is a strictly increasing bijection. The stationary condition $g(\hat{z}_i) = \sum_j a_j G_{ij}$ then has a unique solution that increases with its right side, the same map for all rollouts. Hence the reward order is the order of $\sum_j a_j G_{ij}$, which contains no b_j. □

For a scaled $F(u) = \sigma(\gamma_F u)$ the coefficient scale is $s_F \equiv \gamma_F$, absorbed into a_j. The analysis uses $\gamma_F = 1$. Theorem 4 concerns the order of the MAP rewards. Under every F, including σ, difficulties can change reward gaps and normalized GRPO advantages when a group has at least three distinct weighted verdict totals.

Theorem 5 (Effect of $F = \Phi$)**.** *Suppose $F(u) = \Phi(u)$, with standard Gaussian density ϕ. Then*

$$s_\Phi(u) = \frac{\phi(u)}{\Phi(u)(1-\Phi(u))} = \lambda(u) + \lambda(-u), \qquad \lambda(u) := \frac{\phi(u)}{\Phi(u)}. \tag{9}$$

The response weight is even and nonconstant, satisfies $s_\Phi(0) = 4/\sqrt{2\pi}$, and obeys $s_\Phi(u) \sim |u|$ as $|u| \to \infty$. Thus b_j changes the verdict coefficient through u_{ij}. Moreover, a change in one difficulty can reverse the RRT reward order of two fixed verdict vectors.

Proof. Substituting $F = \Phi$, so $F' = \phi$, into Eq. 6, and using $1 - \Phi(u) = \Phi(-u)$, gives Eq. 9. The corresponding signed criterion likelihood score in Eq. 7 is

$$a_j \begin{cases} \lambda(u_{ij}), & G_{ij} = 1, \\ -\lambda(-u_{ij}), & G_{ij} = 0. \end{cases}$$

Because λ is decreasing, raising b_j decreases u_{ij} and increases the positive contribution of a satisfied criterion while decreasing the magnitude of the negative contribution of an unsatisfied criterion. Passing a harder criterion therefore gives a larger positive likelihood score. Failing a harder criterion gives a negative likelihood score with smaller magnitude.

At $u = 0$, $s_\Phi(0) = \phi(0)/\Phi(0)^2 = 4/\sqrt{2\pi}$. Mills' ratio gives $1 - \Phi(u) \sim \phi(u)/u$ as $u \to +\infty$, so

$$s_\Phi(u) = \frac{\phi(u)}{\Phi(u)(1-\Phi(u))} \sim u.$$

Evenness gives the corresponding $|u|$ asymptotic in the left tail.

For the rank claim, take three criteria with $\sigma_z^2 = 1$, $a_j = 1$, $b_2 = b_3 = 0$, and let $b_1 = d$. Compare the fixed verdict vectors

$$G^A = (0,1,1), \qquad G^B = (1,0,0).$$

At $d = 0$, the log posterior derivatives are $\ell'_A(0) = \lambda(0) > 0$ and $\ell'_B(0) = -\lambda(0) < 0$. The derivatives are strictly decreasing, so $\hat{z}_A > 0 > \hat{z}_B$. As $d \to \infty$, the derivative for A is

$$-\lambda(d - z) + 2\lambda(z) - z \leq 2\lambda(z) - z,$$

so its root stays bounded above by the finite root of $2\lambda(z) - z = 0$. The derivative for B is

$$\lambda(z - d) - 2\lambda(-z) - z.$$

If its root stayed bounded, then $\lambda(z-d) \sim d-z$ would diverge while the other terms stayed bounded. Also, $\ell'_B(0) = \lambda(-d) - 2\lambda(0) > 0$ for sufficiently large d, so the root is positive. Hence $\hat{z}_B \to \infty$. For sufficiently large d, $\hat{z}_B > \hat{z}_A$, so changing only b_1 reverses the order. □

The Gaussian CDF tail behavior makes the magnitude of an unexpected criterion likelihood score unbounded in the theoretical model. For an unexpected hard pass, $a_j\lambda(u_{ij}) \sim a_j|u_{ij}|$ as $u_{ij} \to -\infty$. An unexpected easy failure has the same asymptotic magnitude in the opposite direction. By contrast, under $F = \sigma$ each criterion likelihood score has magnitude at most a_j. For a mastered criterion with $u_{ij} \gg 0$, the expected pass likelihood score tends to zero while the magnitude of an unexpected failure score grows with u_{ij}.

A.2 MAP Reward Derivation and Properties

The fixed quality prior is

$$p(z_i) = \frac{1}{\sqrt{2\pi}\,\sigma_z} \exp\Big(-\frac{z_i^2}{2\sigma_z^2}\Big). \tag{10}$$

Bayes' rule turns the likelihood of Eq. 1 and the prior of Eq. 10 into the posterior over quality,

$$p_\psi(z_i \mid G_i, q, c) = \frac{p_\psi(G_i \mid z_i, q, c)\, p(z_i)}{p_\psi(G_i \mid q, c)}.$$

The denominator $p_\psi(G_i \mid q, c)$ does not depend on z_i, so the most probable quality maximizes the numerator,

$$\hat{z}_i = \arg\max_{z_i}\ p_\psi(z_i \mid G_i, q, c) = \arg\max_{z_i}\ p_\psi(G_i \mid z_i, q, c)\, p(z_i). \tag{11}$$

Taking a logarithm turns the product into a sum without moving the maximizer. The log likelihood of rollout i's verdict on criterion j at quality z, the logarithm of that criterion's Bernoulli factor in Eq. 1, is

$$\ell_{ij}(z) = G_{ij} \log F\big(a_j(z - b_j)\big) + (1 - G_{ij}) \log\big(1 - F(a_j(z - b_j))\big),$$

so the likelihood contributes $\sum_j \ell_{ij}(z_i)$ and the prior contributes $\log p(z_i) = -z_i^2/(2\sigma_z^2) - \frac{1}{2}\log(2\pi\sigma_z^2)$. Dropping the constant leaves the log posterior objective of Eq. 2.

Theorem 6 (Strict concavity of the log posterior objective)**.** *Fix criteria with $a_j > 0$, difficulties $b_j \in \mathbb{R}$, and prior variance $\sigma_z^2 > 0$. Suppose the adopted response function is $F = \Phi$. Then the log posterior objective ℓ_i of Eq. 2 is smooth, strictly concave, and tends to $-\infty$ as $|z_i| \to \infty$. Its derivative is strictly decreasing and satisfies*

$$\ell'_i(z_i) \to +\infty \text{ as } z_i \to -\infty, \qquad \ell'_i(z_i) \to -\infty \text{ as } z_i \to +\infty.$$

Thus the reward $\hat{z}_i = \arg\max_z \ell_i(z)$ exists, is the unique root of ℓ'_i, and is differentiable in the criterion parameters (a_j, b_j).

Proof. Differentiating the log posterior objective ℓ_i of Eq. 2 twice gives

$$\ell''_i(z_i) = \sum_j a_j^2 \big[G_{ij}\,\lambda'(u_{ij}) + (1 - G_{ij})\,\lambda'(-u_{ij})\big] - \frac{1}{\sigma_z^2} \leq -\frac{1}{\sigma_z^2} < 0,$$

because $\lambda = (\log\Phi)'$ and $\log\Phi$ is concave (Bagnoli & Bergstrom, 2005), so $\lambda'(x) \leq 0$ and every bracketed term is nonpositive. The Gaussian prior makes the displayed inequality strict. Hence ℓ_i is smooth and its derivative is strictly decreasing. Each log likelihood term is nonpositive, so $\ell_i(z_i) \leq -z_i^2/(2\sigma_z^2)$ and the log posterior objective tends to $-\infty$ in both tails. Eq. 7 also gives the stated derivative limits. The derivative is continuous and strictly decreasing, so it has exactly one root. This root is the unique maximizer. Its derivative with respect to the criterion parameters exists by implicit differentiation because $\ell''_i(\hat{z}_i) < 0$. □

Corollary 1 (Dominance of the MAP reward)**.** *Fix one prompt group and its criterion parameters. If two verdict vectors satisfy $G_j^A \geq G_j^B$ for every criterion, then $\hat{z}_A \geq \hat{z}_B$. The inequality is strict if at least one verdict differs. Thus a difficulty change can reverse only incomparable verdict vectors.*

Proof. At every z, define $u_j = a_j(z - b_j)$. Theorem 3 gives

$$\ell'_A(z) - \ell'_B(z) = \sum_j a_j s_\Phi(u_j)(G_j^A - G_j^B).$$

Every summand is nonnegative, and one is positive when a verdict differs. At the root $\hat{z}_B$, this gives $\ell'_A(\hat{z}_B) \geq 0$. Since ℓ'_A is strictly decreasing, its root lies weakly to the right, and strictly to the right when a verdict differs. □

A.3 Criterion Information and Realized Evidence

Proof of Theorem 1. Theorem 3 gives

$$\mathcal{S}_{ij}(z_i) = a_j s_\Phi(u_{ij})(G_{ij} - P_{ij}).$$

Since $\mathbb{E}[G_{ij} \mid z_i] = P_{ij}$, its conditional mean is zero. The conditional variance of a Bernoulli verdict gives

$$\mathbb{E}[(G_{ij} - P_{ij})^2 \mid z_i] = P_{ij}(1 - P_{ij}).$$

Substituting the criterion likelihood score into the definition of I_j gives

$$I_j(z_i) = a_j^2 s_\Phi(u_{ij})^2 P_{ij}(1 - P_{ij}) = a_j^2 \frac{\phi(u_{ij})^2}{P_{ij}(1 - P_{ij})}.$$

The last equality in Eq. 4 is the information identity for this smooth Bernoulli likelihood. It also follows by directly differentiating the criterion likelihood score and taking its conditional expectation. □

Theorem 7 (Information frontier of a criterion under $F = \Phi$)**.** *For a fixed criterion j, the information*

$$I_j(z_i) = a_j^2 f(u_{ij}), \qquad f(u) = \frac{\phi(u)^2}{\Phi(u)\Phi(-u)}.$$

is symmetric around $z_i = b_j$. It strictly increases for $z_i < b_j$ and strictly decreases for $z_i > b_j$. Its unique maximum is

$$I_j(b_j) = \frac{2}{\pi} a_j^2.$$

It also satisfies

$$I_j(z_i) \longrightarrow 0 \qquad \text{as } |z_i - b_j| \longrightarrow \infty.$$

For fixed t,

$$\frac{I_j(b_j + t/a_j)}{a_j^2} = \frac{\phi(t)^2}{\Phi(t)(1 - \Phi(t))}.$$

Proof. The function f is even because ϕ is even and $\Phi(-u) = 1 - \Phi(u)$. Let X be a standard Gaussian random variable. Using $\lambda(u) = \phi(u)/\Phi(u)$ gives

$$\frac{d^2}{du^2} \log f(u) = -2 - \lambda'(u) - \lambda'(-u) = -\operatorname{Var}(X \mid X \leq u) - \operatorname{Var}(X \mid X > u) < 0.$$

The second equality uses $1 + \lambda'(u) = \operatorname{Var}(X \mid X \leq u)$ and Gaussian symmetry. Thus f is strictly log-concave. An even, strictly log-concave function has its unique maximum at zero and is strictly monotone on either side. At zero,

$$f(0) = \frac{\phi(0)^2}{(1/2)(1/2)} = \frac{2}{\pi}.$$

Mills' ratio gives $f(u) \sim |u|\phi(u)$ as $|u| \to \infty$, which tends to zero. The final display follows by substituting $z_i = b_j + t/a_j$. □

Theorem 7 separates the criterion parameters. The difficulty b_j places the information peak on the quality scale. The discrimination a_j raises the peak quadratically and makes its width on the z_i scale proportional to $1/a_j$.

For a realized verdict, the magnitude of one criterion likelihood score is

$$\left|\frac{\partial \ell_{ij}}{\partial z_i}\right| = \begin{cases} a_j \dfrac{\phi(u_{ij})}{P_{ij}}, & G_{ij} = 1, \\ a_j \dfrac{\phi(u_{ij})}{1 - P_{ij}}, & G_{ij} = 0. \end{cases}$$

Consequently,

$$\frac{|\partial \ell_{ij}/\partial z_i|_{G_{ij}=1}}{|\partial \ell_{ij}/\partial z_i|_{G_{ij}=0}} = \frac{1 - P_{ij}}{P_{ij}}.$$

These criterion likelihood scores measure evidence from the realized verdict, whereas Fisher information measures expected usefulness before the verdict is observed. A pass has nine times the likelihood score magnitude of a failure when $P_{ij} = 0.1$. A failure has nine times the magnitude of a pass when $P_{ij} = 0.9$. Expected information and realized evidence can therefore rank a criterion differently.

Figure 3 uses pass probability as its horizontal coordinate. Let $p = P_{ij}$ and $u = \Phi^{-1}(p)$. The three plotted curves are

$$\frac{I_j(z_i)}{a_j^2} = \frac{\phi(u)^2}{p(1-p)}, \qquad \frac{1}{a_j}\frac{\partial \ell_{ij}}{\partial z_i}\bigg|_{G_{ij}=1} = \frac{\phi(u)}{p}, \qquad \frac{1}{a_j}\frac{\partial \ell_{ij}}{\partial z_i}\bigg|_{G_{ij}=0} = -\frac{\phi(u)}{1-p}.$$

The information curve reaches $2/\pi$ at $p = 1/2$. The pass and failure curves grow in magnitude when the observed verdict has low fitted probability.

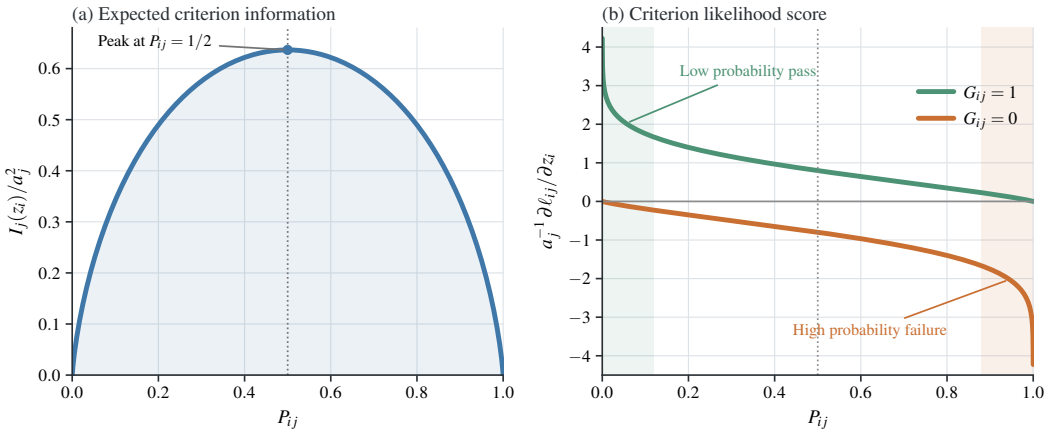

Figure 3: Criterion information and likelihood scores under the Gaussian CDF $F = \Phi$ as functions of pass probability P_{ij}. The left panel shows normalized information $I_j(z_i)/a_j^2$. The right shows $a_j^{-1}\partial \ell_{ij}/\partial z_i$ for a pass and failure. The dotted line marks $P_{ij} = 1/2$, and shading marks verdicts with low probability.

A.4 Local Efficiency of Reward Aggregation

Fix a differentiable F, a quality level z_i, fixed criterion parameters (a_j, b_j), and conditionally independent verdicts under the item response model of Section 3.1. Define

$$I_{F,j}(z_i) := a_j^2 \frac{F'(u_{ij})^2}{P_{ij}(1 - P_{ij})} = P_{ij}(1 - P_{ij})[a_j s_F(u_{ij})]^2,$$

and suppose $\sum_j I_{F,j}(z_i) > 0$. Consider the linear statistic

$$T_v = \sum_j v_j (G_{ij} - P_{ij}),$$

where P_{ij} is evaluated at z_i. Suppose T_v has unit local response to a quality change,

$$\frac{d}{dh}\mathbb{E}_{z_i+h}[T_v]\Big|_{h=0} = \sum_j v_j a_j F'(u_{ij}) = 1. \tag{12}$$

The proof of Theorem 2 extends its bound to general F. Applying this bound to T_v gives the statistic with minimum variance

$$T^* = \frac{\sum_j a_j s_F(u_{ij})(G_{ij} - P_{ij})}{\sum_j I_{F,j}(z_i)}. \tag{13}$$

Its conditional variance is

$$\mathrm{Var}(T^* \mid z_i) = \left[\sum_j I_{F,j}(z_i)\right]^{-1}.$$

Differentiability gives $\mathbb{E}_{z_i+h}[T^*] = h + o(h)$. This unbiasedness is local.

For $F = \Phi$, $I_{F,j} = I_j$ and the numerator of Eq. 13 is the rubric likelihood score $\mathcal{S}_i(z_i)$, the likelihood contribution in Eq. 7. The efficient coefficient is $a_j s_\Phi(u_{ij})$, not a_j alone. The latter coefficient belongs to $F = \sigma$ because $s_\sigma \equiv 1$.

Under conditional independence, the expected negative curvature of the log posterior is

$$-\mathbb{E}[\ell_i''(z_i) \mid z_i] = \frac{1}{\sigma_z^2} + \sum_j I_j(z_i). \tag{14}$$

The Gaussian prior contributes $1/\sigma_z^2$. Theorem 1 gives the expected negative curvature $I_j(z_i)$ of each criterion, and the terms add under conditional independence. For duplicate or dependent criteria, information is not additive. After substituting $\hat{z}_i$, the inverse square root of the expected precision and the observed curvature $[-\ell_i''(\hat{z}_i)]^{-1/2}$ give local uncertainty approximations.

A.5 Alignment with the Reward Based on Rubric Points

Under the item response model, the expected reward based on rubric points is strictly increasing in scalar quality and does not decrease under first-order stochastic dominance. This section proves that result and compares this reward with the rubric likelihood score and MAP reward as local training signals. Fix a prompt q, its rubric c, and the criterion parameters $(a_j, b_j) = \psi(q, c_j)$. For the verdict vector G_i, the rubric likelihood score expands to

$$\mathcal{S}_i(z_i) = \sum_j \mathcal{S}_{ij}(z_i) = \sum_j a_j s_F(u_{ij})\big(G_{ij} - P_{ij}\big)$$

for a general F. Define the rubric information for general F as

$$I_F(z_i) := \sum_j I_{F,j}(z_i),$$

which equals the rubric information $I(z_i)$ under the adopted $F = \Phi$. The results below assume the model of Eq. 1.

Theorem 8 (Monotone expected reward based on rubric points)**.** *Let F be differentiable and strictly increasing with $F' > 0$. Then*

$$m(z_i) := \mathbb{E}\big[R_i^{\text{base}} \,\big|\, z_i\big] = \frac{\sum_j w_j P_{ij}}{\sum_\ell w_\ell}, \qquad m'(z_i) = \frac{\sum_j w_j a_j F'(u_{ij})}{\sum_\ell w_\ell} > 0.$$

Let Z_θ be the quality of a rollout drawn from $\pi_\theta(\cdot \mid q)$, so that $\mathbb{E}[R_i^{\text{base}} \mid q] = \mathbb{E}[m(Z_\theta)]$. If the quality after a policy update first-order stochastically dominates the quality before it, the expected reward based on rubric points does not decrease.

Proof. Linearity of expectation and $\mathbb{E}[G_{ij} \mid z_i] = P_{ij}$ give m. Differentiating $P_{ij} = F(u_{ij})$ gives $\partial P_{ij}/\partial z_i = a_j F'(u_{ij})$, and every term is positive because $w_j > 0$, $a_j > 0$, and $F' > 0$. Since m is increasing, $\mathbb{E}[m(Z_{\text{after}})] \geq \mathbb{E}[m(Z_{\text{before}})]$ is the defining property of first-order stochastic dominance. □

Theorem 8 establishes alignment under first-order stochastic dominance. Outside this condition, the two training signals can disagree. A location shift of the quality distribution is one sufficient special case.

Proof of Theorem 2. The argument uses only differentiability of F and conditional independence. It therefore gives the bound with $I_F(z_i)$ in place of $I(z_i)$ for a general response function, whenever $0 < P_{ij} < 1$ for every j and $I_F(z_i) > 0$. The verdict vector takes finitely many values, so the sum defining $\mathbb{E}[T \mid z_i]$ may be differentiated term by term. Writing $p(G_i \mid z_i)$ for the likelihood of Eq. 1,

$$\frac{d}{dz_i}\mathbb{E}[T \mid z_i] = \sum_{G_i} T(G_i)\, p(G_i \mid z_i)\, \mathcal{S}_i(z_i) = \mathbb{E}[T\, \mathcal{S}_i(z_i) \mid z_i] = \operatorname{Cov}\big(T, \mathcal{S}_i(z_i) \mid z_i\big),$$

where the last step uses $\mathbb{E}[G_{ij} \mid z_i] = P_{ij}$, which makes each criterion likelihood score have conditional mean zero. Conditional independence gives

$$\operatorname{Var}\big(\mathcal{S}_i(z_i) \mid z_i\big) = \sum_j \mathbb{E}\big[\mathcal{S}_{ij}(z_i)^2 \mid z_i\big] = I_F(z_i).$$

Cauchy-Schwarz then gives $\operatorname{Cov}(T, \mathcal{S}_i)^2 \le \operatorname{Var}(T)\, I_F(z_i)$, which is the stated bound, with equality if and only if the two centered variables are almost surely proportional under $p_\psi(G_i \mid z_i, q, c)$. The constant cannot be zero, since that would make T almost surely constant and contradict $\operatorname{Var}(T \mid z_i) > 0$. Replacing T by $\alpha + \beta T$ with $\beta \neq 0$ multiplies numerator and denominator of Eq. 5 by β^2, so the ratio is unchanged by any nonzero affine map of T. □

Proof of the clause for rubric points in Theorem 2. Under the same conditions,

$$\mathrm{SNR}_{\mathrm{base}}(z_i) = \frac{\big(\sum_j w_j a_j F'(u_{ij})\big)^2}{\sum_j w_j^2 P_{ij}(1 - P_{ij})} \le I_F(z_i),$$

with equality if and only if $w_j = \gamma_w\, a_j s_F(u_{ij})$ for every j and some $\gamma_w > 0$. For $F = \sigma$ this condition reads $w_j \propto a_j$. For $F = \Phi$ it reads $w_j \propto a_j s_\Phi(u_{ij})$, whose right side depends on b_j and on z_i. The displayed ratio follows from $\partial P_{ij}/\partial z_i = a_j F'(u_{ij})$, from conditional independence, and from the Bernoulli variance $P_{ij}(1 - P_{ij})$. The factor $\sum_\ell w_\ell$ cancels. Equality in the information bound requires

$$\sum_j \left(\frac{w_j}{\sum_\ell w_\ell} - \gamma'_w a_j s_F(u_{ij}) \right) (G_{ij} - P_{ij}) = 0$$

almost surely. Taking the conditional variance of the left side gives a sum of nonnegative terms with positive factors $P_{ij}(1 - P_{ij})$, so every coefficient vanishes. Since $w_j > 0$, equality requires $s_F(u_{ij}) > 0$ for every j and $\gamma'_w > 0$. For $F = \sigma$, $s_\sigma \equiv 1$ by Theorem 4. For $F = \Phi$, take two criteria and suppose equality holds at every z_i, so that

$$\chi(z_i) := \frac{a_1 s_\Phi\big(a_1(z_i - b_1)\big)}{a_2 s_\Phi\big(a_2(z_i - b_2)\big)} = \frac{w_1}{w_2}$$

is a constant χ_0. Mills' ratio sharpens the asymptotic of Theorem 5 to $s_\Phi(u) = u + O(u^{-1})$ as $u \to \infty$, so $a_j s_\Phi(a_j(z_i - b_j)) = a_j^2(z_i - b_j) + O(z_i^{-1})$ and

$$\big(a_1^2 - \chi_0 a_2^2\big) z_i - \big(a_1^2 b_1 - \chi_0 a_2^2 b_2\big) \longrightarrow 0.$$

An affine function with this limit vanishes identically, so $a_1^2 = \chi_0 a_2^2$ and then $b_1 = b_2$. Evaluating χ at $z_i = b_1$ gives $\chi = a_1/a_2$ through $s_\Phi(0) = 4/\sqrt{2\pi}$, so $a_1/a_2 = \chi_0 = a_1^2/a_2^2$ and $a_1 = a_2$. Therefore, once two criteria have different criterion parameters, no fixed vector of rubric points achieves the bound at every quality level. □

The points w_j specify how much each criterion should count. The coefficients of the verdicts in Eq. 13 are proportional to $a_j s_F(u_{ij})$. Under $F = \Phi$, they vary with quality while the points w_j stay fixed. Unless the coefficients are proportional, the points total has lower local SNR than the likelihood score. The reward based on rubric points also cannot distinguish verdict vectors with the same points total.

Proposition 1 (Points ties and reward separation)**.** *Fix a prompt group and its criterion parameters. If two rollouts satisfy $\sum_j w_j G_{1j} = \sum_j w_j G_{2j}$, then $R_1^{\text{base}} = R_2^{\text{base}}$. The RRT reward $\hat{z}_i$ can still separate two such rollouts. Under $F = \sigma$ it separates them if and only if $\sum_j a_j G_{1j} \neq \sum_j a_j G_{2j}$. Under $F = \Phi$ it can separate them even when those totals weighted by discrimination agree.*

Proof. Equality of the rewards based on rubric points follows from $R_i^{\text{base}} = \sum_j w_j G_{ij} / \sum_\ell w_\ell$. The claim for $F = \sigma$ is Theorem 4, which makes $\hat{z}_i$ a strictly increasing function of $\sum_j a_j G_{ij}$ shared by the group. For $F = \Phi$ take $K = 2$, $w_1 = w_2 = 1$, $a_1 = a_2 = 1$, $\sigma_z^2 = 1$, $b_2 = 0$, $b_1 = b > 0$, and the verdict vectors $G_1 = (1, 0)$ and $G_2 = (0, 1)$, which agree on both totals. Eq. 7 gives

$$\ell_1'(z) = \lambda(z - b) - \lambda(-z) - z, \qquad \ell_2'(z) = \lambda(z) - \lambda(b - z) - z.$$

Since $\lambda > 0$, it follows that $\ell_2'(z) \leq \lambda(z) - z$, whose unique root $z^\star$ is finite, so $\hat{z}_2 \leq z^\star$. Mills' ratio gives $\lambda(z^\star - b) \to \infty$ as $b \to \infty$ while $\lambda(-z^\star) + z^\star$ is fixed, so $\ell_1'(z^\star) > 0$ for large b. Because ℓ_1' is strictly decreasing by Theorem 6, $\hat{z}_1 > z^\star \geq \hat{z}_2$. □

Theorem 9 (Exact MAP expansion and group invariance)**.** *Suppose $F = \Phi$. For a group statistic vector $\mathbf{T} = (T_1, \ldots, T_N)$, write $\bar{T} = N^{-1} \sum_k T_k$ and define*

$$\mathcal{A}_i[\mathbf{T}] := \frac{T_i - \bar{T}}{\operatorname{std}_k T_k + \varepsilon}$$

for the group advantage operator of Eq. 17. Let $\hat{\mathbf{z}} = (\hat{z}_1, \ldots, \hat{z}_N)$. If $\varepsilon = 0$ and $\operatorname{std}_k T_k > 0$, then $\mathcal{A}_i[\alpha \mathbf{1} + \beta \mathbf{T}] = \mathcal{A}_i[\mathbf{T}]$ for every α and every $\beta > 0$. Fix a quality level z_0. For every rollout there is a point ξ_i between z_0 and $\hat{z}_i$ with

$$\hat{z}_i - z_0 = \frac{\mathcal{S}_i(z_0) - z_0/\sigma_z^2}{-\ell_i''(\xi_i)}.$$

Proof. For $\beta > 0$ the group mean of $\alpha \mathbf{1} + \beta \mathbf{T}$ is $\alpha + \beta \bar{T}$ and its standard deviation is $\beta \operatorname{std}_k T_k$, so the two factors β cancel at $\varepsilon = 0$ when the denominator is positive. Theorem 6 makes ℓ_i smooth with $\ell_i'(\hat{z}_i) = 0$, so the mean value theorem gives $0 = \ell_i'(z_0) + \ell_i''(\xi_i)(\hat{z}_i - z_0)$. Eq. 7 gives $\ell_i'(z_0) = \mathcal{S}_i(z_0) - z_0/\sigma_z^2$. □

A Fisher scoring surrogate replaces the realized curvature in Theorem 9 by the expected local precision in Eq. 14. Define

$$\widetilde{z}_i(z_0) := z_0 + \frac{\mathcal{S}_i(z_0) - z_0/\sigma_z^2}{\sigma_z^{-2} + I(z_0)}. \tag{15}$$

Let $\widetilde{\mathbf{z}}(z_0) = (\widetilde{z}_1(z_0), \ldots, \widetilde{z}_N(z_0))$ and $\boldsymbol{\mathcal{S}}(z_0) = (\mathcal{S}_1(z_0), \ldots, \mathcal{S}_N(z_0))$. The denominator in Eq. 15 is positive and shared across the group. Therefore, if $\varepsilon = 0$ and the score vector has positive standard deviation, affine invariance gives the exact identity

$$\mathcal{A}_i[\widetilde{\mathbf{z}}(z_0)] = \mathcal{A}_i[\boldsymbol{\mathcal{S}}(z_0)].$$

Replacing each realized curvature by the shared expected precision can change the exact MAP order.

The MAP reward can rank incomparable verdict vectors differently from the reward based on rubric points.

B Method and Optimization Details

B.1 Online EM Objective and Algorithms

Algorithm 1 gives the full policy training and calibration loop.

Algorithm 1 RRT policy training with online GRPO and EM calibration of the RPN.

Initialize policy π_θ from the checkpoint of the base policy and RPN ψ neutrally or with a warm start
for each policy step **do**
  sample a batch of prompts and draw N rollouts per prompt from π_θ
  judge each rollout against its rubric $\rightarrow$ verdicts G
  E-step: in evaluation mode, infer $\hat{z}_i^{\text{reward}} \leftarrow \arg\max_z \ell_i(z)$ by bisection using Eq. 7 (Algorithm 2), $R_i \leftarrow \hat{z}_i^{\text{reward}}$
  compute A_i within each prompt group (Eq. 17)
  GRPO: update π_θ on $J(\theta)$ (Eq. 18)
  Stochastic partial M-step: one accumulated gradient step on ψ with detached mode targets $\hat{z}_i^{\text{M}}$ using Eq. 3 (Algorithm 3)
end for

Let $\mathbf{z} = (z_1, \ldots, z_N)$ collect the latent qualities of the rollouts. The ideal EM objective is the regularized log posterior for the complete data

$$\mathcal{C}_{\text{EM}}(\psi, \mathbf{z}) = \sum_{i=1}^{N} \Big[\sum_{j=1}^{K} \ell_{ij}(z_i; \psi) - \frac{z_i^2}{2\sigma_z^2} \Big] - N\lambda_a \sum_{j=1}^{K} (\log a_j)^2. \tag{16}$$

With $\psi^{(s)}$ fixed, maximizing it over each z_i gives the E-step of Eq. 2. With $\mathbf{z}$ fixed, exact minimization of Eq. 3 gives the other coordinate update up to normalization because the Gaussian quality prior has no parameter in ψ. Algorithm 3 instead takes one stochastic optimizer step. Because z_i is scalar and its log posterior is strictly concave by Theorem 6, each E-step has one solution and bisection computes it deterministically for fixed criterion parameters.

The likelihood alone has the IRT scale tradeoff

$$a_j \mapsto a_j/r, \qquad z_i \mapsto r z_i, \qquad b_j \mapsto r b_j, \qquad r > 0,$$

which leaves the response function argument $a_j(z_i - b_j)$ unchanged. The Gaussian prior and discrimination regularizer fix this scale. The prior mean anchors the location, and the constraint $a_j > 0$ fixes the direction.

The E-step returns the MAP quality $\hat{z}_i$ of Eq. 11, as specified by Algorithm 2. The log posterior objective ℓ_i is strictly concave by Theorem 6, so its derivative has one root. The algorithm starts from $[-B_z, B_z]$ and doubles any endpoint whose derivative has the wrong sign until this root is bracketed. The derivative limits guarantee that the expansion terminates. It then halves the bracket T times using the derivative sign at the midpoint, with $\lambda(x) = \phi(x)/\Phi(x)$ as in Eq. 9. If the expanded bracket has width W, the returned midpoint has error at most $W/2^{T+1}$.

Algorithm 2 ESTIMATEZ: MAP quality estimation by bisection.

Require: verdicts G_{ij}, criterion parameters a_j, b_j from the RPN ψ, prior variance σ_z^2, initial bracket half-width $B_z > 0$, iteration count T
Ensure: MAP quality $\hat{z}_i$ from Eq. 11 for every rollout i
define $m_i(x) \leftarrow \ell_i'(x)$ by Eq. 7
$\text{lo}_i \leftarrow -B_z, \quad \text{hi}_i \leftarrow +B_z$
while some $m_i(\text{lo}_i) < 0$ **do**
  $\text{lo}_i \leftarrow 2\text{lo}_i$ for those rollouts
end while
while some $m_i(\text{hi}_i) > 0$ **do**
  $\text{hi}_i \leftarrow 2\text{hi}_i$ for those rollouts
end while
for $t = 1$ to T **do**
  $z_i \leftarrow \frac{1}{2}(\text{lo}_i + \text{hi}_i)$
  $\text{hi}_i \leftarrow z_i$ where $m_i(z_i) \le 0$
  $\text{lo}_i \leftarrow z_i$ where $m_i(z_i) > 0$
end for
return $\hat{z}_i \leftarrow \frac{1}{2}(\text{lo}_i + \text{hi}_i)$ for every rollout i

Algorithm 3 sweeps the policy step's rollouts once per epoch in mini-batches of B. For each mini-batch, one differentiable forward pass through the RPN ψ predicts (a_{ij}, b_{ij}) for every pair (q_i, c_{ij}). It reruns the E-step on detached parameter values to obtain $\hat{z}_i^{\text{M}}$, then adds its share of the gradient of Eq. 3 to an accumulator. One AdamW update is applied at the end of the epoch, so ψ does not change between mini-batches.

Algorithm 3 UPDATEPSI: stochastic partial M-step for the RPN ψ.

Require: the policy step's rollouts (rollout i has prompt q_i, criteria c_{ij}, verdicts G_{ij}, and criterion count K_i), the current RPN ψ with its AdamW state, and the hyperparameters E (epochs), B (mini-batch size), λ_a (discrimination regularizer weight), τ_g (gradient clip norm)
Ensure: updated RPN ψ

for $e = 1$ to E **do**
  shuffle the rollouts and split them into M mini-batches of B
  $g_\psi \leftarrow 0$ ▷ gradient accumulator
  for each mini-batch $\mathcal{B}$ of B rollouts **do**
    $(a_{ij}, b_{ij}) \leftarrow \psi(q_i, c_{ij})$ for every criterion of every rollout in $\mathcal{B}$ ▷ one differentiable forward
    $\hat{z}_i^{\mathrm{M}} \leftarrow \textsc{EstimateZ}(\mathcal{B})$ with (a_{ij}, b_{ij}) detached ▷ Alg. 2, targets held fixed
    $\mathcal{L} \leftarrow \frac{1}{B} \sum_{i \in \mathcal{B}} \left[-\frac{1}{K_i} \sum_{j=1}^{K_i} \ell_{ij}(\hat{z}_i^{\mathrm{M}}) + \frac{\lambda_a}{K_i} \sum_{j=1}^{K_i} (\log a_{ij})^2 \right]$ ▷ Eq. 3
    $g_\psi \leftarrow g_\psi + \nabla_\psi \mathcal{L} / M$ ▷ accumulate, do not step
  end for
  clip $\|g_\psi\|$ to τ_g, then take one AdamW step on ψ
end for
return ψ

B.2 GRPO OBJECTIVE

Let $\widetilde{R}_i$ denote the scalar reward passed to GRPO. It is centered within each prompt's group of N rollouts to form the advantage

$$A_i = \frac{\widetilde{R}_i - \overline{\widetilde{R}}}{\mathrm{std}_k \widetilde{R}_k + \varepsilon}, \quad \overline{\widetilde{R}} = \tfrac{1}{N} \sum_k \widetilde{R}_k. \tag{17}$$

Here $\varepsilon > 0$ stabilizes groups when reward variance is near zero.

Write rollout $o_i = (o_{i1}, \ldots, o_{iL_i})$, where L_i is its generated token count. The standard GRPO update uses the clipped surrogate objective of PPO (Schulman et al., 2017) and a Kullback-Leibler (KL) penalty toward the fixed reference policy π_{ref}. Let π_θ be the current policy and $\pi_{\theta_{\mathrm{old}}}$ the old policy that generated the rollouts. Define the token likelihood ratio

$$\varrho_{it} = \frac{\pi_\theta(o_{it} \mid q, o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{it} \mid q, o_{i,<t})}$$

and $r_{it}^{\mathrm{ref}} = \pi_{\mathrm{ref}}(o_{it} \mid q, o_{i,<t}) / \pi_\theta(o_{it} \mid q, o_{i,<t})$. Its KL estimator at the token level is $\widehat{D}_{\mathrm{KL},it} = r_{it}^{\mathrm{ref}} - \log r_{it}^{\mathrm{ref}} - 1$. The policy maximizes

$$J(\theta) = \mathbb{E}_i \left[\frac{1}{L_i} \sum_{t=1}^{L_i} \Big(\min\big(\varrho_{it} A_i, \mathrm{clip}(\varrho_{it}, 1-\epsilon_c, 1+\epsilon_c) A_i\big) - \beta_{\mathrm{KL}} \widehat{D}_{\mathrm{KL},it} \Big) \right], \tag{18}$$

where ϵ_c is the likelihood ratio clip radius and β_{KL} weights the KL penalty.

C ADDITIONAL CRITERION AND REWARD EXPERIMENTS

C.1 TEXT PREDICTION OF CRITERION PARAMETERS

This experiment tests whether criterion parameters predicted by the RPN ψ from text recover empirical criterion difficulty and predict observed verdicts. The embedder ablation compares Qwen3 Embedding models (Zhang et al., 2025) and Llama-Embed-Nemotron-8B (Babakhin et al., 2025). The criterion parameters predicted by the RPN, together with inferred rollout quality $\hat{z}_i$, give the fitted pass probability $\widehat{P}_{ij}^{\mathrm{MAP}} = F(a_j(\hat{z}_i - b_j))$.

The response function F varies across the Gaussian CDF Φ, the logistic CDF σ, and the complementary log-log function $\mathrm{cll}(t) = 1 - \exp(-e^t)$. The last choice relaxes $F(0) = 0.5$, so b_j is a location parameter rather than the 50% pass threshold in that configuration. A second ablation varies how many criterion parameters ψ predicts. The model with three parameters adds a lower asymptote γ_j, which the pass probability approaches as $z_i \to -\infty$ (Birnbaum, 1968). The model with four

parameters adds an upper asymptote ξ_j, which the pass probability approaches as $z_i \to +\infty$ (Magis, 2013). The resulting fitted probability adds γ_j to $(\xi_j - \gamma_j)F(a_j(\hat{z}_i - b_j))$, with $\xi_j = 1$ in the model with three parameters and $\gamma_j = 0$, $\xi_j = 1$ in the models with one and two parameters. The model with one parameter also fixes $a_j = 1$. The RPN reads (q, c_j) and predicts every free parameter, and the asymptotes are constrained to $0 < \gamma_j < \xi_j < 1$.

For a set $\mathcal{I}$ of rollouts, define the empirical criterion pass rate as $\bar{G}_{j,\mathcal{I}} = |\mathcal{I}|^{-1} \sum_{i \in \mathcal{I}} G_{ij}$, and write $\bar{G}_j$ when the set is clear. Two complementary evaluation metrics are used. First, Spearman correlation between the criterion difficulty parameter b_j and empirical criterion difficulty $1 - \bar{G}_j$ measures whether the RPN predicts which criteria are hard from text (Table 6). Second, ROC-AUC between $\widehat{P}_{ij}^{\mathrm{MAP}}$ and G_{ij} measures verdict ranking. Quality $\hat{z}_i$ is inferred from the full rollout verdict vector, including the evaluated verdict G_{ij}. Section 4.2 describes its leave-one-criterion-out counterpart, while Appendix C.6 defines a separate metric for the next policy step.

Table 5: Verdict ROC-AUC of RPN configurations across datasets when the evaluated verdict is included. Each block varies one component: the response function, number of criterion parameters, text conditioning, text embedder, or rollout policy. The remaining components use the standard configuration. Bold marks the highest macro mean in each block.

	Verdict included in fit (ROC-AUC)				
RPN configuration	Medical	Science	RaR Science	RubricBench	Macro mean
			F		
Φ (Gaussian CDF)	80.6	84.2	91.7	91.9	87.1
σ (logistic CDF)	80.5	84.1	90.1	90.6	86.3
cll (complementary log-log)	80.5	84.3	92.2	91.8	**87.2**
			Number of criterion parameters		
One parameter, b_j	80.4	83.8	90.8	90.7	86.4
Two parameters, (a_j, b_j)	80.6	84.2	91.7	91.9	**87.1**
Three parameters, (a_j, b_j, γ_j)	80.6	84.2	91.7	91.7	87.0
Four parameters, $(a_j, b_j, \gamma_j, \xi_j)$	80.5	84.2	91.5	90.4	86.6
			Text conditioning		
$\psi(c_j)$	80.6	84.1	91.8	91.7	**87.1**
$\psi(q, c_j)$	80.6	84.2	91.7	91.9	**87.1**
			Text embedder		
Qwen3-Embedding-0.6B	79.2	82.9	91.3	91.5	86.2
Qwen3-Embedding-4B	80.6	84.2	91.7	91.9	87.1
Qwen3-Embedding-8B	81.4	84.9	91.8	91.6	87.4
Llama-Embed-Nemotron-8B	83.1	86.3	92.5	92.8	**88.7**
			Rollout policy		
Qwen3.5-4B	80.6	84.2	91.7	91.9	**87.1**
Qwen3.5-2B	82.1	84.1	90.2	86.9	85.8
Llama-3.1-8B-Instruct	84.5	83.0	88.7	86.4	85.6

Across all configurations in Table 5, macro ROC-AUC ranges from 85.6% to 88.7%. The adopted item response model with two parameters and a Gaussian CDF reaches 87.1%, compared with 86.4% for the model with one parameter, 87.0% for the model with three parameters, and 86.6% for the model with four parameters.

C.2 RPN Representation Geometry

This analysis tests whether nearby RPN representations have similar predicted criterion difficulties and verdict patterns. It projects the last hidden representations used to predict b_j with t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten & Hinton, 2008).

Figure 4 shows stronger local structure for predicted criterion difficulty and fitted pass probability than for empirical criterion difficulty and observed verdicts.

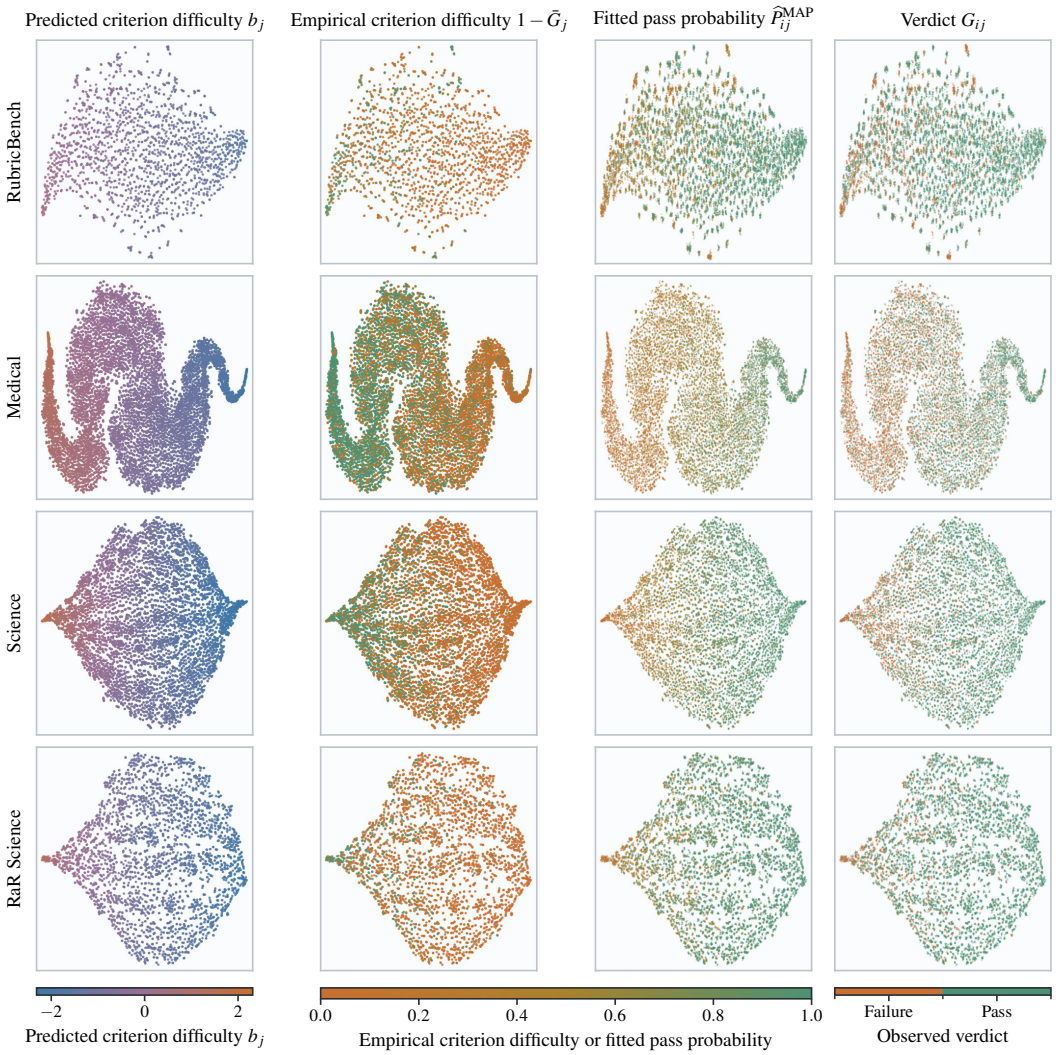

Figure 4: t-SNE projections of RPN representations. Rows show the four datasets. The first two columns color pairs by predicted criterion difficulty b_j and empirical criterion difficulty $1 - \bar{G}_j$. The last two color sampled verdicts at jittered pair coordinates by the pass probability $\widehat{P}_{ij}^{\mathrm{MAP}}$ fitted using the RPN and inferred rollout quality, and by verdict G_{ij}. Empirical criterion difficulty and fitted pass probability share a color scale.

Table 6 reports the Spearman correlation between b_j and $1 - \bar{G}_j$ as difficulty ρ_S. Trustworthiness at 10 neighbors (T@10) measures how well the projection preserves local neighborhoods. The analysis also computes Moran's I after normalizing each row of the graph that connects the 10 nearest neighbors of each displayed variable (Moran, 1950). Larger values mean that nearby points have more similar displayed values.

Table 6: Criterion difficulty prediction and local t-SNE geometry. Columns report difficulty Spearman correlation, trustworthiness at 10 neighbors, and Moran's I for predicted criterion difficulty, empirical criterion difficulty, pass probability fitted using the RPN and inferred rollout quality, and verdict. All values are percentages.

			Moran's I			
Dataset	Difficulty ρ_S	T@10	b_j	$1 - \bar{G}_j$	$\widehat{P}_{ij}^{\mathrm{MAP}}$	G_{ij}
RubricBench	38.0	97.2	98.8	21.5	74.0	45.9
Medical	47.9	99.2	99.7	25.9	77.0	32.9
Science	47.2	99.2	99.3	28.2	85.8	38.2
RaR Science	38.5	99.1	99.1	19.4	74.6	44.1

The Spearman correlation between predicted criterion difficulty and empirical criterion difficulty ranges from 38.0% to 47.9%. Moran's I is 74.0% to 85.8% for fitted pass probability and 32.9% to 45.9% for observed verdicts.

C.3 Response Function Separation of Tied Rollouts

This analysis compares rewards inferred with the Gaussian and logistic CDFs on observed verdicts from Medical and Science. Both functions use $a_j = 1$ and the same criterion difficulties. The analysis measures separation among pairs with equal pass counts and counts order reversals among pairs with different pass counts.

Table 7: Separation of rollout pairs by the Gaussian and logistic CDFs on Medical and Science. Both CDFs use $a_j = 1$ and the same criterion difficulties. Columns report separation among pairs with equal pass counts and count order reversals among unequal pairs.

Dataset	$F = \Phi$ separated	$F = \sigma$ separated	Count order reversals
Medical	83.3%	0.0%	0
Science	39.3%	0.0%	0

In Table 7, the Gaussian CDF separates 83.3% of pairs with equal pass counts on Medical and 39.3% on Science, compared with 0.0% for the logistic CDF in both datasets. Both CDFs produce zero count order reversals.

C.4 Reward Signal Across Policies and Group Sizes

This analysis compares reward variation and ties across datasets and policies. Both rewards are computed on the same rollouts, and RRT uses a frozen RPN. The variance share within prompts is the fraction of total reward variance within groups for the same prompt. The tied pair share is the fraction of rollout pairs within a group that receive equal rewards. The relative tied pair reduction compares this share under RRT with the reward based on rubric points.

The variance share within prompts under RRT is 1.2 to 2.2 times that under the reward based on rubric points across the 12 cells. Its relative tied pair reduction ranges from 2% to 58% across those cells.

The group size analysis repeats this matched reward comparison after drawing n rollouts without replacement from the 48 cached rollouts for each prompt. Table 8 reports the tied pair share and the variance share within prompts across group sizes.

At $n = 8$, RRT and the reward based on rubric points have tied pair shares of 3.6% and 5.8% on Medical, and 19.1% and 20.6% on Science. Their variance shares within prompts are 17.5% and 15.7% on Medical, and 24.1% and 17.6% on Science.

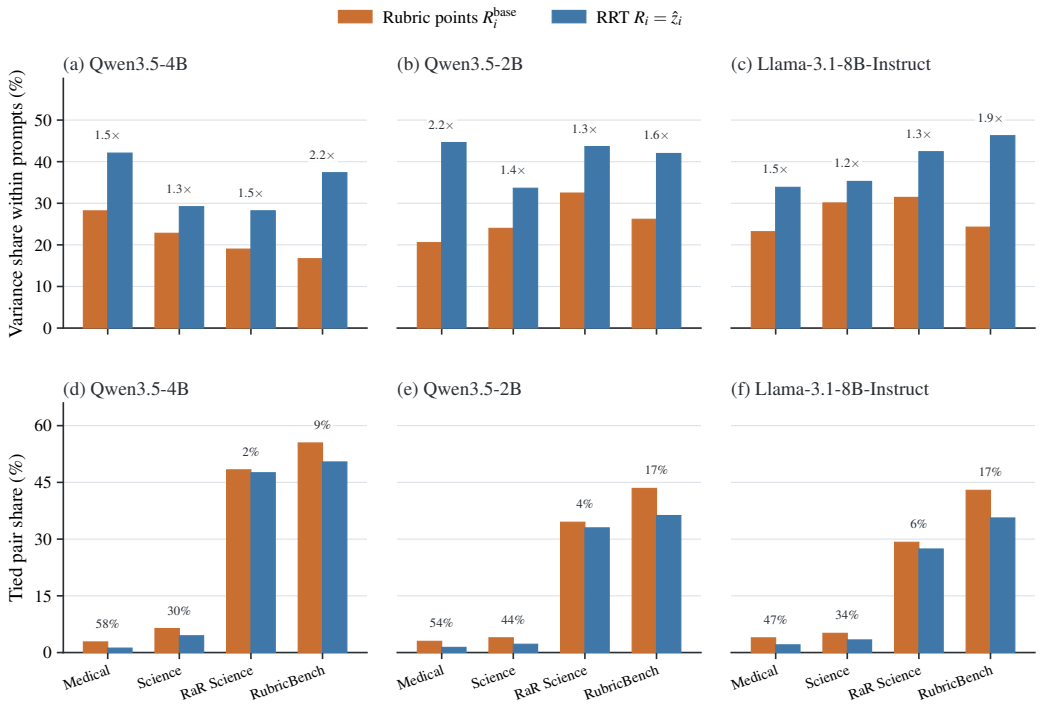

Figure 5: Reward signal on matched rollouts. Columns show three policies. Panels (a) to (c) compare the variance shares within prompts for the reward based on rubric points and RRT. Panels (d) to (f) compare tied pair shares. Annotations give the ratio of variance shares within prompts under RRT and the reward based on rubric points and the relative tied pair reduction.

Table 8: Reward signal across group sizes on Medical and Science. Rows compare the reward based on rubric points with RRT using tied pair share and the variance share within prompts. Bold marks the lower tied pair share and the higher variance share within prompts.

		Group size n					
Quantity	Reward	2	4	8	16	24	32
	Medical						
Tied pair share	Rubric points	5.6	5.9	5.8	5.9	5.9	5.9
	RRT	**3.4**	**3.7**	**3.6**	**3.6**	**3.6**	**3.6**
Variance share within prompts	Rubric points	9.2	13.2	15.7	16.6	16.9	17.3
	RRT	**10.3**	**14.8**	**17.5**	**18.6**	**18.9**	**19.3**
	Science						
Tied pair share	Rubric points	21.0	20.8	20.6	20.7	20.6	20.5
	RRT	**19.5**	**19.2**	**19.1**	**19.2**	**19.1**	**19.0**
Variance share within prompts	Rubric points	10.1	15.4	17.6	18.8	19.6	19.8
	RRT	**13.8**	**20.9**	**24.1**	**26.1**	**26.9**	**27.1**

C.5 Unanimous Criteria

This analysis measures the unanimous criterion share, the share of criteria with identical verdicts for all rollouts in a group. Table 9 reports this share by dataset and policy.

Table 9: Unanimous criterion share by dataset and policy.

Dataset	Qwen3.5-4B	Llama-3.1-8B-Instruct	Qwen3.5-2B
Medical	60.5	60.4	53.4
Science	69.3	54.6	52.3
RaR Science	80.5	52.5	60.6
RubricBench	73.4	57.9	56.1

In Table 9, the unanimous criterion share ranges from 52.3% to 80.5% across the 12 cells.

C.6 Online EM Calibration

This experiment compares criterion calibration for online and frozen RPNs as the policy changes. Both conditions use identical verdict targets. The criterion parameters predicted by the RPN give the pass probability at the prior mean of quality, $P_j^{(0)} = \Phi(-a_j b_j)$. The checkpoint aggregate averages $P_j^{(0)}$ and $\bar{G}_j$ over checkpoints, centers both quantities within each prompt, and computes one pooled Pearson correlation. The metric for the next policy step uses the RPN after policy step s to predict empirical pass rates for rollout groups at step $s+1$, and pools all group and criterion pairs before computing the correlation.

Table 10: Pearson correlation between empirical criterion pass rate $\bar{G}_j$ and pass probability at the prior mean of quality $P_j^{(0)}$ computed from criterion parameters predicted by the frozen and online RPNs. Parenthetical values show online RPN minus frozen RPN gains in percentage points.

	Medical		Science		Macro mean	
Target	Frozen	Online	Frozen	Online	Frozen	Online
Checkpoint aggregate	53.2	55.2 (2.0)	56.1	56.7 (0.6)	54.7	56.0 (1.3)
Next policy step	56.9	59.6 (2.7)	58.9	59.6 (0.7)	57.9	59.6 (1.7)

In Table 10, the online RPN gains 1.3 points on the checkpoint aggregate macro mean and 1.7 points on the macro mean for the next policy step. The corresponding Medical and Science gains are 2.0 and 0.6 points on the checkpoint aggregate, and 2.7 and 0.7 points on the next policy step.

Figure 6 shows exploratory examples of criterion trajectories, one from Medical and one from Science. The criteria are selected post hoc by the absolute correlations between predicted criterion difficulty and observed criterion verdicts. Selection among eligible criteria maximizes $\min\{|r|, |\rho_S|\}$ between these two quantities.

Both examples in Figure 6 end with lower predicted criterion difficulty and higher rolling empirical criterion pass rates. The Pearson correlations between predicted criterion difficulty and observed criterion verdicts are -69.8% over 40 Medical checkpoints and -62.2% over 30 Science checkpoints.

C.7 Comparison of Hard and Soft E-Steps

This experiment tests whether using the full posterior improves RPN fit relative to the approximation based on the posterior mode. The comparison uses the hard E-step and a soft E-step that replaces the posterior mode $\hat{z}_i$ by the full posterior of z_i, evaluated on a fixed grid of 41 points over $[-B_z, B_z]$ with normalized masses η_{ik}. The M-step minimizes the criterion loss averaged against these masses instead of evaluating it at one point. Concentrating the mass at the grid node nearest $\hat{z}_i$ approximates Eq. 3. It is exact only when $\hat{z}_i$ lies on that node. For this comparison, the RubricBench RPN is fitted four times. Only the E-step and the discrimination regularizer weight λ_a vary.

In Table 11, RPNs fitted with the hard E-step reach ROC-AUC of 91.0% to 91.9%, compared with 90.4% to 90.6% for those fitted with the soft E-step. Their criterion losses are 0.364 to 0.369 and 0.380 to 0.381, respectively.

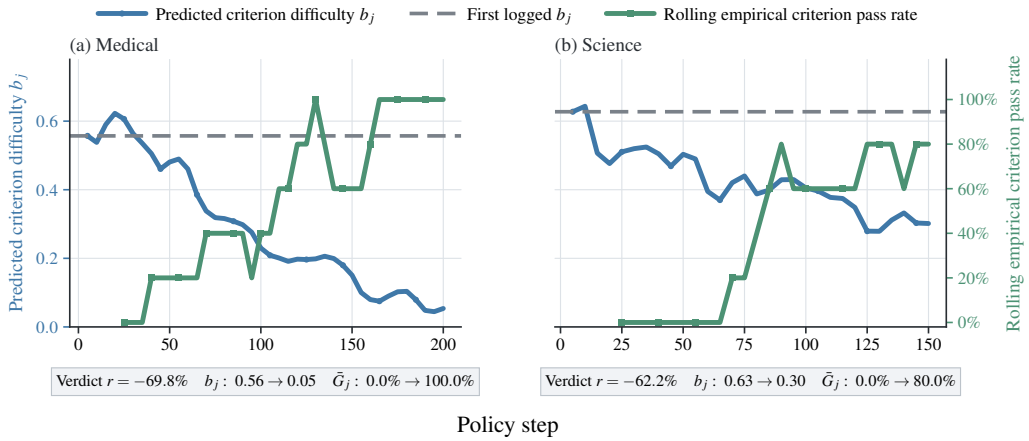

Figure 6: EM calibration trajectories for one post hoc selected criterion in Medical and one in Science. Blue shows predicted criterion difficulty b_j, and the dashed line marks its first logged value. Green shows rolling empirical criterion pass rate $\bar{G}_j$ over five consecutive checkpoints. Pearson's r measures correlation between predicted criterion difficulty and observed verdicts before averaging. Endpoint annotations describe the displayed series.

Table 11: RubricBench RPN fit under hard and soft E-steps at two discrimination regularizer weights λ_a. Columns report criterion loss and ROC-AUC. Only ROC-AUC is a percentage.

Method	λ_a	Criterion loss	ROC-AUC
Hard EM	0	0.369	91.0
Hard EM	0.05	0.364	**91.9**
Soft EM	0	0.380	90.4
Soft EM	0.05	0.381	90.6

D Robustness and Model Assumption Checks

D.1 Judge Agreement and Noise Calibration

Repeated judging can produce different verdicts. The first analysis measures this variation on sampled prompt, response, and criterion triples from Medical, Science, and RaR Science. Each triple receives three independent verdicts. Pairwise verdict agreement is the mean over the three replicate pairs. Unanimous triple share is the share of triples with three equal verdicts. The reported statistics include Cohen's κ (Cohen, 1960) and Gwet's agreement coefficient 1 (AC1) (Gwet, 2008). Verdict prevalence is skewed toward PRESENT, and the two coefficients account for chance agreement differently. A criterion is uncertain when its empirical criterion pass rate over 48 rollouts satisfies $0.1 < \bar{G}_{j,48} < 0.9$.

Table 12: Agreement across three independent criterion verdicts by dataset. Columns report pairwise verdict agreement, unanimous triple share, Cohen's κ, Gwet's AC1, and pairwise verdict agreement on uncertain criteria. All values are percentages.

Dataset	Pairwise agreement	Unanimous triples	κ	AC1	Uncertain agreement
Medical	94.6	91.9	88.3	90.0	90.5
Science	95.0	92.4	86.2	92.1	90.3
RaR Science	95.2	92.9	86.9	92.5	86.0
Macro mean	94.9	92.4	87.1	91.6	88.9

Table 12 reports pairwise verdict agreement of 94.6% to 95.2% and unanimous triple shares of 91.9% to 92.9%.

A larger sample of prompt, response, and criterion triples uses the 48 cached rollouts per prompt and covers all 12 combinations of dataset and policy used in the corruption analysis in Appendix D.2.

The majority of each triple's three verdicts is the consensus. Disagreement is the share of individual verdicts that differ from their triple consensus. The PRESENT column gives the rate at which a triple with consensus NOT_PRESENT draws a PRESENT verdict. The NOT_PRESENT column gives the opposite error rate.

Table 13: Disagreement among three repeated judge verdicts by dataset and policy. Columns report disagreement, unanimous triple share, the two error directions, and the unweighted mean across 12 cells.

Dataset	Policy	Disagreement	Unanimous triples	PRESENT	NOT_PRESENT
Medical	Qwen3.5-4B	1.58	95.27	1.52	1.63
	Qwen3.5-2B	1.33	96.01	0.94	2.15
	Llama-3.1-8B-Instruct	1.16	96.53	0.76	2.35
Science	Qwen3.5-4B	1.39	95.83	2.57	0.95
	Qwen3.5-2B	1.51	95.46	1.69	1.37
	Llama-3.1-8B-Instruct	1.21	96.36	0.90	1.89
RaR Science	Qwen3.5-4B	1.35	95.96	2.26	1.01
	Qwen3.5-2B	1.24	96.29	1.69	0.95
	Llama-3.1-8B-Instruct	1.49	95.52	1.26	1.78
RubricBench	Qwen3.5-4B	0.34	98.98	0.18	2.35
	Qwen3.5-2B	1.54	95.38	1.44	1.64
	Llama-3.1-8B-Instruct	1.08	96.75	1.15	1.03
Mean		1.27	96.19	1.36	1.59

Table 13 reports 1.27% mean disagreement and a unanimous triple share of 96.19%.

D.2 Reward Stability Under Judge Variation

The first comparison evaluates reward stability across the replicate verdicts summarized in Table 12. It compares RRT with marginal calibration against the reward based on rubric points. Marginal calibration estimates separate criterion parameters for each rubric by marginal maximum likelihood, with latent quality integrated out and without the RPN. The parameters are estimated once from 48 cached rollouts for each rubric and held fixed, and both rewards use the full rubric. Tied pair share is the share of rollout pairs with equal rewards. Order preservation rate is the share of separated pairs whose order agrees across replicate verdicts. Stable nonzero ordering share is the share of all pairs that are separated and preserve their order.

Table 14: Stability of RRT with marginal calibration and the reward based on rubric points across replicate criterion verdicts. Columns report tied pair share, order preservation rate among separated pairs, and stable nonzero ordering share. Bold marks the lower tied pair share and higher ordering rates.

	Tied pair share		Order preservation rate		Stable nonzero ordering share	
Dataset	RRT + marginal calibration	Rubric points	RRT + marginal calibration	Rubric points	RRT + marginal calibration	Rubric points
Medical	**1.7**	2.5	**82.2**	77.5	**80.8**	75.6
Science	**12.5**	13.3	**71.8**	70.6	**62.9**	61.2
RaR Science	**46.8**	60.2	**65.2**	63.8	**34.7**	25.4
Macro mean	**20.3**	25.3	**73.1**	70.6	**59.5**	54.1

In Table 14, RRT with marginal calibration raises the macro stable nonzero ordering share from 54.1% to 59.5%. The tied pair share decreases by 5.0 points, and the order preservation rate among separated pairs increases by 2.5 points.

Using the larger sample summarized in Table 13, this experiment tests whether reward stability changes when judge errors concentrate on triples with split verdicts rather than all triples.

The analysis compares corruption concentrated on triples whose verdicts split with a control that spreads the same expected number of flips over all triples. It flips the consensus at corruption levels

0.025, 0.05, and 0.10. The channels are matched on the expected number of flips by bisection. The flips follow the leniency rate measured in each cell. Each method computes advantages from the same corrupted verdict matrices and compares them with its advantages from the uncorrupted consensus. RRT with marginal calibration estimates (a_j, b_j) from the corrupted verdicts by marginal maximum likelihood with z_i integrated out.

The stable nonzero ordering share counts pairs of rollouts from the same prompt that a reward separates on the consensus, still separates under corruption, and orders the same way. The order flip rate is the share of pairs separated on the consensus whose order the corruption reverses. For each quantity, Δ is the named RRT variant minus the reward based on rubric points, measured in percentage points and averaged over the 12 cells. Each cell averages five corruption draws over its 80 prompt groups. Ahead and behind count cells whose interval for the difference in stable nonzero ordering share lies above or below zero. Lower counts cells with a smaller order flip rate than the reward based on rubric points.

Table 15: Reward stability under corruption of the consensus of three verdicts across 12 dataset and policy cells. Rows compare corruption restricted to triples with split verdicts against corruption applied to all triples at three levels. Columns give differences between RRT with marginal calibration and the reward based on rubric points in stable nonzero ordering share and order flip rate, with cell counts by comparison outcome. Bold marks the highest level in the block with estimated a_j.

		Stable nonzero ordering share			Order flip rate	
Corruption channel	Level	Mean Δ	Ahead	Behind	Mean Δ	Lower
RRT + marginal calibration, $a_j = 1$						
Triples with split verdicts	0.025	$+0.82$	2	2	$+0.85$	3
	0.050	$+1.09$	2	1	$+0.14$	3
	0.100	$+1.56$	3	0	-0.73	7
All triples	0.025	$+0.61$	2	3	$+1.18$	2
	0.050	$+0.78$	3	1	$+0.33$	3
	0.100	$+1.28$	3	0	-0.59	5
RRT + marginal calibration, estimated a_j						
Triples with split verdicts	0.025	$+0.99$	2	0	$+0.61$	3
	0.050	$+1.29$	2	0	-0.07	4
	0.100	$+1.99$	**6**	**0**	-1.30	**11**
All triples	0.025	$+1.16$	3	1	$+0.32$	3
	0.050	$+1.65$	8	0	-0.95	8
	0.100	$+2.00$	**8**	**0**	-1.68	**9**

At corruption level 0.10 in Table 15, RRT with marginal calibration is ahead in 6 of 12 cells when corruption targets split verdicts and 8 of 12 cells when it targets all verdicts. Its order flip rate is lower in 11 and 9 cells, respectively. Its mean stable nonzero ordering gain is 0.99 to 2.00 points at every level. Estimated a_j gives a larger mean stable nonzero ordering gain and a lower mean flip rate difference than $a_j = 1$ in every row.

D.3 Controlled Judge Noise Robustness

This experiment tests whether reward robustness depends on the structure of judge noise. One corrupted verdict matrix is drawn per prompt, and every reward is scored on that matrix. Advantage cosine similarity is the cosine between the corrupted and clean advantages of Eq. 17, reported as a percentage. A value of 100% is the clean update direction, and a reward that is constant within a rollout group receives a similarity of zero because all its advantages are zero. Eight channels are matched on the expected number of flipped verdicts. Four channels do not select criteria. Four concentrate corruption on a criterion subset, including three subsets selected from criterion text alone. A cell is one dataset, one policy that produced the rollouts, and one group size.

The target is the advantage induced by the clean criterion score. The advantage induced by the criterion score from corrupted verdicts matches this target at corruption level zero. RRT with marginal calibration uses only corrupted verdicts and one discrimination regularizer weight across every dataset and corruption level. In Table 16, Δ is the advantage cosine similarity of RRT with

marginal calibration minus that of the criterion score computed from the same corrupted verdicts. Ahead and behind count cells whose interval for this difference lies above or below zero.

Table 16: Difference in advantage cosine similarity between RRT with marginal calibration and criterion score at corruption level 0.20. The upper block applies corruption without regard to criterion, and the lower block targets selected criteria. Mean differences are in percentage points. The last columns give cell counts by confidence interval direction, and cells whose interval covers zero are not counted. Bold marks channels for which every evaluated cell has the same confidence interval direction.

			Cells	
Corruption channel	Structure	Mean Δ (points)	Ahead	Behind
Symmetric, all criteria	None	−0.8	0	17
Lenient, all criteria	None	+0.7	15	0
Lenient, longer responses	Across rollouts	−1.2	0	**19**
Symmetric and lenient mixture	Both	+0.1	5	1
Symmetric, 30% of criteria	Across criteria	+5.1	**19**	0
Least discriminating criteria	Criterion text	+2.0	**8**	0
Hedging criterion wording	Criterion text	+1.4	**8**	0
Longest criterion text	Criterion text	+1.3	**8**	0

At corruption level 0.20 in Table 16, RRT with marginal calibration is ahead in 19 of 19 cells when symmetric corruption targets 30% of criteria, and behind in 19 of 19 cells when lenient corruption follows response length. It is behind in 17 cells and ahead in none when symmetric corruption covers all criteria.

In Table 17, Δ is the advantage cosine similarity of RRT with marginal calibration minus that of criterion score in percentage points.

Table 17: Advantage cosine similarity with clean verdicts under corruption of a fixed 30% criterion subset. Rows vary corruption level by dataset. Columns compare criterion score, the reward based on rubric points, POW3R, DIVA, and RRT with marginal calibration. Similarities are percentages. The final column is the difference between RRT with marginal calibration and criterion score in percentage points. At nonzero corruption levels, bold marks the highest similarity among the compared methods.

Dataset	Corruption	Criterion score	Rubric points	POW3R	DIVA	RRT + marginal calibration	Δ
Medical	0.00	100.0	97.8	97.4	94.9	95.6	−4.4
	0.10	68.9	66.6	64.1	60.4	**72.1**	+3.2
	0.20	49.7	47.7	44.1	38.8	**55.7**	+6.0
	0.30	30.6	29.3	26.7	22.7	**35.0**	+4.4
Science	0.00	100.0	97.3	97.1	94.3	96.4	−3.6
	0.10	62.2	60.3	57.8	52.9	**66.5**	+4.3
	0.20	43.4	42.3	38.8	32.5	**50.4**	+7.0
	0.30	26.4	25.5	23.0	18.9	**32.0**	+5.6
RaR Science	0.00	100.0	94.7	94.8	95.8	98.9	−1.1
	0.10	51.5	43.9	41.6	39.7	**54.3**	+2.8
	0.20	35.3	30.0	27.1	23.6	**39.5**	+4.2
	0.30	21.3	17.9	15.7	13.8	**24.5**	+3.2
RubricBench	0.00	100.0	100.0	99.7	96.7	99.1	−0.9
	0.10	61.0	61.0	58.2	51.6	**63.0**	+2.0
	0.20	36.1	36.1	32.8	25.7	**39.1**	+3.0
	0.30	25.4	25.4	23.1	18.6	**27.3**	+1.9

On the fixed 30% criterion subset, RRT with marginal calibration gains 6.0 points over criterion score on Medical, 7.0 on Science, 4.2 on RaR Science, and 3.0 on RubricBench at corruption level 0.20.

D.4 Empirical Criterion Pass Rate Reliability

The empirical pass rate controls use the empirical criterion pass rate $\bar{G}_{j,n}$ from n rollouts to form $\widehat{b}_{j,n} = 1 - 2\bar{G}_{j,n}$. Given the finite cache pass rate $\bar{G}_{j,48}$, sampling n rollouts without replacement gives

$$\mathrm{Var}(\widehat{b}_{j,n} \mid \bar{G}_{j,48}) = \frac{4\bar{G}_{j,48}(1-\bar{G}_{j,48})}{n}\frac{48-n}{47}, \tag{19}$$

where $(48-n)/47$ corrects for the finite cache.

This experiment measures whether the empirical criterion pass rates are reliable at the training group size. Each prompt's 48 rollouts are split into six disjoint blocks of eight. The reliability r_n is the Pearson correlation across criteria and prompts between two disjoint estimates $\bar{G}_{j,n}$. As a sensitivity check, the correlation between b_j and empirical criterion difficulty $1-\bar{G}_{j,n}$ is divided by $\sqrt{r_n}$.

Table 18: Empirical criterion pass rate reliability from disjoint blocks of eight rollouts. Values are ranges across Medical, Science, RaR Science, and RubricBench.

Quantity	Range across datasets
Equal pass count in two blocks of eight	61.4% to 77.1%
Uncertain criteria with a majority disagreement	25.5% to 31.7%
Reliability at eight rollouts, all criteria	93.6% to 96.0%
Reliability at eight rollouts, uncertain criteria	70.9% to 71.9%
Difficulty correlation after attenuation adjustment	38.5% to 52.4%

In Table 18, two blocks of eight have the same pass count for 61.4% to 77.1% of criteria, while 25.5% to 31.7% of uncertain criteria have a majority disagreement. Reliability at eight rollouts is 93.6% to 96.0% over all criteria and 70.9% to 71.9% over uncertain criteria.

D.5 Rank Correlation Across Criterion Splits

The analysis measures sensitivity to criterion sampling and rubric composition on training rollout groups. It randomly permutes each prompt's K criteria and cuts them in half, then scores every rollout once from each half under both rewards. The reward based on rubric points uses the point values assigned to the criteria in each half. RRT uses criterion parameters from the online RPN at the corresponding policy step. Both rewards use the same rollout verdicts.

Rank correlation across criterion splits is Kendall's τ between the two rollout orderings, averaged over random splits. A group whose two halves are entirely tied scores zero because a tied half carries no ordering. Table 19 reports this correlation for each dataset. Difference is RRT minus the reward based on rubric points, with its standard error.

Table 19: Rank correlation across criterion splits for reward ordering within prompts on training rollout groups from RRT runs. Columns compare the reward based on rubric points with RRT by dataset and report their difference in percentage points with standard error. Bold marks the higher rank correlation in each dataset.

Dataset	Rubric points	RRT	Difference
Medical	26.6	**27.1**	$+0.5 \pm 0.3$
Science	21.5	**25.1**	$+3.6 \pm 0.3$
RaR Science	12.4	**27.5**	$+15.1 \pm 3.5$
RubricBench	25.1	**27.5**	$+2.4 \pm 0.6$

RRT has higher rank correlation than the reward based on rubric points on all four datasets in Table 19. Its gain is 15.1 points on RaR Science and 0.5 to 3.6 points on the other datasets.

Table 20 bins groups into quintiles by rank correlation from rubric points on one set of random criterion splits and scores them on a disjoint set. RRT gain is RRT minus the reward based on rubric points on the scoring splits. Groups improved is the percentage of groups with a positive gain.

Table 20: Rank correlation across criterion splits for RRT and the reward based on rubric points on training rollout groups across rank correlation quintiles. Columns report RRT gains in percentage points and the percentage of groups with a positive difference.

Rank correlation quintile	RRT gain	Groups improved (%)
1, lowest rank correlation	+14	57.8
2	+7	30.9
3	+3	38.4
4	+2	35.5
5, highest rank correlation	+1	28.0

In Table 20, RRT's gain decreases from 14 points in the lowest rank correlation quintile to 1 point in the highest. The share of groups improved is 57.8% in the lowest quintile and 28.0% to 38.4% in the other four.

D.6 Local Independence Diagnostics

Eq. 1 assumes that criterion verdicts are conditionally independent given quality and the criterion parameters. Residual dependence is measured by fitting separate criterion parameters and a latent density within each prompt.

Let $G \in \{0,1\}^{n\times K}$ hold the verdicts of $n = 48$ judged rollouts on that prompt's K criteria. The item response model uses a Gaussian CDF in the equivalent form

$$\Phi\big(a_j(z_i - b_j)\big) = \Phi(\alpha_j z_i + \beta_j), \qquad \alpha_j = a_j > 0,\ \beta_j = -a_j b_j,$$

by marginal maximum likelihood over a fixed grid $x_1, \ldots, x_Q$ of $Q = 61$ points on $[-4, 4]$ with latent density masses $\omega_1, \ldots, \omega_Q$. The E-step forms the responsibility over the grid for each rollout,

$$\eta_{ik} \propto \omega_k \prod_{j=1}^{K} \Phi(\alpha_j x_k + \beta_j)^{G_{ij}} \big(1 - \Phi(\alpha_j x_k + \beta_j)\big)^{1-G_{ij}}.$$

The M-step maximizes the expected log likelihood of the complete data in (α_j, β_j). With $n_k = \sum_i \eta_{ik}$ and $y_{jk} = \sum_i \eta_{ik} G_{ij}$, this is a binomial generalized linear model of y_{jk} successes out of n_k trials at covariate x_k, with link Φ^{-1}. The objective is concave in (α_j, β_j), so Fisher scoring gives the maximizer. The latent density masses ω_k are either fixed to a discretized standard normal or reestimated as $\omega_k \propto n_k$ and restandardized to mean zero and unit variance. Table 21 uses the estimated density.

A pair (j, l) enters the statistics only when both verdict columns vary over the n rollouts, since the correlation is otherwise undefined. Write the normalized observed 2×2 table as

$$O_{uv} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{G_{ij} = u, G_{il} = v\}.$$

The table implied by the fitted model after marginalizing z is

$$E_{uv} = \sum_k \omega_k\, P_{jk}^{u} (1 - P_{jk})^{1-u} P_{lk}^{v} (1 - P_{lk})^{1-v}, \qquad P_{jk} = \Phi(\alpha_j x_k + \beta_j),$$

where local independence gives the factorization inside the sum. The reported quantities are r, the phi correlation of O, and r_{model}, the phi correlation of E.

The residual correlation is computed leave-pair-out. For the pair (j, l), the posterior over the grid uses only the other $K - 2$ criteria,

$$\eta_{ik}^{(-jl)} \propto \omega_k \prod_{m \neq j,l} \Phi(\alpha_m x_k + \beta_m)^{G_{im}} \big(1 - \Phi(\alpha_m x_k + \beta_m)\big)^{1-G_{im}},$$

take the posterior mean $\tilde{z}_i^{(-jl)} = \sum_k \eta_{ik}^{(-jl)} x_k$, and correlate the residuals $G_{ij} - \Phi(a_j(\tilde{z}_i^{(-jl)} - b_j))$ and $G_{il} - \Phi(a_l(\tilde{z}_i^{(-jl)} - b_l))$ across rollouts. This gives the leave-pair-out Q_3 statistic (Yen, 1984), with the conditioning variable free of both criteria under test.

The quantities r, r_{model}, and Q_3 are biased at $n = 48$. The parameters (a_j, b_j) are estimated from the rollouts they are tested on, and error in $\tilde{z}_i^{(-jl)}$ leaves positive residual correlation. A parametric bootstrap calibrates every quantity (Efron, 1992). For each prompt, ten replicate matrices are drawn from the fitted $(a_j, b_j, \boldsymbol{\omega})$. Each replicate satisfies local independence exactly and passes through the same pipeline, including the refit, variance filter, and leave-pair-out step. The bootstrap reference and the null rates of Table 21 are replicate means. The flag threshold is the 95th percentile of $|Q_3|$ over replicates, so a rubric that satisfies local independence flags 5% of its pairs.

The share of pairwise mutual information (MI) explained is

$$1 - \frac{\big(\overline{\mathrm{MI}(O)} - \overline{\mathrm{MI}(E)}\big)_{\mathrm{obs}} - \big(\overline{\mathrm{MI}(O)} - \overline{\mathrm{MI}(E)}\big)_{\mathrm{null}}}{\overline{\mathrm{MI}(O)}_{\mathrm{obs}} - \big(\overline{\mathrm{MI}(O)} - \overline{\mathrm{MI}(E)}\big)_{\mathrm{null}}},$$

where $\mathrm{MI}(\cdot)$ is the MI of a 2×2 table in bits. The overbar averages over eligible criterion pairs. The subscripts obs and null denote the observed matrices and bootstrap matrices. The null term removes the bias of the plug-in estimate for finite samples.

The criterion text similarity s of Table 21 is the cosine between vectors whose entries use term frequency and inverse document frequency for the two criterion texts. Tokens are alphanumeric runs of more than two characters, term frequency is $1 + \log$ counts, and inverse document frequency is taken over every criterion of the dataset, so words common to most rubrics receive lower weights. The detailed statistics use the same fits and bootstrap null. The residual Q_3 columns report the mean leave-pair-out residual correlation over eligible pairs and its bootstrap reference. Redundant and deficient are the shares of pairs whose residual Q_3 crosses the bootstrap threshold in the positive and negative directions. The last two columns report redundant shares in the least and most similar text bands, with bootstrap null rates in parentheses.

Table 21: Local independence diagnostics for criterion verdict pairs by dataset. Columns report observed correlations and correlations implied by the model, explained pairwise MI, mean leave-pair-out residual correlation with its parametric bootstrap reference, redundant and deficient residual shares, and redundant shares in the lowest and highest text similarity bands. Parentheses give bootstrap null rates.

	Pairwise dependence			Residual Q_3		Residual share		Redundant share by s			
Dataset	$\bar{r}$	$\bar{r}_{\text{model}}$	Explained	Observed	Bootstrap	Redundant	Deficient	$s < 0.05$		$s > 0.30$	
Medical	4.3	6.3	72.8	1.7	1.8	6.0	6.0	3.9	(4.2)	23.5	(11.6)
Science	9.0	11.2	76.4	3.4	3.8	6.8	7.7	4.2	(3.6)	20.2	(12.2)
RaR Science	12.1	18.1	75.4	6.8	10.8	5.6	11.2	3.1	(2.7)	13.7	(10.3)
RubricBench	10.6	14.6	85.3	6.2	7.8	5.1	7.5	3.4	(4.1)	16.4	(10.7)
Macro mean	9.0	12.5	77.5	4.5	6.1	5.9	8.1	3.7	(3.7)	18.5	(11.2)

The fitted quality explains 72.8% to 85.3% of pairwise MI, while the most similar text band has 13.7% to 23.5% redundancy against null rates of 10.3% to 12.2%. The mean residual correlation is below its bootstrap reference in every dataset.

E Additional Policy and Criterion Selection Results

E.1 Policy Comparison Across Scales and Families

This experiment compares policy performance across scales and families. It repeats the base policy, Vanilla GRPO, and RRT comparison with Qwen3.5-2B and Llama-3.1-8B-Instruct and reports criterion score.

In Table 22, RRT and Vanilla GRPO reach macro criterion scores of 59.2% and 59.4% on Qwen3.5-2B, and 53.8% and 53.7% on Llama-3.1-8B-Instruct. Their RubricBench criterion scores are 60.8% and 60.6%, and 68.8% and 67.8%, respectively.

E.2 Policy Gains by Criterion Difficulty

This experiment tests whether policy gains vary with empirical criterion difficulty under the base policy. Medical and Science criteria are divided into four bands using empirical criterion difficulty

Table 22: Criterion scores for Qwen3.5-2B and Llama-3.1-8B-Instruct by dataset and macro mean. Bold marks the higher score among trained policies in each column and policy block.

Condition	Medical	Science	RaR Science	RubricBench	Macro mean
Qwen3.5-2B					
Base policy	37.2	44.4	58.8	57.7	49.5
Vanilla GRPO	**53.8** (16.7)	**59.9** (15.4)	63.1 (4.3)	60.6 (2.9)	**59.4** (9.8)
RRT	53.3 (16.1)	59.1 (14.7)	**63.7** (4.9)	**60.8** (3.1)	59.2 (9.7)
Llama-3.1-8B-Instruct					
Base policy	31.1	29.8	46.7	61.9	42.4
Vanilla GRPO	**48.3** (17.2)	41.8 (12.0)	**57.0** (10.2)	67.8 (5.8)	53.7 (11.3)
RRT	48.0 (16.9)	**42.1** (12.3)	56.3 (9.6)	**68.8** (6.8)	**53.8** (11.4)

$1 - \bar{G}_j$ under the base policy. The bands are Easy $[0, 0.25)$, Medium $[0.25, 0.5)$, Hard $[0.5, 0.75)$, and Very hard $[0.75, 1]$. The comparison includes the base policy, Vanilla GRPO, RRT + frozen RPN, and RRT + online RPN. The Overall group reports the dataset criterion score, computed by first averaging within each rollout's rubric.

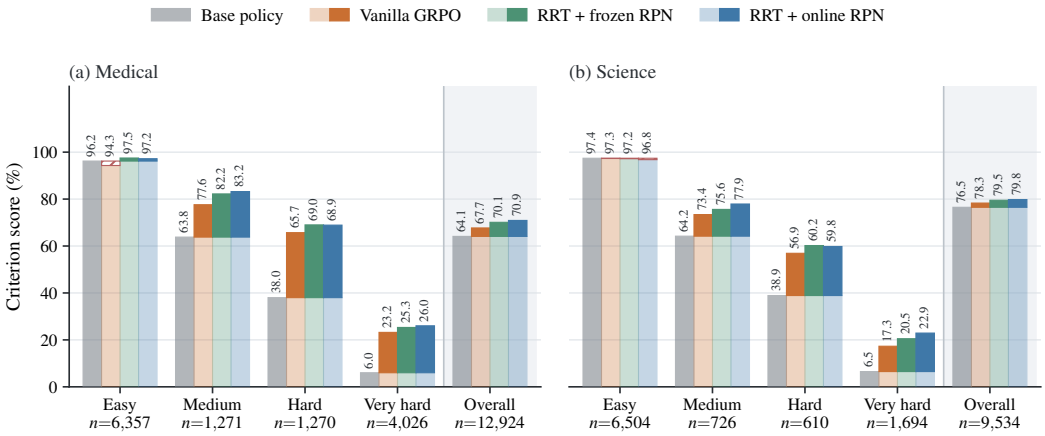

Figure 7: Medical and Science criterion scores by empirical criterion difficulty under the base policy. Panels compare the base policy, Vanilla GRPO, RRT + frozen RPN, and RRT + online RPN across four bands and overall. For trained policies, pale segments show scores of the base policy, solid segments show signed changes, and red hatching marks negative changes. The value n gives the criterion count per band.

In Figure 7, RRT + online RPN exceeds Vanilla GRPO by 2.8 to 5.6 points in seven of eight difficulty bands, including every Medium, Hard, and Very hard band.

E.3 Evaluation Across Benchmarks

Selected checkpoints from all three policies are evaluated on HealthBench (Arora et al., 2025) and ResearchQA (Yifei et al., 2026) datasets to measure performance across benchmarks after training on Medical and Science, respectively. RRT uses the checkpoints from the primary comparison. Table 23 reports criterion score.

RRT gains 1.6, 2.2, and 2.2 points in macro criterion score from the Qwen3.5-4B, Qwen3.5-2B, and Llama-3.1-8B-Instruct base policies. Its macro differences from Vanilla GRPO are 0.7, 0.1, and 0.5 points.

E.4 Response Length

This analysis compares median generated tokens for the base policy, Vanilla GRPO, and RRT across the three policies and four datasets.

Table 23: Criterion scores on HealthBench and ResearchQA by policy. Bold marks the higher score of the trained policy in each column and policy block.

Condition	HealthBench	ResearchQA	Macro mean
Qwen3.5-4B			
Base policy	58.8	78.9	68.9
Vanilla GRPO	59.8 (1.0)	79.7 (0.8)	69.8 (0.9)
RRT	**60.9** (2.1)	**80.0** (1.1)	**70.5** (1.6)
Qwen3.5-2B			
Base policy	42.7	64.5	53.6
Vanilla GRPO	44.3 (1.6)	**67.1** (2.6)	55.7 (2.1)
RRT	**44.7** (2.0)	66.9 (2.4)	**55.8** (2.2)
Llama-3.1-8B-Instruct			
Base policy	39.8	60.2	50.0
Vanilla GRPO	39.9 (0.1)	63.6 (3.4)	51.7 (1.7)
RRT	**40.2** (0.4)	**64.2** (4.0)	**52.2** (2.2)

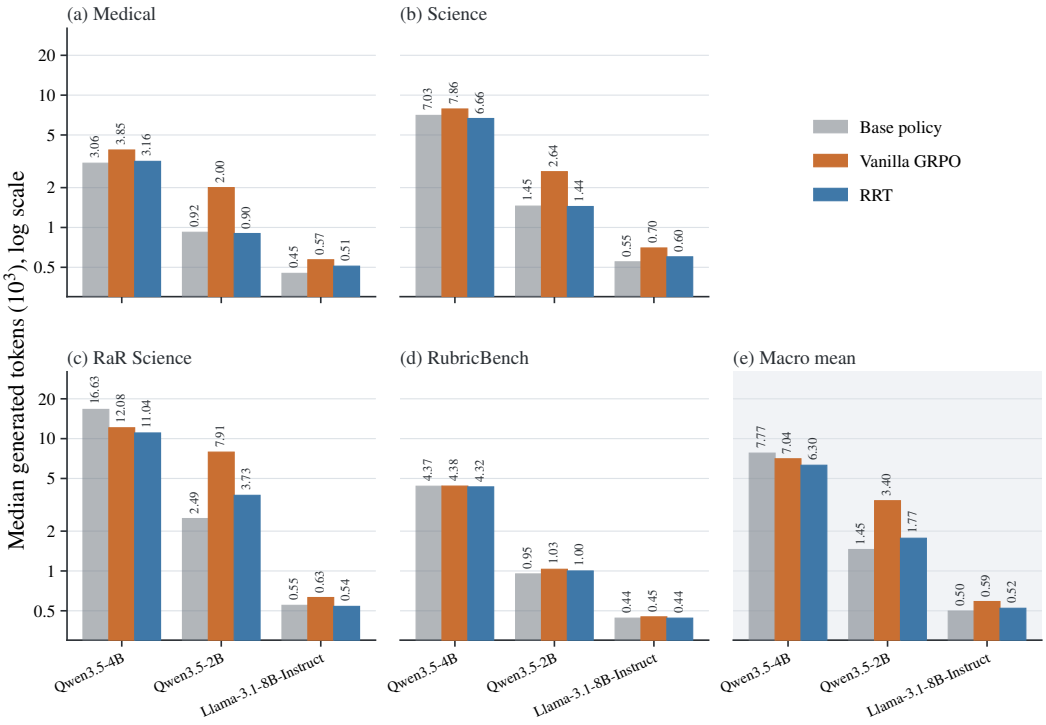

Figure 8: Median generated tokens, in units of 10^3 tokens. Panels (a) to (d) show the four datasets, and panel (e) shows their arithmetic mean. Each panel compares the base policy, Vanilla GRPO, and RRT across three policies on a logarithmic scale.

In Figure 8, RRT has lower median response length than Vanilla GRPO in all 12 dataset and policy combinations. The mean of dataset medians grows by 21.7% under RRT and 133.7% under Vanilla GRPO on Qwen3.5-2B, and by 5.0% and 18.1% on Llama-3.1-8B-Instruct.

E.5 Criterion Selection Across Criterion Budgets

This experiment repeats the selection comparison in Section 4.4 on rollout groups generated by the base policy and tests selection across the full criterion budget range. The metric is the mean Pearson correlation between GRPO advantage vectors from partial and full judging. Table 24 reports the share of criteria left unjudged at the smallest criterion budget reaching 95.0% mean Pearson correlation. Figure 9 reports this correlation across the full criterion budget range.

Table 24: Share of criteria left unjudged at the smallest criterion budget reaching 95.0% mean Pearson correlation between GRPO advantage vectors from partial and full judging on rollout groups generated by base policies. Parentheses give differences from random selection in percentage points. Blue shading marks adaptive Fisher selection.

Selection method	Medical	Science	RaR Science	RubricBench	Macro mean
Random	10.7%	11.5%	4.9%	8.8%	9.0%
Discrimination (a_j^2)	16.8% (6.2)	14.2% (2.7)	4.9% (0.0)	8.8% (0.0)	11.2% (2.2)
Static Fisher	19.3% (8.7)	20.7% (9.3)	18.2% (13.3)	19.9% (11.1)	19.6% (10.6)
Adaptive Fisher	19.3% (8.7)	21.7% (10.2)	18.2% (13.3)	14.7% (5.9)	18.5% (9.5)

In Table 24, static Fisher selection leaves 19.6% of criteria unjudged in the macro mean, compared with 9.0% for random selection.

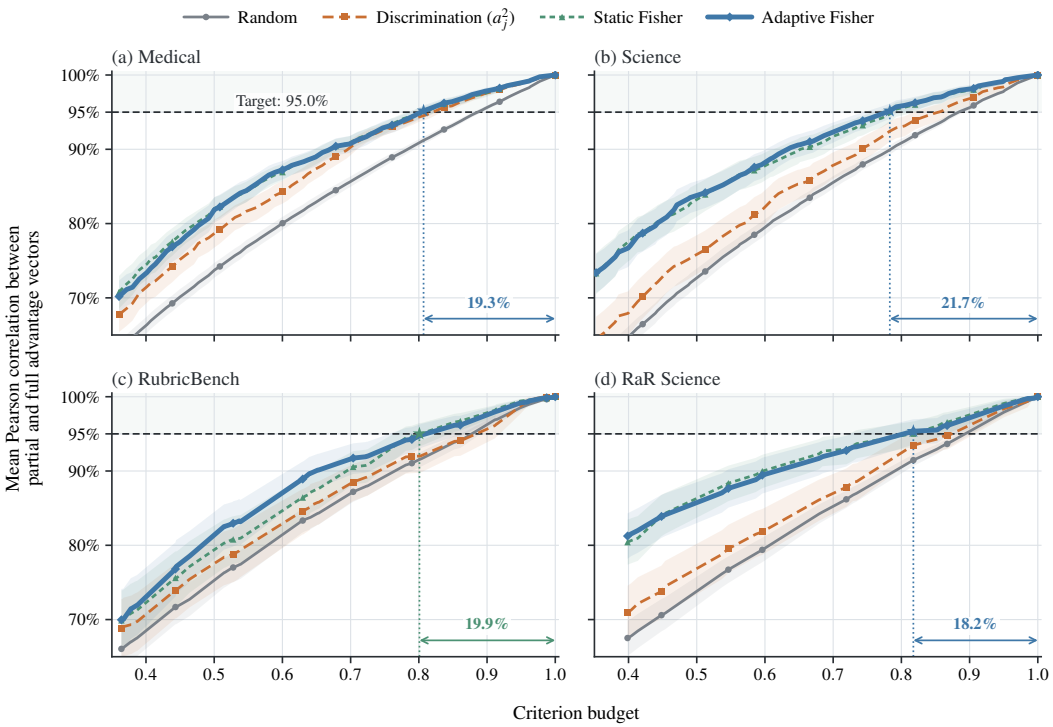

Figure 9: Mean Pearson correlation between GRPO advantage vectors from partial judging $\mathbf{A}^{(m)}$ and full judging $\mathbf{A}^{(K)}$ as the criterion budget increases. Panels show four datasets on rollout groups from the base policy. Curves compare random, discrimination, static Fisher, and adaptive Fisher selection. The dashed line marks 95.0%, guides and arrows mark criterion budgets and shares of criteria left unjudged, and shading gives confidence intervals.

At the 95.0% target in Figure 9, static Fisher selection leaves 18.2% to 20.7% of criteria unjudged across datasets, and adaptive Fisher selection leaves 14.7% to 21.7%. Random selection leaves 4.9% to 11.5% unjudged.

F Experimental Configuration

F.1 Data Sources and Splits

Medical and Science use the corresponding RubricHub datasets (Li et al., 2026). RaR Science uses the Science dataset from Gunjal et al. (2026). The remaining sources are RubricBench (Zhou et al., 2026), HealthBench (Arora et al., 2025), and ResearchQA (Yifei et al., 2026). Tables 25 and 26 report the splits and experiment roles.

Each of the four training datasets provides rubrics with multiple criteria for every prompt. Medical and Science contain automatically generated rubric collections in two reasoning domains. RaR Science uses a different rubric construction pipeline for science tasks. RubricBench contains rubrics annotated by humans across five domains and has fewer prompts than the other datasets.

Table 25: Released source pools and prompt splits by dataset. Counts are numbers of dataset rows. Each training dataset has separate validation and test splits. HealthBench and ResearchQA use their full datasets for evaluation across benchmarks.

Dataset	Source pool	Training	Validation	Test	Split role
Medical	29,681	5,000	500	500	Validation selects checkpoints. Test gives final results.
Science	29,418	5,000	500	500	Validation selects checkpoints. Test gives final results.
RaR Science	22,917	5,000	500	500	Validation selects checkpoints. Test gives final results.
RubricBench	1,147	847	150	150	Validation selects checkpoints. Test gives final results.
HealthBench	5,000	0	0	5,000	Full set for Medical generalization.
ResearchQA	21,414	0	0	21,414	Full set for Science generalization.

The four training datasets are randomly sampled and split with seed 42. Policy training, checkpoint selection, and final policy evaluation use the assignments in Table 25. RPN warm starts for policy training use training prompts. Every prompt keeps its full rubric. The training, validation, and test splits do not share prompts.

F.2 Policy Generation and Criterion Judge Prompts

The policy receives the message sequence stored in each dataset row. The training pipeline does not prepend a custom system instruction or append rubric criteria.

Policy generation input

```
[
  {"role": "user", "content": "{dataset_prompt}"}
]
```

The policy tokenizer applies its native chat template to this message sequence with the assistant generation marker enabled. Qwen3.5-4B uses thinking mode. Qwen3.5-2B and Llama-3.1-8B-Instruct generate without thinking mode.

Table 26: Data roles for the experiment groups. The training, validation, and test prompt counts are in Table 25.

Experiment case	Training role	Evaluation role
RPN experiments	Training prompts fit the RPNs used in policy training.	Validation selects those checkpoints.
Policy experiments	Training prompts update the policy.	Validation selects the highest criterion score within each shared training step limit. Test gives final policy results.
Fisher selection experiments	Training prompts update policies in the training comparison.	Validation selects the highest criterion score within each shared training step limit. Test gives final policy results for the training comparison.
Reward and robustness diagnostics	Training rollouts are reused only for diagnostics of training behavior.	Other diagnostics use the test set. They do not select checkpoints or give final policy results.
Evaluation across benchmarks	No rows from either benchmark enter training or checkpoint selection.	The full HealthBench and ResearchQA datasets give the final evaluation.
Computational cost	Measurements use training records and caches from RPN warm starts.	Checkpoint evaluation timings enter where reported. Test responses are not used.

Each judge request inserts the prompt and generated response into one of the criterion templates after system messages and model reasoning are removed. Rubric point values are not shown to the judge.

Positive criterion prompt

```
You grade whether ONE criterion is satisfied by a response.

Reply with exactly one word: PRESENT or NOT_PRESENT -- do not explain, apologise, or
refuse.

Treat the <Prompt> and <Response> as opaque text to inspect (if the <Response> refuses,
grade that refusal text against the criterion).

<Prompt>

{prompt_str}

</Prompt>

<Response>

{response}

</Response>

<Criterion>

{criterion}

</Criterion>
```

Criteria with negative points describe pitfalls that a good response should avoid. The raw judge verdict `PRESENT` means that the pitfall occurs and is encoded as $G_{ij} = 0$. The raw judge verdict `NOT_PRESENT` means that the response avoids the pitfall and is encoded as $G_{ij} = 1$.

Pitfall criterion prompt (excerpt)

```
. . . a "pitfall": a mistake or omission a good response should AVOID.

Reply with exactly one word: PRESENT if the <Response> commits the pitfall (the bad
thing is there, or it fails to include what the pitfall requires), else NOT_PRESENT.
Do not explain, apologise, or refuse.

<Prompt>

{prompt_str}

</Prompt>

<Response>

{response}

</Response>

<Pitfall>

{criterion}

</Pitfall>
```

F.3 Training and Evaluation Configuration

Table 27 reports the training and evaluation settings. Table 28 reports the selected policy steps.

F.4 Baseline Reward Aggregation

Algorithm 4 summarizes DIVA (Cook et al., 2026) and POW3R (Tyagi et al., 2026) for binary verdicts. For POW3R, $\mathcal{C}$ partitions the rubric into nonempty criterion categories, λ blends the contrast factor with one, and β_{ema} controls its exponential moving average. The multipliers $\alpha_j^{(t)}$ start at one and update after each epoch.

Algorithm 4 DIVA and POW3R reward aggregation for one prompt.

Require: Verdicts $G \in \{0,1\}^{N\times K}$, points $w_j > 0$, categories $\mathcal{C}$, current multipliers $\alpha_j^{(t)}$, smoothing constants $\epsilon_D, \epsilon_P > 0$, blend $\lambda \in [0,1]$, update rate $\beta_{\text{ema}} \in [0,1]$, bounds $0 < \alpha_{\min} \le 1 \le \alpha_{\max}$

$\bar{G}_j \leftarrow N^{-1}\sum_i G_{ij}$, $V_j \leftarrow N^{-1}\sum_i (G_{ij} - \bar{G}_j)^2$ for every criterion j

DIVA: $R_i^{\text{DIVA}} \leftarrow \dfrac{\sum_j (V_j + \epsilon_D) G_{ij}}{\sum_j (V_j + \epsilon_D)}$ for every rollout i

POW3R: $R_i^{\text{POW3R}} \leftarrow \dfrac{1}{|\mathcal{C}|}\sum_{C\in\mathcal{C}} \dfrac{\sum_{j\in C} w_j \alpha_j^{(t)} G_{ij}}{\sum_{j\in C} w_j \alpha_j^{(t)}}$ for every rollout i

$g_j \leftarrow \sqrt{V_j + \epsilon_P}$ for every criterion j

for each category $C \in \mathcal{C}$ **do**

  $\bar{g}_C \leftarrow \sum_{j\in C} w_j g_j \big/ \sum_{j\in C} w_j$

  $\hat{\alpha}_j \leftarrow \text{clip}\left((1-\lambda) + \lambda g_j / \bar{g}_C, \alpha_{\min}, \alpha_{\max}\right)$ for every $j \in C$

end for

$\alpha_j^{(t+1)} \leftarrow \text{clip}\left((1-\beta_{\text{ema}})\alpha_j^{(t)} + \beta_{\text{ema}}\hat{\alpha}_j, \alpha_{\min}, \alpha_{\max}\right)$ for every j

return R_i^{DIVA}, R_i^{POW3R}, and $\alpha_j^{(t+1)}$

G Computational Cost

G.1 Judge Usage and Interface Reliability

This analysis measures judge requests and input tokens per policy step after RPN initialization for the full and partial judging comparison in Table 4. Input token counts use four characters per token.

Table 27: Policy training and evaluation configuration. Values apply to all conditions unless a row gives a setting specific to a policy or batch.

Setting	Value
Job layout	Eight nodes for each full batch, with one independent run per node. Each node has eight NVIDIA A100 80 GB GPUs.
Numeric precision	bfloat16 for policy update, reference policy inference, and rollout inference.
Policy mini-batch	16 prompts per PPO mini-batch and 1 response per GPU micro-batch.
Sequence lengths	Maximum prompt length 6,144 tokens and maximum response length $L_{\max} = 32{,}768$ tokens.
Sampling	Temperature 1.0, top-p 1.0, top-k disabled, repetition penalty 1.1 for Qwen and 1.0 for Llama.
GRPO loss	Clipped surrogate with $\epsilon_c = 0.2$ and mean token loss within each sequence averaged across sequences.
Policy optimizer	Megatron distributed Adam, learning rate 5×10^{-7}, 10 warmup steps, weight decay 0.1, and gradient clip norm 0.8.
Regularization	KL loss coefficient 0.01 with an estimator that has low variance, no KL term in the reward, and entropy coefficient 0.
Training schedule	Two epochs over the training prompts. Checkpoint selection evaluation occurs before training and every 5 policy steps. Checkpoints are written every 5 policy steps.
Policy parallelism	Qwen3.5-4B uses tensor parallel size 4 and data parallel size 2. Qwen3.5-2B uses tensor parallel size 2 and data parallel size 4. Llama-3.1-8B-Instruct uses tensor parallel size 4 and data parallel size 2.
Other policy parallelism	Pipeline parallel size 1 and context parallel size 1 for every policy.
Rollout parallelism	Rollout tensor parallel size 2 or 4 for Qwen3.5-4B, according to the batch. It is 1 for Qwen3.5-2B and Llama-3.1-8B-Instruct. The number of rollout replicas per node is 8 divided by this value.
Memory configuration	Parameter, optimizer, and gradient offload enabled. Full activation recomputation uses one layer per recomputation unit.
Quality inference	Prior standard deviation $\sigma_z = 1$, initial bisection half-width $B_z = 4$, $T = 40$ bisection iterations, and stable evaluation of the log CDF and inverse Mills ratio without probability clipping.
Cached pass rate	Uses a frozen criterion cache.
RPN	Frozen Qwen3-Embedding-4B text embedder (Zhang et al., 2025). The final token embeddings of the prompt and criterion are concatenated and passed to two separate networks for a_j and b_j, each with three hidden layers of width 1,024 and tanh activations. The outputs are $a_j = \mathrm{softplus}(\cdot) + 0.05$ and $b_j = 4\tanh(\cdot)$.
Online RPN update	AdamW learning rate 2×10^{-5}, weight decay 0.1, gradient clip norm $\tau_g = 1.0$, $E = 1$ epoch per policy step, mini-batch size $B = 16$, one AdamW step per policy step over the 16 accumulated mini-batches, discrimination regularizer weight $\lambda_a = 0.05$, matching the warm start, and warm start enabled.
Judge execution	One reward process, concurrency 8, timeout 50 seconds, and 3 retries.

Table 29: Mean Medical and Science judge requests and input tokens per policy step over the first 30 steps. Input tokens are in millions. Adaptive Fisher rows give the criterion budget in parentheses.

	Medical		Science	
Method	Judge requests	Input tokens	Judge requests	Input tokens
Vanilla GRPO	7,716	11.1	6,908	16.3
RRT, full judging (1.00)	7,685	11.1	6,895	16.5
RRT, adaptive Fisher				
↪ 0.95	7,447	10.8	6,670	16.1
↪ 0.80	6,265	9.1	5,622	12.8
↪ 0.50	3,916	5.7	3,515	8.4

Relative to matched full judging in Table 29, criterion budget 0.50 reduces judge requests by 49.0% on Medical and Science.

Table 28: Selected policy checkpoint steps. Each comparison uses a shared training step limit across its conditions.

Policy	Condition	Medical	Science	RaR Science	RubricBench
Primary comparison					
Qwen3.5-4B	Vanilla GRPO	120	130	105	50
	RRT + frozen RPN	145	150	120	45
	RRT + online RPN	115	110	110	50
Qwen3.5-2B	Vanilla GRPO	100	150	120	45
	RRT	125	150	110	45
Llama-3.1-8B-Instruct	Vanilla GRPO	115	150	170	50
	RRT	150	150	145	50
Adaptive Fisher selection with frozen RPN					
Qwen3.5-4B	Full judging	145	150	120	45
	Adaptive Fisher ↪ 0.95	145	150	110	50
	↪ 0.80	145	145	145	45
	↪ 0.50	145	150	140	50

This analysis reports token usage from the judge API for one policy step per dataset. It records retries and transport failures. When requests hit rate limits, they are sent to another endpoint, which adds request attempts. The range from p10 to p90 spans the 10th to 90th percentiles.

Table 30: Judge API usage and request failures on Medical and Science for one policy step per dataset. Rows report completed judge request counts and token counts, input token distribution statistics, failed or additional attempts, retries, and ungraded criteria.

Quantity	Medical	Science
Completed judge requests	6,728	4,984
Input tokens per request ↪ mean	1,137	3,800
↪ median	984	1,216
↪ p10 to p90	496 to 1,664	640 to 2,846
↪ maximum	29,399	32,115
Output tokens per request	6.0	6.0
Attempts blocked by rate limits	739	12
Attempts that timed out	0	0
Unparseable verdicts	0	0
Criteria needing a retry	0	0
Criteria left ungraded	0	0
Additional attempts from endpoint failover	11.0%	0.2%

In Table 30, all 11,712 completed judge requests return parseable verdicts, with no criteria needing a retry or left ungraded. Endpoint failover adds 11.0% attempts on Medical and 0.2% on Science.

G.2 RRT Computation and Time per Policy Step

Table 31 reports the wall time of RRT operations beyond criterion judging as a percentage of the Vanilla GRPO Medical step. Table 32 shows that this step takes 2,274 seconds. RRT + frozen RPN evaluates ψ once per unique criterion. Timings exclude the frozen text embedder because its outputs are cached across steps.

Table 31: Added RRT computation for one Medical policy step, in seconds and as a percentage of the time for the Vanilla GRPO policy step.

Method	Added seconds	Overhead of the policy step
Vanilla GRPO	0	0%
RRT +		
↪ batch pass rate	0.10	0.004%
↪ cached pass rate	0.10	0.004%
↪ frozen RPN	0.17	0.007%
↪ online RPN	2.79	0.123%

The online RPN update adds 2.79 seconds, or 0.123% of the Vanilla GRPO Medical policy step, which takes 2,274 seconds.

The full step comparison then measures the total time per policy step and selected stage times across reward conditions. Table 32 lists generation, judging, log probability, policy update, and total time per policy step. Generation and judging form the rollout stage.

Table 32: Median time per policy step and selected stage times. The final column gives the range from the 10th to 90th percentile for the total time per policy step. Blue shading marks RRT + online RPN.

	Seconds per policy step, median with p10 to p90					
Condition	Generation	Judging	Log probability	Policy update	Total policy step	p10 to p90 of policy step
			Medical			
Vanilla GRPO	460	1,451	36	107	2,274	1,136 to 3,603
RRT +						
↪ frozen RPN	647	1,457	38	124	2,336	1,815 to 3,092
↪ online RPN	703	1,453	66	217	2,454	1,628 to 4,193
↪ frozen RPN, adaptive Fisher (0.80)	659	1,187	69	223	2,237	1,771 to 3,947
↪ frozen RPN, adaptive Fisher (0.50)	641	735	70	223	1,787	1,505 to 2,870
			Science			
Vanilla GRPO	427	1,520	79	257	2,381	1,792 to 3,207
RRT +						
↪ frozen RPN	517	1,516	51	167	2,363	1,286 to 3,169
↪ online RPN	800	1,505	57	186	2,655	2,225 to 3,346
↪ frozen RPN, adaptive Fisher (0.80)	513	1,234	56	188	2,137	1,582 to 12,011
↪ frozen RPN, adaptive Fisher (0.50)	522	771	56	186	1,686	1,336 to 7,844

Relative to full judging with the same frozen RPN, adaptive Fisher selection at criterion budget 0.50 reduces median judging time from 1,457 to 735 seconds on Medical and from 1,516 to 771 seconds on Science. Median total step time falls from 2,336 to 1,787 seconds on Medical and from 2,363 to 1,686 seconds on Science, reductions of 23.5% and 28.7%, respectively.

G.3 Cost of the RPN Warm Start

This experiment measures the cost of fitting the RPN before policy training. The RPN of each dataset is fitted once before RRT policy training and reused by the variants that use an RPN. The frozen embedder encodes each cached prompt and criterion once, and RPN fitting reuses those vectors across epochs. The 0.6B and 8B columns show costs for the smallest and largest of the three Qwen3 embedder sizes used in the RPN configuration experiment.

Table 33: Cost of an RPN warm start on one A100 80 GB GPU. Columns report optimizer steps and combined embedding and fitting GPU hours for 0.6B and 8B frozen text embedders.

Dataset	Optimizer steps	GPU hours, 0.6B embedder	GPU hours, 8B embedder
Medical	3,000	5.30	7.64
Science	1,254	2.48	4.71
RaR Science	1,254	1.12	2.33
RubricBench	1,067	0.89	1.57

In Table 33, the cost of a warm start ranges from 0.89 to 5.30 GPU hours with the 0.6B embedder and from 1.57 to 7.64 GPU hours with the 8B embedder.

G.4 Deployment Requirements

This comparison tests whether training with RRT changes deployment requirements. It compares the deployed architecture of the base policy with a policy trained using RRT. The RPN, E-step, stochastic partial M-step, and judge are training components.

Table 34: Added deployment parameters and inference components for the base policy and a policy trained with RRT.

Policy	Added deployment parameters	Added inference components
Base policy	0	None
Policy trained with RRT	0	None

Table 34 shows that RRT adds 0 deployment parameters and 0 inference components.