RAC-GRPO: RISK-CONSTRAINED ADVANTAGE CALIBRATION FOR FINITE ROLLOUT GROUPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Group Relative Policy Optimization (GRPO) constructs advantages from reward statistics estimated within finite rollout groups. As a result, its training weights are sensitive to finite-group sampling. With binary rewards, this issue is most extreme when all sampled responses are correct or all are wrong, in which case GRPO assigns zero advantage to the entire group. Mixed groups are also affected because their reward means and standard deviations are estimated from only a few samples. We introduce Risk-Constrained Advantage Calibration for GRPO (RAC-GRPO), a stateless calibration algorithm based on the observed success count and group size. It adjusts the reward statistics by adding symmetric virtual success and failure counts. A reference posterior describes uncertainty in the underlying success probability. A mean squared error constraint limits the allowable correction strength. Within this constraint and a fixed cap, RAC-GRPO selects the largest admissible correction. The resulting advantages decompose into an attenuated within-group contrast and a shared offset. This recovers small updates for all-correct and all-wrong groups while also calibrating mixed groups. Under independent on-policy sampling with equally weighted responses, we characterize the expected update. We show that endpoint recovery is aligned with the success gradient. The recovered component's mean and directional standard deviation both scale linearly with recovery magnitude. Training requires only a precomputed lookup table, with no additional responses, reward history, or learned critic. Across seven mathematical reasoning benchmarks, RAC-GRPO improves over GRPO by 2.53, 2.08, and 2.82 percentage points on Qwen3-1.7B, 4B, and 8B, respectively.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) trains language models from automatically checked outcomes (Guo et al., 2025; Team et al., 2025; Hu et al., 2025). For each prompt, the model samples a rollout group and receives binary rewards from a verifier. Group Relative Policy Optimization (GRPO) avoids a learned critic by constructing response advantages from the group's empirical reward mean and standard deviation (Shao et al., 2024). Because these statistics are estimated from only a finite number of responses, the resulting training weights depend directly on the sampled group composition.

When all responses receive the same reward, this group-relative construction loses its reward-derived training signal. For example, if all eight responses are wrong, every reward and the group mean are zero, so every advantage is zero. The same cancellation occurs when all eight responses are correct. We call these *zero-variance groups*. They do not imply that the model always fails or succeeds. With a true success probability of 0.1 and group size eight, a zero-variance group occurs with probability 43.05%.

Mixed groups are affected more gradually by the same finite-sample issue. If only one of eight responses is correct, the positive weight on that response and the seven negative weights depend on the reward mean and standard deviation estimated from those eight samples. Zero-variance collapse is therefore the endpoint case of a broader finite-group calibration problem. This motivates the question: *how strongly should reward statistics be corrected at each observed success count?*

Existing approaches use additional sampling, virtual rewards, or historical statistics to recover signals from zero-variance groups (Yu et al., 2025; He et al., 2026; Kim et al., 2026; Han et al., 2026). We

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.

introduce RAC-GRPO, a stateless calibration algorithm that uses only the current success count K and group size G. It adds the same virtual count, or *pseudo-count*, to successes and failures when computing reward statistics. This moves the estimated success rate toward $1/2$ and produces nonzero advantages even when all observed rewards agree. A fixed pseudo-count ties endpoint recovery to the adjustment applied in mixed groups. RAC-GRPO instead chooses the pseudo-count from (K, G), allowing the correction strength to vary with the observed reward composition.

The remaining choice is the size of the pseudo-count. A reference posterior describes uncertainty about the underlying success probability. A mean squared error constraint limits the admissible correction strength. Within this constraint and a fixed cap, RAC-GRPO selects the largest admissible pseudo-count. The resulting advantages give all-correct groups small positive weights and all-wrong groups small negative weights. They also rescale and shift the weights in mixed groups. The probability-estimation risk serves only as a constraint on correction strength. Whether the calibrated advantages improve policy optimization is determined empirically.

Recovery strength also affects the scale of stochastic variation in the recovered update component. We therefore analyze both the expected update and its variation under independent on-policy sampling with equally weighted responses.

Our contributions are:

1. **Calibration under risk constraints.** We introduce a stateless count-dependent calibration algorithm that adjusts GRPO advantages across all observed success counts. A probability-estimation risk constraint limits the correction strength, without requiring additional rollouts, reward history, or a critic.
2. **Theory of calibrated updates.** We derive a contrast–offset decomposition of the calibrated advantages, characterize their expected on-policy update, and analyze the second moment of the recovered zero-variance component.
3. **Evaluation across model scales.** Across three seeds on Qwen3-1.7B, 4B, and 8B, RAC-GRPO improves the average over seven benchmarks compared with GRPO. Ablations show contributions from both mixed-group calibration and zero-variance recovery, and compare the count-dependent calibration with fixed correction strengths.

2 Related Work

Group objectives and reward normalization. Several alternatives to PPO derive training signals from sampled responses without a learned critic: ReMax uses greedy completions, RLOO uses leave-one-out response baselines, and GRPO centers and normalizes rewards within each sampled group (Schulman et al., 2017; Li et al., 2023; Ahmadian et al., 2024; Shao et al., 2024). Recent work studies how normalization, aggregation, and prompt weighting affect policy optimization with response groups, including Dr. GRPO, GSPO, and analyses of finite-group weighting and policy gradient structure (Liu et al., 2025; Zheng et al., 2025; Bay & Yearick, 2026; Mishra et al., 2026). SoftmaxGRPO replaces reward normalization with centered softmax weighting, but zero-variance groups still receive no reward-derived advantage (Hernandez et al., 2026). RAC-GRPO instead calibrates GRPO's reward mean and standard deviation as a function of the observed success count.

Recovering signals from zero-variance groups. Several methods recover learning signals from groups with zero empirical reward variance. DAPO, AR3PO, VIGOR, and AERO modify rollout selection or allocation to obtain more informative samples (Yu et al., 2025; Zhang et al., 2025b; Jiang et al., 2026; Zhang et al., 2026). AERO additionally applies a fixed $\mathrm{Beta}(1, 1)$ posterior mean to residual zero-variance groups together with iterative rescue and rejection sampling (Zhang et al., 2026). EDGE-GRPO changes response generation, while AVSPO augments normalization statistics with virtual rewards when collapse is detected (Zhang et al., 2025a; He et al., 2026). Nie et al. (2026) use a fixed sign-based advantage in every group and connect recovery in all-wrong groups to reducing pass@G failure probability. These methods primarily address zero-variance groups through fixed recovery rules, collapse-triggered corrections, or changes to sampling. In contrast, RAC-GRPO applies count-dependent calibration at every observed success count, treating zero-variance recovery as the endpoint case of the same finite-group calibration problem.

Statistical reward estimation and calibration. Statistical approaches estimate reward uncertainty or success probabilities to improve reward normalization or adaptive weighting. BNPO fits a Beta density to the distribution of success probabilities across prompts and uses the resulting statistics for normalization (Xiao et al., 2025). DBB maintains discounted statistics for each prompt, while EBPO constructs empirical priors from accumulated policy statistics; both connect reward estimation with recovery from zero-variance groups (Kim et al., 2026; Han et al., 2026). PSPO and RePO further exploit historical statistics or additional responses (Zhu et al., 2026; Li et al., 2025). CAPO targets predictive calibration, while OAR focuses on finer credit assignment (Wang et al., 2026; Li et al., 2026). RAC-GRPO instead determines a count-dependent correction from the current group alone, using reference uncertainty only to constrain its strength.

3 Problem Setup: Learning from Finite Rollout Groups

Let $\pi_{\mathrm{old}}(y \mid x)$ denote the model's probability of generating response y to prompt x when collecting training samples. For a fixed prompt, let $p_x = \Pr_{y\sim\pi_{\mathrm{old}}(\cdot|x)}[R(x,y)=1]$. Assume G conditionally independent responses sampled from $\pi_{\mathrm{old}}(\cdot \mid x)$ with binary verifier rewards $r_i \in \{0,1\}$. The observed success count and empirical rate are

$$K = \sum_{i=1}^{G} r_i \sim \mathrm{Binomial}(G, p_x), \qquad \widehat{p}_K = K/G. \tag{1}$$

Zero-variance groups correspond to $K \in \{0, G\}$, and mixed groups to $0 < K < G$. A mixed group is balanced when $K = G/2$ (for even G) and unbalanced otherwise. Here, zero variance refers to the observed group and does not imply $p_x \in \{0,1\}$.

For the analysis, we use the population standard-deviation convention for binary rewards (the implementation convention is recorded in Appendix C) and write GRPO as

$$\widehat{A}_i^{\mathrm{GRPO}} = \frac{r_i - \widehat{p}_K}{\widehat{s}_K + \epsilon_{\mathrm{num}}}, \qquad \widehat{s}_K = \sqrt{\widehat{p}_K(1-\widehat{p}_K)}. \tag{2}$$

For $0 < K < G$, successful responses receive positive advantages and failed responses receive negative advantages. For example, with $G = 8$, $K = 1$, and negligible ϵ_{num}, Equation 2 gives approximately $+2.65$ to the correct response and -0.38 to each wrong response. These magnitudes depend on the observed count through both the estimated reward mean and standard deviation. For $K \in \{0, G\}$, the reward residuals vanish, so every advantage and the resulting reward-derived update are zero under either standard-deviation convention, regardless of $\epsilon_{\mathrm{num}} > 0$.

The probability that a rollout group has zero empirical reward variance is

$$q_0(p_x; G) = (1-p_x)^G + p_x^G. \tag{3}$$

At $G = 8$, this probability reaches 16.78% at $p_x = 0.2$ and 43.05% at $p_x = 0.1$, with symmetric values near one. Zero-variance groups can therefore arise from finite-sample fluctuations even when $0 < p_x < 1$, and occur deterministically at the endpoints $p_x \in \{0,1\}$. These effects of finite rollout groups have been analyzed in recent studies of advantage collapse and optimization with group comparisons (He et al., 2026; Bay & Yearick, 2026). Appendix B illustrates the probability and a training trace. This motivates calibrating the reward statistics at every count, with zero-variance recovery as the endpoint case $K \in \{0, G\}$. Section 4.4 characterizes the expected update.

4 RAC-GRPO: Advantage Calibration under Risk Constraints

RAC-GRPO changes how rewards become training weights: count the correct responses, look up a correction strength, and use the adjusted reward mean and standard deviation to compute advantages. Figure 1 summarizes sampling, calibration, and advantage construction for three reward compositions; Figure 2 shows the mapping from counts to advantages used at $G = 8$.

4.1 Advantage Construction from Success Counts

We add the same pseudo-count $c_K = c(K, G) > 0$ to both reward outcomes, while each observed reward contributes a count of one. This changes only the statistics and requires no additional responses.

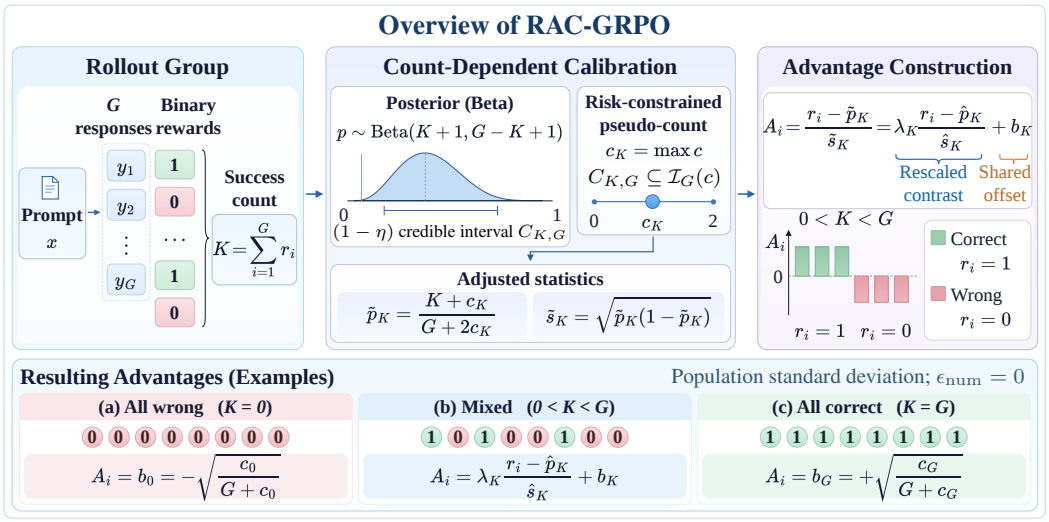


Figure 1: **Overview of RAC-GRPO. Top:** Count correct responses, choose the pseudo-count for successes and failures ($c_K \leq 2$), and compute adjusted reward statistics and training weights. **Bottom: (a,c)** All-wrong and all-correct groups receive negative and positive shared offsets, respectively; **(b)** mixed groups receive a rescaled contrast within each group plus a shared offset. The illustration uses the population standard deviation and $\epsilon_{\text{num}} = 0$.

The adjusted statistics and advantages are

$$\widetilde{p}_K = \frac{K + c_K}{G + 2c_K}, \qquad \widetilde{s}_K = \sqrt{\widetilde{p}_K(1 - \widetilde{p}_K)}, \qquad \widehat{A}_i^{\text{RAC}} = \frac{r_i - \widetilde{p}_K}{\widetilde{s}_K + \epsilon_{\text{num}}}. \tag{4}$$

For binary rewards, $\widetilde{p}_K$ estimates both the success probability and the reward mean. It lies in $(0, 1)$, keeping $\widetilde{s}_K > 0$ at every count. The pseudo-count controls how far the mean moves from K/G toward $1/2$. For c fixed before sampling, $\widetilde{p}_K$ is the posterior mean under a $\text{Beta}(c, c)$ prior. Reward normalization uses the Bernoulli plug-in variance $\widetilde{p}_K(1 - \widetilde{p}_K)$. For calibration, a separate reference posterior under a fixed uniform prior constrains c_K at each observed count (Section 4.3).

4.2 Decomposition into Contrast and Offset

Changing the reward standard deviation rescales differences between response weights; changing the mean shifts every weight in the group. Using the population standard deviation in Equation 2, we obtain a *decomposition into contrast and offset* by adding and subtracting $\widehat{p}_K$, with $D_K = \widetilde{s}_K + \epsilon_{\text{num}}$:

$$\widehat{A}_i^{\text{RAC}} = \underbrace{\frac{\widehat{s}_K + \epsilon_{\text{num}}}{D_K}}_{\lambda_K} \widehat{A}_i^{\text{GRPO}} + \underbrace{\frac{\widehat{p}_K - \widetilde{p}_K}{D_K}}_{b_K}. \tag{5}$$

The first term rescales the centered GRPO advantages; the second shifts every response weight by the same amount b_K. Figure 1 shows this decomposition for mixed groups in the upper right panel ($\epsilon_{\text{num}} = 0$, $A_i = \widehat{A}_i^{\text{RAC}}$) and for zero-variance groups in panels (a,c). Since shrinkage moves $\widehat{p}_K$ toward $1/2$, $0 < \lambda_K \leq 1$. The offset has the same sign as $2K - G$, agreeing with the majority reward in an unbalanced group. At a balanced count, the empirical mean is already $1/2$, so the calibrated advantages coincide with the GRPO advantages in Equation 2. For an unbalanced mixed group (Figure 1b), the shared offset can increase the magnitude of advantages for responses with the majority reward even as the relative contrast is attenuated. Responses with rewards of one and zero receive positive and negative advantages, respectively, because $0 < \widetilde{p}_K < 1$.

At endpoint counts (Figure 1a,c), the contrast component vanishes and each advantage equals the shared offset. Ignoring the numerical safeguard,

$$K = 0: \quad \widehat{A}_i^{\text{RAC}} = -\sqrt{\frac{c_0}{G + c_0}}, \qquad K = G: \quad \widehat{A}_i^{\text{RAC}} = +\sqrt{\frac{c_G}{G + c_G}}. \tag{6}$$

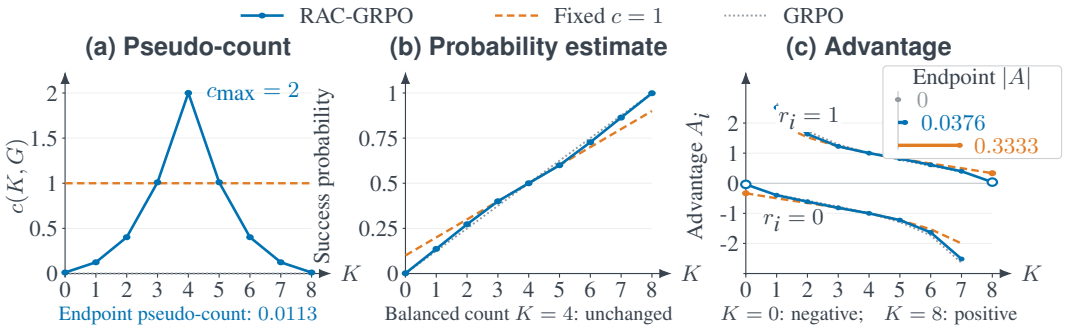


Figure 2: **Calibration at $G = 8$. (a)** Pseudo-counts, **(b)** probability estimates, and **(c)** advantages for RAC-GRPO, fixed $c = 1$, and GRPO, with endpoint magnitudes 0.0376, 0.3333, and 0 in the inset, respectively. Only observed reward classes are shown at each count, using the population standard deviation and $\epsilon_{\text{num}} = 0$; lines connect counts.

With the $G = 8$ lookup table, all-wrong and all-correct responses receive approximately -0.0376 and $+0.0376$, respectively, versus zero under GRPO and magnitudes of 0.3333 for fixed $c = 1$.

4.3 Correction Strength under a Risk Constraint

The shared offset in Equation 5 has magnitude

$$|b_K(c)| = \frac{|2K - G|c}{G\{\sqrt{(K + c)(G - K + c)} + (G + 2c)\epsilon_{\text{num}}\}}. \tag{7}$$

At unbalanced counts, $|b_K(c)|$ strictly increases and λ_K decreases with c (Appendix A.6).

We constrain correction strength using the mean squared error (MSE) of success probability estimates. Under a fixed uniform prior, the reference posterior $\text{Beta}(K + 1, G - K + 1)$ describes uncertainty after K successes in G responses. Its equal-tailed credible interval $C_{K,G} = [L_{\eta/2}, U_{1-\eta/2}]$ has mass $1 - \eta$. For fixed c, define $\mathcal{I}_G(c)$ as the set of success probabilities for which the estimator's repeated-sampling MSE is no greater than that of K/G. We require $C_{K,G} \subseteq \mathcal{I}_G(c)$, making this comparison hold throughout the reference interval. The cap $c \leq 2$ also guarantees no greater conditional risk than K/G under the reference posterior (Proposition 2). The uniform prior provides a fixed, symmetric reference for calibration without estimating dataset-level prior parameters. It does not assume that success probabilities are uniformly distributed across training prompts.

We select the largest pseudo-count satisfying these constraints, maximizing the shared offset for an unbalanced group (Equation 7). The risk constraint limits correction strength; it does not optimize the training objective or guarantee higher accuracy.

The constraint yields a closed-form calibration rule. Define

$$\ell_{K,G} = \min\{L_{\eta/2}, 1 - U_{1-\eta/2}\}, \quad q_{K,G} = \ell_{K,G}(1 - \ell_{K,G}), \quad d_{K,G} = G - 4(G + 1)q_{K,G}. \tag{8}$$

The quantity $\ell_{K,G}$ is the distance from the reference credible interval to the nearer endpoint of $[0, 1]$. The main setting uses $1 - \eta = 0.95$ and defines the ideal calibration rule

$$c^\star(K, G) = \begin{cases} \min\{2, 4Gq_{K,G}/d_{K,G}\}, & d_{K,G} > 0, \\ 2, & d_{K,G} \leq 0. \end{cases} \tag{9}$$

The rule is symmetric in K and $G - K$; intervals near zero or one allow less correction. At $G = 8$, endpoint pseudo-counts are small and interior values larger (Figure 2). Appendix A.4 derives the rule.

Implementation. We precompute the $G + 1$ pseudo-count lookup entries $\bar{c}(K, G)$ for each group size and retrieve one per prompt during training (Algorithm 1). Lookup precision and preserved risk guarantees are detailed in Appendix A.4. Training retains the baseline objective (Appendix C.1).

Algorithm 1 RAC-GRPO training with advantages calibrated by success count

Require: Initial policy π_θ, prompt set $\mathcal{D}$, binary verifier R, group size G
Ensure: Trained policy π_θ
Precompute $\{\bar{c}(K,G)\}_{K=0}^{G}$ using Equation 9 and Appendix C.
for each training iteration **do**
  Set $\theta_{\text{old}} \leftarrow \theta$; sample a prompt batch $\mathcal{B}$ from $\mathcal{D}$.
  for all $x \in \mathcal{B}$ **do**
    Sample $y_{1:G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$; evaluate $r_i = R(x, y_i) \in \{0, 1\}$.
    $K \leftarrow \sum_{i=1}^{G} r_i$; $c_K \leftarrow \bar{c}(K, G)$.
    $\widetilde{p}_K \leftarrow (K + c_K)/(G + 2c_K)$; $\widetilde{s}_K \leftarrow \sqrt{\widetilde{p}_K(1 - \widetilde{p}_K)}$.
    $\widehat{A}_i^{\text{RAC}} \leftarrow \text{stopgrad}((r_i - \widetilde{p}_K)/(\widetilde{s}_K + \epsilon_{\text{num}}))$, $i = 1, \ldots, G$.
  end for
  Update θ with the batch advantages and the baseline clipped objective (Appendix C.1).
end for
return π_θ

4.4 Expected Updates and Stochastic Variation

We analyze the expected update direction and the random variation introduced by recovery. For a fixed prompt, consider independent responses from the current policy at $\theta = \theta_{\text{old}}$, with success probability $p = p_\theta(x)$. The score function $s_\theta(y) = \nabla_\theta \log \pi_\theta(y \mid x)$ points toward increasing the sampled response's log probability. Writing $\overline{s}_\theta = G^{-1} \sum_i s_\theta(y_i)$, the equally weighted update is $\widehat{g} = G^{-1} \sum_i \widehat{A}_i^{\text{RAC}} s_\theta(y_i)$: each advantage weights its response's update direction. Equation 5 expresses this update as a rescaled GRPO contribution plus $b_K \overline{s}_\theta$. Although $\mathbb{E}[\overline{s}_\theta] = 0$, the offset depends on rewards and is correlated with the sampled score vectors, so its expected contribution need not vanish. Prior work also studies alignment of endpoint updates using a fixed reference or virtual rewards (Nie et al., 2026; He et al., 2026).

Proposition 1 (Expected update and zero-variance recovery) *Fix $G \geq 2$ and a prompt with $p \in (0, 1)$. Assume i.i.d. on-policy sampling, a binary verifier independent of θ, standard regularity conditions for the score function, and stop-gradient advantages. Then $\mathbb{E}[\widehat{g}] = \alpha_{\text{RAC}}(p) \nabla_\theta p$ for a positive coefficient $\alpha_{\text{RAC}}(p)$. Let $\widehat{g}_{\text{mix}}$ use identical advantages in mixed groups, zero endpoint advantages, and the same $1/G$ normalization. For symmetric endpoint advantages of common magnitude $a_{0,\epsilon} > 0$,*

$$\mathbb{E}[\widehat{g} - \widehat{g}_{\text{mix}}] = a_{0,\epsilon}[p^{G-1} + (1 - p)^{G-1}] \nabla_\theta p, \tag{10}$$

so the recovered contribution is aligned with the success gradient whenever that gradient is nonzero.

Appendix A.3 proves the result and derives α_{RAC} and $a_{0,\epsilon}$. The comparison fixes advantages in mixed groups, as in the *Mixed groups only* ablation (Table 2).

Recovery magnitude and stochastic variation. Write the isolated zero-variance group component as $H_a = a(\mathbf{1}_{\{K=G\}} - \mathbf{1}_{\{K=0\}})\overline{s}_\theta$, with $a = a_{0,\epsilon}$. At a fixed policy where score vectors have finite second moments, its mean and directional standard deviation scale linearly in a, while its second moment and covariance scale as a^2 (Appendix A.7). Positive rescaling preserves its signal-to-noise ratio whenever defined; the full covariance also involves mixed groups and cross-covariances.

At a locally saturated prompt, assume the verifier is constant over the policy support in a parameter neighborhood. Then $\nabla_\theta p = 0$ and

$$\mathbb{E}[H_a] = 0, \qquad \text{Cov}(H_a) = \frac{a^2}{G} F_x, \qquad F_x = \mathbb{E}[s_\theta(y) s_\theta(y)^\top]. \tag{11}$$

The recovered component is pure sampling noise when $F_x \neq 0$. A smaller endpoint magnitude limits its absolute size, while also reducing any nonzero expected recovery signal at nonsaturated prompts. An observed zero-variance group alone does not establish local saturation.

Scope. The analysis assumes equally weighted on-policy score vectors for one prompt. GRPO and variants with fixed positive pseudo-counts also have positive update coefficients; different coefficients reweight prompts and can change the aggregate gradient direction. Averaging over tokens, clipping, and repeated optimization are excluded (Appendix C.1).

Table 1: **Performance across Qwen3 model scales (%).** Rows showing only the model name report performance before RL. AIME24/25/26 and AMC23 use Avg@32 at temperature 0.2; the remaining benchmarks use greedy decoding. RLVR benchmark cells are means across three seeds; Average reports mean ± sample SD of the unweighted mean across seven benchmarks for each seed. Table 9 gives SDs for individual benchmarks.

Model / method	AIME24	AIME25	AIME26	AMC23	MATH500	Minerva	Olympiad	Average
Qwen3-1.7B	11.98	10.10	7.08	44.30	72.00	29.41	39.76	30.66
+ GRPO	31.32	29.53	21.21	70.83	85.60	35.42	57.47	47.34 ± 1.04
+ AVSPO	27.41	25.49	18.47	63.54	83.40	33.09	54.60	43.72 ± 0.57
+ DBB	31.74	27.99	22.36	72.63	87.27	36.64	59.99	48.37 ± 0.75
+ RAC-GRPO	**33.47**	**29.57**	**25.42**	**75.55**	**87.87**	**36.76**	**60.44**	$\mathbf{49.87 \pm 0.98}$
Qwen3-4B	19.79	18.85	14.79	65.08	85.20	39.34	51.93	42.14
+ GRPO	53.51	47.22	49.90	89.77	93.60	47.59	70.03	64.52 ± 0.47
+ AVSPO	46.53	41.46	42.01	87.16	92.80	45.47	65.68	60.16 ± 0.10
+ DBB	52.54	44.86	46.36	89.45	94.07	46.57	70.57	63.49 ± 0.80
+ RAC-GRPO	**57.88**	**47.78**	**52.71**	**90.73**	**95.18**	**49.06**	**72.85**	$\mathbf{66.60 \pm 0.44}$
Qwen3-8B	25.10	19.06	14.69	67.11	84.40	40.07	54.01	43.49
+ GRPO	60.52	47.15	50.21	91.17	94.33	46.81	70.42	65.80 ± 0.28
+ AVSPO	58.09	46.11	51.11	90.65	94.73	47.67	71.51	65.70 ± 0.50
+ DBB	58.16	46.08	50.97	90.16	94.47	48.28	72.06	65.74 ± 0.93
+ RAC-GRPO	**62.53**	**49.93**	**53.82**	**92.92**	**95.87**	**50.61**	**74.63**	$\mathbf{68.62 \pm 0.50}$

5 Experiments

5.1 Experimental Setup

We train Qwen3-1.7B, 4B, and 8B (Yang et al., 2025) for two epochs on deduplicated DAPO-Math-17K (Yu et al., 2025), using $G = 8$ responses per prompt unless otherwise specified. Paired runs share the initialization, prompt order, rollout budget, verifier, decoding constraints, optimization settings, and training seed. Training uses VERL (Sheng et al., 2024), with details in Appendix C.

Baselines and evaluation. We compare RAC-GRPO with GRPO, AVSPO, and Discounted Beta–Bernoulli (DBB). Our comparisons focus on advantage estimation under a common rollout budget, with sampling and optimization held fixed. We select AVSPO (He et al., 2026) and DBB (Kim et al., 2026) as recent baselines representing two complementary strategies: recovery with virtual rewards and statistical reward estimation from discounted history. AVSPO and DBB are independently reimplemented under the shared setup, with configurations and deviations from their original protocols recorded in Appendix C.4. Evaluation covers seven mathematical reasoning benchmarks: AIME24, AIME25, AIME26, AMC23, MATH500, Minerva, and Olympiad (Hendrycks et al., 2021; Lewkowycz et al., 2022; He et al., 2024). AIME24/25/26 and AMC23 report the fraction of correct responses averaged over 32 generations per problem at temperature 0.2 (Avg@32); the other benchmarks use greedy decoding. Average is the unweighted mean across the seven benchmarks, computed separately for each training seed. The main comparisons, Qwen3-8B ablations, fixed-c sweep, $G = 4$ comparison, and GSPO comparison use three training seeds per method. The credible-level experiments are exploratory and use one seed per non-default setting. Summaries across three seeds report mean ± sample standard deviation (SD).

5.2 Performance Across Model Scales

Table 1 shows that RAC-GRPO achieves the highest seven-benchmark average at all three model scales under the shared rollout budget. It improves on GRPO by 2.53, 2.08, and 2.82 percentage points on Qwen3-1.7B, 4B, and 8B, respectively. On Qwen3-4B, its average (66.60) also exceeds those of all three Qwen3-8B baselines (65.70–65.80).

The average gains are accompanied by improvements across all seven benchmarks: RAC-GRPO attains the highest mean in each of the 21 model–benchmark comparisons. The margins vary across tasks. For example, on Qwen3-8B, the gains over GRPO are 4.21 points on Olympiad and 3.80

points on Minerva, compared with 1.54 points on MATH500. Appendix D reports variability across benchmarks and training seeds.

5.3 Component Ablations and Calibration Sensitivity

On Qwen3-8B ($G = 8$), we study component contributions, compare fixed and count-dependent pseudo-counts, and isolate the effect of endpoint recovery strength.

Table 2: **Component ablations on Qwen3-8B ($G = 8$, %).** *Endpoint* denotes zero-variance groups; *Mixed* denotes groups with both reward outcomes. Three-seed means; Average reports mean ± sample SD. Bold marks column maxima; evaluation follows Table 1.

Method / component	AIME24	AIME25	AIME26	AMC23	MATH500	Minerva	Olympiad	Average
GRPO	60.52	47.15	50.21	91.17	94.33	46.81	70.42	65.80 ± 0.28
Endpoint groups only	61.88	47.67	53.26	91.41	95.20	47.67	73.24	67.19 ± 0.70
Mixed groups only	61.28	48.19	53.13	92.71	95.27	46.45	72.55	67.08 ± 0.47
Mixed + all-wrong	59.34	49.90	51.53	91.85	95.47	49.26	74.08	67.35 ± 0.58
Mixed + all-correct	62.40	49.72	52.69	92.11	95.73	49.14	73.00	67.83 ± 0.45
RAC-GRPO (full)	**62.53**	**49.93**	**53.82**	**92.92**	**95.87**	**50.61**	**74.63**	$\mathbf{68.62 \pm 0.50}$

Component contributions. The *Endpoint groups only* variant retains GRPO advantages in mixed groups and uses $c_0 = c_G = 0.0113$ at the endpoints. The *Mixed groups only* variant calibrates mixed groups and sets both endpoint advantages to zero; the two one-endpoint variants additionally recover either all-wrong or all-correct groups. All variants retain the same response mask and token-mean loss denominator, including valid tokens from groups with zero advantages. Using recovery alone or calibration in mixed groups alone yields averages of 67.19 and 67.08, respectively, versus 65.80 for GRPO. The full method reaches 68.62; its advantage over *Endpoint groups only* shows that mixed-group calibration contributes beyond restoring nonzero endpoint advantages. Its higher mean than either one-endpoint variant further supports recovery at both endpoints.

Table 3: **Sensitivity to correction strength on Qwen3-8B ($G = 8$, %).** Fixed-c variants use the same pseudo-count for all success counts. Three-seed means; Average reports mean ± sample SD. Bold marks column maxima; evaluation follows Table 1.

Method / setting	AIME24	AIME25	AIME26	AMC23	MATH500	Minerva	Olympiad	Average
GRPO	60.52	47.15	50.21	91.17	94.33	46.81	70.42	65.80 ± 0.28
$c = 0.01$	61.22	49.13	53.78	92.39	95.53	48.16	72.95	67.59 ± 0.52
$c = 0.1$	61.01	48.37	51.56	91.28	95.33	48.29	72.60	66.92 ± 0.41
$c = 0.5$	59.24	47.29	52.22	90.16	94.93	48.16	72.65	66.38 ± 0.48
$c = 1$	58.23	47.29	50.28	90.57	94.67	46.77	71.27	65.58 ± 0.76
$c = 2$	57.99	45.62	49.41	90.31	94.80	48.90	70.57	65.37 ± 0.10
RAC-GRPO	**62.53**	**49.93**	**53.82**	**92.92**	**95.87**	**50.61**	**74.63**	$\mathbf{68.62 \pm 0.50}$

Fixed versus count-dependent calibration. Among the tested fixed pseudo-counts in Table 3, mean performance decreases as c increases. The best fixed setting, $c = 0.01$, reaches 67.59 ± 0.52, compared with 68.62 ± 0.50 for RAC-GRPO, a gain of 1.03 points. The $c = 1$ and $c = 2$ variants fall below GRPO despite nonzero endpoint advantages, indicating sensitivity to correction strength. This sweep changes calibration in mixed groups and recovery at endpoints together.

Endpoint recovery strength. Figure 3 varies only the endpoint magnitude a, keeping RAC-GRPO calibration in mixed groups fixed. The calibrated $a \approx 0.0376$ yields the highest mean among the five tested settings (68.62 ± 0.50), 1.54 points above no recovery ($a = 0$). At $a \approx 0.3333$—the magnitude induced by $c = 1$ only at the endpoints—the average falls to 66.05 ± 0.39, below the $a = 0$ reference. Stronger endpoint recovery is therefore not uniformly beneficial. This is consistent with Section 4.4: at a fixed policy, increasing a scales the endpoint component's mean and standard deviation together, so expected alignment alone does not ensure better training performance.

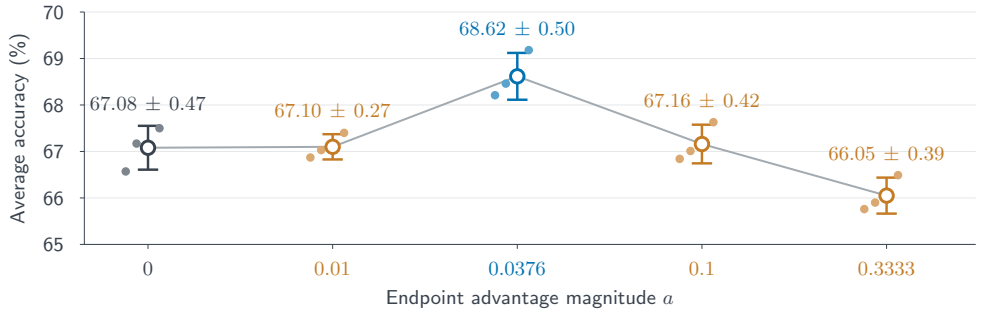


Figure 3: **Endpoint recovery strength on Qwen3-8B ($G = 8$).** Mixed-group calibration is fixed; all-wrong/all-correct advantages are $-a/+a$. Dots show individual seeds; circles and bars show mean ± sample SD of the seven-benchmark average over three seeds. Blue denotes RAC-GRPO; horizontal spacing is categorical.

Reference credible level. On Qwen3-4B, the 80%, 85%, and 90% settings score 63.58, 64.88, and 65.03, below 66.60 ± 0.44 at the default 95% (Table 12). Lower levels narrow the interval over which the MSE bound must hold, permitting larger pseudo-counts and stronger offsets at unbalanced counts (Appendix E.1). These exploratory single-seed comparisons are consistent with Table 3 and Figure 3: stronger correction need not improve performance.

5.4 Group Size and Transfer to GSPO

Smaller rollout groups increase the probability of zero-variance groups at a fixed interior success probability (Equation 3). At $G = 4$, RAC-GRPO improves the Qwen3-8B average by 2.70 points across three seeds, compared with 2.82 at $G = 8$ (Tables 4 and 1). Gains occur on all seven benchmarks at both tested group sizes; Appendix D reports the individual $G = 4$ runs. Using calibrated advantages in GSPO (Zheng et al., 2025) raises the three-seed average from 65.40 ± 1.27 to 67.20 ± 0.99, a gain of 1.80 percentage points, with higher means on all seven benchmarks.

Table 4: **Smaller rollout groups and transfer to GSPO on Qwen3-8B (%).** Three-seed means; Average reports mean ± sample SD. Bold marks the higher mean within each comparison; evaluation follows Table 1.

Method	AIME24	AIME25	AIME26	AMC23	MATH500	Minerva	Olympiad	Average
GRPO ($G = 4$)	54.76	41.46	47.33	87.55	93.67	45.71	69.53	62.86 ± 0.35
RAC-GRPO ($G = 4$)	**56.53**	**47.08**	**50.42**	**91.48**	**94.53**	**47.55**	**71.31**	$\mathbf{65.56 \pm 0.24}$
GSPO	57.12	46.50	51.67	90.49	93.93	47.43	70.67	65.40 ± 1.27
GSPO + calibrated advantages	**60.31**	**48.82**	**52.67**	**92.66**	**95.27**	**49.14**	**71.51**	$\mathbf{67.20 \pm 0.99}$

6 Conclusion

RAC-GRPO calibrates GRPO reward statistics across success counts using only the composition of the current rollout group. This adjusts advantages in mixed groups and recovers small training weights at the zero-variance endpoints. Our analysis shows that zero-variance recovery is aligned with the success gradient in expectation, while its signal and stochastic variation scale together with recovery magnitude. Across three Qwen3 model scales, RAC-GRPO improves the average over seven benchmarks compared with GRPO in three-seed experiments under a shared rollout budget. Component ablations show gains from both mixed-group calibration and zero-variance recovery, with the full method achieving the highest mean among the tested variants.

Limitations. The calibration rule assumes binary rewards and conditionally independent rollouts. The gradient analysis uses equally weighted score vectors from on-policy responses; averaging over tokens, clipping, and repeated optimization introduce additional effects. Evaluation is limited to mathematical reasoning with Qwen3 models and a single training corpus.

Reproducibility statement

The calibrated advantage estimator and pseudo-count calibration rule are specified in Section 4, with the $G = 8$ lookup table in Table 6 and the $G = 4$ lookup in Appendix C. Appendix A gives full proofs of the calibration and update analyses, with exact risk expressions. Appendix C records the objective, training settings, and baseline implementations; Tables 2 and 3 report component ablations and sensitivity to correction strength; Table 4 reports three-seed results for smaller rollout groups and transfer to GSPO. Appendices D and E report results for each seed and sensitivity analyses.

AI use statement

Generative AI tools assisted with polishing the manuscript's language and prototyping some figures. AI tools were also used to search for and identify related work. The authors reviewed all content and are responsible for the final claims, citations, proofs, and artifacts.

References

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Yong Yi Bay and Kathleen A. Yearick. GRPO, Dr. GRPO, and DAPO are three operations on one number: The group-standard-deviation identity. *arXiv preprint arXiv:2607.00152*, 2026. URL https://arxiv.org/abs/2607.00152.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

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*, 2026.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

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. In *International Conference on Machine Learning*, 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. *arXiv preprint arXiv:2103.03874*, 2021.

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.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An open source approach to scaling up reinforcement learning on the base model. *Advances in Neural Information Processing Systems*, 38:162239–162262, 2025.

Heyang Jiang, Henry Liu, and Baharan Mirzasoleiman. Learning as reasoning unfolds: Progressive rollout allocation for efficient reinforcement learning. *arXiv preprint arXiv:2607.22002*, 2026.

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.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Siheng Li, Zhanhui Zhou, Wai Lam, Chao Yang, and Chaochao Lu. RePO: Replay-enhanced policy optimization. *arXiv preprint arXiv:2506.09340*, 2025.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24681–24693, 2026.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*, 2023.

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.

Amritansh Mishra, Supriyo Chakraborty, and Berkcan Kapusuzoglu. On the policy gradient foundations of group relative policy optimization: Credit assignment, gradient sparsity, and rank collapse. *arXiv preprint arXiv:2606.29238*, 2026.

Wenhua Nie, Jianan Wu, Junlin Liu, Ziwei Li, Zheng Lin, Zijian Zhang, Yilong Fan, Haoran Zheng, and Jyh-Shing Roger Jang. Gradient starvation in binary-reward GRPO: Why group-mean centering fails and why the simplest fix works. *arXiv preprint arXiv:2605.07689*, 2026.

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, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.

Ziqi Wang, Xingzhou Lou, Meiqi Wu, Zhengqi Wen, and Junge Zhang. Calibration-aware policy optimization for reasoning LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18375–18390, 2026.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. BNPO: Beta normalization policy optimization. *arXiv preprint arXiv:2506.02864*, 2025.

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

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2025.

Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. EDGE-GRPO: Entropy-driven GRPO with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*, 2025a.

Yuheng Zhang, Wenlin Yao, Changlong Yu, Yao Liu, Qingyu Yin, Bing Yin, Hyokun Yun, and Lihong Li. Improving sampling efficiency in RLVR through adaptive rollout and response reuse. *arXiv preprint arXiv:2509.25808*, 2025b.

Zhi Zhang, Zhen Han, Costas Mavromatis, Qi Zhu, Yunyi Zhang, Sheng Guan, Dingmin Wang, Xiong Zhou, Shuai Wang, Soji Adeshina, et al. Train less, learn more: Adaptive efficient rollout optimization for group-based reinforcement learning. *arXiv preprint arXiv:2602.14338*, 2026.

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

Xinxin Zhu, Ying He, Haowen Hou, Ruichong Zhang, Nianbo Zeng, Yulin Peng, Jiongfeng Fang, and F Richard Yu. PSPO: Prompt-level prioritization and experience-weighted smoothing for efficient policy optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 29186–29194, 2026.

A Proofs and Additional Analysis

A.1 Proof of the Conditional Posterior Risk Bound

Conditional posterior risk. The reference posterior defines the uncertainty interval for the advantage-calibration risk constraint. We use the uniform reference prior and Bernoulli likelihood to define $p_x \mid K, G \sim \mathrm{Beta}(K+1, G-K+1)$. For an estimate a, write its conditional squared-error risk as $\mathcal{R}_K(a) = \mathbb{E}[(a-p_x)^2 \mid K, G]$. The following bound holds for the probability estimator induced by a count-dependent pseudo-count lookup.

Proposition 2 (Conditional posterior risk comparison) *For every $G \geq 1$, $K \in \{0, \ldots, G\}$, and $0 < c_K \leq 2$,*

$$\mathcal{R}_K\left(\frac{K+c_K}{G+2c_K}\right) \leq \mathcal{R}_K\left(\frac{K}{G}\right), \tag{12}$$

with strict inequality unless $K = G/2$. Consequently, the induced probability estimator has strictly smaller uniform-prior integrated risk than K/G. Two is the largest cap independent of G that guarantees this comparison for all group sizes.

Probability-estimation risk and the calibration constraint. For a count-based probability estimator δ_K,

$$\underbrace{\int_0^1 \mathbb{E}_{K|p}[(\delta_K - p)^2]\, dp}_{\mathcal{B}(\delta)} = \frac{1}{6(G+2)} + \frac{1}{G+1}\sum_{K=0}^{G}\left(\delta_K - \frac{K+1}{G+2}\right)^2. \tag{13}$$

Thus fixed $c = 1$ is optimal for uniform-prior integrated squared-error estimation among count-based estimators. The corresponding fixed-c integrated risk is $\mathcal{B}(c) = (c^2/3 + G/6)/(G+2c)^2$. For the $G = 8$ lookup table used in training,

$$0.01667 \quad (c=1) \;<\; 0.01989 \quad (\text{adaptive}) \;<\; 0.02083 \quad (\text{empirical}). \tag{14}$$

Appendix A.9 proves the identity and gives the exact finite sum. The probability estimator with fixed $c = 1$ has the lowest risk under this reference measure, whereas RAC-GRPO achieves a higher RL score than the training variant with fixed $c = 1$ (Table 5). The calibration rule uses probability-estimation risk to constrain correction strength; benchmarks assess downstream policy optimization.

Proof of Proposition 2. Under the reference posterior,

$$m_K = \mathbb{E}[p_x \mid K, G] = \frac{K+1}{G+2}.$$

For any point estimate a, conditional squared-error risk decomposes as

$$\mathcal{R}_K(a) = \mathrm{Var}(p_x \mid K, G) + (a - m_K)^2. \tag{15}$$

The empirical and shrunk estimates satisfy

$$\frac{K}{G} - m_K = \frac{2K - G}{G(G+2)}, \tag{16}$$

$$\frac{K + c_K}{G + 2c_K} - m_K = \frac{(2K - G)(1 - c_K)}{(G + 2c_K)(G+2)}. \tag{17}$$

If $K = G/2$, both deviations vanish. Otherwise their absolute ratio is

$$\frac{|(K + c_K)/(G + 2c_K) - m_K|}{|K/G - m_K|} = \frac{G|1 - c_K|}{G + 2c_K} < 1 \qquad (0 < c_K \le 2). \tag{18}$$

The strict inequality follows directly for $c_K \le 1$; for $1 < c_K \le 2$, its numerator is at most G while its denominator exceeds G. Equation 15 completes the risk comparison. More generally, at an unbalanced count, the risk bound holds for all $c > 0$ when $G \le 2$, and for $c \le 2G/(G-2)$ when $G > 2$ (with equality giving equal risk). Moreover, for any constant cap $c_{\max} > 2$, choosing an integer $G > 2c_{\max}/(c_{\max} - 2)$ makes $G(c_{\max} - 1) > G + 2c_{\max}$. The ratio in equation 18 then exceeds one at any unbalanced count. Thus two is the largest group-size-independent cap that guarantees the comparison for every G.

The conditional statement also yields an integrated-risk guarantee for the count-dependent probability estimator. Let $\delta_K = (K + c_K)/(G + 2c_K)$ and let R_{dyn} be defined by equation 37. Under the uniform reference prior, the prior-predictive mass of every count is

$$w_K = \int_0^1 \binom{G}{K} p^K (1-p)^{G-K}\, dp = \frac{1}{G+1}.$$

Multiplying the conditional inequalities by w_K and summing gives

$$\int_0^1 R_{\mathrm{dyn}}(p)\, dp < \int_0^1 \frac{p(1-p)}{G}\, dp = \frac{1}{6G}. \tag{19}$$

Strictness follows from positive prior-predictive mass at every count, including unbalanced counts.

A.2 Fixed-c MSE Calibration

This section derives the fixed-c risk region used in Equation 33. The empirical maximum likelihood estimator (MLE) is $\widehat{p}_K = K/G$. For $c > 0$ fixed before sampling, define $\widetilde{p}_{K,c} = (K + c)/(G + 2c)$. Then

$$\mathrm{MSE}_{\mathrm{MLE}}(p) = \frac{p(1-p)}{G}, \qquad \mathrm{MSE}_c(p) = \frac{c^2(1-2p)^2 + Gp(1-p)}{(G+2c)^2}. \tag{20}$$

Proposition 3 (Fixed-c risk) *For any fixed $c > 0$, $\mathrm{MSE}_c(p) < \mathrm{MSE}_{\mathrm{MLE}}(p)$ if and only if*

$$p(1-p) > \frac{Gc}{4(G + c + Gc)}. \tag{21}$$

Equivalently, fixed-c shrinkage lowers pointwise MSE exactly when $p_-(G,c) < p < p_+(G,c)$, where

$$p_\pm(G,c) = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{Gc}{Gc + G + c}}. \tag{22}$$

Proof. For $K \sim \mathrm{Binomial}(G,p)$ and fixed $c > 0$,

$$\mathrm{Bias}(\widetilde{p}_{K,c}) = \frac{c(1-2p)}{G+2c}, \qquad \mathrm{Var}(\widetilde{p}_{K,c}) = \frac{Gp(1-p)}{(G+2c)^2}, \tag{23}$$

which gives equation 20. Comparing with $p(1-p)/G$ and multiplying by $G(G+2c)^2$ yields

$$Gc^2(1-2p)^2 < 4c(G+c)p(1-p), \tag{24}$$

$$Gc\{1 - 4p(1-p)\} < 4(G+c)p(1-p). \tag{25}$$

Collecting terms in the second line proves $p(1-p) > Gc/[4(G+c+Gc)]$. At equality, use $p(1-p) = 1/4 - (p-1/2)^2$ to obtain equation 22.

Proposition 3 quantifies the bias–variance trade-off in estimating the success probability, which motivates the calibration constraint. Near $p = 1/2$, variance reduction can outweigh the bias introduced by shrinkage. Near either endpoint, the empirical estimator has small variance, and shrinkage toward $1/2$ increases MSE. For example, $\mathrm{MSE}_c(0) = c^2/(G+2c)^2 > 0$ whereas $\mathrm{MSE}_{\mathrm{MLE}}(0) = 0$, and symmetrically at $p = 1$. For $G = 8$ and $c = 1$, shrinkage has lower MSE than the empirical estimator when $0.1362 < p < 0.8638$.

A.3 Proof of Expected Update and Zero-Variance Recovery

Exact update expressions. Let $\widehat{g}_{\mathrm{GRPO}} = G^{-1}\sum_i \widehat{A}_i^{\mathrm{GRPO}} s_\theta(y_i)$ under the assumptions of Proposition 1. Equation 5 gives

$$\widehat{g} = \lambda_K \widehat{g}_{\mathrm{GRPO}} + b_K \overline{s}_\theta, \qquad \overline{s}_\theta = \frac{1}{G}\sum_i s_\theta(y_i). \tag{26}$$

Writing $A^+(K)$ and $A^-(K)$ for success and failure advantages, the expected coefficient is

$$\mathbb{E}[\widehat{g}] = \alpha_{\mathrm{RAC}}(p)\nabla_\theta p, \qquad \alpha_{\mathrm{RAC}}(p) = \mathbb{E}_{S\sim\mathrm{Binomial}(G-1,p)}[A^+(S+1) - A^-(S)] > 0. \tag{27}$$

For a pseudo-count lookup with symmetric endpoint values, the common advantage magnitude is

$$a_{0,\epsilon} = \frac{c_0}{\sqrt{c_0(G+c_0)} + (G+2c_0)\epsilon_{\mathrm{num}}} > 0. \tag{28}$$

Proof of Proposition 1. Let $S = \sum_{j\neq i} R(y_j)$. It is independent of $(R(y_i), s_\theta(y_i))$ and distributed as $\mathrm{Binomial}(G-1, p)$. The score identity gives

$$\mathbb{E}[R(y_i)s_\theta(y_i)] = \nabla_\theta p, \qquad \mathbb{E}[(1-R(y_i))s_\theta(y_i)] = -\nabla_\theta p. \tag{29}$$

Conditioning on S therefore yields

$$\mathbb{E}[A_i s_\theta(y_i) \mid S] = A^+(S+1)\mathbb{E}[R s_\theta(y_i)] + A^-(S)\mathbb{E}[(1-R)s_\theta(y_i)] \tag{30}$$

$$= \{A^+(S+1) - A^-(S)\}\nabla_\theta p. \tag{31}$$

Averaging over S proves equation 27; the coefficient is strictly positive because $A^+ > 0 > A^-$. Exchangeability gives the same expectation for the group average.

RAC-GRPO and the *Mixed groups only* variant differ in exactly two terms of this expectation. For $S = G-1$, the reward-one branch is an all-correct group and contributes $A^+(G)p^{G-1}$ to the coefficient. For $S = 0$, the reward-zero branch is an all-wrong group and contributes $-A^-(0)(1-p)^{G-1}$. The *Mixed groups only* variant sets both terms to zero, proving equation 10. Finally, symmetry and equation 4 give $A^+(G) = -A^-(0) = a_{0,\epsilon}$ from equation 28, which yields the stated expression.

The proof uses the sampling and weighting assumptions in Proposition 1. Section 4.4 discusses aggregation across prompts and the relation to the implemented loss.

A.4 Proof of Maximal Calibration and Exact Adaptive Risk

Correction strength under a risk constraint. Let $C_{K,G} = [L_{\eta/2}, U_{1-\eta/2}]$ be the equal-tailed reference-posterior interval of mass $1-\eta$; the main setting uses $\eta = 0.05$. For c fixed before sampling, the probability estimator $\widetilde{p}_{K,c}$ has no larger pointwise MSE than K/G on the interval

$$\mathcal{I}_G(c) = [p_-(G,c), p_+(G,c)], \qquad p_\pm(G,c) = \frac{1}{2} \pm \frac{1}{2}\sqrt{1 - \frac{Gc}{Gc+G+c}}. \tag{32}$$

Appendix A.2 derives this bias–variance comparison. Larger c narrows $\mathcal{I}_G(c)$, reflecting the cost of a stronger correction when the true success probability is near zero or one. We select

$$c^\star(K,G) = \max\{c \in (0,2] : C_{K,G} \subseteq \mathcal{I}_G(c)\}. \tag{33}$$

The uniform reference prior, credible level, and cap define the feasible family at every count. Maximizing the pseudo-count within this family specifies all $G+1$ correction strengths.

Proposition 4 (Closed-form calibration) *Define* $\ell_{K,G} = \min\{L_{\eta/2}, 1 - U_{1-\eta/2}\}$, $q_{K,G} = \ell_{K,G}(1-\ell_{K,G})$, *and* $d_{K,G} = G - 4(G+1)q_{K,G}$. *Then*

$$c^\star(K,G) = \begin{cases} \min\{2, 4Gq_{K,G}/d_{K,G}\}, & d_{K,G} > 0, \\ 2, & d_{K,G} \le 0. \end{cases} \tag{34}$$

The calibration rule is symmetric in K and $G-K$, and the reference posterior assigns at least $1-\eta$ mass to $\mathcal{I}_G(c^\star(K,G))$.

The proof below solves the boundary constraint in Equation 32. The posterior-mass guarantee concerns the fixed-c risk region associated with the selected pseudo-count. The repeated-sampling risk of the induced count-dependent probability estimator is the squared error averaged over the sampling distribution of K (Equation 37).

Proof of Proposition 4. Proposition 3 defines $\mathcal{I}_G(c)$ as the interval where fixed c has no larger MSE than the empirical estimator. This interval is symmetric around $1/2$. The reference credible interval is contained in $\mathcal{I}_G(c)$ exactly when its distance to the nearer endpoint of $[0,1]$ is at least $p_-(G,c)$. Define this distance and the corresponding Bernoulli variance by

$$\ell_{K,G} = \min\{L_{\eta/2}, 1 - U_{1-\eta/2}\}, \qquad q_{K,G} = \ell_{K,G}(1-\ell_{K,G}). \tag{35}$$

Matching $p_-(G,c)$ to $\ell_{K,G}$ and solving for c yields

$$d = G - 4(G+1)q_{K,G}, \qquad c_{\mathrm{raw}}(K,G) = \frac{4Gq_{K,G}}{d}. \tag{36}$$

When $d > 0$, c_{raw} is the largest uncapped feasible pseudo-count. When $d \le 0$, the boundary remains covered for every finite c. Indeed, the threshold $Gc/[4\{G+(G+1)c\}]$ defining $\mathcal{I}_G(c)$ has derivative $G^2/[4\{G+(G+1)c\}^2] > 0$, so $\mathcal{I}_G(c)$ narrows monotonically with c. Imposing the cap at two therefore gives exactly $c^\star$ in equation 9 and preserves containment. The reference posterior assigns probability $1-\eta$ to $C_{K,G}$; hence containment gives the final probability statement in Proposition 4. The calibration rule is symmetric, $c^\star(K,G) = c^\star(G-K,G)$, with $G+1$ entries per group size.

Finite-precision implementation. The $G = 8$ experiments used the ideal pseudo-counts truncated to four decimal places (Table 6). Each entry satisfies $0 < \bar{c}(K,8) \le c^\star(K,8)$ and $c^\star(K,8) - \bar{c}(K,8) < 10^{-4}$. Since $\mathcal{I}_G(c)$ narrows with c, truncation preserves equal-tailed interval containment; the cap preserves Proposition 2. Direct enumeration gives $\min_K \Pr_{p_x|K,G}[p_x \in \mathcal{I}_8(\bar{c}(K,8))] \approx 0.9686$. Proposition 4 establishes exact maximality for unrounded $c^\star$.

Exact repeated-sampling risk. For a deterministic lookup $c_K = c(K,G)$, define

$$R_{\mathrm{dyn}}(p) = \sum_{K=0}^{G} \binom{G}{K} p^K (1-p)^{G-K} \left(\frac{K + c_K}{G + 2c_K} - p \right)^2. \tag{37}$$

When c is selected from K, Equation 37 is the appropriate repeated-sampling risk and can be evaluated exactly for any fixed G. For comparison, the uniformly weighted integrated risk of the fixed-c probability estimator has the closed form

$$\int_0^1 \mathrm{MSE}_c(p)\, dp = \frac{c^2/3 + G/6}{(G+2c)^2}. \tag{38}$$

For $G = 8$, we use the fixed-c formula above for comparison and Equation 45 to evaluate the probability estimator induced by the pseudo-count lookup table truncated to four decimal places. The resulting integrated risks under uniform weighting satisfy

$$0.01667\ (c=1) < 0.01989\ (\text{adaptive}) < 0.02083\ (\text{MLE}).$$

The count-dependent probability estimator reduces uniform-prior integrated risk relative to the MLE, as established in equation 19; fixed $c = 1$ has the smallest integrated risk. Table 5 reports lower RL performance for the training variant with fixed $c = 1$ than for RAC-GRPO, although the fixed-c probability estimator has lower integrated MSE. The two orderings assess probability-estimation

Table 5: Probability-estimation risk and downstream performance ($G = 8$). Integrated MSE uses uniform weighting over $p \in [0, 1]$; endpoint $|A|$ assumes $\epsilon_{\text{num}} = 0$. RLVR averages are Qwen3-8B seven-benchmark means (%), reported as mean ± sample SD over three seeds.

Probability estimator	Integrated MSE	Endpoint $\lvert A\rvert$	RLVR average
Empirical (GRPO)	0.02083	0	65.80 ± 0.28
Fixed $c = 1$	**0.01667**	0.3333	65.58 ± 0.76
Count-dependent (RAC-GRPO)	0.01989	0.0376	$\mathbf{68.62 \pm 0.50}$

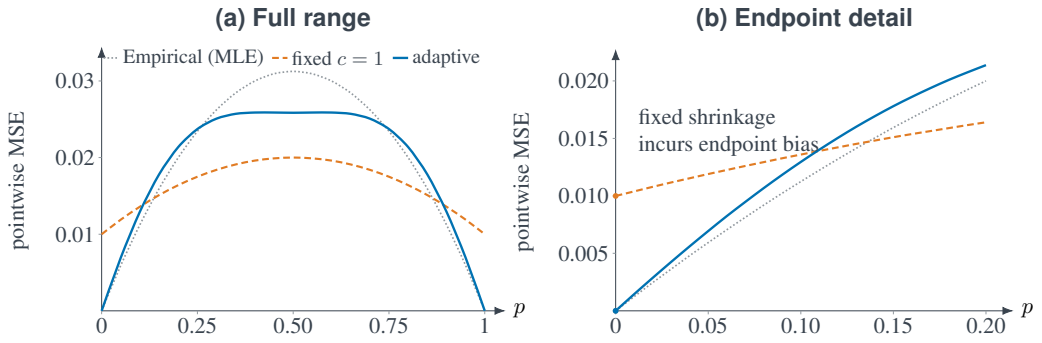


Figure 4: Pointwise probability-estimation MSE for $G = 8$. **(a)** The empirical estimator (MLE), fixed $c = 1$, and the probability estimator induced by the adaptive lookup table over $p \in [0, 1]$. **(b)** Zoom near $p = 0$; the behavior near $p = 1$ is symmetric. Table 5 gives integrated risks and training scores.

accuracy and downstream training performance, respectively. Here, probability-estimation risk sets the admissible correction range. Uniform weighting gives a reference comparison over $p \in [0, 1]$; the training prompt distribution may assign different weights.

Figure 4 shows the corresponding pointwise risks. Fixed $c = 1$ pays a large bias cost at $p = 0$ and $p = 1$ but achieves lower risk across much of the interior. The count-dependent probability estimator remains close to the empirical rate at endpoint counts because $\bar{c}(0, 8) = \bar{c}(8, 8) = 0.0113$; relative to the MLE, it has higher risk near the probability boundaries and lower risk over part of the interior. The curves cross, and their areas give the integrated ordering above.

Table 6 gives the $G = 8$ lookup table derived from the ideal calibration rule and used in training. The cap is active at $K = 4$, where the uncapped pseudo-count is 2.6921. At this balanced count, symmetric shrinkage gives $\widetilde{p} = 1/2$ for every c.

Table 6: The $G = 8$ lookup table used in training, obtained by truncating the ideal pseudo-counts to four decimal places, and the resulting probability estimates.

K	$\bar{c}(K, 8)$ used	$\widetilde{p}_{K,\bar{c}}$
0	0.0113	0.0014
1	0.1247	0.1363
2	0.4024	0.2729
3	1.0105	0.4002
4	2.0000	0.5000
5	1.0105	0.5998
6	0.4024	0.7271
7	0.1247	0.8637
8	0.0113	0.9986

A.5 Decomposition into Contrast and Offset

Exact decomposition. Fix a count K, let $\widehat{p}_K = K/G$, and let $\widetilde{p}_K = (K + c_K)/(G + 2c_K)$ for $c_K > 0$. Write

$$\widehat{s}_K = \sqrt{\widehat{p}_K(1-\widehat{p}_K)}, \qquad \widetilde{s}_K = \sqrt{\widetilde{p}_K(1-\widetilde{p}_K)}, \qquad D_K = \widetilde{s}_K + \epsilon_{\text{num}},$$

where $\epsilon_{\text{num}} > 0$. Under the population standard-deviation convention for GRPO,

$$\widehat{A}_i^{\text{RAC}} = \lambda_K \widehat{A}_i^{\text{GRPO}} + b_K, \qquad \lambda_K = \frac{\widehat{s}_K + \epsilon_{\text{num}}}{D_K}, \qquad b_K = \frac{\widehat{p}_K - \widetilde{p}_K}{D_K}. \tag{39}$$

For every count, $0 < \lambda_K \leq 1$ and

$$\frac{1}{G}\sum_{i=1}^{G} \widehat{A}_i^{\text{RAC}} = b_K, \qquad \operatorname{sign}(b_K) = \operatorname{sign}(2K - G). \tag{40}$$

At a balanced count, $\lambda_K = 1$ and $b_K = 0$. At an endpoint count, every GRPO advantage vanishes and every RAC-GRPO advantage equals b_K.

Proof. Adding and subtracting $\widehat{p}_K$ gives

$$\frac{r_i - \widetilde{p}_K}{D_K} = \frac{r_i - \widehat{p}_K}{D_K} + \frac{\widehat{p}_K - \widetilde{p}_K}{D_K},$$

which is the stated identity. Symmetric shrinkage satisfies

$$\widetilde{p}_K - \frac{1}{2} = \frac{G}{G + 2c_K}\left(\widehat{p}_K - \frac{1}{2}\right).$$

Consequently $\widetilde{s}_K \geq \widehat{s}_K$, with equality only at a balanced count, proving the bound on λ_K. Finally,

$$\widehat{p}_K - \widetilde{p}_K = \frac{c_K(2K - G)}{G(G + 2c_K)}.$$

This proves the sign claim; averaging over observed rewards gives the group-mean identity.

In mixed groups, the decomposition gives a rescaled contrast and a shared offset whose sign agrees with the majority reward. At the endpoints, the GRPO contrast vanishes and the offset remains. Responses with equal rewards receive equal calibrated advantages. Their score vectors can sum to a nonzero vector in a finite sample, allowing the shared offset to affect the sampled update; the expected offset contribution depends on the reward–score relationship in Proposition 1.

A.6 Correction Strength and Shared Offset Magnitude

For fixed K, G, let $Q(c) = (K + c)(G - K + c)$ and $D(c) = \sqrt{Q(c)} + (G + 2c)\epsilon_{\text{num}}$. Equation 7 follows by substituting the adjusted reward mean and standard deviation into b_K. Since $Q'(c) = G + 2c$,

$$\frac{d}{dc}\frac{c}{D(c)} = \frac{[K(G-K) + Gc/2]/\sqrt{Q(c)} + G\epsilon_{\text{num}}}{D(c)^2} > 0 \qquad (c > 0).$$

This proves strict monotonicity of $|b_K|$ for $K \neq G/2$. Meanwhile, $\widetilde{p}_K$ moves toward $1/2$ as c increases, so its Bernoulli standard deviation increases and λ_K decreases at every unbalanced count. At $K = G/2$, $b_K = 0$ and $\lambda_K = 1$ for every c. The largest feasible pseudo-count therefore maximizes the magnitude of the shared offset under the risk constraint.

A.7 Endpoint Second Moments and Saturated-Prompt Noise

Assume Proposition 1 applies, with finite second moments of the score vectors conditional on the reward. For $p \in (0, 1)$ write $v = \nabla_\theta p$ and $\Sigma_r = \operatorname{Cov}(s_\theta(y) \mid R = r)$. The score identities give

$$\mu_1 = \mathbb{E}[s_\theta \mid R = 1] = v/p, \qquad \mu_0 = \mathbb{E}[s_\theta \mid R = 0] = -v/(1-p).$$

Conditional on an all-correct or all-wrong group, the G score vectors remain independent, each following the score distribution conditional on the corresponding reward. Therefore, for $H_a = a(\mathbf{1}_{\{K=G\}} - \mathbf{1}_{\{K=0\}})\overline{s}_\theta$,

$$\mathbb{E}[H_a] = a\{p^{G-1} + (1-p)^{G-1}\}v, \tag{41}$$

$$\mathbb{E}[H_a H_a^\top] = a^2\left[\frac{p^G\Sigma_1 + (1-p)^G\Sigma_0}{G} + \{p^{G-2} + (1-p)^{G-2}\}vv^\top\right]. \tag{42}$$

To verify the second identity, condition on the two disjoint endpoint events and use $\mathbb{E}[\overline{s}\,\overline{s}^\top \mid R_{1:G} = r] = \Sigma_r/G + \mu_r\mu_r^\top$. Subtracting $\mathbb{E}[H_a]\mathbb{E}[H_a]^\top$ gives the covariance, also proportional to a^2. The standard deviation in any fixed direction therefore scales linearly in a. Consequently any directional signal-to-standard-deviation ratio of the isolated zero-variance group component is invariant to positive rescaling of a, whenever its denominator is nonzero.

For a numerical comparison of endpoint magnitudes, hold the policy and rollout law fixed and omit ϵ_{num}. At $G = 8$,

$$\frac{a_{c=1}}{a_{\text{adaptive}}} = \frac{1/3}{\sqrt{0.0113/8.0113}} \approx 8.88, \qquad \left(\frac{a_{c=1}}{a_{\text{adaptive}}}\right)^2 \approx 78.77.$$

These factors compare endpoint components under a fixed policy. The full update $g = g_{\text{mix}} + H_a$ also has mixed-group and cross-covariance terms:

$$\text{Cov}(g) = \text{Cov}(g_{\text{mix}}) + \text{Cov}(H_a) + \text{Cov}(g_{\text{mix}}, H_a) + \text{Cov}(H_a, g_{\text{mix}}).$$

For a locally saturated prompt, assume the verifier is constant over the policy support in a parameter neighborhood, with the usual score regularity and finite $F_x = \mathbb{E}[s_\theta s_\theta^\top]$. Then p is locally constant, $\mathbb{E}[s_\theta] = 0$, and every sampled group has zero reward variance, so $H_a = \pm a\overline{s}_\theta$. Independence gives $\text{Cov}(H_a) = a^2F_x/G$ and $\mathbb{E}[H_a] = 0$, proving Equation 11. The covariance can be nonzero whenever the policy has nonzero score variance, even though its success gradient vanishes. Both adaptive and fixed shrinkage introduce this sampling noise whenever the endpoint magnitude is positive and the score variance is nonzero. A smaller magnitude limits absolute stochastic updates at locally saturated prompts; at a fixed nonsaturated policy, it also reduces any nonzero expected endpoint contribution.

A.8 How Zero-Variance Recovery Scales with Group Size

Let $1 - \eta \in (0, 1)$ be a fixed reference credible level and define $t_\eta = -\log(1 - \eta/2) > 0$. For the ideal capped calibration and either endpoint,

$$c^\star(0, G) = c^\star(G, G) = \frac{4t_\eta}{G} + O(G^{-2}). \tag{43}$$

With $\epsilon_{\text{num}} = 0$, the resulting endpoint advantage magnitude and probability displacement satisfy

$$|A_{\text{end}}| = \frac{2\sqrt{t_\eta}}{G} + O(G^{-2}), \qquad \widetilde{p}_0 = 1 - \widetilde{p}_G = \frac{4t_\eta}{G^2} + O(G^{-3}). \tag{44}$$

These statements concern the ideal calibration rule before finite-precision storage.

Derivation. At $K = 0$, the reference posterior is $\text{Beta}(1, G + 1)$, whose quantile at level u is $1 - (1 - u)^{1/(G+1)}$. The lower endpoint of the equal-tailed interval is

$$L_{\eta/2} = 1 - \exp\left(-\frac{t_\eta}{G+1}\right) = \frac{t_\eta}{G} + O(G^{-2}),$$

whereas $1 - U_{1-\eta/2} = (\eta/2)^{1/(G+1)} \longrightarrow 1$. Thus, for all sufficiently large G, the boundary governing the symmetric fixed-c risk interval is $\ell_{0,G} = L_{\eta/2}$ and $q_{0,G} = \ell_{0,G}(1 - \ell_{0,G}) = t_\eta/G + O(G^{-2})$. Substitution into the uncapped solution gives

$$\frac{4Gq_{0,G}}{G - 4(G+1)q_{0,G}} = \frac{4t_\eta + O(G^{-1})}{G - 4t_\eta + O(G^{-1})} = \frac{4t_\eta}{G} + O(G^{-2}).$$

The cap is inactive for sufficiently large G because this expression tends to zero. Posterior symmetry gives the same result at $K = G$. The remaining statements follow by substituting this expansion into $|A_{\text{end}}| = \sqrt{c^\star/(G + c^\star)}$ and $\widetilde{p}_0 = c^\star/(G + 2c^\star)$.

For a 95% interval, $2\sqrt{t_{.05}} \approx 0.3182$. Thus ideal calibrated endpoint advantages decay as G^{-1}, whereas fixed positive pseudo-counts give endpoint advantages of order $G^{-1/2}$ when the numerical safeguard is omitted. For fixed $\epsilon_{\text{num}} > 0$, the ideal calibrated magnitude eventually scales as $4t_\eta/(\epsilon_{\text{num}}G^2)$ as $G \to \infty$; the G^{-1} comparison applies when the safeguard is negligible.

Endpoint contribution as a scalar objective. For a pseudo-count lookup with symmetric endpoint values, let $a_{0,\epsilon} > 0$ be its endpoint advantage magnitude, fixed with respect to θ for a given G and lookup. Under the assumptions of Proposition 1, the expected update with these endpoint advantages minus the expected update with both endpoint advantages set to zero equals

$$\nabla_\theta \Phi_G(p_\theta(x)), \qquad \Phi_G(p) = \frac{a_{0,\epsilon}}{G}\{p^G - (1-p)^G\}.$$

Indeed, $\Phi'_G(p) = a_{0,\epsilon}\{p^{G-1} + (1-p)^{G-1}\} > 0$. This identity expresses the endpoint contribution as a gradient of Φ_G. Shared model parameters combine these prompt-specific gradients with different coefficients, as discussed in Section 4.4.

A.9 An Exact Finite Sum for Integrated Risk

Under the uniform reference prior, the integrated risk can be computed exactly as a finite sum. Numerical integration is therefore unnecessary. For any deterministic lookup $c_K > 0$, set $\delta_K = (K + c_K)/(G + 2c_K)$ and $m_K = (K + 1)/(G + 2)$. Then

$$\int_0^1 R_{\text{dyn}}(p)\, dp = \frac{1}{6(G+2)} + \frac{1}{G+1}\sum_{K=0}^{G}\left\{\frac{(2K-G)(1-c_K)}{(G+2c_K)(G+2)}\right\}^2. \tag{45}$$

Derivation. The uniform reference prior induces predictive mass $1/(G + 1)$ for each count. The posterior-risk decomposition therefore yields

$$\int_0^1 R_{\text{dyn}}(p)\, dp = \frac{1}{G+1}\sum_{K=0}^{G}\left\{\text{Var}(p \mid K, G) + (\delta_K - m_K)^2\right\}.$$

For the reference posterior,

$$\text{Var}(p \mid K, G) = \frac{(K+1)(G-K+1)}{(G+2)^2(G+3)},$$

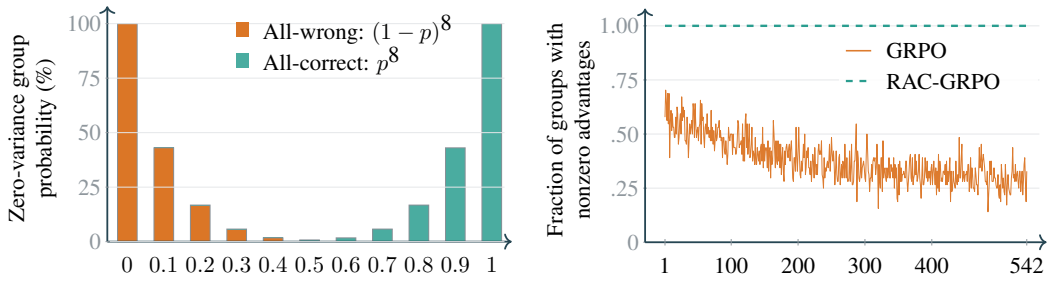
and $\sum_{K=0}^{G}(K + 1)(G - K + 1) = (G + 1)(G + 2)(G + 3)/6$. The mean posterior variance is consequently $1/[6(G + 2)]$. Finally,

$$\delta_K - m_K = \frac{(2K-G)(1-c_K)}{(G+2c_K)(G+2)},$$

giving the finite-sum expression. The first term is the integrated risk of the reference posterior mean. Thus fixed $c = 1$ minimizes this criterion over all estimators that depend only on (K, G); different pseudo-counts at a balanced count represent the same estimate. Count-dependent calibration controls the induced advantages subject to the probability-estimation constraint in Equation 33.

B Finite-Group Collapse Diagnostic

Figure 5 shows the probability of sampling zero-variance groups and the observed fraction of groups with nonzero advantages in one Qwen3-8B GRPO run. The run spans two training epochs on the deduplicated DAPO-Math-17K training set. The horizontal line for RAC-GRPO at a value of one is an analytical reference derived from the calibrated advantage estimator.



(a) Zero-variance group probability

(b) Fraction of groups with nonzero advantages

Figure 5: **Finite-group collapse at $G = 8$. (a)** $q_0(p; G) = (1-p)^G + p^G$, decomposed into all-wrong and all-correct events. **(b)** Fraction of groups with nonzero advantages in a single-seed Qwen3-8B GRPO run (542 steps, two epochs). The RAC-GRPO line is the analytical reference at one.

C Implementation and Evaluation Details

C.1 Objective Integration

With the token-level importance ratio

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})},$$

the per-group form of the clipped token-mean surrogate is

$$J_{\text{RAC}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, \{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{\sum_{j=1}^{G}|y_j|} \sum_{i=1}^{G}\sum_{t=1}^{|y_i|} \min\left(\rho_{i,t}(\theta)\widehat{A}_i^{\text{RAC}}, \text{clip}\left(\rho_{i,t}(\theta), 1-\epsilon_{\text{clip}}, 1+\epsilon_{\text{clip}}\right)\widehat{A}_i^{\text{RAC}}\right)\right]. \tag{46}$$

Here, $\widehat{A}_i^{\text{RAC}}$ is defined in Equation 4, broadcast to the valid response tokens of y_i, and treated as a stop-gradient quantity. Equation 46 gives the per-group form. In training, only the advantages change; baseline importance ratios, clipping, token masks, and token-mean aggregation are retained.

C.2 Training Configuration

Table 7 gives the main $G = 8$ configuration shared by Qwen3-1.7B, 4B, and 8B; the group-size and credible-level experiments vary their designated settings. Paired comparisons follow Section 5.1.

Training is implemented in VERL (Sheng et al., 2024) with vLLM rollouts on a single node with eight NVIDIA H200 GPUs. Dynamic token batching and gradient checkpointing are enabled.

Comparison methods. AVSPO restores a learning signal by augmenting normalization statistics with virtual rewards when its online collapse monitor triggers. DBB provides a statistical baseline based on discounted Beta–Bernoulli history.

C.3 Advantage Estimator Implementation

After the binary verifier returns one reward per response, RAC-GRPO computes $K = \sum_i r_i$, retrieves the precomputed $\bar{c}(K, G)$, and forms

$$\widetilde{p} = \frac{K + \bar{c}(K, G)}{G + 2\bar{c}(K, G)}. \tag{47}$$

Each response then receives

$$\widehat{A}_i = \frac{r_i - \widetilde{p}}{\sqrt{\widetilde{p}(1-\widetilde{p})} + \epsilon_{\text{num}}}. \tag{48}$$

Table 7: Main training configuration shared across the three Qwen3 model scales.

Hyperparameter	Value
Training	
Training dataset	DAPO-Math-17K (deduplicated)
Number of epochs	2
Training batch size	64
Mini-batch size	64
Group size G	8
Maximum prompt length	2,048
Maximum response length	8,192
Thinking mode	Disabled
Sampling	
Training temperature	1.0
Top-p	1.0
Top-k	-1
Optimization	
Optimizer	AdamW
Learning rate	1×10^{-6}
Weight decay	0.1
Warmup steps	0
Gradient clipping norm	1.0
PPO clip $\varepsilon_{\mathrm{l}}/\varepsilon_{\mathrm{h}}$	0.2 / 0.2
Numerical safeguard ϵ_{num}	1×10^{-6}
Loss aggregation	Token mean
KL coefficient	0
Entropy coefficient	0
RAC-GRPO calibration	
Reference prior	$\mathrm{Beta}(1,1)$
Reference credible level	95%
Pseudo-count cap	$c_{\max} = 2$

The GRPO baseline computes the group standard deviation with `torch.std`. With its default sample correction, this standard deviation is $\widehat{s}_K^{\mathrm{sample}} = \sqrt{G/(G-1)}\,\widehat{s}_K$ for $G > 1$. The analysis uses the population standard deviation in Equation 2. For the baseline using the sample standard deviation, the same decomposition into contrast and offset holds with $\lambda_K^{\mathrm{sample}} = (\widehat{s}_K^{\mathrm{sample}} + \epsilon_{\mathrm{num}})/D_K$ and unchanged b_K. Thus equality at balanced counts and the bound $\lambda_K \leq 1$ hold for the reference using the population standard deviation; zero-variance recovery and its moment analysis are unchanged.

These response-level advantages are passed directly to the existing GRPO or GSPO objective. Fixed-c ablations replace $\bar{c}(K,G)$ with the designated constant. Zero-variance group ablations set the designated endpoint advantages to zero without changing the response mask; all valid response tokens therefore remain in the token-mean loss denominator. Reference posterior quantiles and lookup values are precomputed, requiring no additional rollouts or model evaluations. For $G = 4$, evaluating Equation 9 with the corresponding reference posterior gives the lookup:

$$\big(c^\star(0,4), c^\star(1,4), c^\star(2,4), c^\star(3,4), c^\star(4,4)\big) \approx (0.0206, 0.2664, 1.3371, 0.2664, 0.0206).$$

The training configuration uses the pseudo-counts computed from the formula and truncated to four decimal places, as in the $G = 8$ experiments. The reported three-seed $G = 4$ results use this lookup. The $G = 8$ values are reported in Table 6.

C.4 Baseline Reimplementation Details

We independently implemented AVSPO and DBB from their published algorithm descriptions. Executable method implementations were unavailable to us when these experiments were conducted. All reimplemented baselines use the same model initialization, training data, rollout budget, optimization settings, and evaluation protocol as RAC-GRPO unless stated below.

Table 8: Method-specific hyperparameters for independently reimplemented baselines.

Method	Hyperparameter	Value
AVSPO	Initial trigger $\tau_{\text{adapt}}^{(0)}$	0.5
	Sensitivity α	0.5
	Threshold learning rate η	0.01
	Collapse threshold τ	10^{-6}
	Trigger bounds $[\tau_{\min}, \tau_{\max}]$	$[0.1, 0.9]$
	r_{anchor}	0.1
DBB	λ (Qwen3-1.7B)	0.5
	λ (Qwen3-4B, Qwen3-8B)	0.75

For AVSPO, let ACR denote the batch advantage-collapse rate and set

$$K_v = \max\{1, \min[G, \lceil G\,\mathrm{ACR}^{\alpha}\rceil]\}.$$

We create these K_v virtual rewards only when both the adaptive batch trigger condition and the per-group collapse criterion are met, and use them only in the normalization statistics for real responses. For $k = 1, \ldots, K_v$, set

$$r_k^v = \begin{cases} r_{\text{obs}}\left(1 - \dfrac{k}{K_v + 1}\right), & r_{\text{obs}} > 0, \\ r_{\text{anchor}}\dfrac{K_v - k + 1}{K_v}, & r_{\text{obs}} = 0, \end{cases}$$

where r_{obs} is the common binary reward. The adaptive threshold is updated by the published signed-improvement rule and clipped to the bounds in Table 8.

For DBB, the parameters for each prompt start at $(\alpha_0, \beta_0) = (1, 1)$ and are updated at each visit as

$$\alpha_\tau = \lambda\alpha_{\tau-1} + \sum_i r_i, \qquad \beta_\tau = \lambda\beta_{\tau-1} + G - \sum_i r_i.$$

The resulting estimates

$$\widehat{p}_\tau = \frac{\alpha_\tau}{\alpha_\tau + \beta_\tau}, \qquad \widehat{s}_\tau = \frac{\sqrt{\alpha_\tau\beta_\tau}}{\alpha_\tau + \beta_\tau} = \sqrt{\widehat{p}_\tau(1 - \widehat{p}_\tau)}$$

replace the empirical group mean and standard deviation. The DBB paper reports $\lambda = 0.5$ for Qwen3-1.7B and 0.75 for Qwen3-8B. For Qwen3-4B, which is absent from that evaluation, we use the 8B value 0.75. The DBB comparison measures performance under our shared two-epoch budget, a shorter horizon than the original four-epoch protocol.

Table 8 lists additional hyperparameters for the reimplemented baselines.

AVSPO uses our dataset and framework; DBB uses our epoch budget and evaluation protocol.

C.5 Evaluation Details

For each training run, we select the checkpoint with the best performance on the validation set. The test benchmarks are reserved for offline evaluation and excluded from checkpoint selection and hyperparameter tuning.

Decoding and metrics. AIME24/25/26 and AMC23 report Avg@32 from 32 generations per problem at $T = 0.2$; MATH500, Minerva, and Olympiad use greedy decoding.

D Three-Seed Results

Table 10 reports three pairs of matched runs for Qwen3-8B at $G = 8$. Each run index identifies a matched seed pair; the GRPO and RAC-GRPO runs in a pair share the initialization, prompt order, rollout budget, verifier, decoding constraints, and optimization settings. Average is the per-run

Table 9: Per-benchmark variability for Table 1 (%). Base denotes pre-RL checkpoints; RLVR cells report mean ± sample SD over three matched seeds. Average is the per-seed unweighted seven-benchmark mean. Evaluation follows Section 5.1.

Benchmark	Base	GRPO	AVSPO	DBB	RAC-GRPO
Qwen3-1.7B					
AIME24	11.98	31.32 ± 2.32	27.41 ± 1.09	31.74 ± 1.93	$\mathbf{33.47 \pm 2.30}$
AIME25	10.10	29.53 ± 0.85	25.49 ± 2.38	27.99 ± 0.83	$\mathbf{29.57 \pm 1.17}$
AIME26	7.08	21.21 ± 1.24	18.47 ± 0.94	22.36 ± 1.84	$\mathbf{25.42 \pm 1.68}$
AMC23	44.30	70.83 ± 0.84	63.54 ± 0.83	72.63 ± 0.38	$\mathbf{75.55 \pm 2.25}$
MATH500	72.00	85.60 ± 1.25	83.40 ± 0.35	87.27 ± 0.50	$\mathbf{87.87 \pm 0.64}$
Minerva	29.41	35.42 ± 0.42	33.09 ± 0.74	36.64 ± 0.76	$\mathbf{36.76 \pm 1.68}$
Olympiad	39.76	57.47 ± 1.07	54.60 ± 0.26	59.99 ± 1.05	$\mathbf{60.44 \pm 0.37}$
Average	30.66	47.34 ± 1.04	43.72 ± 0.57	48.37 ± 0.75	$\mathbf{49.87 \pm 0.98}$
Qwen3-4B					
AIME24	19.79	53.51 ± 2.91	46.53 ± 0.97	52.54 ± 1.75	$\mathbf{57.88 \pm 0.94}$
AIME25	18.85	47.22 ± 1.59	41.46 ± 1.87	44.86 ± 1.12	$\mathbf{47.78 \pm 0.22}$
AIME26	14.79	49.90 ± 1.56	42.01 ± 1.24	46.36 ± 0.93	$\mathbf{52.71 \pm 0.52}$
AMC23	65.08	89.77 ± 0.16	87.16 ± 1.13	89.45 ± 1.96	$\mathbf{90.73 \pm 3.61}$
MATH500	85.20	93.60 ± 0.92	92.80 ± 0.87	94.07 ± 0.12	$\mathbf{95.18 \pm 0.37}$
Minerva	39.34	47.59 ± 0.92	45.47 ± 0.57	46.57 ± 0.77	$\mathbf{49.06 \pm 1.23}$
Olympiad	51.93	70.03 ± 0.20	65.68 ± 0.91	70.57 ± 0.91	$\mathbf{72.85 \pm 1.04}$
Average	42.14	64.52 ± 0.47	60.16 ± 0.10	63.49 ± 0.80	$\mathbf{66.60 \pm 0.44}$
Qwen3-8B					
AIME24	25.10	60.52 ± 0.75	58.09 ± 3.84	58.16 ± 1.60	$\mathbf{62.53 \pm 1.96}$
AIME25	19.06	47.15 ± 0.73	46.11 ± 1.93	46.08 ± 2.38	$\mathbf{49.93 \pm 0.99}$
AIME26	14.69	50.21 ± 1.03	51.11 ± 0.81	50.97 ± 1.74	$\mathbf{53.82 \pm 0.42}$
AMC23	67.11	91.17 ± 1.17	90.65 ± 1.30	90.16 ± 0.08	$\mathbf{92.92 \pm 0.90}$
MATH500	84.40	94.33 ± 0.64	94.73 ± 0.50	94.47 ± 0.12	$\mathbf{95.87 \pm 0.23}$
Minerva	40.07	46.81 ± 0.85	47.67 ± 1.39	48.28 ± 1.12	$\mathbf{50.61 \pm 0.56}$
Olympiad	54.01	70.42 ± 0.08	71.51 ± 0.97	72.06 ± 1.64	$\mathbf{74.63 \pm 1.18}$
Average	43.49	65.80 ± 0.28	65.70 ± 0.50	65.74 ± 0.93	$\mathbf{68.62 \pm 0.50}$

Table 10: Qwen3-8B matched-seed results (accuracy in %). Each run index pairs GRPO with RAC-GRPO under the shared conditions in Section 5.1.

Method	Run	AIME24	AIME25	AIME26	AMC23	MATH500	Minerva	Olympiad	Average
GRPO	1	60.31	47.92	49.06	89.84	94.60	46.32	70.47	65.50
GRPO	2	59.90	47.08	50.52	91.64	93.60	47.79	70.47	65.86
GRPO	3	61.35	46.46	51.04	92.03	94.80	46.32	70.33	66.05
RAC-GRPO	1	61.35	49.90	53.44	93.52	96.00	50.00	73.29	68.21
RAC-GRPO	2	61.46	48.96	54.27	91.88	96.00	51.10	75.52	68.46
RAC-GRPO	3	64.79	50.94	53.75	93.36	95.60	50.74	75.07	69.18

seven-benchmark mean, rounded to two decimals; aggregate means and sample standard deviations use these Average values. Reported gains are differences between the means displayed in the tables.

Table 11 reports the three training runs per method for Qwen3-8B at $G = 4$. Average is the per-run unweighted seven-benchmark mean, rounded to two decimals; its aggregate mean and sample standard deviation use the reported per-run Average values.

E Additional Sensitivity Results

Tables 2 and 3 report the full benchmark results for component ablations and sensitivity to correction strength; Table 4 reports the group-size and GSPO comparisons. This appendix reports benchmark-level sensitivity to the reference credible level.

Table 11: Qwen3-8B results across three seeds per method ($G = 4$, %). Run indices identify seeds; SD denotes the sample standard deviation. Evaluation follows Table 1.

	GRPO				RAC-GRPO			
Benchmark	Run 1	Run 2	Run 3	Mean ± SD	Run 1	Run 2	Run 3	Mean ± SD
AIME24	55.10	54.38	54.79	54.76 ± 0.36	56.88	55.94	56.77	56.53 ± 0.51
AIME25	41.04	41.15	42.19	41.46 ± 0.63	48.02	49.27	43.96	47.08 ± 2.78
AIME26	48.13	46.77	47.08	47.33 ± 0.71	48.23	50.83	52.19	50.42 ± 2.01
AMC23	86.41	87.89	88.36	87.55 ± 1.02	92.03	90.86	91.56	91.48 ± 0.59
MATH500	92.80	94.20	94.00	93.67 ± 0.76	94.60	94.40	94.60	94.53 ± 0.12
Minerva	45.59	46.32	45.22	45.71 ± 0.56	46.69	48.16	47.79	47.55 ± 0.76
Olympiad	68.25	69.88	70.47	69.53 ± 1.15	71.51	71.36	71.07	71.31 ± 0.22
Average	62.47	62.94	63.16	62.86 ± 0.35	65.42	65.83	65.42	65.56 ± 0.24

E.1 Sensitivity to the Credible Level

We vary the credible level of the reference posterior interval used to calibrate the count-dependent pseudo-count on Qwen3-4B. The main experiments use the 95% level. Table 12 compares that three-seed result with single-seed runs at 80%, 85%, and 90%.

At fixed K and G, lowering the credible level produces a nested, narrower reference interval. If $C'_{K,G} \subseteq C_{K,G}$, every pseudo-count satisfying $C_{K,G} \subseteq \mathcal{I}_G(c)$ also satisfies $C'_{K,G} \subseteq \mathcal{I}_G(c)$: the fixed-$c$ MSE comparison is required over fewer success probabilities. The feasible set therefore expands, and its largest pseudo-count cannot decrease under the same cap (Equation 9); some entries can remain unchanged. Since $|b_K(c)|$ increases with c at unbalanced counts (Equation 7), this permits stronger corrections, including larger endpoint advantages. The lower scores at 80–90% are consistent with sensitivity to correction strength, as also observed in Table 3 and Figure 3. The single-seed runs remain exploratory and do not establish why performance changes.

Table 12: Qwen3-4B sensitivity to the reference posterior credible level (%). Average is the unweighted mean over all seven benchmarks. GRPO and the 95% setting are three-seed means from Table 1; each 80–90% setting uses one exploratory run.

Method / credible level	AIME24	AIME25	AIME26	AMC23	MATH500	Minerva	Olympiad	Average
GRPO	53.51	47.22	49.90	89.77	93.60	47.59	70.03	64.52 ± 0.47
RAC-GRPO (80%)	50.21	45.52	48.75	89.45	94.20	47.06	69.88	63.58
RAC-GRPO (85%)	53.02	**48.54**	47.50	**91.80**	**95.20**	46.32	71.81	64.88
RAC-GRPO (90%)	54.90	47.08	47.81	90.86	94.60	47.43	72.55	65.03
RAC-GRPO (95%)	**57.88**	47.78	**52.71**	90.73	95.18	**49.06**	**72.85**	$\mathbf{66.60 \pm 0.44}$