DIRICHLET-GRPO: POSTERIOR-CALIBRATED ADVANTAGE TRANSPORT FOR MULTIMODAL RLVR

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal reinforcement learning with verifiable rewards converts small groups of discrete outcomes into policy-update weights, introducing error through uncertain outcome frequencies and approximate representatives of the resulting probability regions. To address both sources, we propose Dirichlet-GRPO, which combines Jeffreys–Dirichlet posterior mean probabilities with exact Gaussian interval means so that probabilities determine each ordered region's area and its mean determines the update weight. We establish the optimal representative for a fixed partition, characterize the bias–variance trade-off of probability smoothing, and bound how probability error affects the initial policy update, while exact controlled experiments isolate the two corrections and identify conditions where smoothing is unfavorable. Across the 2B, 4B, and 8B scales of Qwen3-VL, our Dirichlet-GRPO improves upon GRPO by 2.77, 1.37, and 0.76 points on average. On 4B scale, it surpasses both GRPO and G^{2}RPO across all 11 evaluated benchmarks, notably boosting WeMath, MMMU, and MMStar. To disentangle these gains, exact Bernoulli ablations isolate the impact of posterior calibration and Gaussian interval integration, showing up to a 37.0% reduction in initial gradient MSE over tied ranks.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) improves multimodal reasoning through verifiable signals, with visual perturbations and self-verification providing richer feedback (Wang et al., 2025; Qiu et al., 2025; Sathish et al., 2026). How that feedback changes the policy depends on the *advantage*, the weight assigned to each sampled response. Group Relative Policy Optimization (GRPO) estimates these weights from responses to the same prompt without learning a value critic (Shao et al., 2024), calculating relative advantages against the mean reward of the peer outputs. Small groups provide uncertain evidence, since no right rollout does not imply that success is impossible, while a lone success gives an imprecise frequency estimate. Equal rewards nevertheless produce zero-centered advantages, and rare successes can receive large relative weights (Liu et al., 2025; He et al., 2026). Recent methods address these effects by pooling evidence across prompts or training history (Zeng et al., 2025; Han et al., 2026a; Kim et al., 2026), or by adapting rollout allocation and difficulty-dependent weighting (Zhang et al., 2026; Gan et al., 2026). Instead, we separate uncertainty about outcome probabilities from the rule that converts those probabilities into advantages.

Recently, Gaussian GRPO (G^{2}RPO) (Hu et al., 2026) makes this distinction concrete by assigning Gaussian quantiles to ranked feedback and averaging within ties. For a fixed rollout group, each observed outcome occupies a Gaussian interval whose area equals its empirical frequency. However, estimated outcome probabilities determine the interval boundaries, whereas averaging finitely many quantiles approximates the mean within each interval, so more accurate integration cannot account for unseen outcomes, and better probability estimates alone cannot eliminate integration error.

We propose Dirichlet-GRPO to address these two sources of error by combining Jeffreys–Dirichlet posterior mean probabilities with exact Gaussian interval means. Specifically, adding half a count to each valid verifier outcome (Jeffreys, 1946) reserves probability for unobserved outcomes, while exact integration converts the resulting intervals into outcome-specific advantages. As shown in Figure 1, this yields negative advantages for sampled responses in all-failure binary groups without introducing fictitious successes, while mixed groups illustrate the distinct effects of probability

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

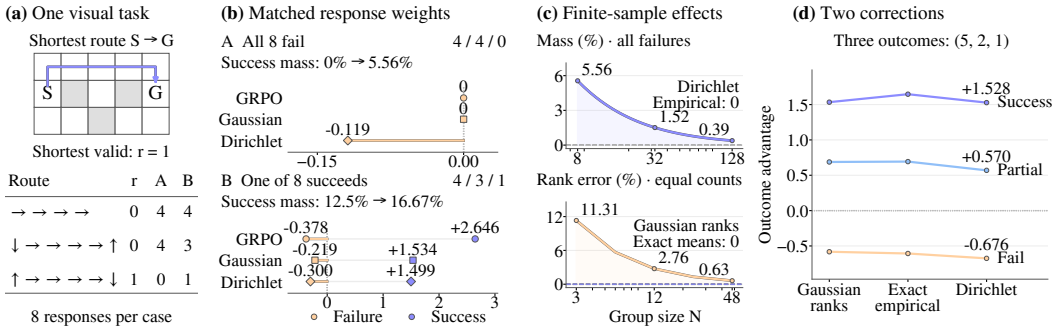


Figure 1: Two corrections from one task. (a) Route counts in two eight-rollout groups, A and B group. (b) Probabilities and advantages under different methods. (c) Reserved mass and rank approximation error relative to exact interval means. (d) Integration and calibration corrections for counts $(5, 2, 1)$.

calibration and exact integration. The Gaussian provides an ordered reference without imposing a Gaussian distribution on rewards, and nonzero advantages alone do not guarantee a useful update.

To understand when and why the method helps, we establish the optimal advantage for fixed intervals, examine the bias–variance trade-off introduced by probability smoothing, and analyze how probability estimation errors propagate to the initial policy update, complementing existing finite-group gradient analyses (Zhou et al., 2026). Our controlled analyses therefore evaluate both probability and update errors, including cases where smoothing is unfavorable, since more accurate probabilities do not necessarily produce more accurate updates. Moreover, outcome counts alone cannot distinguish grounded reasoning from guessing when both receive the same verifier label, highlighting the need for complementary credit assignment and verification signals (Guo et al., 2026a; Sathish et al., 2026; Yao et al., 2026). Comparisons of Qwen3-VL (Bai et al., 2025) show average gains of 2.77, 1.37, and 0.76 points over GRPO at the 2B, 4B, and 8B scales, respectively, as reported in Table 1. The 4B model shows the most consistent improvements, outperforming both GRPO and G^{2}RPO on all benchmarks. Specifically, compared with GRPO, it improves WeMath and MMMU by 1.78% each, MMStar by 1.60%, and MathVision by 1.58%, covering math and general multimodal understanding.

Overall, our contributions are threefold. First, we separate probability estimation error from finite-rank approximation error and introduce Dirichlet-GRPO to address them through Jeffreys–Dirichlet posterior probabilities and exact Gaussian interval means. Second, we establish the optimality of these means for fixed intervals and characterize how probability smoothing affects estimation and the initial policy update, clarifying both its benefits and limitations. Third, comprehensive experiments show that our method outperforms GRPO and G^{2}RPO approaches at the 2B, 4B, and 8B scales, with particularly clear improvements over two baseline methods on the Qwen3-VL 4B model.

2 Motivating

Our goal is to turn a small group of verifier outcomes into useful response weights. Following G^{2}RPO, we place better outcomes farther to the right on a standard Gaussian scale, with each outcome occupying an interval whose probability equals its frequency. The Gaussian is a chosen reference, not an assumption about rewards. We first ask which single weight should represent an interval, then how to estimate its probability, and finally how estimation errors affect the update.

2.1 Choosing One Weight for Each Outcome

Fix a prompt and the policy used to sample N independent rollouts. The verifier has K valid outcomes, ordered as $v_1 < \cdots < v_K$, including outcomes absent from this group. Write n_k for outcome k's count, p_k for its unknown true probability, and $\widehat{p}_k = n_k/N$ for its observed frequency. Suppose the interval probabilities $\mathbf{w}$ are known. Every response with the same outcome must receive one weight, even though its interval contains many Gaussian values. We choose that weight to minimize average squared distance to those values. Let Z be a standard Gaussian variable, let V label the interval containing Z, and let $T_k(\mathbf{w})$ be the mean within interval k. The following identity separates the error that any single weight must incur from the extra error caused by choosing the wrong weight, and appendices A.1 and B provide the formal setup and detailed proofs.

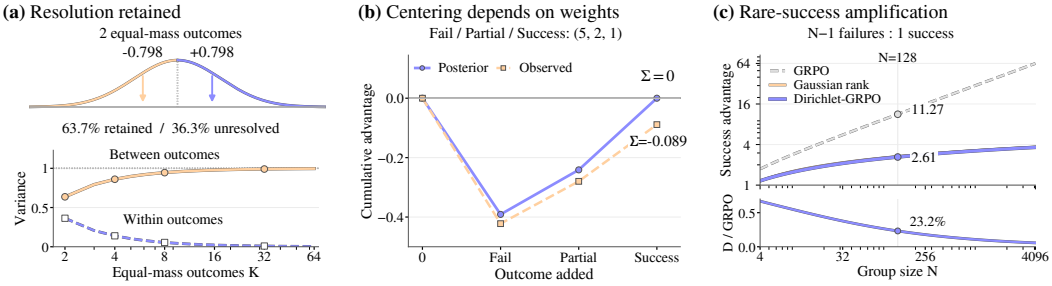


Figure 2: Analytic properties of interval advantages. (a) Retained and unresolved Gaussian variance. (b) Posterior versus observed centering. (c) Rare-success weights and Dirichlet/GRPO ratio, with one success among N rollouts. Where smaller weights do not imply lower training-gradient variance.

Theorem 1 (Resolution and representation error)**.** *Fix positive masses* $\mathbf{w}$ *and their ordered Gaussian partition. If* $a_V = a_k$ *when* $V = v_k$*, then every one-value-per-outcome representation satisfies*

$$\underbrace{\mathbb{E}[(Z - a_V)^2]}_{\text{total error}} = \underbrace{\mathbb{E}[\mathrm{Var}(Z \mid V)]}_{\text{resolution loss } D(\mathbf{w})} + \underbrace{\sum w_k \big(a_k - T_k(\mathbf{w})\big)^2}_{\text{representative error}} \tag{1}$$

Under the Gaussian reference, the unique minimizers are the order-preserving weights $a_k = T_k(\mathbf{w})$*, with zero weighted mean and variance below one, or zero for a single outcome.*

Where the first term is the variation left inside each interval, which one weight cannot preserve, the second vanishes when we use the interval mean. Thus, exact means remove avoidable approximation error, and rescaling them to unit variance would change this optimum. Figure 2(a) shows that splitting a fixed interval can preserve more Gaussian variation, without establishing that adding verifier labels improves learning. We next address unknown probabilities that determine these intervals.

2.2 Estimating Probabilities from a Small Group

Empirical frequencies collapse unobserved outcomes to empty intervals, yet small samples cannot preclude their possibility. To preserve this uncertainty, we use a categorical Jeffreys prior (Jeffreys, 1946), which adds a half-count to every declared outcome, yielding the posterior mean:

$$\widetilde{p}_k = \frac{n_k + 1/2}{N + K/2}\,. \tag{2}$$

Like other shrinkage estimators (Zeng et al., 2025; Zhang et al., 2026; Kim et al., 2026), it reduces sampling fluctuations by moving toward a reference, here the uniform vector $\mathbf{u} = (1/K, \ldots, 1/K)$ with weight $\lambda_N = K/(2N + K)$. To determine when this helps, we compare squared probability errors averaged over repeated groups from the same population.

Proposition 1 (When probability smoothing reduces error)**.** *For fixed* K *and independent samples with outcome probabilities* $\mathbf{p}$*, the count vector is multinomial and the posterior mean satisfies*

$$\mathbb{E}\|\widetilde{\mathbf{p}} - \mathbf{p}\|_2^2 = (1 - \lambda_N)^2 \frac{1 - \|\mathbf{p}\|_2^2}{N} + \lambda_N^2 \|\mathbf{u} - \mathbf{p}\|_2^2. \tag{3}$$

The empirical estimator has error $(1 - \|\mathbf{p}\|_2^2)/N$*, so smoothing improves on it exactly when the variance reduction exceeds the added squared bias.*

Equation 3 explains why smoothing can reduce average error at small sample sizes but hurt nearly deterministic populations, where observed frequencies already have little variance. Figure 3(a,b) displays both regimes, and Appendix B.1 derives their boundary. The half-count is therefore a deliberate bias–variance trade-off, not a guarantee of more accurate probabilities for every prompt.

2.3 From Probability Error to Update Error

Shifting interval boundaries alters their means and response weights. To isolate this error, we compare identical responses under estimated versus true probabilities. At the sampling policy without clipping,

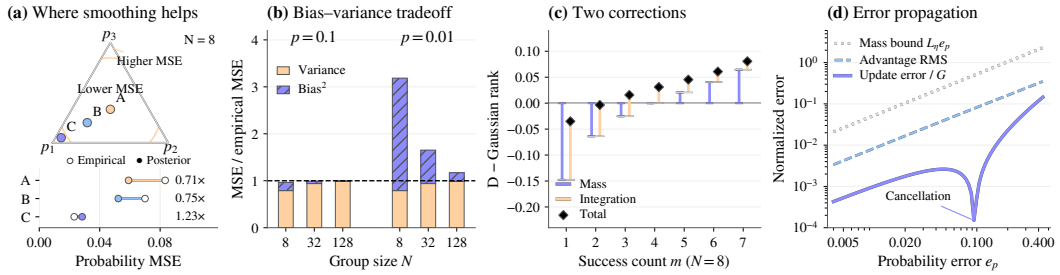


Figure 3: Comparisons of (a) empirical (hollow) and posterior (filled) MSE, (b) variance and squared bias relative to empirical risk, (c) additive mass/integration corrections to Gaussian ranks, and (d) update error and its bounds versus probability error. In (b), totals below one indicate improvement.

the gradient is $g(A) = N^{-1}\sum_i A_i s_i$, where $s_i = T_i^{-1}\nabla \log \pi_\theta(y_i \mid x)$ is the length-normalized score. With $G = (N^{-1}\sum_i \|s_i\|_2^2)^{1/2}$ bounding score magnitudes, and the next bound holds these directions fixed and controls only weight changes.

Proposition 2 (Calibration error and the initial update)**.** *Fix the sampled group and suppose every component of* $\mathbf{p}$ *and* $\widetilde{\mathbf{p}}$ *is at least* $\eta > 0$*, with* $\eta < 1/2$*. Let* $k(i)$ *index response* i*'s outcome,* $A_i^\star = T_{k(i)}(\mathbf{p})$*, and* $A_i^{\mathrm{D}} = T_{k(i)}(\widetilde{\mathbf{p}})$*. Treating these advantages as constants during differentiation, the initial gradients satisfy*

$$\|g(A^{\mathrm{D}}) - g(A^\star)\|_2 \le G L_\eta \|\widetilde{\mathbf{p}} - \mathbf{p}\|_1, \tag{4}$$

where L_η *bounds how much interval means change per unit probability error.*

This bound connects probability estimation to the initial update, complementing finite-group gradient analyses (Zhou et al., 2026). It weakens near-zero probabilities due to sensitive tail means and loosens when opposing response gradients cancel, as in Figure 3(d). Thus, lower probability error implies neither lower update error, smaller gradient variance, nor better training performance (Appendix B.7).

3 Dirichlet-GRPO

Following §2 analysis, for each rollout group, Dirichlet-GRPO converts outcome counts into calibrated probabilities and assigns each response the exact Gaussian interval mean of its outcome.

3.1 Posterior Probability Calibration

We add a half-count to each of the K valid outcomes and normalize via Eq. 2, keeping unobserved intervals non-empty under a declared support while fixing total prior strength, which decays with group size. As evidence concentrates the posterior toward the empirical proportions shown in Figure 4, we use this deterministic posterior mean solely, without sampling probability vectors.

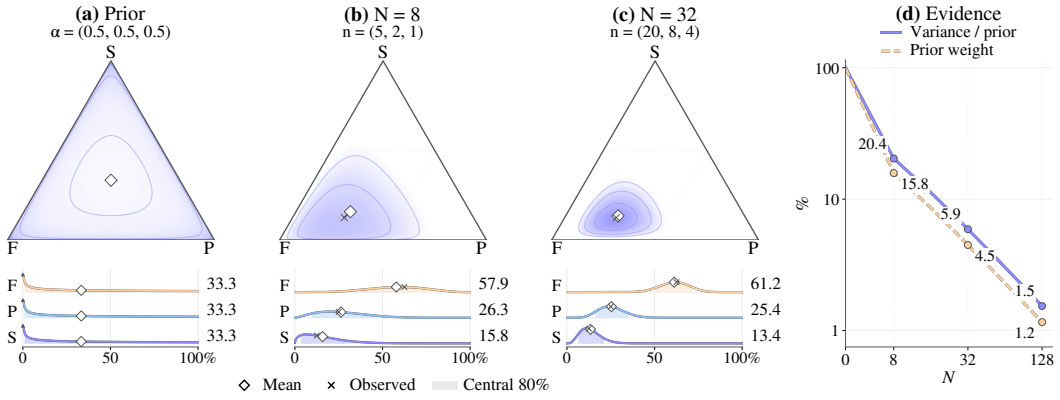


Figure 4: Posterior concentration at fixed observed proportions. F/P/S denote failure/partial success/success. (a–c) Joint and Beta marginal densities, where diamonds show means and crosses show observations. (d) Posterior/prior variance and prior weight. These describe conditional uncertainty.

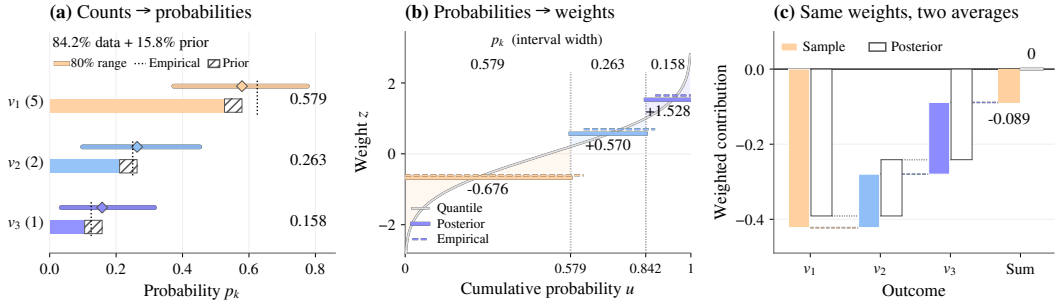


Figure 5: Analytic example with counts $(5, 2, 1)$. (a) Marginal 80% ranges and empirical/prior mass contributions with hatched prior mass. (b) Exact Gaussian interval means (dashed: empirical-mass). (c) Identical advantages weighted by observed frequencies (filled) or posterior probabilities (outlined).

Ordering outcomes from worst to best along the Gaussian axis, each boundary $q_k = \Phi^{-1}(\sum_{j \le k} \tilde{p}_j)$ captures the cumulative mass to its left (with $q_0 = -\infty, q_K = +\infty$), where Φ and φ denote the standard Gaussian CDF and density. Per Theorem 1, averaging within each resulting interval yields:

$$A_k^{\mathrm{D}} = T_k(\tilde{\mathbf{p}}) = \frac{\varphi(q_{k-1}) - \varphi(q_k)}{\tilde{p}_k}, \qquad A_i = A_{k(i)}^{\mathrm{D}}. \tag{5}$$

The numerator is the integral of $z\varphi(z)$ over the interval, and dividing by its probability gives the conditional mean. In the running $(5, 2, 1)$ example, the three advantages are $(-0.676, 0.570, 1.528)$, as shown in Figure 5. Appendix C gives stable evaluation with zero density at infinite endpoints.

We use the calibrated advantages via the Dirichlet posterior mean directly as fixed weights in the GRPO objective, without additional centering or rescaling, while keeping the remaining policy-update procedure unchanged. Compared with G^{2}RPO, Dirichlet-GRPO uses posterior probabilities to adjust the Gaussian interval boundaries and computes each interval mean exactly instead of averaging finitely many quantiles. Importantly, our strategy addresses probability estimation and numerical approximation, respectively, with the exact decomposition and proofs given in Appendix B.4.

To separate our changes from G^{2}RPO, let A_k^{G} be its average of finitely many Gaussian quantiles for outcome k. The intermediate rule $T_k(\hat{\mathbf{p}})$ uses the same observed frequencies but computes an exact mean. For each observed outcome,

$$\underbrace{A_k^{\mathrm{D}} - A_k^{\mathrm{G}}}_{\text{total change}} = \underbrace{T_k(\tilde{\mathbf{p}}) - T_k(\hat{\mathbf{p}})}_{\text{probability correction}} + \underbrace{T_k(\hat{\mathbf{p}}) - A_k^{\mathrm{G}}}_{\text{integration correction}}. \tag{6}$$

The first term changes the interval boundaries, whereas the second replaces finite quantile averaging with exact integration. Zero-count outcomes have no empirical interval mean and are excluded from this comparison. Both corrections vanish as N grows for fixed positive population probabilities.

4 Experiment

4.1 Experimental Setup

We compare GRPO, G^{2}RPO, and our proposed Dirichlet-GRPO across the 2B, 4B, and 8B variants of Qwen3-VL (Bai et al., 2025). We evaluate on ten benchmarks covering mathematical reasoning, general multimodal understanding, and hallucination. WeMath, MathVista, MathVerse, MathVision, and MMK12 assess mathematical reasoning, while MME, AI2D, MMMU, and MMStar assess broader visual understanding and reasoning, and HallusionBench. Table 1 reports overall performance.

Training is conducted on the `ViRL39K`[1] multimodal reasoning dataset using the AdamW optimizer with a cosine learning rate scheduler decaying to 10% of its peak value. We employ an initial learning rate of 1×10^{-6} for policy parameters and train with a global prompt batch size of 128 for 1 epoch (around 150 optimization steps). For each prompt, we sample $G = 8$ rollouts at a temperature of 1.0, capping the maximum response length at 2048 tokens. Policy updates adopt a PPO clipping radius of $\epsilon = 0.2$ and a KL divergence penalty coefficient of $\beta = 0.04$ against the reference behavior policy.

[1] https://huggingface.co/datasets/TIGER-Lab/ViRL39K

Table 1: Benchmark performance across Qwen3-VL model scales. We report scores (%) on 11 multimodal benchmarks covering math, general multimodal understanding, and hallucination alongside gains (Δ) of Dirichlet-GRPO over GRPO and G^{2}RPO. Best results in each benchmark are in **bold**.

Method (-2B)	We Math	MME	Hallu aACC	Hallu fACC	AI2D	Math Vista	Math Verse	MM MU	MM Star	Math Vision	MM K12	Mean
GRPO	49.94	38.75	64.77	39.42	71.99	59.90	39.06	40.33	**58.27**	**23.39**	66.49	50.21
G^{2}RPO	46.15	36.10	65.40	38.86	71.57	56.90	39.26	40.89	56.73	22.24	68.17	49.30
Dirichlet-GRPO	**51.21**	**55.98**	**65.62**	**40.17**	**72.09**	**61.50**	**39.87**	**42.22**	57.73	23.29	**71.10**	**52.80**
Δ vs. GRPO	+1.27	+17.23	+0.85	+0.75	+0.10	+1.60	+0.81	+1.89	-0.54	-0.10	+4.61	+2.59
Δ vs. G^{2}RPO	+5.06	+19.88	+0.22	+1.31	+0.52	+4.60	+0.61	+1.33	+1.00	+1.05	+2.93	+3.50

Method (-4B)	We Math	MME	Hallu aACC	Hallu fACC	AI2D	Math Vista	Math Verse	MM MU	MM Star	Math Vision	MM K12	Mean
GRPO	74.43	86.77	65.93	42.19	80.38	72.80	59.80	55.33	68.33	44.31	66.49	65.16
G^{2}RPO	75.23	86.06	64.98	41.91	80.60	71.60	59.37	54.89	66.67	44.51	68.17	64.91
Dirichlet-GRPO	**76.21**	**87.53**	**66.14**	**42.20**	**81.80**	**73.80**	**60.08**	**57.11**	**69.93**	**45.89**	**71.10**	**66.53**
Δ vs. GRPO	+1.78	+0.76	+0.21	+0.01	+1.42	+1.00	+0.28	+1.78	+1.60	+1.58	+4.61	+1.37
Δ vs. G^{2}RPO	+0.98	+1.47	+1.16	+0.29	+1.20	+2.20	+0.71	+2.22	+3.26	+1.38	+2.93	+1.62

Method (-8B)	We Math	MME	Hallu aACC	Hallu fACC	AI2D	Math Vista	Math Verse	MM MU	MM Star	Math Vision	MM K12	Mean
GRPO	78.33	**89.47**	**75.92**	**54.38**	82.42	77.40	62.84	**63.56**	**71.07**	47.49	66.49	69.94
G^{2}RPO	**80.57**	**89.47**	75.39	53.64	81.38	78.10	61.80	63.33	71.00	**48.82**	68.17	70.15
Dirichlet-GRPO	80.00	**89.47**	75.24	53.91	**83.00**	**78.30**	**63.12**	63.33	70.80	48.26	**71.10**	**70.59**
Δ vs. GRPO	+1.67	+0.00	-0.68	-0.47	+0.58	+0.90	+0.28	-0.23	-0.27	+0.77	+4.61	+0.65
Δ vs. G^{2}RPO	-0.57	+0.00	-0.15	+0.27	+1.62	+0.20	+1.32	+0.00	-0.20	-0.56	+2.93	+0.44

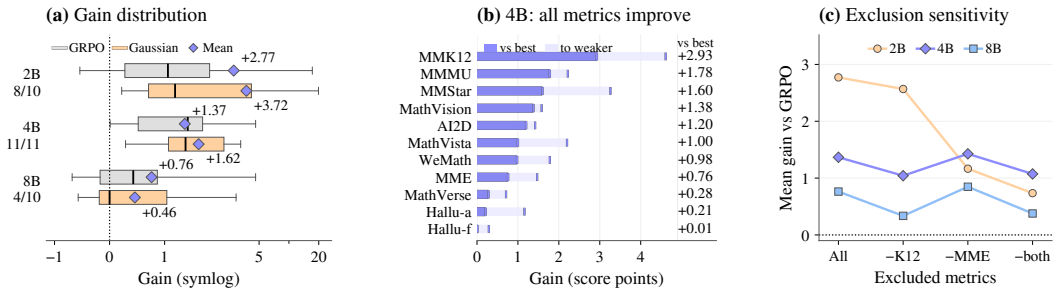


Figure 6: (a) Gain distributions across benchmarks (boxes: quartiles, bars: medians, diamonds: means, left ratios denote win rates over baselines). (b) Ranked 4B improvements against the best baseline and weaker baseline. (c) Robustness of mean gains over GRPO under metric exclusions. Error bars reflect cross-benchmark variance, not seed variance. -a/f: aACC/fACC; K12: MMK12.

4.2 Main Results

Table 1 and Figure 6(a)(b) highlight the clear overall advantage of Dirichlet-GRPO across scales, with the most decisive gains emerging at the 4B scale. Here, our method achieves a clean sweep—outperforming both GRPO and G^{2}RPO across all 11 evaluated benchmarks, lifting the overall average to 66.53 (compared to 65.16 for GRPO and 64.91 for G^{2}RPO). Notably, these improvements span diverse domains, delivering substantial boosts on challenging math reasoning (*e.g.*, WeMath +1.78) and general visual understanding (*e.g.*, MMMU +1.78), while consistently edging ahead even on stringent fine-grained hallucination checks (HalluBench-fACC +1.31). At the 2B scale, Dirichlet-GRPO delivers substantial gains, raising the overall average by +2.77 points over GRPO. At the largest 8B scale, our method continues to maintain a positive average margin of +0.76 points, holding the top mean score across all competing methods (72.26). While performance across individual 8B tasks displays natural domain variance, the aggregate gains across scales confirm that posterior calibration reliably enhances the policy update. More details are in Appendix D.

To confirm these gains are not driven by specific influential entries, such as the cross-scale identical MMK12 scores or the sharp MME surge at the smallest scale, we conduct an exclusion sensitivity analysis in Figure 6 (c) and Table 4. Dirichlet-GRPO maintains substantial margins across all model sizes: excluding MMK12 preserves gains of up to $+3.808$ against G^{2}RPO and $+2.568$ against GRPO, while leaving out MME yields peak advantages of $+1.633$ (4B vs. G^{2}RPO) and $+0.848$ (8B vs. GRPO). Remarkably, even under the simultaneous exclusion of both outliers, our method retains unbroken positive margins over GRPO ($+0.735$, $+1.073$, and $+0.378$ from smallest to largest scale), with gains over G^{2}RPO reaching up to $+1.799$ and $+1.489$. These persistent advantages confirm that our gains reflect genuine, model-wide improvements rather than isolated metric artifacts.

4.3 Which Part of the Advantage Rule Helps?

The model results motivate a more specific question about the mechanism. Does the benefit come from computing Gaussian interval means exactly, from estimating their probabilities differently, or from both? To separate these effects, we give every rule the same response outcomes. G^{2}RPO averages finitely many Gaussian quantiles, the exact-empirical variant integrates the same empirical intervals exactly, and Dirichlet-GRPO additionally replaces empirical frequencies with posterior probabilities. A fourth variant uses Dirichlet-GRPO but sets all advantages to zero whenever every response has the same outcome, isolating the contribution of these otherwise inactive groups.

We use a categorical policy whose true outcome probabilities are known, so an ideal reference update can be computed from the population Gaussian interval means. Each finite response group produces an estimated update, and its squared distance from that reference measures update error. Averaging over all possible count vectors gives the exact mean squared error (MSE), without sampling uncertainty. Figure 7(a,b) divides each method's MSE by G^{2}RPO's MSE for the same task, making 1 the baseline and values below 1 improvements. This diagnostic asks how well a finite group recovers the intended Gaussian update, rather than whether that update maximizes reward.

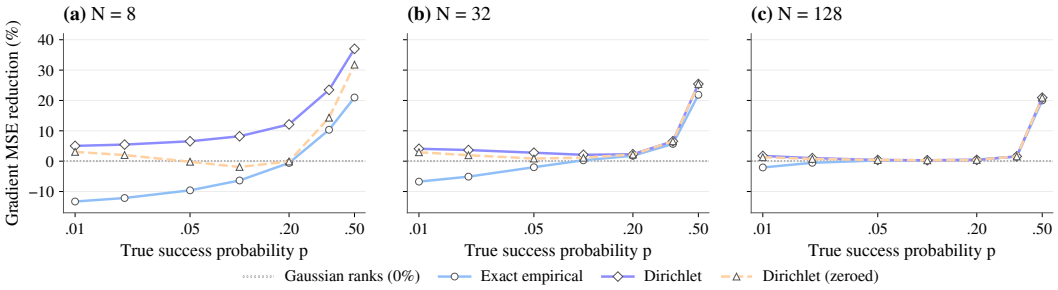


Figure 7: Exact component and prior ablations at $N = 8$. Relative update error uses G^{2}RPO as 1, and lower is better. P1–P5 are defined in Appendix E.2.

For binary feedback, we test seven success probabilities from 0.01 to 0.50 at group sizes 8, 32, and 128. Dirichlet-GRPO lowers update MSE relative to G^{2}RPO in all 21 settings. At the training group size of eight, the reductions are 6.55%, 12.11%, and 37.00% for success probabilities 0.05, 0.20, and 0.50, respectively. It also improves on exact empirical integration in 19 of the 21 settings, showing that integration alone does not explain the result. The two exceptions occur at $N = 128$, where the posterior rule increases MSE by less than 0.04%. Appendix Table 6 includes every setting, so both the improvements and exceptions remain visible.

For three ordered outcomes, we enumerate five probability profiles ranging from balanced outcomes to rare success or a dominant middle outcome, together with two reversed profiles that check symmetry. Across the three group sizes $N \in \{8, 16, 32\}$, Dirichlet-GRPO reduces MSE against G^{2}RPO by 0.96% to 29.07% and improves on exact empirical integration throughout. Figure 7(b) shows the five profiles at $N = 8$, while Appendix E.2 and Figure 10 report the full grid. The result extends the mechanism beyond binary correctness, where the position and probability of a middle outcome also affect the Gaussian intervals.

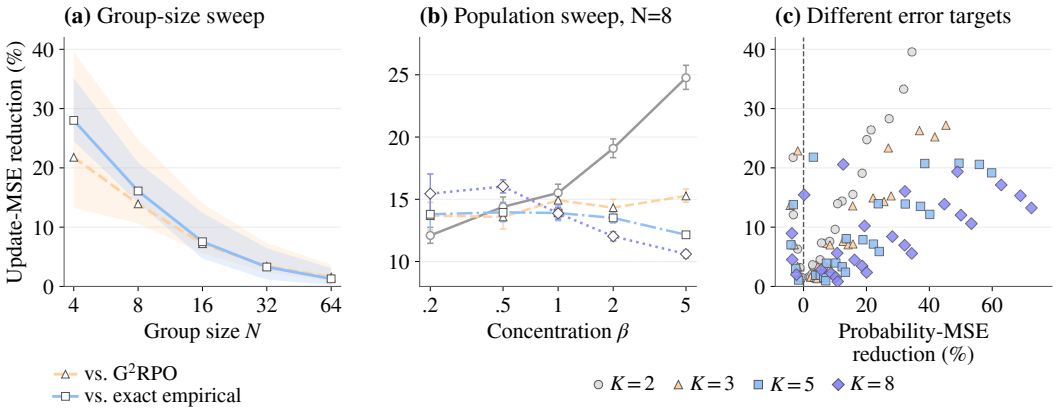


Figure 8: Calibration across 100 configurations. Shading gives full ranges in (a), and error bars give pointwise 95% intervals in (b).

4.4 Does the Benefit Persist across Feedback Distributions?

A few fixed examples cannot show how sensitive the result is to the choice of outcome probabilities. We therefore evaluate a larger, prespecified grid that varies the number of outcomes, their imbalance, and the response-group size. We use $K \in \{2, 3, 5, 8\}$ outcomes and draw populations from symmetric Dirichlet distributions with concentration $\beta \in \{0.2, 0.5, 1, 2, 5\}$. Small β produces more uneven probabilities, while larger β produces more balanced ones. This parameter generates test tasks and is distinct from the estimator's fixed half-count prior.

For each of the 20 combinations of K and β, we draw 512 populations and sample 128 response groups per population at each $N \in \{4, 8, 16, 32, 64\}$. The study therefore covers 10,240 populations, 100 configurations, and 6,553,600 response groups. All methods share the same populations and sampled groups. We first average squared update error across tasks, then normalize by the average squared reference-update magnitude, avoiding excessive weight on nearly deterministic tasks with very small reference updates. Paired bootstrap intervals resample 32 independent population blocks and retain the dependence between methods (Appendix E.3).

Dirichlet-GRPO lowers update MSE in all 100 configurations against both G^{2}RPO and exact empirical integration. The reductions range from 0.84% to 39.56% against G^{2}RPO and from 0.23% to 35.15% against the exact-empirical variant, with positive pointwise 95% intervals for every configuration. At $N = 8$, the reduction over G^{2}RPO is 10.60% to 24.76% across all twenty combinations of outcome count and imbalance (Figure 8(b)). The comparison with exact empirical integration is particularly useful because it attributes an additional gain to probability calibration after integration error has already been removed.

Figure 8(a) also shows when calibration matters most. The median MSE reduction over G^{2}RPO falls from 21.78% at four responses to 13.93%, 7.21%, 3.42%, and 1.56% at eight, sixteen, thirty-two, and sixty-four responses. Larger groups provide more reliable frequencies and more quantile representatives, leaving less error for either correction to remove. The strongest measured benefit therefore occurs in the small-group regime that motivates the method.

Probability accuracy and update accuracy nevertheless remain different quantities. Analytic probability MSE improves in 82 configurations, worsens in 17, and is unchanged in one, although update MSE improves across the grid. Figure 8(c) compares probability error with empirical frequencies and update error with G^{2}RPO. Changing an interval boundary affects both the assigned advantage and its contribution to the policy update, so a larger probability error need not produce a larger update error. The pattern agrees with Proposition 2, which bounds error propagation but does not rank all probability estimators.

4.5 Prior Strength and All-Equal Groups

The preceding results use the default half-count without fitting it to the test populations. To check its sensitivity, we repeat the exact three-outcome calculation at $N = 8$ with prior counts $\alpha \in$

$\{0.1, 0.25, 0.5, 1\}$. Figure 7(c) shows that $\alpha = 1$ has the lowest update MSE on each of the five tested profiles. The half-count is therefore a fixed theoretically motivated choice, rather than an empirically optimal value established by these experiments. Larger α places more weight on a uniform distribution, so its effect depends on both the available data and the true outcome imbalance. Moreover, using the same per-outcome count with different K changes the total prior mass, which should be considered when transferring the method to a new verifier.

All-equal groups provide a second useful check because exact empirical integration alone leaves their advantages at zero. At binary $N = 8, p = 0.05$, suppressing the posterior rule's all-equal-group updates increases update MSE by 7.22%. In the random-population study, retaining these groups gives lower MSE in 81 of 100 configurations, so they contribute in many settings but are not uniformly beneficial. The result also depends on which outcome is repeated. A group containing only the middle outcome can still have zero advantage under a symmetric three-outcome prior, whereas an all-failure binary group receives a negative weight.

Finally, we test whether the correction changes repeated learning in a small tabular policy, using 30 seeds and the same total response budget for five advantage rules. At $N = 4$, Dirichlet-GRPO improves final expected reward over G^{2}RPO by 0.00763, with a paired 95% interval of $[0.00747, 0.00778]$, while the benefit falls to 0.00098 at $N = 8$ and is not resolved at $N = 16$ or 32. These runs support the same small-group pattern, but GRPO attains higher reward in all eight tabular settings. Appendix E.5, Figure 11, and Table 8 retain the full curves, noise and learning-rate sweeps. Thus more accurate Gaussian-update estimation should be read as a mechanism result alongside the model benchmarks, rather than as a universal learning guarantee.

4.6 Computational Cost and Scope

Once verifier outcomes have been indexed, Dirichlet-GRPO requires $O(N + K)$ time and $O(K)$ auxiliary storage to count outcomes, evaluate interval means, and assign advantages. It introduces no learned critic or extra responses, although end-to-end training throughput has not been measured. Appendix E.4 further separates update directions from step sizes, showing why a nonzero advantage by itself is insufficient evidence of a better policy step. Appendix E.6 specifies the remaining comparisons, including matched rollout budgets, fixed verifier support, and prior-strength sweeps.

5 Related Work

GRPO estimates relative advantages from group rewards, while Dr.GRPO, DAPO, and GSPO modify their weighting or policy objective (Shao et al., 2024; Liu et al., 2025; Yu et al., 2025; Zheng et al., 2025). G^{2}RPO maps ordered feedback to Gaussian quantiles (Hu et al., 2026), motivating our probability calibration and exact interval integration. Bayesian and shrinkage methods use current outcomes or reward histories to estimate probabilities and baselines (Zhang et al., 2026; Han et al., 2026a; Kim et al., 2026; Zeng et al., 2025), while GUPO uses Dirichlet evidence to aggregate gradients (Guo et al., 2026b). Our posterior instead determines the probability intervals used to weight individual responses. Methods that enrich sparse feedback through virtual rewards, additional sampling, or visual evidence address complementary sources of information (He et al., 2026; Yan et al., 2026; Wang et al., 2025; Qiu et al., 2025). Appendix F gives the full comparison.

6 Conclusion

Dirichlet-GRPO addresses the fundamental pathology of small-group advantage estimation in discrete RLVR by coupling Jeffreys–Dirichlet posterior probabilities with exact Gaussian interval means, thereby decoupling and reducing both probability estimation bias and geometric discretization error. Theoretically, we formalize the conditions under which interval expectations achieve minimal squared approximation error, while delineating the boundary where prior smoothing transitions from mitigating variance to introducing structural bias. Controlled synthetic validations confirm that this formulation yields consistently lower estimation error relative to true Gaussian policy updates across diverse reward topologies, delivering the largest calibration margins at compact group sizes. On large-scale benchmarks with Qwen3-VL, Dirichlet-GRPO translates these theoretical advantages into tangible downstream gains across the 2B, 4B, and 8B parameter scales, notably surpassing standard GRPO and Dr. GRPO across all 11 evaluated benchmarks at the 4B scale.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Xiaoyuan Cheng, Wenxuan Yuan, Yiming Yang, Yuanzhao Zhang, Sibo Cheng, Yi He, and Zhuo Sun. Information shapes Koopman representation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.13025.

Fei Ding, Yongkang Zhang, Yuhao Liao, Zijian Zeng, and Huiming Yang. On the Impossibility of Unbiased and Length-Invariant Policy Optimization with Outcome Rewards. *arXiv preprint arXiv:2607.23364*, 2026. URL https://arxiv.org/abs/2607.23364.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. Onethinker: All-in-one reasoning model for image and video. *arXiv preprint arXiv:2512.03043*, 2025.

Siyuan Gan, Yuhan Li, Xiran Wang, Linjian Meng, Boyan Wang, Zhen Zhao, Jing Huo, Lei Bai, and Yang Gao. Learning from Hard Prompts: Difficulty-aware Advantage Amplification in Dynamic Sampling. *arXiv preprint arXiv:2608.27982*, 2026. URL https://arxiv.org/abs/2608.27982.

Jiani Guo, Junjie Wang, Jie Wu, Pengxiang Zhao, Dongdong Zhang, Shaohan Huang, Yujiu Yang, and Furu Wei. Learning where outcomes change: Credit-addressable reasoning for multimodal geometry. *arXiv preprint arXiv:2608.30457*, 2026a.

Peizheng Guo, Jianqi Zhang, Xingyu Zhang, Yun Fan, Jiahuan Zhou, Changwen Zheng, and Wenwen Qiang. GUPO: Gradient Uncertainty-aware Policy Optimization for Post-Training Large Language Models. *arXiv preprint arXiv:2608.17411*, 2026b. URL https://arxiv.org/abs/2608.17411.

Kevin Han, Yuhang Zhou, Mingze Gao, Gedi Zhou, Serena Li, Abhishek Kumar, Xiangjun Fan, Weiwei Li, and Lizhu Zhang. EBPO: Empirical Bayes Shrinkage for Stabilizing Group-Relative Policy Optimization. *arXiv preprint arXiv:2602.05165*, 2026a. URL https://arxiv.org/abs/2602.05165.

Yuhang Han, Yuyang Wu, Zhengbo Jiao, Yiyu Wang, Xuyang Liu, Shaobo Wang, Hanlin Xu, Xuming Hu, and Linfeng Zhang. Bridging Visual Representation and Reinforcement Learning from Verifiable Rewards in Large Vision-Language Models. *arXiv preprint arXiv:2603.27375*, 2026b. URL https://arxiv.org/abs/2603.27375.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage Collapse in Group Relative Policy Optimization: Diagnosis and Mitigation. *arXiv preprint arXiv:2605.21125*, 2026. URL https://arxiv.org/abs/2605.21125.

Jefferson Hernandez, Jaywon Koo, Zilin Xiao, Chen Wei, and Vicente Ordonez. SoftmaxGRPO: Learning to Reason using Softmax Advantage Group Estimation. *arXiv preprint arXiv:2608.09271*, 2026. URL https://arxiv.org/abs/2608.09271.

Wenbo Hu, Xin Chen, Yan Gao-Tian, Yihe Deng, Nanyun Peng, and Kai-Wei Chang. Openvl-thinkerv2: A generalist multimodal reasoning model for multi-domain visual tasks. *arXiv preprint arXiv:2604.08539*, 2026.

Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

Bingqing Jiang and Difan Zou. Structured Role-Aware Policy Optimization for Multimodal Reasoning. *arXiv preprint arXiv:2605.07274*, 2026. URL https://arxiv.org/abs/2605.07274.

Haechan Kim, Soohyun Ryu, Gyouk Chu, Doohyuk Jang, and Eunho Yang. Discounted Beta-Bernoulli Reward Estimation for Sample-Efficient Reinforcement Learning with Verifiable Rewards. *arXiv preprint arXiv:2603.18444*, 2026. URL https://arxiv.org/abs/2603.18444.

Yu Li, Rui Miao, Tian Lan, and Zhengling Qi. OPPO: Bayesian Value Recursion for Token-Level Credit Assignment in LLM Reasoning. *arXiv preprint arXiv:2605.21851*, 2026. URL https://arxiv.org/abs/2605.21851v1.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo

Molchanov. GDPO: Group reward-Decoupled Normalization Policy Optimization for Multi-reward RL Optimization. *arXiv preprint arXiv:2601.05242*, 2026. URL https://arxiv.org/abs/2601.05242.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Kirill Pavlenko, Alexander Golubev, Simon Karasik, and Boris Yangel. Blockwise Advantage Estimation for Multi-Objective RL with Verifiable Rewards. *arXiv preprint arXiv:2602.10231*, 2026. URL https://arxiv.org/abs/2602.10231.

Longtian Qiu, Shan Ning, Jiaxuan Sun, and Xuming He. NoisyGRPO: Incentivizing Multimodal CoT Reasoning via Noise Injection and Bayesian Estimation. *arXiv preprint arXiv:2510.21122*, 2025. URL https://arxiv.org/abs/2510.21122.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Mohammad Mahdi Salmani-Zarchi, Zahra Rahimi, Heshaam Faili, and Mohammad Javad Dousti. MDP-GRPO: Stabilized Group Relative Policy Optimization for Multi-Constraint Instruction Following. *arXiv preprint arXiv:2606.06058*, 2026. URL https://arxiv.org/abs/2606.06058.

Vishwas Sathish, Viresh Ranjan, Xinliang Zhu, Arnab Dhua, and Douglas Gray. Eliciting Self-Verification in Multimodal Reasoning Agents with Reinforcement Learning. *arXiv preprint arXiv:2609.08025*, 2026. URL https://arxiv.org/abs/2609.08025.

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

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

Siddarth Venkatraman, Matthieu Dinot, and Laurence Aitchison. Le Critique: Privileged Value Functions for LLM Reinforcement Learning. *arXiv preprint arXiv:2608.16739*, 2026. URL https://arxiv.org/abs/2608.16739.

Cédric Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

Jiamian Wang, Samyadeep Basu, Koustava Goswami, Tong Yu, and Zhiqiang Tao. Spurious Advantage Hidden in GRPO. *arXiv preprint arXiv:2609.04063*, 2026a. URL https://arxiv.org/abs/2609.04063.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-Aware Policy Optimization for Multimodal Reasoning. *arXiv preprint arXiv:2507.06448*, 2025. URL https://arxiv.org/abs/2507.06448.

Ziqi Wang, Xingzhou Lou, Meiqi Wu, Zhengqi Wen, and Junge Zhang. Calibration-Aware Policy Optimization for Reasoning LLMs. *arXiv preprint arXiv:2604.12632*, 2026b. URL https://arxiv.org/abs/2604.12632.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. BNPO: Beta Normalization Policy Optimization. *arXiv preprint arXiv:2506.02864*, 2025. URL https://arxiv.org/abs/2506.02864.

Feng Xiong, Leyan Xue, and Hongyu Lin. Correcting What You Cannot See: Credit Assignment for Perception Distillation in Multimodal Reasoners. *arXiv preprint arXiv:2607.28336*, 2026. URL https://arxiv.org/abs/2607.28336.

Zhengqing Yan, Xinyang Liu, Yi Zhang, Fan Guo, ChengXun Jia, Junchen Wan, Yao Liu, Qi Liu, Jihao Huang, and Kang Song. GDEPO: Group Dual-dynamic and Equal-right Advantage Policy Optimization with Enhanced Training Data Utilization for Sample-Constrained Reinforcement

Learning. *arXiv preprint arXiv:2601.06795*, 2026. URL https://arxiv.org/abs/2601.06795.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, Yaodong Yang, Jianxin Li, and Yikun Ban. Your Group-Relative Advantage Is Biased. *arXiv preprint arXiv:2601.08521*, 2026. URL https://arxiv.org/abs/2601.08521.

Jiayu Yao, Yiwei Wang, Anmeng Zhang, Zhe Sun, Songsong Wang, Lingrui Mei, Yuyao Ge, and Shenghua Liu. Multimodal Reward Hacking in Reinforcement Learning. *arXiv preprint arXiv:2607.09492*, 2026. URL https://arxiv.org/abs/2607.09492.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *arXiv preprint arXiv:2503.14476*, 2025. URL https://arxiv.org/abs/2503.14476.

Song Yu, Li Li, Wenwen Zhao, and Zhisheng Yang. EP-GRPO: Entropy-Progress Aligned Group Relative Policy Optimization with Implicit Process Guidance. *arXiv preprint arXiv:2605.04960*, 2026. URL https://arxiv.org/abs/2605.04960.

Guanning Zeng, Zhaoyi Zhou, Daman Arora, and Andrea Zanette. Shrinking the Variance: Shrinkage Baselines for Reinforcement Learning with Verifiable Rewards. *arXiv preprint arXiv:2511.03710*, 2025. URL https://arxiv.org/abs/2511.03710. Revised February 2026.

Xianzhou Zeng, Jing Huang, Chunmei Xie, Gongrui Nan, Siye Chen, Mengyu Lu, Weiqi Xiong, Qixuan Zhou, Junhao Zhang, Qiang Zhu, Yadong Li, and Xingzhong Xu. UCPO: Uncertainty-Aware Policy Optimization. *arXiv preprint arXiv:2601.22648*, 2026. URL https://arxiv.org/abs/2601.22648.

Zhi Zhang, Zhen Han, Costas Mavromatis, Qi Zhu, Yunyi Zhang, Sheng Guan, Dingmin Wang, Xiong Zhou, Shuai Wang, Soji Adeshina, Vassilis Ioannidis, and Huzefa Rangwala. Train Less, Learn More: Adaptive Efficient Rollout Optimization for Group-Based Reinforcement Learning. *arXiv preprint arXiv:2602.14338*, 2026. URL https://arxiv.org/abs/2602.14338.

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

Hongyi Zhou, Kai Ye, Erhan Xu, Jin Zhu, Ying Yang, Shijin Gong, and Chengchun Shi. Demystifying Group Relative Policy Optimization: Its Policy Gradient is a U-Statistic. *arXiv preprint arXiv:2603.01162*, 2026. URL https://arxiv.org/abs/2603.01162.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

APPENDIX

A Notation and Reading Guide . . . 14

A.1 Formal setup and baseline definitions . . . 14

B Detailed Theory and Complete Proofs . . . 16

B.1 Categorical Jeffreys prior and exact mass risk . . . 16

B.2 Quadratically optimal interval coupling . . . 18

B.3 Projection, ordering, and finite feedback resolution . . . 18

B.4 Rank decomposition and asymptotic scope . . . 19

B.5 Mass perturbations and interior condition . . . 20

B.6 Clipped surrogate and initial gradient . . . 21

B.7 Probability estimation and the oracle update . . . 22

B.8 Posterior centering versus empirical moments . . . 22

B.9 Rare-event scaling . . . 23

B.10 Binary population alignment . . . 23

C Implementation and Numerical Details . . . 24

D Checkpoint Data Provenance . . . 25

E Empirical Protocol Needed to Test the Mechanism . . . 26

Dr.GRPO comparison and draft placeholders . . . 26

E.1 Complete controlled Bernoulli ablation . . . 27

E.2 Prospective training ablations . . . 33

F Extended Related Work . . . 33

G AI Use Statement . . . 35

A NOTATION AND READING GUIDE

Table 2 distinguishes population probabilities, sampled frequencies, and posterior estimates, whose different roles determine the probability and representation errors analyzed below.

Table 2: **Notation used consistently in the text, proofs, and figures.** All population quantities condition on a fixed prompt and behavior policy unless stated otherwise.

Symbol	Meaning
x, y_i, T_i	Prompt, response i, and its number of generated tokens.
N, K	Responses per group and atoms in the declared verifier support.
$\mathcal{V}$, v_k, r_i	Ordered support, its kth value, and response i's verifier feedback.
$\mathbf{n}$, $\mathbf{p}$, $\widehat{\mathbf{p}}$, $\widetilde{\mathbf{p}}$	Counts, population masses, empirical masses, and posterior mean masses.
$\mathbf{u}$, λ_N	Uniform mass vector and the shrinkage weight $K/(2N+K)$.
$C_k(\mathbf{w})$	Cumulative mass through atom k for a generic mass vector $\mathbf{w}$.
U, Z, V	Uniform quantile, Gaussian target variable, and coupled verifier atom.
q_k, I_k	Gaussian quantile boundary and interval under $\widetilde{\mathbf{p}}$.
$T_k(\mathbf{w})$	Exact mean of the Gaussian interval occupied by mass w_k.
A_k^{z}, A_k^{G}, A_k^{D}	Atom advantages: z-score, tied Gaussian rank, and Dirichlet transport.
A_i, $k(i)$	Response advantage and the index of its observed reward atom.
$A_i^\star$, $g(A)$	Oracle interval advantage and the initial empirical token-surrogate gradient.
G, L_η	Root-mean-square response-score norm and interval sensitivity when masses are at least η.
$\rho_{i,t}$, ϵ, $\varepsilon_{\mathrm{num}}$	Token likelihood ratio, clipping radius, and numerical z-score stabilizer.
$D(\mathbf{w})$	Within-atom Gaussian resolution loss, $\mathbb{E}[\mathrm{Var}(Z \mid V)]$.
p, m, $A_{\pm}^{\mathrm{D}}(m)$	In the binary case only: true success probability, success count, and the two advantages at count m.

How to read the theory. The argument proceeds by estimating how often each verifier outcome occurs, representing each outcome by a Gaussian interval mean, and tracing errors in that representation into the initial policy update. To distinguish these steps, Eq. B.21 identifies exactly what changes relative to tied Gaussian ranks, Proposition 1 studies probability error across independently sampled groups, and Theorem 1 studies representation error once those probabilities are fixed. Proposition 2 then connects probability error to update error when all relevant masses are bounded away from zero, while the final binary calculation gives a population update identity under stronger assumptions, including full-sequence scores without length normalization. Together, these results explain the estimator from complementary perspectives without asserting a general training guarantee.

A.1 FORMAL SETUP AND BASELINE DEFINITIONS

Sampling and notation. For a fixed prompt x, possibly containing an image, the behavior policy independently samples N responses $y_i \sim \pi_{\theta_{\mathrm{old}}}(\cdot \mid x)$, and a fixed verifier returns $r_i = R(x, y_i)$ in a known, ordered support $\mathcal{V} = \{v_1 < \cdots < v_K\}$. Conditional on x, the population probabilities, counts, and empirical masses are

$$p_k = \mathbb{P}(r_i = v_k \mid x), \qquad n_k = \sum_{i=1}^{N} \mathbf{1}\{r_i = v_k\}, \qquad \widehat{p}_k = n_k/N. \tag{A.1}$$

Here $\mathbf{p}$ describes the underlying outcome frequencies, $\widehat{\mathbf{p}}$ records their frequencies in one sampled group, and $\widetilde{\mathbf{p}}$ below is the posterior predictive estimate. Throughout, an *atom* denotes a possible verifier outcome and its *mass* denotes its probability, with population statements keeping both the prompt and behavior policy fixed. The sampling model covers a deterministic verifier or independent verifier randomness, whereas correlated responses require a separate analysis.

The token-level policy objective. Let $T_i = |y_i|$ and $\rho_{i,t}(\theta) = \pi_\theta(y_{i,t} \mid x, y_{i,<t}) / \pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})$. With stop-gradient sequence advantage A_i and clipping radius $0 < \epsilon < 1$, define

$$\ell_\epsilon(\rho, A) = \min\{\rho A, \text{clip}(\rho, 1-\epsilon, 1+\epsilon)A\},$$
$$\mathcal{J}(\theta; A) = \mathbb{E}_{x,\{y_i\}\sim\pi_{\theta_{\text{old}}}}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\ell_\epsilon\big(\rho_{i,t}(\theta), A_i\big)\right]. \tag{A.2}$$

A reference-policy KL penalty, if used by the host trainer, is added separately and held fixed across estimators. Because Dirichlet-GRPO changes only A_i in this objective, the weighting effects introduced by response-length normalization and clipping remain part of the host trainer (Schulman et al., 2017; Liu et al., 2025).

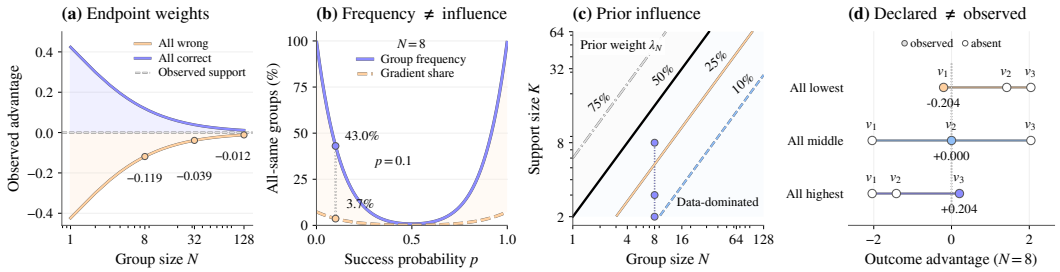


Figure 9: **Support and sample frequency affect different update properties.** Analytic diagnostics of (a) all-wrong/correct advantages, (b) homogeneous-group frequency versus its share of the expected binary update under sequence-score assumptions, (c) prior weight across group and support sizes, and (d) observed (filled) versus absent (hollow) outcomes. Observed-support normalization yields zero for homogeneous groups.

Moment normalization. GRPO uses $A_i^{\text{z}} = (r_i - \bar{r})/(s_r + \varepsilon_{\text{num}})$, where $s_r^2 = N^{-1}\sum_i (r_i - \bar{r})^2$. For binary rewards and empirical success fraction $\widehat{p} = m/N \in (0,1)$, ignoring numerical stabilization,

$$A_+^{\text{z}} = \sqrt{\frac{1-\widehat{p}}{\widehat{p}}}, \qquad A_-^{\text{z}} = -\sqrt{\frac{\widehat{p}}{1-\widehat{p}}}. \tag{A.3}$$

When $m = 1$, the observed success has advantage $\sqrt{N-1}$, whereas all-same groups have zero centered reward. For true success probability p, all-same groups occur with probability $p^N + (1-p)^N$ (Figure 9(b)), with the all-failure event alone accounting for probability $(1-p)^N$; Figure 1(b) compares the weights assigned to such a group with $N = 8$. The exactly-one-minority probability is $Np(1-p)^{N-1} + N(1-p)p^{N-1}$ for $N \geq 3$, while at $N = 2$ those events coincide and must not be counted twice.

Tie-averaged Gaussian ranks. For an observed atom k, set $s_k = \sum_{j<k} n_j$. A finite-group version of the Gaussian rank rule is

$$A_k^{\text{G}} = \frac{1}{n_k}\sum_{j=s_k+1}^{s_k+n_k}\Phi^{-1}\left(\frac{j-1/2}{N}\right), \qquad n_k > 0, \tag{A.4}$$

where Φ is the standard normal CDF. The rule follows the tied-quantile averaging in Hu et al. (2026), which generally differs from averaging ranks *before* applying the inverse CDF. Because its original task-level application and a prompt-group application use different reference populations, our estimator comparisons fix the same group to isolate the reward-to-advantage map; the archived benchmark labels alone do not establish which normalization pool their training code used.

An exact interval operator. For a probability vector $\mathbf{w}$, write $C_k(\mathbf{w}) = \sum_{j\leq k} w_j$, with $C_0 = 0$ and $C_K = 1$. Define

$$T_k(\mathbf{w}) = \frac{1}{w_k}\int_{C_{k-1}(\mathbf{w})}^{C_k(\mathbf{w})}\Phi^{-1}(u)\,du, \qquad w_k > 0. \tag{A.5}$$

The cumulative masses divide the unit interval into adjacent pieces, one per outcome, and T_k averages the Gaussian quantiles within piece k. Equation A.4 approximates $T_k(\widehat{\mathbf{p}})$ by the midpoint rule, taking one quantile from each probability cell of width $1/N$. Since both constructions assign equal advantages to tied outcomes, their comparison separates the probabilities used to define the pieces from the method used to compute each piece's mean.

B Detailed Theory and Complete Proofs

We first derive the posterior probabilities and their estimation risk, then establish the interval construction's coupling and projection properties to connect probability perturbations to policy updates. The remaining results explain empirical centering and rare outcomes before addressing the restricted binary population case.

B.1 Categorical Jeffreys prior and exact mass risk

Posterior distribution and shrinkage. Under the Jeffreys half-count prior, the posterior distribution and its mean are

$$\mathbf{p} \mid \mathbf{n} \sim \mathrm{Dirichlet}(n_1 + 1/2, \ldots, n_K + 1/2), \qquad \widetilde{p}_k = \mathbb{E}[p_k \mid \mathbf{n}] = \frac{n_k + 1/2}{N + K/2}. \tag{B.1}$$

Writing $\mathbf{u} = (1/K, \ldots, 1/K)$ for the uniform vector, the same estimator is

$$\widetilde{\mathbf{p}} = (1 - \lambda_N)\widehat{\mathbf{p}} + \lambda_N \mathbf{u}, \qquad \lambda_N = \frac{K}{2N + K}. \tag{B.2}$$

Thus the half-count mixes the observed frequencies with a uniform reference whose influence decreases as N grows. We apply the interval transform to this posterior mean rather than averaging transformed posterior draws, since these operations generally differ, *i.e.*, $T_k(\mathbb{E}[\mathbf{p} \mid \mathbf{n}]) \neq \mathbb{E}[T_k(\mathbf{p}) \mid \mathbf{n}]$.

Reading the Dirichlet distributions. Figure 4(a–c) shows the prior $\mathrm{Dirichlet}(0.5, 0.5, 0.5)$ and posteriors $\mathrm{Dirichlet}(5.5, 2.5, 1.5)$ and $\mathrm{Dirichlet}(20.5, 8.5, 4.5)$, obtained from counts $(5, 2, 1)$ and $(20, 8, 4)$, so both observed groups have the same proportions but the larger group provides more evidence. The three panels share a triangle geometry and density scale with respect to (p_1, p_2). Because the Jeffreys density is unbounded at the boundary, the rendering evaluates interior masses of at least 0.001 and saturates colors outside $[0.5, 64]$ solely for display, with diamonds marking distribution means rather than modes.

Total probability variance is $(1 - \|\widetilde{\mathbf{p}}\|_2^2)/(N + K/2 + 1)$, using the prior mean at $N = 0$, and equals 0.2667, 0.0543, and 0.0157 for the three states, falling to 20.4% and 5.9% of its prior value. To show how additional evidence changes this concentration, panel (d) connects the exact settings $N = 0, 8, 16, \ldots, 128$ with counts $N(5, 2, 1)/8$, using a solid curve for total variance divided by the prior value $4/15$ and a dashed curve for the prior's share $1.5/(N + 1.5)$ in the posterior mean. Both curves are percentages, with markers at $N = 8, 32, 128$, where the prior shares are 15.8%, 4.5%, and 1.2%. The sample-size axis is linear through eight responses and logarithmic thereafter, while the percentage axis is logarithmic; this fixed-proportion comparison illustrates posterior concentration and shrinkage rather than repeated-sampling estimation error.

Below Figure 4(a–c), the Beta marginal density curves shade the equal-tailed 80% intervals between the 0.1 and 0.9 quantiles, with all nine curves sharing one vertical density scale capped at 8 for display. Only the unbounded prior $\mathrm{Beta}(1/2, 1)$ exceeds this cap, for $p < 1/256$, so upward arrows indicate its continuation while its central interval $[0.01, 0.81]$ remains unaffected. The mean interval widths across outcomes are 0.800, 0.345, and 0.184 for the prior and two posteriors, describing individual outcome probabilities rather than a joint credible region. Independent marginal integration verifies the stored means, variances, and interval probabilities, and Figure 5(a) places the eight-response ranges alongside the probability construction.

Derivation of the prior. For $K \geq 2$, parameterize the simplex interior by $p_1, \ldots, p_{K-1}$, with $p_K = 1 - \sum_{j<K} p_j$. The Fisher information of one categorical observation is

$$\mathcal{I}(\mathbf{p}) = \mathrm{diag}(p_1^{-1}, \ldots, p_{K-1}^{-1}) + p_K^{-1}\mathbf{1}\mathbf{1}^\top. \tag{B.3}$$

The matrix determinant lemma gives

$$\det \mathcal{I}(\mathbf{p}) = \left(\prod_{j<K} p_j^{-1} \right) \left(1 + p_K^{-1} \sum_{j<K} p_j \right) = \prod_{k=1}^{K} p_k^{-1}. \tag{B.4}$$

Thus the Jeffreys density is proportional to $\prod_k p_k^{-1/2}$, a proper $\mathrm{Dirichlet}(1/2, \ldots, 1/2)$ density. Multiplication by the multinomial likelihood $\prod_k p_k^{n_k}$ gives the posterior in Eq. B.1; its mean is $(n_k + 1/2)/(N + K/2)$. For $K = 1$, the simplex is a single point with mass one, and its interval advantage is zero. The prior is invariant to smooth reparameterization of a fixed categorical model; changing the support or K changes that model.

Posterior decision rule for the masses. For any fixed candidate vector $\mathbf{w}$, expansion around $\widetilde{\mathbf{p}} = \mathbb{E}[\mathbf{p} \mid \mathbf{n}]$ gives

$$\mathbb{E}[\|\mathbf{p} - \mathbf{w}\|_2^2 \mid \mathbf{n}] = \mathbb{E}[\|\mathbf{p} - \widetilde{\mathbf{p}}\|_2^2 \mid \mathbf{n}] + \|\mathbf{w} - \widetilde{\mathbf{p}}\|_2^2. \tag{B.5}$$

The cross term vanishes by conditional centering, so $\widetilde{\mathbf{p}}$ uniquely minimizes posterior squared mass error. The posterior variance of a component is

$$\mathrm{Var}(p_k \mid \mathbf{n}) = \frac{\widetilde{p}_k(1 - \widetilde{p}_k)}{N + K/2 + 1}. \tag{B.6}$$

This decision rule justifies the mass estimate in Figure 5(a), while the component variances describe the posterior concentration in Figure 4. The optimality statement concerns squared probability error before the nonlinear transform T.

Exact comparison with empirical risk. To compare estimators across repeated samples, we keep the population $\mathbf{p}$ fixed and average over possible sampled groups, rather than conditioning on the observed group as in the posterior decision rule above. Proposition 1 shows that smoothing has no greater squared probability risk than the empirical estimator, whose risk is $(1 - \|\mathbf{p}\|_2^2)/N$, exactly when

$$\lambda_N \|\mathbf{u} - \mathbf{p}\|_2^2 \le (2 - \lambda_N) \frac{1 - \|\mathbf{p}\|_2^2}{N}. \tag{B.7}$$

Proof of Proposition 1. From multinomial sampling,

$$\mathbb{E}[\widehat{\mathbf{p}}] = \mathbf{p}, \qquad \mathrm{Cov}(\widehat{\mathbf{p}}) = \frac{\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top}{N}, \qquad \mathbb{E}\|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2 = \frac{1 - \|\mathbf{p}\|_2^2}{N}. \tag{B.8}$$

Using Eq. B.2,

$$\widetilde{\mathbf{p}} - \mathbf{p} = (1 - \lambda_N)(\widehat{\mathbf{p}} - \mathbf{p}) + \lambda_N(\mathbf{u} - \mathbf{p}). \tag{B.9}$$

The first term is sampling noise with zero mean and the second is the bias introduced by shrinking toward uniform probabilities, so their cross term vanishes in expectation and taking the expected squared norm proves Eq. 3. Subtracting the empirical estimator's risk and using $\lambda_N > 0$ then gives Eq. B.7. The proof also covers boundary populations, including a deterministic population for which empirical estimation risk is zero while smoothing has strictly positive risk if $K \ge 2$. □

Reading the three-outcome risk diagram. Since $\|\mathbf{u} - \mathbf{p}\|_2^2 = \|\mathbf{p}\|_2^2 - 1/K$, Eq. B.7 depends only on the concentration $\|\mathbf{p}\|_2^2$. At $K = 3, N = 8$, equal risk occurs at $\|\mathbf{p}\|_2^2 = 43/59$, so Figure 3(a) marks the less concentrated interior as lower risk and the near-deterministic corners as higher risk. Its three labeled populations have posterior/empirical risk ratios 0.71, 0.75, and 1.23, which enumeration of all 45 multinomial count vectors verifies along with the boundary. The triangle describes repeated-sampling mass risk, whereas Figure 4(b) describes a posterior conditional on one observed group. The component bars in Figure 3(b) divide Eq. 3 by the empirical mass risk, so the reference level one has the same meaning for both populations.

B.2 WHY THE INTERVAL COUPLING IS QUADRATICALLY OPTIMAL

Once the probabilities are fixed, we associate each ordered outcome with part of the Gaussian reference by introducing auxiliary variables V and Z, without assuming that the verifier has a latent Gaussian mechanism. Using the same uniform quantile for both variables gives an increasing coupling, *i.e.*, lower outcomes receive lower Gaussian quantiles. For the posterior mean masses, let $U \sim \mathrm{Unif}(0,1)$ and set $V = v_k$ when $U \in (C_{k-1}(\widetilde{\mathbf{p}}), C_k(\widetilde{\mathbf{p}})]$. With $Z = \Phi^{-1}(U)$, this gives

$$V \sim \sum_k \widetilde{p}_k \delta_{v_k}, \quad Z \sim \mathcal{N}(0,1), \quad Z \mid V = v_k \text{ lies in } I_k = (q_{k-1}, q_k], \quad q_k = \Phi^{-1}(C_k(\widetilde{\mathbf{p}})). \tag{B.10}$$

For any strictly positive $\mathbf{w}$, the same construction sets $V = v_k$ on $U \in (C_{k-1}(\mathbf{w}), C_k(\mathbf{w})]$ and $Z = \Phi^{-1}(U)$, thus producing the required marginals. We use a coupling because a deterministic function of a finite-valued variable cannot itself have a continuous Gaussian distribution, and the following argument establishes that this coupling minimizes squared cost.

For any other coupling of the same marginals,

$$\mathbb{E}[(V - Z)^2] = \mathbb{E}[V^2] + \mathbb{E}[Z^2] - 2\mathbb{E}[VZ]. \tag{B.11}$$

The first two terms are fixed by the marginals, so minimizing cost amounts to maximizing $\mathbb{E}[VZ]$. Decompose the ordered discrete variable as

$$V = v_1 + \sum_{j=1}^{K-1} (v_{j+1} - v_j)\mathbf{1}\{V > v_j\}. \tag{B.12}$$

The event $E_j = \{V > v_j\}$ has probability $1 - C_j(\mathbf{w})$, so setting $t_j = \Phi^{-1}(C_j(\mathbf{w}))$ gives an event $H_j = \{Z > t_j\}$ with the same probability and the pointwise inequality

$$(Z - t_j)(\mathbf{1}_{E_j} - \mathbf{1}_{H_j}) \le 0. \tag{B.13}$$

Above t_j the second factor is nonpositive and below it the second factor is nonnegative, so equal event probabilities imply $\mathbb{E}[Z\mathbf{1}_{E_j}] \le \mathbb{E}[Z\mathbf{1}_{H_j}]$. Since every coefficient $v_{j+1} - v_j$ is positive and all inequalities are attained simultaneously by the nested thresholds $E_j = H_j$, the stated quantile coupling maximizes $\mathbb{E}[VZ]$ and thus minimizes the squared cost. All expectations exist because V has finite support and Z is Gaussian, completing the discrete-source specialization of increasing rearrangement (Villani, 2003).

Closed form. For $I = (\ell, u]$ with positive normal probability, including infinite endpoints,

$$\mathbb{E}[Z \mid Z \in I] = \frac{\int_\ell^u z\varphi(z)\,dz}{\Phi(u) - \Phi(\ell)} = \frac{\varphi(\ell) - \varphi(u)}{\Phi(u) - \Phi(\ell)}, \tag{B.14}$$

using $\varphi'(z) = -z\varphi(z)$, so substituting the cumulative-mass boundaries proves Eq. 5. The implementation uses the exact zero-density limit at either infinite endpoint.

B.3 PROJECTION, ORDERING, AND FINITE FEEDBACK RESOLUTION

Although the coupling assigns a Gaussian interval to each outcome, the policy update needs one number per response, which the interval mean provides with the smallest squared error. The remaining error is the within-interval Gaussian variation that one representative per outcome cannot retain. For Theorem 1, write $a_V = a_k$ when $V = v_k$, $a_k^\star = T_k(\mathbf{w})$, and $D(\mathbf{w}) = \mathbb{E}[\mathrm{Var}(Z \mid V)]$. The mean and variance consequences are

$$\sum_k w_k a_k^\star = 0, \qquad \sum_k w_k (a_k^\star)^2 = 1 - D(\mathbf{w}) < 1. \tag{B.15}$$

For $K \ge 2$, the second moment is strictly positive; for $K = 1$, it is zero. Under a refinement that splits intervals without changing their combined masses, $D(\mathbf{w})$ cannot increase.

Proof of Theorem 1. Write $a_V = a_k$ and $a_V^\star = a_k^\star$ on $V = v_k$, and set $M = \mathbb{E}[Z \mid V] = a_V^\star$. Expanding $Z - a_V = (Z - M) + (M - a_V)$ yields

$$\mathbb{E}[(Z - a_V)^2] = \mathbb{E}[(Z - M)^2] + \mathbb{E}[(M - a_V)^2] + 2\mathbb{E}[(Z - M)(M - a_V)]. \tag{B.16}$$

The cross term is zero because $M - a_V$ is measurable with respect to V and $\mathbb{E}[Z - M \mid V] = 0$, leaving the first term $D(\mathbf{w})$ and the second term $\sum_k w_k(a_k - a_k^\star)^2$. Since every $w_k > 0$, the latter vanishes only when $\mathbf{a} = \mathbf{a}^\star$, establishing unique conditional optimality.

Total expectation gives $\sum_k w_k a_k^\star = \mathbb{E}[Z] = 0$. Total variance gives

$$1 = \mathrm{Var}(Z) = \mathbb{E}[\mathrm{Var}(Z \mid V)] + \mathrm{Var}(\mathbb{E}[Z \mid V]) = D(\mathbf{w}) + \sum_k w_k (a_k^\star)^2. \tag{B.17}$$

Every positive-probability Gaussian interval has strictly positive conditional variance, including unbounded endpoint intervals, so $D(\mathbf{w}) > 0$ and the second moment is strictly below one. For $K \geq 2$, adjacent interval means lie strictly on opposite sides of their common finite boundary because each conditional distribution puts positive mass away from it; the means are therefore strictly ordered and not all zero, giving a positive second moment. Samples sharing an atom necessarily share its mean, which also establishes tie consistency.

For a mass-preserving refinement, let $\mathcal{F}$ and $\mathcal{F}'$ be the coarse and fine interval sigma-algebras, with $\mathcal{F} \subseteq \mathcal{F}'$. The tower property and conditional Jensen inequality give

$$\mathbb{E}[\mathbb{E}[Z \mid \mathcal{F}]^2] = \mathbb{E}[\mathbb{E}[\mathbb{E}[Z \mid \mathcal{F}'] \mid \mathcal{F}]^2] \leq \mathbb{E}[\mathbb{E}[Z \mid \mathcal{F}']^2]. \tag{B.18}$$

Subtracting both sides from one proves that refinement cannot increase resolution loss when the combined masses remain fixed, while re-estimating those masses on a different support is a separate operation. □

Order invariance and optimality scope. Because Eq. A.5 depends on outcome order and probabilities rather than the numerical gaps between v_k, any strictly increasing recoding $h(v_k)$ gives the same intervals and representatives, and the increasing coupling remains quadratically optimal for the recoded source. The coupling result establishes how to pair outcomes with Gaussian quantiles, whereas the projection result establishes how to choose one representative per outcome; both take the Gaussian reference as given, so choosing the best reference for downstream reward remains an empirical question.

B.4 Exact rank decomposition and its asymptotic scope

To compare with tied Gaussian ranks, we first hold the empirical probabilities fixed so that the numerical integration change is isolated from the posterior probability correction. For an observed atom, its empirical interval is the union of cells $((j-1)/N, j/N]$ for $j = s_k + 1, \ldots, s_k + n_k$. Let $f = \Phi^{-1}$ and define $f_N(u) = f((j - 1/2)/N)$ in cell j. Then

$$T_k(\widehat{\mathbf{p}}) = \frac{1}{\widehat{p}_k} \int_{s_k/N}^{(s_k+n_k)/N} f(u)\, du, \qquad A_k^{\mathrm{G}} = \frac{1}{\widehat{p}_k} \int_{s_k/N}^{(s_k+n_k)/N} f_N(u)\, du. \tag{B.19}$$

Adding and subtracting $T_k(\widehat{\mathbf{p}})$ gives Eq. B.21, separating the posterior correction from the midpoint integration error between tied-quantile averaging and exact empirical interval means, both of which preserve ties.

For fixed K and population masses $p_k > 0$, the strong law gives $\widehat{\mathbf{p}} \to \mathbf{p}$ almost surely, with all atoms eventually observed. Since $\lambda_N \to 0$, we also have $\widetilde{\mathbf{p}} \to \mathbf{p}$, so continuity of T on the simplex interior implies $T_k(\widetilde{\mathbf{p}}) - T_k(\widehat{\mathbf{p}}) \to 0$ almost surely.

To complete the comparison, we establish $\|f_N - f\|_{L^1(0,1)} \to 0$ despite the infinite endpoint quantiles. On every $[a, 1-a]$, $a > 0$, continuity gives uniform convergence of cell midpoint approximations, while in a tail where f has constant sign, monotonicity of its absolute value bounds each midpoint cell's absolute contribution by twice the integral over the half-cell farther into that tail. The total tail contribution of f_N is therefore bounded by twice the corresponding integral of $|f|$, with at most one

vanishing boundary-cell correction. Since $\int_0^1 |f(u)|\,du = \mathbb{E}|Z| < \infty$, taking $N \to \infty$ on the interior and then $a \downarrow 0$ proves the claimed convergence and yields

$$|T_k(\widehat{\mathbf{p}}) - A_k^{\mathrm{G}}| \le \frac{\|f - f_N\|_{L^1}}{\widehat{p}_k} \longrightarrow 0. \tag{B.20}$$

The result establishes pointwise consistency for fixed positive population masses without providing a uniform rate over rare atoms or growing supports. If every observed reward is distinct and observed-support mode is used, $K = N$ and $\widetilde{p}_k = (1+1/2)/(N+N/2) = 1/N$ exactly, so the mass-correction term is already zero at finite N.

Equal outcomes receive equal advantages, and strictly increasing recodings preserve the rule because it uses outcome order rather than reward gaps. Transforming posterior means differs from averaging advantages over posterior draws, so the separate estimation and representation optima do not establish a Bayes-optimal RL objective.

Inserting exact empirical interval means isolates the two corrections for every observed outcome,

$$\underbrace{A_k^{\mathrm{D}} - A_k^{\mathrm{G}}}_{\text{total change}} = \underbrace{T_k(\widetilde{\mathbf{p}}) - T_k(\widehat{\mathbf{p}})}_{\text{probability correction}} + \underbrace{T_k(\widehat{\mathbf{p}}) - A_k^{\mathrm{G}}}_{\text{integration correction}}. \tag{B.21}$$

The first term changes interval probabilities, whereas the second removes finite integration error. Both vanish for fixed positive population masses as N grows, but their signs can differ.

B.5 Mass perturbations and a finite-sample interior condition

To determine how probability error affects the advantages, we next bound the change in an interval mean when its input probabilities change. Since an interval near a Gaussian tail is more sensitive than one well inside the distribution, the bound depends on the smallest outcome probability and assumes a common positive lower bound on all masses.

Proposition 3 (Local mass-to-advantage stability)**.** *Let $\mathbf{w}, \mathbf{w}'$ have the same $K \ge 2$ atoms and every mass at least η, where $0 < \eta < 1/2$. Set $Q_\eta = |\Phi^{-1}(\eta)|$, $B_\eta = \varphi(\Phi^{-1}(\eta))/\eta$, and $L_\eta = (Q_\eta + B_\eta/2)/\eta$. Then*

$$\|T(\mathbf{w}) - T(\mathbf{w}')\|_\infty \le L_\eta \|\mathbf{w} - \mathbf{w}'\|_1. \tag{B.22}$$

In particular, for $\mathbf{p}, \widetilde{\mathbf{p}}$ in this region, the oracle advantage error is at most $L_\eta\big((1-\lambda_N)\|\widehat{\mathbf{p}} - \mathbf{p}\|_1 + \lambda_N\|\mathbf{u} - \mathbf{p}\|_1\big)$.

To control the numerator and denominator of each interval mean separately, define $H(u) = \varphi(\Phi^{-1}(u))$ for $u \in (0,1)$, with continuous endpoint values $H(0) = H(1) = 0$. Since $H'(u) = -\Phi^{-1}(u)$, changing an interior boundary by d changes H by at most $Q_\eta d$ on $[\eta, 1-\eta]$.

Proof of Proposition 3. Let $d = \frac{1}{2}\|\mathbf{w} - \mathbf{w}'\|_1$ be the total variation distance, so differences in any coordinate mass or any prefix mass are bounded by d. Because each atom has mass at least η, every nontrivial cumulative boundary of either vector is in $[\eta, 1-\eta]$, while the endpoint boundaries 0 and 1 remain fixed, giving

$$|[H(C_{k-1}(\mathbf{w})) - H(C_k(\mathbf{w}))] - [H(C_{k-1}(\mathbf{w}')) - H(C_k(\mathbf{w}'))]| \le 2Q_\eta d. \tag{B.23}$$

The absolute normal mean on any interval of mass at least η is at most B_η, because the threshold argument in Appendix B.2 shows that, among events of a given probability, selecting the upper tail maximizes the conditional mean and selecting the lower tail minimizes it. Since the upper-tail mean decreases as its probability increases, an interval of probability at least η has mean in $[-B_\eta, B_\eta]$.

Dividing the preceding numerator difference by $w_k \ge \eta$ and accounting separately for the denominators gives

$$|T_k(\mathbf{w}) - T_k(\mathbf{w}')| \le \frac{2Q_\eta d}{w_k} + \frac{|T_k(\mathbf{w}')|\,|w_k - w_k'|}{w_k} \tag{B.24}$$

$$\le \frac{2Q_\eta + B_\eta}{\eta} d = L_\eta \|\mathbf{w} - \mathbf{w}'\|_1. \tag{B.25}$$

The final claim follows by the triangle inequality in Eq. B.2, and the proof also applies to the closure of the restricted region of masses at least η, since no nontrivial boundary approaches 0 or 1 there. □

When is the interior restriction satisfied? Let $\delta \in (0,1)$ and

$$t_{N,\delta} = \sqrt{\frac{\log(2K/\delta)}{2N}}, \qquad h_{N,\delta} = (1-\lambda_N)t_{N,\delta} + \lambda_N \|\mathbf{u}-\mathbf{p}\|_\infty. \tag{B.26}$$

Applying Hoeffding's inequality to the K indicator averages and taking a union bound gives $\|\widehat{\mathbf{p}} - \mathbf{p}\|_\infty \le t_{N,\delta}$ with probability at least $1-\delta$, which also ensures $\|\widetilde{\mathbf{p}} - \mathbf{p}\|_\infty \le h_{N,\delta}$. If $\min_k p_k \ge 2\eta$ and $h_{N,\delta} \le \eta$, both vectors therefore have masses at least η, yielding

$$\|T(\widetilde{\mathbf{p}}) - T(\mathbf{p})\|_\infty \le L_\eta K h_{N,\delta} \qquad \text{with probability at least } 1-\delta. \tag{B.27}$$

Although this sufficient finite-sample condition may exclude very sparse populations, its dependence on η explains why lower probability risk alone does not imply lower expected advantage error, since the nonlinear transform can amplify probability errors differently across the simplex.

B.6 Perturbation of the clipped surrogate and initial gradient

To quantify how changing advantages alters the weights in the policy objective for fixed sampled responses, we first bound the change in surrogate value and then the gradient difference at the behavior policy. In the gradient bound, s_i is the response's average token score vector, which describes how a parameter change alters its log probability per generated token, while G summarizes the sizes of these vectors across the group.

Proposition 4 (From advantages to the policy update)**.** *Fix a sampled group and denote its empirical version of Eq. A.2 by $\widehat{\mathcal{J}}$. If all its token ratios are at most R_ρ and $M_\rho = \max\{R_\rho, 1+\epsilon\}$, then*

$$|\widehat{\mathcal{J}}(\theta; A) - \widehat{\mathcal{J}}(\theta; B)| \le \frac{M_\rho}{N} \sum_i |A_i - B_i|. \tag{B.28}$$

At $\theta = \theta_{\text{old}}$ and $\epsilon > 0$, let $s_i = T_i^{-1} \sum_t \nabla_\theta \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})$ and $G = (N^{-1} \sum_i \|s_i\|_2^2)^{1/2}$. For fixed advantages,

$$\|\nabla\widehat{\mathcal{J}}(A) - \nabla\widehat{\mathcal{J}}(B)\|_2 \le G\sqrt{\frac{1}{N}\sum_i (A_i - B_i)^2}. \tag{B.29}$$

Proof of Proposition 4. For a fixed nonnegative likelihood ratio ρ, let $c = \text{clip}(\rho, 1-\epsilon, 1+\epsilon)$. The surrogate equals $\ell_\epsilon(\rho, A) = \min(\rho, c)A$ when $A \ge 0$ and $\max(\rho, c)A$ when $A < 0$. For the usual clipping radius $0 < \epsilon < 1$, both slopes are nonnegative and at most $\max\{\rho, 1+\epsilon\}$; because the function is continuous at zero, its absolute slope is therefore bounded by M_ρ for all token ratios in the fixed group. Applying this Lipschitz bound token by token and using $T_i^{-1} \sum_t 1 = 1$ proves Eq. B.28, with an identical fixed KL penalty canceling from the difference.

At the behavior policy, each ratio equals one and lies in the interior of the clipping interval, so differentiating with detached advantages gives

$$g(A) := \nabla\widehat{\mathcal{J}}(A) = \frac{1}{N}\sum_i A_i s_i. \tag{B.30}$$

Consequently, the triangle inequality and Cauchy–Schwarz imply

$$\|g(A) - g(B)\|_2 \le \frac{1}{N}\sum_i |A_i - B_i|\,\|s_i\|_2 \le \left(\frac{1}{N}\sum_i \|s_i\|_2^2\right)^{1/2} \left(\frac{1}{N}\sum_i (A_i - B_i)^2\right)^{1/2}. \tag{B.31}$$

The inequality holds for each fixed group, even though its advantages and score vectors arise from the same responses. □

Equation B.28 requires a ratio bound that PPO clipping alone does not supply because the surrogate retains an unclipped branch, whereas the gradient identity applies at the behavior policy, before subsequent optimizer steps can activate clipping.

B.7 From probability estimation to the oracle update

Combining the previous two bounds connects probability error to advantage error and then to the initial gradient difference. To isolate this effect, the oracle comparison in Proposition 2 uses the same sampled responses and Gaussian reference for both gradients, replacing only the estimated probability vector with the true probabilities of the fixed prompt and behavior policy. Let $A_i^\star = T_{k(i)}(\mathbf{p})$ and $A_i^{\mathrm{D}} = T_{k(i)}(\widetilde{\mathbf{p}})$, treating both advantages as constants when differentiating.

Proof of Proposition 2. Under the positive-mass condition, Proposition 3 bounds every outcome-advantage difference by $L_\eta \|\widetilde{\mathbf{p}} - \mathbf{p}\|_1$. Therefore,

$$\left(\frac{1}{N}\sum_i (A_i^{\mathrm{D}} - A_i^\star)^2\right)^{1/2} \le \|T(\widetilde{\mathbf{p}}) - T(\mathbf{p})\|_\infty \le L_\eta \|\widetilde{\mathbf{p}} - \mathbf{p}\|_1. \tag{B.32}$$

Applying Proposition 4 at the behavior policy yields Eq. 4, without requiring independence between the sampled counts, advantages, and policy-score vectors because the argument holds pointwise. An identical reference-policy penalty cancels from the gradient difference. □

Finite-sample interpretation. On the high-probability event in Appendix B.5, provided $\min_k p_k \ge 2\eta$ and $h_{N,\delta} \le \eta$, the same reasoning yields

$$\|g(A^{\mathrm{D}}) - g(A^\star)\|_2 \le G L_\eta K h_{N,\delta}. \tag{B.33}$$

Because the score norm G is evaluated on the sampled group and can vary between groups, this statement is conditional on the stated event rather than an unconditional expected-gradient-risk bound. Although the oracle uses known probabilities within the Gaussian representation, its gradient need not be the optimal reward gradient for multi-level feedback; the separate binary population calculation below examines reward-gradient alignment under full-sequence scores and more restrictive assumptions.

Controlled diagnostic in Figure 3(d). To illustrate the two successive inequalities and their slack, we use a one-token Bernoulli policy with scalar logit, success probability p, and fixed counts $(6, 2)$, whose score values are $-p$ and $1-p$ and whose posterior success mass is $\widetilde{p}_+ = 2.5/9$. We vary $p = \widetilde{p}_+ + h$ over $h \in [0.002, 0.22]$, so both population and posterior masses exceed $\eta = 0.25$, and compute both gradients on the same eight responses with detached advantages. The plot compares $|g(A^{\mathrm{D}}) - g(A^\star)|/G$, the empirical RMS advantage difference, and $L_\eta\|\widetilde{\mathbf{p}} - \mathbf{p}\|_1$; because signed response-score contributions can cancel, the actual gradient error need not increase monotonically with mass error. The script independently differentiates the clipped surrogate at the behavior logit to verify the gradients, providing an analytic sensitivity diagnostic rather than a training trajectory or a distribution over sampled groups.

B.8 Posterior centering versus empirical moments

The interval means are centered under posterior probabilities, whereas the training objective weights the responses that actually appeared, so the half-counts introduce a difference between the two weighting schemes. To characterize its effect, the next result makes the empirical mean explicit and bounds the empirical second moment.

Corollary 1 (Empirical moments and rare observed successes)**.** *For the advantages in Eq. 5,*

$$\frac{1}{N}\sum_i A_i = -\frac{1}{2N}\sum_k A_k^{\mathrm{D}}, \qquad \frac{1}{N}\sum_i A_i^2 < 1 + \frac{K}{2N}. \tag{B.34}$$

Hence $\|\nabla \widehat{\mathcal{J}}(A^{\mathrm{D}})\|_2 \le G\sqrt{1 + K/(2N)}$ *at the behavior policy. If a top atom has mass* $w \to 0$*, its interval mean is* $\Theta(\sqrt{\log(1/w)})$*. For a* single observed binary success *and* $N \to \infty$*, this gives*

$$A_+^{\mathrm{D}}(m=1) = \Theta(\sqrt{\log N}), \qquad A_+^{\mathrm{z}}(m=1) = \sqrt{N-1}. \tag{B.35}$$

Proof of the empirical-moment part of Corollary 1. Rearranging $\sum_k (n_k + 1/2)A_k^{\mathrm{D}} = 0$ and dividing by N proves the empirical mean identity, while the second moment satisfies

$$\frac{1}{N}\sum_i A_i^2 = \sum_k \frac{n_k}{N}(A_k^{\mathrm{D}})^2 \tag{B.36}$$

$$\leq \frac{N + K/2}{N}\sum_k \widetilde{p}_k (A_k^{\mathrm{D}})^2 < 1 + \frac{K}{2N}, \tag{B.37}$$

where $n_k \leq n_k + 1/2$ and Theorem 1 supplies the strict upper bound. Taking $B = 0$ in Eq. B.29 yields the gradient-norm bound, while the second-moment bound also controls empirical advantage variance because the latter cannot exceed the former. These within-group statements do not determine the bias or variance of the policy gradient across sampled groups. □

Support-aware endpoint groups. In an all-wrong binary group, let $w_N = 1/[2(N+1)]$ be the posterior mass of unobserved success. Then

$$A_-^{\mathrm{D}}(0) = -\frac{w_N}{1 - w_N}\frac{\varphi(\Phi^{-1}(1 - w_N))}{w_N} < 0. \tag{B.38}$$

With both outcomes declared, every response in an all-wrong group receives the same negative weight, so the support supplies a direction without distinguishing among those responses; all-correct behavior follows by symmetry. The tail result below gives a magnitude of $\Theta(\sqrt{\log N}/N)$ as $N \to \infty$, although the behavior depends on the support because a symmetric three-atom all-middle group has an observed interval mean of exactly zero. In observed-support mode, every all-same group also receives zero because its sole interval covers the entire normal distribution.

B.9 Rare-event scaling without conflating mass and sample size

To distinguish a Gaussian tail interval whose mass tends to zero from a sampled binary group with one success as N grows, we first derive the tail scaling and then substitute the posterior mass for the single-success case. For $q > 0$, integration by parts gives

$$1 - \Phi(q) = \frac{\varphi(q)}{q} - \int_q^\infty \frac{\varphi(z)}{z^2}\, dz. \tag{B.39}$$

Because the integral is positive and strictly below $(1 - \Phi(q))/q^2$, rearranging yields the Mills bounds

$$\frac{q}{1 + q^2}\varphi(q) < 1 - \Phi(q) < \frac{\varphi(q)}{q}. \tag{B.40}$$

For an upper atom of mass $w = 1 - \Phi(q)$, its mean therefore satisfies

$$q < \frac{\varphi(q)}{w} < q + \frac{1}{q}. \tag{B.41}$$

Taking logarithms in the Mills bounds shows $\log(1/w) = q^2/2 + \log q + \frac{1}{2}\log(2\pi) + O(q^{-2})$, so $q \sim \sqrt{2\log(1/w)}$ and the mean is $\Theta(\sqrt{\log(1/w)})$. The lower-tail case follows by symmetry.

For binary groups containing exactly one observed success, substituting $\widetilde{p}_+ = 3/[2(N+1)]$ gives Eq. B.35, while Eq. A.3 gives $\sqrt{N-1}$ exactly when numerical stabilization is ignored. Because the tied Gaussian rank for that lone success is also $\Phi^{-1}(1 - 1/(2N)) = \Theta(\sqrt{\log N})$, mild tail growth is a shared property of the Gaussian target, with the separate posterior correction quantified by Eq. B.21.

At fixed N, K, every posterior atom has mass at least $1/(2N + K)$ and every observed atom has mass at least $3/(2N + K)$, so no infinite rare-mass limit occurs at a fixed rollout size. All constants remain finite, although changing N or the declared support changes the resulting finite geometry.

B.10 Binary population alignment under sequence-score updates

Having compared the estimated and oracle Gaussian representations, we now examine when the expected update under binary feedback points along the gradient of success probability. The identity below establishes alignment for an unclipped, full-sequence score update, under assumptions that deliberately differ from those of the token-normalized host objective.

Binary rule and population statement. Let m denote the success count, $\widetilde{p}_+ = (m+1/2)/(N+1)$, and $q = \Phi^{-1}(1-\widetilde{p}_+)$. The two outcome advantages are

$$A_+^{\mathrm{D}} = \frac{\varphi(q)}{\widetilde{p}_+}, \qquad A_-^{\mathrm{D}} = -\frac{\varphi(q)}{1-\widetilde{p}_+}. \tag{B.42}$$

We write $A_\pm^{\mathrm{D}}(m)$ for these values at count m. For one fixed prompt, independent groups, and an unclipped, detached-advantage full-sequence score update without response-length normalization, the following identity holds under the differentiability and interchange conditions stated below,

$$\mathbb{E}\left[\frac{1}{N}\sum_i A_i^{\mathrm{D}} \nabla \log \pi_\theta(y_i \mid x)\right] = c_N(p)\nabla p, \qquad c_N(p) = \mathbb{E}_{B\sim \mathrm{Bin}(N-1,p)}[A_+^{\mathrm{D}}(B+1) - A_-^{\mathrm{D}}(B)] > 0. \tag{B.43}$$

Assume the policy is differentiable at the behavior parameters, the reward partition is fixed, and differentiation can pass through its probability integrals. Let $s(y) = \nabla \log \pi_\theta(y \mid x)$ be the *full sequence* score. For $p = \mathbb{P}(R(x,y)=1) \in (0,1)$,

$$\mathbb{E}[s(y) \mid R=1] = \frac{\nabla p}{p}, \qquad \mathbb{E}[s(y) \mid R=0] = -\frac{\nabla p}{1-p}. \tag{B.44}$$

These follow by differentiating the integrals over the success and failure sets.

For response i, let $B = \sum_{j\neq i} r_j$. Independent sampling implies $B \sim \mathrm{Bin}(N-1,p)$ independently of y_i. Conditional on success, the total group count is $B+1$ and its advantage is $A_+^{\mathrm{D}}(B+1)$; conditional on failure it is B and the advantage is $A_-^{\mathrm{D}}(B)$. Taking expectations,

$$\mathbb{E}[A_i^{\mathrm{D}} s(y_i)] = p\mathbb{E}_B[A_+^{\mathrm{D}}(B+1)]\frac{\nabla p}{p} - (1-p)\mathbb{E}_B[A_-^{\mathrm{D}}(B)]\frac{\nabla p}{1-p} \tag{B.45}$$

$$= \mathbb{E}_B[A_+^{\mathrm{D}}(B+1) - A_-^{\mathrm{D}}(B)]\nabla p. \tag{B.46}$$

By exchangeability, averaging over i leaves the expectation unchanged, while positive posterior mass for both binary atoms ensures $A_+^{\mathrm{D}}(m) > 0$ and $A_-^{\mathrm{D}}(m) < 0$ for every m; thus $c_N(p) > 0$, completing the proof of Eq. B.43.

Contribution of all-same groups. Write $a_N = A_+^{\mathrm{D}}(N) = -A_-^{\mathrm{D}}(0) > 0$. The all-correct event has probability p^N and conditional score mean $\nabla p/p$; the all-wrong event has probability $(1-p)^N$ and conditional score mean $-\nabla p/(1-p)$. Their combined coefficient is therefore $a_N[p^{N-1}+(1-p)^{N-1}]$. Subtracting it from $c_N(p)$ gives the mixed-group contribution, which is nonnegative by the same conditional-score argument. Figure 9(b) divides the all-same coefficient by the total coefficient and compares that share with the probability of an all-same group. At $N=8, p=0.1$, all-same groups occur with probability 43.0% but supply only 3.7% of this coefficient, showing that frequency alone does not determine their update contribution; direct binomial count enumeration verifies the split under the same restricted assumptions.

The expectation here averages detached-advantage score updates, so differentiating through the advantage or sampling law would define a different quantity. Dividing scores by response-dependent lengths can invalidate the conditional-score identities, while averaging prompts can reweight their reward gradients because $c_N(p)$ depends on each prompt's success probability; thus multi-level rewards and the length-normalized implementation require analyses beyond this binary setting.

C Implementation and Numerical Details

Table 3 implements the probability-then-integration rule while retaining unseen valid outcomes in the support; assigning the resulting weights only to observed responses introduces no synthetic samples.

Once each reward has a support index, counting outcomes and assigning advantages back to responses take $O(N)$ time, while computing interval means takes $O(K)$ time and $O(K)$ auxiliary storage besides the output. Mapping arbitrary reward values to sorted support indices requires binary search,

Table 3: **Algorithm: one prompt group, fixed ordered support.** The routine returns detached response advantages for the host policy objective.

Step	Computation
1	Obtain N verified responses, using the task's declared ordered support $\mathcal{V}$.
2	Count all K atoms, including zero counts for valid but unobserved outcomes.
3	Compute $\widetilde{p}_k = (n_k + 1/2)/(N + K/2)$.
4	Form cumulative masses and evaluate $H(C_k) = \varphi(\Phi^{-1}(C_k))$, with exact endpoint values $H(0) = H(1) = 0$.
5	Compute $A_k^{\mathrm{D}} = [H(C_{k-1}) - H(C_k)]/\widetilde{p}_k$.
6	Map atom advantages back to observed responses and use Eq. A.2. Do not add empirical group re-centering in the default mode.

with cost $O(N \log K)$, or a prebuilt lookup; discovering the observed support additionally requires finding and sorting unique values. These bounds describe the estimator's computational cost, whereas training-speed gains have not been measured.

The reference implementation in `scripts/estimator.py` uses float64 for probabilities, cumulative sums, inverse-normal evaluation, and densities. To preserve precision near the tails, it evaluates each interior density from the smaller cumulative tail using $H(u) = H(1 - u)$, sums right-tail probabilities in reverse rather than computing $1 - u$, and sets endpoint densities to their exact value of zero. These choices avoid arbitrary clipping of valid intervals for practical small K and N, although very large supports or counts require further numerical checks, especially when two interior densities are nearly equal. Integration into a trainer should therefore retain this precision until advantages have been computed, casting to the trainer's dtype only afterward and keeping advantages detached from the policy graph.

Choose the support from the verifier specification, using either numeric bin centers or ordered bin labels because the method depends on their order, while recognizing that changing bin edges changes which responses share an outcome. Unordered labels require a justified ordering before the method applies, and combining reward components into a continuous total may remove the small discrete support on which the analysis relies.

Task sampling or fixed task weights can be added separately, while replacing the reference by $\mathcal{N}(0, \sigma^2)$ multiplies all advantages by σ and all second moments by σ^2. Because policy score vectors differ across tasks, their gradient magnitudes can still differ, so empirical re-centering, task-wise renormalization, and changes to loss aggregation should be recorded as explicit variants.

D Checkpoint Results and Data Sources

The current benchmark table is the score source for all reported aggregates. It contains eleven metrics at each scale, with HallusionBench aACC and fACC recorded separately. The supplied revision identifies the former generic Hallu entries at 2B and 8B as aACC and adds fACC values of $(39.42, 38.86, 40.17)$ at 2B and $(54.38, 53.64, 53.91)$ at 8B, ordered as GRPO, G^{2}RPO, and Dirichlet-GRPO. These six additions come from the supplied LaTeX table rather than the older workbook. The other scores agree with `model_benchmark_comparison.xlsx`, whose original cell references remain in `data/benchmark_results.json`. The consolidated values and their provenance are recorded in `data/current_results_20260926.json`.

The means equally weight the eleven reported metrics, so HallusionBench contributes two entries. All means and differences are computed before display rounding. Values supplied only to two decimal places are retained at that precision, rather than assigned artificial extra precision. MME retains the score normalization used in the supplied table. The repeated MMK12 triples and large 2B MME difference remain unchanged, with their influence evaluated through the exclusions in Table 4. Table 5 separately compares the same four named math benchmarks, showing why a positive overall mean can coexist with a negative mean against the stronger per-metric baseline on a subset.

Training configuration and remaining records. The supplied manuscript specifies ViRL39K, AdamW, one epoch, a global prompt batch size of 128, a learning rate of 10^{-6} with cosine decay to 10%, eight rollouts per prompt, temperature 1.0, a 2,048-token response limit, clipping radius 0.2,

Table 4: Mean gains after excluding MME and MMK12. All rows use the current eleven-metric score record.

Scale	Entries retained	Count	ΔGRPO	ΔG^2
2B	All entries	10	+2.772	+3.720
2B	Exclude MMK12	9	+2.568	+3.808
2B	Exclude MME	9	+1.166	+1.924
2B	Exclude both	8	+0.735	+1.799
4B	All entries	11	+1.366	+1.618
4B	Exclude MMK12	10	+1.042	+1.487
4B	Exclude MME	10	+1.427	+1.633
4B	Exclude both	9	+1.073	+1.489
8B	All entries	10	+0.763	+0.459
8B	Exclude MMK12	9	+0.336	+0.184
8B	Exclude MME	9	+0.848	+0.510
8B	Exclude both	8	+0.378	+0.208

Table 5: Shared math results on WeMath, MathVista, MathVerse, and MathVision. Δ_{best} uses the stronger baseline for each metric. LOMO is the range after leaving out one metric, not a confidence interval.

Scale	GRPO	G^2RPO	Dirichlet-GRPO	Δ_{best}	Wins	LOMO range
2B	43.07	41.14	43.97	+0.85	3/4	[+0.59, +1.16]
4B	62.84	62.68	64.00	+0.91	4/4	[+0.75, +1.12]
8B	66.52	67.32	67.42	-0.16	2/4	[-0.31, -0.03]

and KL coefficient 0.04. The older workbook also records two 4B paths ending at step 148, but this alone does not determine the retained training-set size, the number of optimizer steps per rollout batch, or the schedule used at other scales. We therefore omit the previous approximate step count. The package does not contain raw trainer configurations, checkpoint hashes, evaluation predictions, or seed-level runs, so these settings are documented as supplied and no seed uncertainty is inferred from them. Hardware, AdamW coefficients and weight decay, evaluator versions, image preprocessing, and checkpoint-selection details remain unspecified.

E Controlled Studies and Training Protocol

Although the analytic results isolate particular changes in the estimator, testing whether those changes improve trained models requires matched runs and records that the checkpoint archive does not provide. To guide that evaluation, this section specifies the necessary comparisons while distinguishing them from the controlled calculations already completed.

Matched ablations. To isolate the estimator, use the same initialization, prompt groups, verifier, support, token budget, optimizer, clipping, KL settings, response-length normalization, and evaluation harness when comparing (i) tied Gaussian midpoint ranks; (ii) exact empirical interval means; (iii) Jeffreys posterior interval means with declared support; and (iv) the same posterior estimator with all-same groups assigned zero. An observed-support variant is also useful, although it changes K and the prior and therefore differs from merely dropping all-same groups. Keep normalization-pool scope matched, treating task-level weighting as a separate experiment.

Dr.GRPO comparison. The native Dr.GRPO baseline (Liu et al., 2025) subtracts the prompt group's mean reward, including the response itself, without dividing by the group standard deviation, and its loss sums valid-token clipped terms, divides each response sum by the fixed maximum generation length, and then averages responses. A comparison with Dr.GRPO should preserve both changes rather than impose the response-length averaging used in the estimator-only ablations, while matching the initial checkpoint, prompt split, image processing, verifier, group size, sampling settings, maximum length, clipping/KL settings, total generated tokens, and evaluation/selection rules. Record the loss constant and learning rate, provide equal tuning budgets, and report response

lengths and gradient norms, starting at 4B with at least three independent seeds before evaluating the same metrics at 2B and 8B as runs become available. Because the archive contains no matched Dr.GRPO checkpoint or prediction file, its stored metric slots and means remain unmeasured in `data/additional_baseline_drgrpo.json`; scores from the source paper's language-only tasks do not supply those missing multimodal evaluations.

No Dr.GRPO scores appear in the result tables because the archive contains no matched evaluations for that method. Historical interpolated layout placeholders are excluded from every reported comparison.

Comparisons to recent estimators. To isolate the transform in binary feedback, compare posterior-normalized rewards and Gaussian interval means under the same prior and posterior mean, because otherwise changes in probability estimation are mixed with changes in advantage shape; AERO, EBPO, and DBB motivate this control (Zhang et al., 2026; Han et al., 2026a; Kim et al., 2026). History-based estimators should receive the history their methods require, while adaptive sampling should be compared at a matched total rollout budget. SoftmaxGRPO offers an alternative advantage shape and GUPO changes how group gradients are aggregated (Hernandez et al., 2026; Guo et al., 2026b), although these comparisons remain outside the existing checkpoint archive.

Observed feedback topology. For each group and task, record support size, N, atom counts, entropy, and whether all rewards are equal. Define the tie-sample fraction as

$$\mathrm{TieFrac}(\mathbf{n}) = \frac{1}{N}\sum_{k} n_k \mathbf{1}\{n_k \geq 2\}. \tag{E.1}$$

The tie-sample fraction counts every response belonging to a tie, including the first member, whereas $1 - K_{\mathrm{obs}}/N$ counts only repeated observations after the first occurrence of each outcome. To avoid conflating these quantities, name both explicitly when reporting them and follow the field definitions in the included topology script. In an all-same binary group, valid but unobserved outcomes count toward the declared K, but not toward the observed K_{obs}.

Optimization diagnostics. Record empirical mean and second moment of advantages, maximum observed absolute advantage, token-ratio distributions, clip fractions, response lengths, and parameter-gradient norms. To test Eq. B.21, compute all three estimators on the same saved groups and report both correction terms, stratified by minimum observed atom count and all-same status. For the mass-risk mechanism, controlled distributions have known $\mathbf{p}$; actual rollout logs do not, so oracle mass error should not be reported without an independent high-sample reference and its uncertainty.

Statistical reporting. Report the mean and standard deviation across independent training seeds, preferably at least three, and compute paired evaluation intervals from raw predictions, keeping evaluation uncertainty separate from training-seed variability because resampling benchmark entries cannot recover the latter. Match checkpoint-selection rules and record evaluator versions, prompt formats, decoding parameters, and metric normalizations, none of which are supplied in the current archive.

Current analytic coverage. To verify the analytic quantities in the mechanism plots, `scripts/verify_math.py` checks Gaussian means and posterior intervals by integration, risk boundaries and moments by count enumeration, and binary alignment on a finite response space. Finite differences additionally check the calibration-to-gradient calculation on 256 response groups and all 80 plotted Bernoulli cases, establishing agreement between the numerical illustrations and the theory, while checkpoint provenance is assessed separately in Appendix D.

E.1 Complete controlled Bernoulli ablation

For a success count m, let $q = m/N$, so the detached, one-token group gradient is $g = (1 - q)A_0(-p) + qA_1(1 - p)$, with unobserved outcomes contributing zero and therefore requiring no evaluation of their interval representatives. All four variants are evaluated without clipping or length weighting against the common reference $g_\star = \varphi(\Phi^{-1}(p))$, obtained from population Gaussian interval means at the same policy, which differs in scale from the raw success-reward gradient

$p(1-p)$. Table 6 below gives every setting, while the full-precision record additionally separates squared bias, variance, and homogeneous-group contributions.

Enumerating every count $m = 0, \ldots, N$ and weighting it by its binomial probability attributes the difference between D and D_0 entirely to the two homogeneous counts. To check the calculation, `scripts/compute_bernoulli_ablation.py` verifies probability normalization, independent Gaussian integration, closed-form gradients, and the bias–variance identity, while central finite differences of the detached importance-weighted objective check 4,788 count/method cases with maximum absolute gradient discrepancy 1.80×10^{-11}. Full-precision values and per-count contributions are stored in `data/bernoulli_ablation.json`, with the resulting bias and variance describing this controlled sampling model rather than variation across language-model training seeds.

Table 6: Complete exact binary ablation. Rank, Exact, D, and D_0 report normalized update MSE. Mass ratio compares posterior with empirical probability MSE, and P_{hom} is the probability of an all-equal group.

N	p	Rank	Exact	D	D_0	Mass ratio	P_{hom}
8	0.01	4.07586	4.61769	3.87062	3.951	3.18543	0.9227
8	0.02	2.35948	2.64583	2.23057	2.31197	1.95112	0.8508
8	0.05	1.0996	1.20521	1.02754	1.10171	1.21118	0.6634
8	0.10	0.556854	0.592273	0.511185	0.567457	0.96571	0.4305
8	0.20	0.223259	0.224552	0.196227	0.223481	0.84568	0.1678
8	0.35	0.0745205	0.0668147	0.0570204	0.0638529	0.79989	0.03209
8	0.50	0.0412274	0.0325802	0.0259714	0.0281234	0.79012	0.007813
32	0.01	1.8285	1.95176	1.75386	1.7748	1.65297	0.725
32	0.02	0.975675	1.02544	0.939947	0.9566	1.28573	0.5239
32	0.05	0.376718	0.384219	0.366229	0.373445	1.06558	0.1937
32	0.10	0.155547	0.155097	0.152296	0.153797	0.99255	0.03434
32	0.20	0.0486281	0.0477583	0.0475197	0.047563	0.95684	0.0007923
32	0.35	0.0102392	0.00965714	0.00956722	0.00956729	0.94322	1.031e-06
32	0.50	0.00236275	0.00184614	0.00176185	0.00176185	0.94031	4.657e-10
128	0.01	0.575239	0.587177	0.56532	0.567712	1.17110	0.2763
128	0.02	0.276369	0.27785	0.273596	0.274314	1.07497	0.07532
128	0.05	0.0950723	0.0948364	0.0946886	0.0947044	1.01735	0.001408
128	0.10	0.0378311	0.0377157	0.0377191	0.0377192	0.99823	1.39e-06
128	0.20	0.0114977	0.0114453	0.0114495	0.0114495	0.98888	3.94e-13
128	0.35	0.00207573	0.00204343	0.00204317	0.00204317	0.98532	1.13e-24
128	0.50	0.000141929	0.000113525	0.000112235	0.000112235	0.98456	5.877e-39

E.2 Exact multinomial ablation

This experiment uses a one-token categorical policy with three ordered verifier outcomes and softmax logits θ, so a response in outcome k has score vector $e_k - \mathbf{p}$. For each fixed positive population vector $\mathbf{p}$, we enumerate all count vectors $\mathbf{n} = (n_1, n_2, n_3)$ with total N, using $\widehat{\mathbf{p}} = \mathbf{n}/N$ and the exact multinomial weight

$$P(\mathbf{n} \mid \mathbf{p}) = \frac{N!}{n_1! n_2! n_3!} \prod_{k=1}^{3} p_k^{n_k}. \tag{E.2}$$

There are 45, 153, and 561 count vectors for $N = 8, 16, 32$. We compare tied Gaussian ranks, exact empirical interval means, posterior interval means with $\alpha = 1/2$, and the posterior rule with all-equal groups set to zero. Unobserved outcomes retain posterior mass but contribute no sampled advantage terms, while the score correction still retains all policy coordinates. For any detached advantage vector $A(\mathbf{n})$, the group update, population reference, and normalized risk are

$$g(\mathbf{n}; \mathbf{p}) = \widehat{\mathbf{p}} \odot A(\mathbf{n}) - \mathbf{p}\big[\widehat{\mathbf{p}}^{\top} A(\mathbf{n})\big], \tag{E.3}$$

$$g_\star(\mathbf{p}) = \mathbf{p} \odot T(\mathbf{p}), \tag{E.4}$$

$$\mathcal{R}(A; \mathbf{p}, N) = \frac{\sum_{\mathbf{n}} P(\mathbf{n} \mid \mathbf{p}) \|g(\mathbf{n}; \mathbf{p}) - g_\star(\mathbf{p})\|_2^2}{\|g_\star(\mathbf{p})\|_2^2}. \tag{E.5}$$

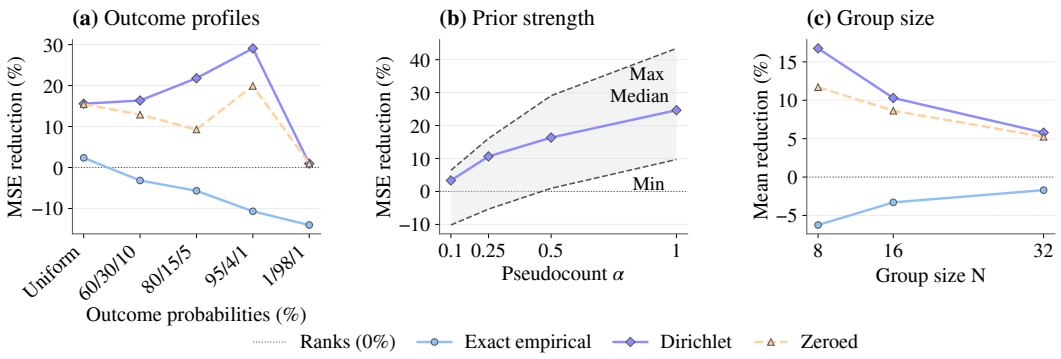


Figure 10: Exact three-outcome ablation. (a) Update-MSE reductions at $N = 8$. (b) Prior-strength sensitivity at $N = 8$ across five distinct profiles, showing their median and range. (c) Mean reductions as group size increases. All panels use tied Gaussian ranks as the reference.

The target follows from $\sum_k p_k T_k(\mathbf{p}) = 0$. All comparisons use the same logit coordinates, sampled count vectors, and target, so the risk measures recovery of the Gaussian-interval vector update rather than convergence or downstream reward improvement.

The five primary profiles are $(1/3, 1/3, 1/3)$, $(.60, .30, .10)$, $(.80, .15, .05)$, $(.95, .04, .01)$, and $(.01, .98, .01)$. Two additional profiles, $(.10, .30, .60)$ and $(.01, .04, .95)$, reverse the second and fourth vectors. Reflection leaves the squared risk unchanged and is used as a consistency check, so these are seven tested profiles but only five distinct profiles up to reversal. Figure 10 uses the five distinct profiles in its summaries to avoid counting a symmetry twice.

At $N = 8$, we also use posterior probabilities $(n_k + \alpha)/(N + 3\alpha)$ for $\alpha \in \{.1, .25, .5, 1\}$. The same interval transform, score vectors, and target are used throughout, so this sweep changes only the probability estimate. The default half-count improves probability MSE in 12 of 21 configurations and update MSE in all 21, further demonstrating that these criteria need not rank estimators identically. The smaller counts .1 and .25 worsen update MSE against tied ranks for the middle-concentrated profile, while $\alpha = 1$ is best among the tested choices. This finite sweep does not imply a universally optimal prior.

The multiclass computation script stores each setting's population vector, count-space size, squared bias, variance, update MSE, and probability MSE in `data/multiclass_ablation.json`. Numerical checks verify probability normalization, direct score-vector construction of the target, the bias–variance decomposition, reversal symmetry, and finite differences of detached importance-weighted objectives. The experiment uses exact count enumeration and contains no sampled training runs or inferred benchmark scores.

E.3 Random-task calibration audit

The fixed-profile ablations evaluate interpretable cases, while the random-task study tests whether the conclusions persist across a broader range of discrete-feedback distributions. Before computing results, we fix $K \in \{2, 3, 5, 8\}$, $\beta \in \{.2, .5, 1, 2, 5\}$, and $N \in \{4, 8, 16, 32, 64\}$. For each (K, β) pair, we draw 512 independent populations $\mathbf{p} \sim \mathrm{Dirichlet}(\beta \mathbf{1}_K)$ and reuse them across the five group sizes, drawing 128 multinomial groups per population and group size. There are 10,240 distinct populations and 6,553,600 sampled groups. The data-generating concentration β changes the prevalence of imbalanced populations and is separate from the fixed estimator pseudocount $\alpha = 1/2$. Since the total prior strength is $K/2$, changing K tests the specified method as a whole rather than isolating support size at fixed total smoothing.

We compare tied Gaussian ranks, exact empirical interval means, posterior interval means, and posterior means with all-equal-group updates removed. Their sampled gradient is $g(\mathbf{n}; \mathbf{p}) = \widehat{\mathbf{p}} \odot A(\mathbf{n}) - \mathbf{p}[\widehat{\mathbf{p}}^\top A(\mathbf{n})]$, with population reference $g_\star(\mathbf{p}) = \mathbf{p} \odot T(\mathbf{p})$. Unobserved outcomes contribute no sampled advantage terms, while the score correction retains all policy coordinates. For each

configuration, the reported normalized update risk is

$$\mathcal{R}_{K,\beta,N}(A) = \frac{\mathbb{E}_{\mathbf{p},\mathbf{n}}\|g(\mathbf{n};\mathbf{p}) - g_\star(\mathbf{p})\|_2^2}{\mathbb{E}_{\mathbf{p}}\|g_\star(\mathbf{p})\|_2^2}. \tag{E.6}$$

Normalizing after averaging avoids dividing individual errors by almost-zero target energies. Relative improvements compare mean squared errors on the same populations and response groups. GRPO is excluded from this particular risk comparison because it targets standardized rewards rather than the Gaussian population update, and comparing it against a different rule's target would not establish policy effectiveness.

We use 32 independent blocks of 16 populations, retaining every within-population group and every method comparison in its block. Pointwise 95% intervals use 2,000 paired block-bootstrap resamples. Resampling whole blocks preserves the dependence between methods and between group sizes induced by reusing populations. These intervals quantify uncertainty in the specified random-task ensemble and are not simultaneous confidence statements over 100 hypotheses or estimates of model-training variation. Figure 8(a) instead shows medians and full ranges across the 20 support/concentration regimes, which describe regime variation rather than confidence intervals.

All 100 update-risk point estimates improve against both tied ranks and exact empirical intervals, and every pointwise interval excludes zero. Median reductions across all configurations are 7.29% against tied ranks and 7.54% against exact empirical intervals. At $N = 8$, improvements over tied ranks span 10.60%–24.76% across all four support sizes and five concentrations. These comparisons support finite-group recovery of the Gaussian update over the declared ensemble, including distributions that differ from the estimator's prior, without asserting a universally optimal probability estimate or reward objective.

Probability risk provides an independent analytic check. For populations drawn from $\mathrm{Dirichlet}(\beta\mathbf{1}_K)$ and posterior pseudocount α, averaging Proposition 1 over populations yields

$$\mathbb{E}\|\widehat{\mathbf{p}} - \mathbf{p}\|_2^2 = \frac{(K-1)\beta}{N(K\beta+1)}, \tag{E.7}$$

$$\mathbb{E}\|\widetilde{\mathbf{p}} - \mathbf{p}\|_2^2 = \frac{(K-1)(N\beta + K\alpha^2)}{(K\beta+1)(N+K\alpha)^2}. \tag{E.8}$$

With $\alpha = 1/2$, probability risk improves precisely when $\beta > N/(4N+K)$. Across this grid, the analytic comparison gives 82 improvements, 17 degradations, and one equality at $K = 8, \beta = .2, N = 8$. The sampled positive estimate in that equality case has an interval containing zero, so it is not an additional established probability-risk improvement. Figure 8(c) retains these mismatched-prior cases and shows why probability risk alone does not determine the error after the Gaussian transform.

The all-equal-group ablation is also retained in full. The complete rule has lower point-estimate update MSE than its zeroed variant in 81 configurations, with 74 positive intervals and one negative interval. The largest positive reduction is 33.83%, while the most negative point estimate is -0.040%, so the dominant effect is favorable but is neither universal nor large in every regime. Raw integration error, probability MSE, squared Hellinger error, homogeneous-group frequency, and every method's update risk are saved for all 100 configurations rather than only those used in summary plots.

The vectorized implementation agrees with the independent scalar estimator to 8.89×10^{-16}, and finite differences of the detached surrogate agree with its gradients to 9.31×10^{-12}. Independent checks reproduce all bootstrap endpoints exactly and compare probability risks with both their conditional and hierarchical closed forms. The source package includes `scripts/random_task_audit.py`, `scripts/verify_random_task_audit.py`, the fixed design, all configuration results, and the 32-block statistics. The local design record precedes the computation and is not an external preregistration.

E.4 Exact one-step policy audit

Nonzero advantages alone do not show that a group provides a useful policy update, so we evaluate the actual reward change from every possible response group. We use binary success probabilities $p \in \{.001, .01, .05, .2, .5, .9\}$ and the five three-outcome profiles from the learning experiment, each

Table 7: Exact one-step diagnostics, averaged equally over all 44 configurations. Reward gains and mean KL are multiplied by 10^3. The KL cap rescales nonzero steps and leaves zero steps unchanged.

	Native step		KL-capped diagnostic		
Method	ΔJ	KL	ΔJ	KL	Zero step (%)
GRPO	6.130	0.291	8.466	0.602	39.75
G^{2}RPO	4.536	0.165	8.438	0.602	39.75
Exact empirical	4.723	0.179	8.443	0.602	39.75
Dirichlet-GRPO	4.740	0.172	10.750	0.931	6.91
Dirichlet-GRPO (zeroed)	4.673	0.172	8.480	0.602	39.75

at $N \in \{4, 8, 16, 32\}$. All 44 configurations are fixed before calculation. For ordered rewards $v_k = (k-1)/(K-1)$, we enumerate multinomial count vectors exactly and compare GRPO, tied ranks, exact empirical intervals, the posterior rule, and its zeroed variant. The native update is

$$\theta' = \log \mathbf{p} + 0.1 g(\mathbf{n}; \mathbf{p}), \qquad \Delta J = \mathrm{softmax}(\theta')^\top v - \mathbf{p}^\top v. \tag{E.9}$$

We record the exact expectation of ΔJ, actual forward KL, and the probability of a zero, positive, or negative reward change.

A separate direction-availability diagnostic rescales each nonzero gradient by a positive coefficient c until $D_{\mathrm{KL}}(\mathbf{p} \| \mathrm{softmax}(\log \mathbf{p} + cg)) = 10^{-3}$, while zero gradients leave the policy unchanged. This imposes the same per-group KL cap rather than equal average realized KL, because stalled baselines use no budget. It also changes the native learning rate, sometimes substantially for tiny rare-event gradients, so its result must not be presented as the performance of the unchanged algorithm.

Table 7 places both update modes side by side. At the common native learning rate, the posterior rule improves mean reward gain over tied ranks in 30 of 44 configurations, but GRPO gives larger gains in every configuration. Retaining homogeneous groups gives small positive native reward contributions in all binary cases, without establishing faster learning. Under the common KL cap, the posterior rule uses more of the available budget and has a larger average reward gain than tied ranks. Homogeneous extreme-outcome groups account for 97.5%–99.2% of the additional reward gain over tied ranks across group sizes, identifying update availability as the main source rather than a general improvement in nonzero-gradient direction. Middle-only groups remain stalled under a symmetric prior.

All individual configurations, effective scaling factors, homogeneous contributions, and actual KL values are stored in `data/one_step_policy_audit.json`. The exact multinomial probabilities sum to one within 1.22×10^{-14}, and an independent finite-difference check agrees with the detached gradient within 2.09×10^{-11}. These results complement the update-risk study by separating estimation accuracy, available directions, and applied step size.

E.5 Synthetic multi-step learning

To test repeated optimization rather than a single gradient estimate, we simulate independent three-action softmax policies $\pi_c = \mathrm{softmax}(\theta_c)$ with true rewards $v = (0, \frac{1}{2}, 1)$. Each of the five profiles in Appendix E.2 initializes four separate contexts, giving 20 contexts with independent parameters and sampling streams. The initial mean expected reward is 0.281. At every optimizer step, each context samples $M = 32$ actions and partitions them into groups of size $N \in \{4, 8, 16, 32\}$ to compute advantages. For sampled action $a_{c,i}$ and observed verifier label $\widetilde{r}_{c,i}$, the update is

$$\theta_c \leftarrow \theta_c + \frac{\eta}{M} \sum_{i=1}^{M} A(\widetilde{r}_{c,i}) \big(e_{a_{c,i}} - \pi_c \big). \tag{E.10}$$

Here the advantages are held fixed within an update, and their group composition is determined by the observed labels. All 32 score terms are averaged before updating, so changing N changes group composition without changing the optimizer batch, update count, or response budget. We use 200 SGD-ascent updates and 6,400 responses per context in every run, without clipping, KL regularization, or additional advantage normalization. The Dirichlet variants use $\alpha = 1/2$ for each of the three declared categories, including unobserved ones. The zeroed variant differs only by setting all advantages to zero when every observed label in a group is equal.

Table 8: Mean final synthetic reward over 30 seeds. Every setting uses 200 updates and 6,400 responses per context. D_0 zeroes all-equal-group advantages.

N	ϵ	η	GRPO	Rank	Exact	D	D_0
8	0	0.1	0.8067	0.7829	0.7855	0.7839	0.7821
8	0.1	0.1	0.7884	0.7710	0.7748	0.7723	0.7709
8	0.2	0.1	0.7733	0.7405	0.7492	0.7440	0.7421
4	0	0.1	0.7868	0.7680	0.7750	0.7756	0.7653
16	0	0.1	0.8757	0.7897	0.7920	0.7898	0.7893
32	0	0.1	0.9519	0.7966	0.8003	0.7966	0.7964
8	0	0.03	0.5969	0.5242	0.5350	0.5302	0.5280
8	0	0.3	0.9960	0.9848	0.9890	0.9845	0.9835

For verifier-noise experiments, each observed label is replaced with probability $\epsilon \in \{0, .1, .2\}$ by one of the other two labels with equal probability. Thus ϵ is the actual mislabeling rate. Advantages depend on the corrupted label, but the score vector in Eq. E.10 still uses the original sampled action. Evaluation uses the exact clean reward $J_c = \pi_c^\top v$, avoiding evaluation sampling noise and separating learning from the corruption used for training.

All runs use 30 seeds and common action, corruption, and replacement uniforms across methods. Policies may diverge, so common uniforms produce method-specific on-policy actions rather than forcing identical actions. For each seed, rewards are first averaged across the 20 fixed contexts, and two-sided 95% Monte Carlo intervals are then computed from the 30 seed means using a Student-t critical value with 29 degrees of freedom. Paired differences are formed within seed before computing their intervals. These pointwise intervals describe sampling variation for the fixed synthetic contexts and do not estimate variability across language-model training runs or task distributions.

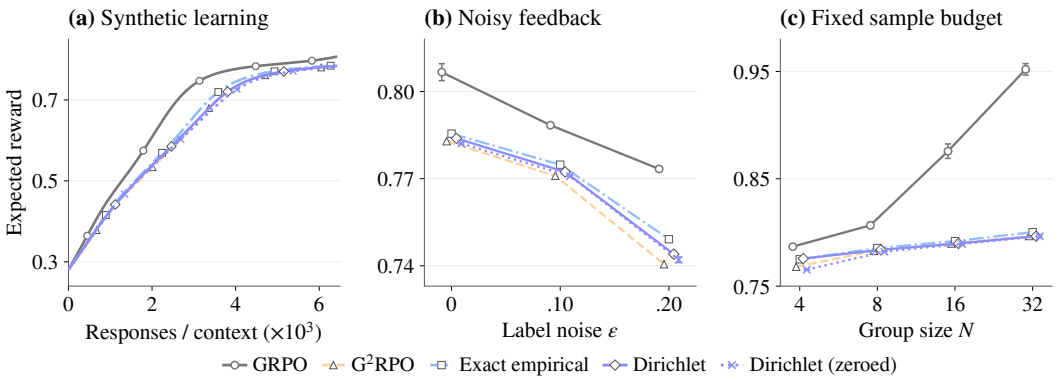


Figure 11: Synthetic learning with a fixed response budget. (a) Clean learning at $N = 8$. (b) Final reward under label errors at $N = 8$. (c) Group-size sensitivity. Bands and error bars show 95% Monte Carlo intervals across 30 seeds.

Figure 11 shows the complete learning curves, noise sweep, and group-size comparison. Table 8 reports all eight prespecified settings. At the primary setting $N = 8$, $\epsilon = 0$, and $\eta = .1$, the paired reward difference between Dirichlet-GRPO and tied ranks is 0.000982 with 95% interval $[0.000896, 0.001069]$. Removing all-equal-group updates lowers reward by 0.001758 with interval $[0.001690, 0.001825]$, while the difference from GRPO is -0.022792 with interval $[-0.025616, -0.019968]$. At $N = 4$, the benefit over tied ranks increases to 0.007626 with interval $[0.007473, 0.007779]$, and the benefit over exact empirical intervals is 0.000650 with interval $[0.000561, 0.000740]$. At $N = 16$ and 32, the intervals for the difference from tied ranks contain zero. The Gaussian-family comparison therefore favors the correction most clearly at the smallest tested group size, while GRPO has the highest mean reward in every setting.

A shared learning rate compares both the shape and scale of the advantages, so Figure 12 includes the complete clean-$N = 8$ sweep over $\eta \in \{.03, .1, .3\}$. All methods improve at .3, but Dirichlet-GRPO reaches 0.9845 compared with 0.9848 for tied ranks, 0.9890 for exact empirical intervals, and 0.9960 for GRPO. The difference from tied ranks is -0.000329 with interval $[-0.000636, -0.000022]$.

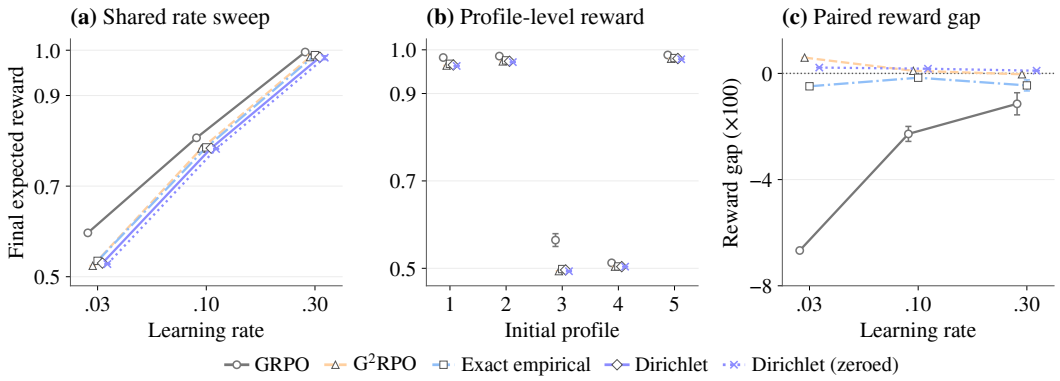


Figure 12: Synthetic learning sensitivity at $N = 8$ without label noise. (a) Shared learning rates. (b) Initial profiles at $\eta = .1$. (c) Dirichlet minus each baseline, multiplied by 100. Intervals use 30 seeds.

In Figure 12(b), profiles 1–5 denote $(.8, .15, .05)$, $(.6, .3, .1)$, $(.95, .04, .01)$, $(.01, .98, .01)$, and $(1/3, 1/3, 1/3)$, respectively. Their rewards at $\eta = .1$ show that the aggregate hides slower learning from the rare-success and middle-concentrated initializations. These results demonstrate why lower error against a Gaussian population update is insufficient to establish a reward-optimal learning rule. The sweep is a common sensitivity analysis, not separate optimizer tuning for each method.

The experiment script writes its design manifest before running the fixed scenario grid, retains all outcomes, and saves complete per-seed reward trajectories and final policy probabilities. This is a local prespecification rather than an external preregistration. Numerical checks verify all-equal-group behavior, probability normalization, action-score gradients under corrupted feedback, and deterministic reuse of random streams. The archive contains `scripts/simulate_learning.py`, `data/learning_simulation_design.json`, `data/learning_simulation.json`, and `data/learning_simulation_seeds.npz`; no measured language-model benchmark scores are generated by this simulation.

E.6 Remaining training comparisons

Model-level ablations should compare tied ranks, exact empirical means, posterior means, and the posterior rule with all-equal-group updates removed, using identical prompts, verifier support, initialization, loss aggregation, and rollout budgets. Prior-count and group-size sweeps should likewise use matched generated-token budgets, since additional sampling changes the comparison. At least three independent training seeds per setting, raw evaluation predictions, and fixed checkpoint-selection rules would allow training variation to be separated from evaluation uncertainty. These model-training results remain unmeasured. The exact calculations and tabular learning simulations provide controlled evidence at the estimator and small-policy levels, while those matched model-training runs remain necessary.

F Extended Related Work

We distinguish related methods by their evidence, credit granularity, and host objective to identify suitable controls.

Feedback-based policy optimization. RLHF and preference optimization use human feedback or preferences, whereas RLVR uses checked outcomes and GRPO estimates group advantages in a clipped surrogate (Ziegler et al., 2019; Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024; Schulman et al., 2017). Dr.GRPO examines length and difficulty weighting, DAPO changes sampling, clipping, and token-loss aggregation, and GSPO weights and clips sequence ratios (Liu et al., 2025; Yu et al., 2025; Zheng et al., 2025). Length-only weighting trades trajectory-gradient correctness against length neutrality under the model of Ding et al. (2026), motivating our fixed host objective to separate aggregation from probability estimation.

Gaussian ranks and interval representatives. G^2RPO assigns Gaussian quantiles to ordered feedback and averages them within ties, potentially using a task-level reference pool (Hu et al., 2026). For a fixed response pool, the empirical outcome frequencies define probability intervals, and the tied quantiles approximate their means by finite midpoint integration. Equation B.21 therefore separates the removal of integration error through exact interval means from changes in interval boundaries through posterior probability estimates. Our controlled comparison holds the pool fixed to isolate these changes, while matched training must additionally align the reference pool and normalization used by the published baseline. The resulting tied advantages remain discrete despite their Gaussian reference.

Advantage shape, difficulty, and reward dimensions. BNPO adapts a Beta-density normalizer for expected binary rewards, SoftmaxGRPO uses temperature-scaled softmax advantages, HA-DW weights historical difficulty, and GDPO normalizes reward dimensions before batch normalization (Xiao et al., 2025; Hernandez et al., 2026; Yang et al., 2026; Liu et al., 2026). Our theorem instead fixes a Gaussian reference and one ordered scalar outcome, leaving target selection and multi-objective aggregation open.

What finite-group analysis establishes. The U-statistic representation of Zhou et al. (2026) studies an unnormalized leave-one-out estimator, with further asymptotic conditions needed for practical normalization, whereas we study categorical risk and initial-update sensitivity for fixed responses. SIGNBALANCE uses verifier signs, global scaling, and classwise balancing against composition-sensitive reinforcement of guessed correct answers (Wang et al., 2026a); our estimator cannot distinguish guessing from grounded reasoning with the same outcome.

Probability estimates and baseline shrinkage. AERO uses a Beta prior for rollout allocation, EBPO learns its prior from global history, DBB discounts prompt-specific history, and EMA-GRPO tracks running rewards (Zhang et al., 2026; Han et al., 2026a; Kim et al., 2026; Feng et al., 2025). Frequentist shrinkage combines within- and across-prompt baselines (Zeng et al., 2025), while LeCritique/TETHER mixes group-relative and learned-value baselines according to critic accuracy (Venkatraman et al., 2026). Our fixed half-counts instead estimate the ordered outcome probabilities of one prompt and behavior policy, using the posterior mean to define the Gaussian intervals before exact integration. The contribution is the connection between probability risk, interval representation, and update sensitivity, rather than the standard extension from Beta to Dirichlet estimation; shrinkage can still increase repeated-sampling error when its bias exceeds its variance reduction.

Homogeneous groups and limited samples. DA3PO filters homogeneous groups and adjusts rare-success weights under dynamic sampling, while AVSPO normalizes equal-reward groups with virtual rewards (Gan et al., 2026; He et al., 2026). GDEPO combines capped resampling with correctness-anchored updates, and MDP-GRPO combines temperature-diverse sampling, absolute goal anchors, and bounded shaping (Yan et al., 2026; Salmani-Zarchi et al., 2026). Our half-counts reserve probability for unseen valid outcomes without new responses, yielding homogeneous-group shifts whose usefulness depends on the responses and objective; nonzero weights are neither unique nor sufficient evidence of improvement.

Different meanings of calibration. CAPO ranks confidence against correctness through a logistic AUC surrogate, GUPO aggregates queries using Dirichlet gradient-uncertainty evidence (Wang et al., 2026b; Guo et al., 2026b), and UCPO scales expressed-uncertainty advantages by correctness while normalizing deterministic-response rewards within their subset (Zeng et al., 2026). Our fixed-prompt, fixed-policy outcome probabilities establish neither verbal calibration nor prompt-gradient agreement, and abstention still requires a justified reward ordering. Table 9 contrasts these statistical objects, showing why a shared Bayesian or Dirichlet formulation need not imply the same response weights or policy update.

Credit within a response. Blockwise estimation assigns objective-specific advantages to text blocks, OPPO estimates success from oracle-conditioned prefix likelihoods, and EP-GRPO uses entropy progress and policy/reference divergence (Pavlenko et al., 2026; Li et al., 2026; Yu et al., 2026). CE-GRPO compares continuations at shared reasoning prefixes, while PCD combines downstream failure with teacher disagreement for perception distillation (Guo et al., 2026a; Xiong

et al., 2026). These additional signals localize credit beyond the one-value-per-outcome representation analyzed here.

Visual evidence and verifier reliability. PAPO contrasts original and masked inputs, KAWHI and SRPO use saliency or perception/reasoning roles, and NoisyGRPO combines visual-noise priors with trajectory rewards (Wang et al., 2025; Han et al., 2026b; Jiang & Zou, 2026; Qiu et al., 2025). SVRL trains agents to verify retrieved evidence and rewards informative search (Sathish et al., 2026). These methods enrich feedback, whereas ours reweights fixed outcome counts and cannot repair unreliable verification, whose optimization may raise proxy rewards without improving performance (Yao et al., 2026).

Scope and comparisons still needed. The archive covers one family at three scales, so further comparisons require repeated runs matched on evidence, support, rollout budgets, and prefix/visual costs (Appendix E). Ordered outcomes remain necessary and larger supports increase prior influence, while neither local bounds nor restricted binary alignment guarantees training gains.

Table 9: **Related methods intervene on different statistical objects.** Entries summarize the cited constructions, not a performance ranking. The checkpoint sheet compares only GRPO, G^{2}RPO, and the proposed-method label.

Method	Object or evidence	How it changes the update
G^{2}RPO (Hu et al., 2026)	Empirical ranks in a task reference pool	Average Gaussian quantiles within ties
BNPO / SoftmaxGRPO (Xiao et al., 2025; Hernandez et al., 2026)	Binary reward expectations / sampled scalar rewards	Adapt the normalizer / apply a softmax advantage rule
AERO (Zhang et al., 2026)	Beta posterior for binary success probability	Combine posterior estimation with adaptive rollout selection
EBPO / DBB (Han et al., 2026a; Kim et al., 2026)	Global reward history / discounted per-prompt history	Shrink the baseline / estimate evolving reward probabilities
NoisyGRPO (Qiu et al., 2025)	Visual-noise information and trajectory rewards	Form a posterior trajectory advantage
AVSPO (He et al., 2026)	Observed advantage collapse and virtual rewards	Modify normalization in homogeneous groups
CAPO (Wang et al., 2026b)	Confidence and correctness ordering	Use an AUC surrogate in advantage construction
GUPO (Guo et al., 2026b)	Gradient uncertainty mapped to Dirichlet evidence	Weight group gradients during aggregation
Dirichlet-GRPO	Categorical counts on a declared ordered support	Use posterior masses and exact Gaussian interval means

Statistical and representation foundations. Jeffreys' prior and ordered transport supply established ingredients (Jeffreys, 1946; Villani, 2003), while Eq. B.21 isolates corrections to G^{2}RPO (Hu et al., 2026). These optima imply neither Bayes-optimal training nor equal task gradients, and concern categorical estimation and finite verifier resolution rather than the Koopman setting of Cheng et al. (2026).

G AI Use Statement

We disclose the use of GPT-6-Astra to assist with language refinement and manuscript preparation, including literature and code analysis, mathematical formulation and proof drafting, interpretation of supplied results, and visualization preparation. The measured training traces and evaluation results were supplied by the authors, and AI-assisted analysis did not replace model training or evaluation. The authors retain responsibility for the Dirichlet-GRPO method and implementation, verification of mathematical arguments and empirical claims, and all research decisions, as well as the accuracy, originality, and integrity of the final manuscript and accompanying artifacts.