WITHIN-CLASS SCORE VARIATION IN STANDARDIZED GROUP-RELATIVE POLICY GRADIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Success rates determine how often a binary-reward group contains a learning signal, but they do not determine the variability of its policy-gradient update. For the on-policy, unclipped standardized estimator used in Group Relative Policy Optimization (GRPO), we derive an exact finite-group covariance decomposition into success-count variation and score variation within each reward class. A four-outcome policy shows that the latter can change while success probability and the expected update remain fixed. We also derive the estimator's influence function: estimating the reward standard deviation contributes to the leading covariance, even though that estimate is consistent. In an audit of 3,072 responses from two Qwen checkpoints, within-class variation accounts for median covariance fractions of 97.8% and 82.4% at group size eight, measured in fixed last-layer LoRA coordinates and conditional on mixed empirical response pools. Validation with 4,096 independent responses separates this algebraic identity from moment-estimation accuracy: 32-response calibration pools underpredict aggregate covariance by about 16% for the smaller model. A separate nine-run sampling-allocation study gives mean held-out accuracy of 57.67% for covariance-informed allocation and 57.11% for uniform allocation; a descriptive paired interval for the difference includes zero. Success rates therefore omit a substantial source of update variation in the measured score space, but using this additional information for sampling has yet to yield a reliable accuracy gain.

1 INTRODUCTION

Two prompts can have the same success rate and the same expected update, yet produce different update variability. This distinction matters for GRPO, which replaces a learned critic with relative rewards within a sampled group (Shao et al., 2024; Guo et al., 2025). The reward counts reveal whether the group is mixed. The likelihood-score vectors determine how those rewards translate into a parameter update. Summarizing the former does not recover the latter.

Our analysis separates success-count and within-class score variation at finite group sizes, then derives their large-group limits. We evaluate the result in two stages: resampling saved responses checks the conditional identity, and fresh model draws test the estimated moments. Independent responses reveal calibration error even when resampling agrees with the formula. Finally, nine online training runs test whether a frozen allocation based on these covariance estimates improves held-out accuracy. Together, these experiments assess both the diagnostic and its use in sampling.

Relation to normalization analyses. Non-uniform group weighting and its interaction with optimizer dynamics are analyzed by Fontana et al. (2026). The group-standard-deviation identity, finite-group attenuation, and silent-group probabilities are treated by Bay & Yearick (2026). Those studies establish the objective change induced by normalization and the probability of binary group degeneracy. We focus on a covariance decomposition and the distinction between a scalar success-count diagnostic and vector-valued update variation. The covariance describes the raw estimator before optimizer transformations. Zhou et al. (2026) develop a U-statistic analysis with finite-sample and optimization results, and extend it to normalized rewards in their Appendix A. Their sample-standard-deviation and leave-one-out conventions differ from ours by the deterministic factor $\sqrt{G/(G-1)}$ in the on-policy, unregularized case. That factor tends to one and does not remove denominator

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.

fluctuations. We distinguish a probability-limit substitution from a first-order covariance calculation: a random denominator can be consistent while still contributing at order $G^{-1/2}$. The binary example below makes that contribution exactly computable. Our result is not a general analysis of their off-policy or regularized optimization setting.

2 Related work and estimator distinctions

Policy gradients and control variates. REINFORCE (Williams, 1992) and the policy-gradient theorem with function approximation (Sutton et al., 1999) provide the likelihood-score basis for the estimator studied here. A baseline independent of the sampled action preserves its expectation, while the variance-minimizing baseline depends on score magnitudes (Greensmith et al., 2004). Generalized advantage estimation instead controls the bias–variance tradeoff in temporal credit assignment (Schulman et al., 2016). Our binary terminal-reward calculation concerns within-prompt sample normalization, not an alternative temporal-credit rule. REINFORCE-style learning from human feedback, including leave-one-out baselines (Ahmadian et al., 2024), supplies a relevant comparison because changing the baseline convention does not have the same effect as dividing by a random reward standard deviation.

Surrogate losses and implementation choices. Trust-region policy optimization (Schulman et al., 2015) and proximal policy optimization (Schulman et al., 2017) address policy updates under distributional change. The on-policy derivative of an unclipped surrogate is a useful local object, but its covariance alone does not characterize several clipped optimization epochs. Human-feedback training (Ouyang et al., 2022) and direct preference optimization (Rafailov et al., 2023) also involve reference-policy regularization; their objectives and data distributions differ from the clean, freshly sampled binary-reward groups considered here. These differences in clipping, regularization, and data collection require separate analyses of each estimator and its optimization dynamics.

Group-relative reasoning optimization. DeepSeekMath introduces GRPO (Shao et al., 2024), and DeepSeek-R1 demonstrates its role in large-scale reasoning training (Guo et al., 2025). Dr. GRPO examines length and difficulty biases in the reduction and normalization rules (Liu et al., 2025). DAPO combines changes including dynamic sampling and decoupled clipping in a full training system (Yu et al., 2025). Our calculation fixes three implementation choices: the sample-variance divisor, zero-variance convention, and completion reduction. It describes within-prompt covariance under those choices. Dynamic prompt selection and the resulting changes to training performance require additional analysis.

Finite-sample statistics and measured coordinates. U-statistic theory (Hoeffding, 1948) and the delta method (van der Vaart, 1998) distinguish exact finite-sample identities from first-order approximations. The random reward denominator is part of the statistic, so its first-order contribution must be tracked along with the numerator. In the model audit, LoRA (Hu et al., 2022) defines a tractable score coordinate space. The resulting covariance trace depends on that choice; it is not an intrinsic measure of functional policy change. In particular, Adam's adaptive preconditioning (Kingma & Ba, 2015) would transform an applied update differently from the raw gradient whose covariance we report.

Empirical scope and uncertainty. The reproducibility issues documented by Henderson et al. (2018) and the statistical evaluation analysis of Agarwal et al. (2021) motivate separating training-seed variation, prompt variation, and within-prompt sampling. A bootstrap (Efron, 1979) over stored responses samples from an empirical distribution; it does not add fresh model evidence. Accordingly, the resampling agreement in our audit checks the conditional moment calculation. It is not evidence that 32 responses accurately estimate every prompt's population covariance, especially when a rare reward class is absent.

3 Estimator and assumptions

Fix a prompt x and a policy parameter θ. Let (r_i, s_i) be i.i.d. samples, where $r_i = r(o_i, x)$ and $s_i = \nabla_\theta \log \pi_\theta(o_i|x)$. Assume the usual score identity $\mathbb{E}s_i = 0$ and differentiation under the

expectation, so that

$$\mu = \mathbb{E}r, \quad \sigma^2 = \mathrm{Var}(r) > 0, \quad a = \mathbb{E}[(r-\mu)s] = \nabla_\theta \mathbb{E}[r \mid x].$$

For the asymptotic result, assume bounded rewards and scores, $|r| \le R$ and $\|s\| \le S$, in fixed parameter dimension. These are mathematical conditions for the theorem; the model audit does not certify them globally. Define

$$\bar{r} = G^{-1}\sum_i r_i, \quad \hat{\sigma}^2 = G^{-1}\sum_i (r_i - \bar{r})^2, \qquad \hat{g}_G = \frac{G^{-1}\sum_i (r_i - \bar{r})s_i}{\hat{\sigma}}.$$

When $\hat{\sigma} = 0$, define $\hat{g}_G = 0$. We analyze exact standardization; keeping a fixed positive stabilizer in the denominator changes the population denominator and requires its own expansion. Clipping, KL penalties, off-policy ratios, multi-epoch updates, and token-specific reductions are outside this model.

4 Exact finite-group covariance for binary rewards

Responses with the same binary reward need not have the same score vector. To distinguish reward-count variation from this within-class variation, let $p = \Pr(r = 1) \in (0, 1)$ and define

$$\mu_+ = \mathbb{E}[s \mid r = 1], \quad \mu_- = \mathbb{E}[s \mid r = 0], \quad \Delta = \mu_+ - \mu_-, \quad \Sigma_\pm = \mathrm{Cov}(s \mid r = 1 \text{ or } 0),$$

where Σ_+ conditions on one and Σ_- on zero. Finite second moments suffice for the result below. This formulation also applies to a fixed empirical distribution over saved response–score pairs, although such a distribution does not in general satisfy the population score identity.

Theorem 1 (Exact conditional-geometry decomposition)**.** *For an i.i.d. group of size $G \ge 2$, put*

$$m_G(p) = \sum_{k=1}^{G-1} \binom{G}{k} p^k (1-p)^{G-k} \sqrt{\frac{k}{G}\left(1 - \frac{k}{G}\right)}.$$

With the same zero-variance convention as before,

$$\mathbb{E}\hat{g}_G = m_G(p)\Delta, \tag{1}$$

$$\mathrm{Cov}(\hat{g}_G) = v_G(p)\Delta\Delta^\top + a_G(p)\Sigma_+ + b_G(p)\Sigma_-, \tag{2}$$

$$v_G(p) = p(1-p)(1-G^{-1}) - m_G(p)^2,$$

$$a_G(p) = \frac{(1-p) - (1-p)^G}{G}, \qquad b_G(p) = \frac{p - p^G}{G}.$$

At $p = 0$ or 1, the estimator is identically zero; no conditional moment of an empty reward class is needed.

Proof. Condition on K, the number of correct responses. For $0 < K < G$, elementary algebra gives

$$\hat{g}_G = h(K)(\bar{s}_+ - \bar{s}_-), \qquad h(K) = \sqrt{(K/G)(1 - K/G)}.$$

The class means are conditionally independent, with conditional covariance $\Sigma_+/K + \Sigma_-/(G-K)$. The conditional expectation is $h(K)\Delta$. The laws of total expectation and covariance therefore give the result once three scalar moments are evaluated. The binomial second moment gives $\mathbb{E}h(K)^2 = p(1-p)(1-G^{-1})$. Also,

$$\mathbb{E}\left[\frac{h(K)^2}{K}\mathbf{1}_{\{0<K<G\}}\right] = \frac{G(1-p) - G(1-p)^G}{G^2} = a_G(p).$$

The symmetric expression is $b_G(p)$. Subtracting $(\mathbb{E}h)^2$ gives v_G. All formulas include the all-correct and all-incorrect groups through their zero contribution. □

The underlying identity $\hat{g}_G = h(K)(\bar{s}_+ - \bar{s}_-)$ and its expected scalar gain have already been studied by Bay & Yearick (2026); we use them to derive the separate covariance contributions. The covariance expression makes an additional distinction explicit: variation in the number of successes and variation among equally rewarded score vectors contribute separately. A solve-rate histogram alone cannot recover the latter terms. This distinction concerns clean rewards and group sampling, rather than an intervention that redistributes verifier mistakes.

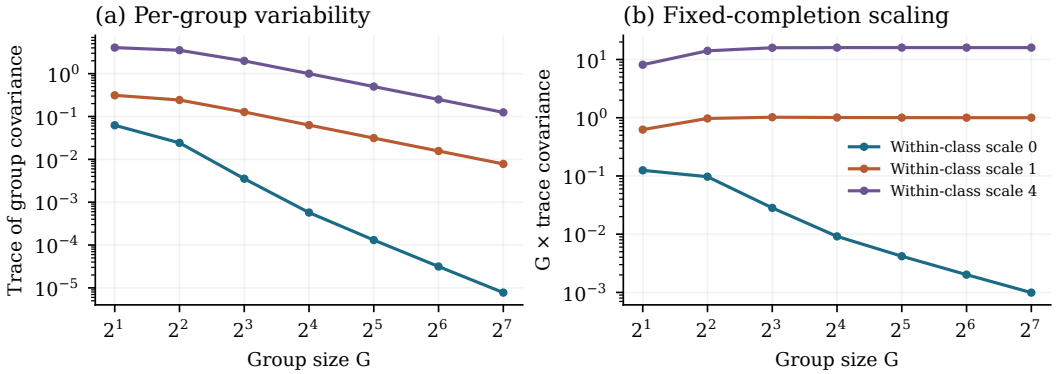


Figure 1: Exact covariance in the four-outcome example with fixed success probability $1/2$ and fixed conditional-mean contrast. Increasing the within-class score scale changes variability while leaving the expected update unchanged. The right panel multiplies by group size to show the variance coefficient when a fixed completion budget is split into independent groups at the same prompt. These are analytical curves, not training results.

Corollary 2 (Binary large-group covariance)**.** *For fixed* $0 < p < 1$ *and finite conditional second moments,*

$$\lim_{G\to\infty} G\mathrm{Cov}(\hat{g}_G) = \frac{(1-2p)^2}{4}\Delta\Delta^\top + (1-p)\Sigma_+ + p\Sigma_-.$$

Proof. The binomial delta method applied to $h(u) = \sqrt{u(1-u)}$ gives $Gv_G(p) \to h'(p)^2 p(1-p) = (1-2p)^2/4$. Moment convergence follows by restricting u to a fixed neighborhood of p, where h is smooth, and bounding its complement with binomial concentration. The two remaining limits follow directly from $Ga_G(p) \to 1-p$ and $Gb_G(p) \to p$. □

At $p = 1/2$, the success-count term has a vanishing first-order coefficient, but the within-class contribution is $(\Sigma_+ + \Sigma_-)/2$. A scalar Bernoulli policy, which has no within-class score variation, is therefore a special case. Its first-order degeneracy does not transfer to arbitrary binary-reward policies.

What matching success rates leaves unidentified. Consider four equally likely outcomes with rewards $(1, 1, 0, 0)$ and score vectors $(1/2, t), (1/2, -t), (-1/2, t), (-1/2, -t)$. They have zero mean score and can be realized as local scores of a finite exponential-family policy at its uniform parameter. Every value of t gives $p = 1/2$ and $\Delta = (1, 0)^\top$, hence the same expected standardized update for every G. Yet $\Sigma_+ = \Sigma_- = \mathrm{diag}(0, t^2)$, so the covariance trace increases by $(a_G + b_G)t^2$. Equal accuracy, group size, and mean update therefore do not identify update variability. This construction isolates within-class variation in a finite policy with exactly known probabilities and scores.

Fixed completion budgets. For B/G independent groups at one fixed prompt, where G divides B, the covariance of their average is $(G/B)\mathrm{Cov}(\hat{g}_G)$. If a fresh prompt is drawn for each group, the covariance additionally contains the between-prompt term $(G/B)\mathrm{Cov}_x(m_G(p_x)\Delta_x)$. This follows directly from total covariance. Increasing group size both changes the target and reduces the number of distinct prompt draws. A reduction in the silent-group rate alone is consequently insufficient to select an accuracy-optimal group size.

Independent numerical checks. We compared Equation (2) with exhaustive enumeration of all ordered groups for 24 four-outcome, three-dimensional distributions and $G \in \{2, 3, 4, 5\}$. Conditional score means were centered to satisfy the score identity. All 96 cases agreed in mean and covariance to within 3×10^{-15} in maximum absolute entry error using double precision. This is a consistency check of an algebraic result, not empirical support for a language-model performance claim.

5 FIRST-ORDER ANALYSIS

Theorem 3 (Fixed-prompt estimand and influence function)**.** *Under the preceding assumptions,* $\hat{g}_G \to a/\sigma$ *in probability and in mean. Define*

$$\psi(r,s) = \frac{(r-\mu)s - a}{\sigma} - \frac{a}{2\sigma^3}\big((r-\mu)^2 - \sigma^2\big).$$

Then

$$\sqrt{G}(\hat{g}_G - a/\sigma) \implies \mathcal{N}(0, \mathrm{Cov}(\psi)), \qquad G\mathrm{Cov}(\hat{g}_G) \to \mathrm{Cov}(\psi).$$

The Gaussian limit may be degenerate. The assertions are pointwise at a fixed $\sigma > 0$ *and do not provide a uniform expansion as* $\sigma \downarrow 0$*.*

The proof in Appendix A expands numerator and denominator jointly, and controls small sample variances using boundedness and concentration.

The denominator contribution in ψ is of the same first order as the centered numerator. Consequently, the leading covariance is not generally $\mathrm{Cov}((r-\mu)s)/(G\sigma^2)$, and $G\sigma^2$ is not a general effective sample size relative to REINFORCE with a baseline.

Proposition 4 (Finite-group norm bound)**.** *If* $\|s_i\| \le S$*, then* $\|\hat{g}_G\| \le S$ *for every realized group, and* $\mathrm{TrCov}(\hat{g}_G) \le S^2$*.*

Proof. For nonzero sample variance, Cauchy–Schwarz gives

$$\left\| G^{-1} \sum_i (r_i - \bar{r}) s_i \right\| \le \sqrt{G^{-1} \sum_i (r_i - \bar{r})^2} \sqrt{G^{-1} \sum_i \|s_i\|^2} \le \hat{\sigma} S.$$

The zero-variance convention satisfies the same bound. Finally, $\mathrm{TrCov}(\hat{g}_G) \le \mathbb{E}\,\|\hat{g}_G\|^2 \le S^2$. □

The Bernoulli-policy calculation and prompt-mixture degeneracy identities are given in Appendix B. They distinguish denominator effects from within-class score variation and show why aggregate success rates can underestimate the frequency of silent groups.

6 A FIXED-POLICY LANGUAGE-MODEL SCORE AUDIT

6.1 QUESTION SELECTION AND RESPONSE GENERATION

We apply the decomposition to saved responses from Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct (Qwen et al., 2024). A fixed hash ordering selects 48 GSM8K test questions (Cobbe et al., 2021), without conditioning on generated rewards. Both checkpoints use these same questions. Each prompt requests brief reasoning and a boxed numerical answer. For each question and checkpoint, we independently sample 32 responses with temperature one, unrestricted top-p and top-k, and a 512-token ceiling. There are 3,072 responses in total. The models receive their own chat templates; this is not a controlled model-size experiment.

A label-independent parser uses the last boxed numerical expression, with a fallback to an explicit answer field. Parsed integers, decimals, and simple fractions are normalized as exact rationals before comparison with GSM8K's reference. An unparsed output receives zero reward. Reward-positive and reward-negative therefore mean outcomes of this numerical checker, rather than independently validated mathematical reasoning. We retain the generated token IDs and full text so this distinction can be examined directly.

The 1.5B checkpoint receives 627 positive rewards among 1,536 outputs, with 510 unparsed responses. The 7B checkpoint receives 1,339 positive rewards, with five unparsed responses. These are outcomes of our sampled protocol, not standard benchmark accuracies. Three 1.5B pools are all negative. For 7B, two pools are all negative and 27 are all positive. Thus 45 and 19 prompt pools, respectively, contain both reward classes. A unanimous pool of 32 samples does not establish that its population success probability is exactly zero or one.

Table 1: Conditional empirical-pool diagnostics for sequence-sum scores. Ratios and fractions are medians across the 45 mixed 1.5B pools or 19 mixed 7B pools. MC is the 5,000-group resampling estimate; asymptotic includes the estimated-denominator term. Unanimous pools have exactly zero empirical-pool covariance and are excluded from ratios.

Model	G	MC / exact	Asymptotic / exact	Within-class fraction
1.5B	4	1.000	1.255	0.964
1.5B	8	0.999	1.038	0.978
1.5B	16	0.999	0.999	0.989
1.5B	32	1.001	0.999	0.991
7B	4	0.999	2.136	0.806
7B	8	1.001	1.299	0.824
7B	16	1.003	1.025	0.871
7B	32	0.996	0.995	0.914

6.2 Score coordinates and exact Gram matrices

For each checkpoint we attach rank-eight LoRA adapters to the query and value projections in its last transformer layer, with scale factor 16 and zero adapter dropout. The A factors use a fixed random initialization and the B factors start at zero, so adding the adapter does not change the base policy. These adapters remain untrained during the fixed-policy measurements. The resulting parameter coordinate spaces have dimensions 38,912 and 90,112. Some coordinates have zero gradient at this initialization; coordinate count is not a claim about intrinsic rank.

The primary score is the gradient of the sum of completion-token log probabilities, including a generated end-of-sequence token when present. Sampling and scoring use the same untempered model distribution. For a response that reaches the ceiling, the score is that of the observed finite prefix; no unobserved continuation is appended. Gradients are computed in the model's bfloat16 execution and converted to double precision before forming the full 32×32 Gram matrix. "Full" here means no random projection within this chosen tangent space; it does not mean a full-model parameter gradient or exact real arithmetic.

As a sensitivity analysis, dividing each saved score by its response length produces token-mean scores. These vectors generally do not satisfy the usual zero-mean likelihood-score identity, so the population policy-gradient interpretation changes. The finite-group moment decomposition itself remains valid for an arbitrary vector attached to each sampled response. We analyze the two reductions separately rather than treating length normalization as a common multiplicative constant.

6.3 Conditional empirical-pool analysis

Each question's 32 saved response–score pairs define a uniform empirical distribution. We compute its reward probability, conditional mean contrast, and within-class covariance traces from the Gram matrix, then evaluate Equation (2) for $G \in \{2, 4, 8, 16, 32\}$. We also draw 5,000 independent groups with replacement from that same pool at each group size. Agreement with this resampling checks the calculation conditional on the saved pool. Section 6.5 tests the estimated moments using fresh model draws.

For comparison, we evaluate the first-order covariance on the empirical centered score vectors, and a version that omits the denominator term. Centering is necessary because a finite pool need not have zero mean score. For pool mean $\bar{s}$, mean reward p, variance $\sigma^2 = p(1-p)$, and $a = \mathbb{E}_{\mathrm{pool}}[(r-p)(s-\bar{s})]$, the correct influence function uses $(r-p)(s-\bar{s})$ in place of $(r-\mu)s$ in Theorem 3. The exact finite-group formula requires no such approximation.

6.4 Observed sources of variability

Within-class score variation accounts for most of the conditional covariance trace in these mixed pools: at $G = 8$, its median fraction is 0.978 for 1.5B and 0.824 for 7B. The scalar success-count fluctuation alone therefore misses most of the variability in these particular score coordinates. The

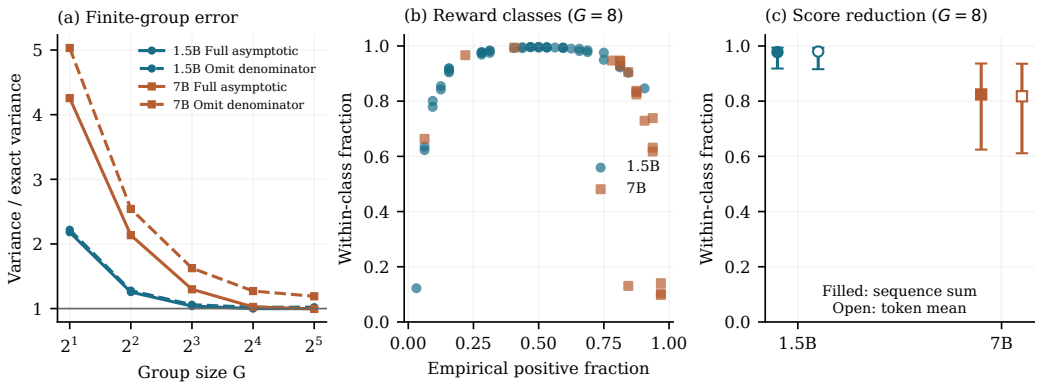


Figure 2: Fixed-policy empirical-pool diagnostics. (a) Median variance ratios relative to the exact finite-pool formula, including and omitting denominator fluctuations. (b) Each point is one mixed pool's within-class contribution at $G = 8$. (c) Medians and interquartile ranges across pools for sequence-sum (filled) and token-mean (open) scores; bars show heterogeneity, not confidence intervals. All panels condition on the observed response pools.

corresponding token-mean fractions are 0.980 and 0.818. This sensitivity retains the qualitative decomposition but does not establish invariance to arbitrary parameterizations or a choice of tangent subspace.

At $G = 4$, the full asymptotic trace is a median 1.255 times the exact finite-pool trace for 1.5B and 2.136 times for 7B. At $G = 32$, these median ratios are near one. Omitting denominator fluctuations gives ratios of 1.278 and 2.542 at $G = 4$, and 1.021 and 1.190 at $G = 32$. These comparisons describe approximation error relative to the empirical distribution; they are not errors relative to an independently measured true-policy covariance. Their different magnitudes also depend on the sampled reward probabilities and the chosen score space.

6.5 Independent response validation

An identity verified by resampling a fixed pool need not accurately predict new model draws when its moments are estimated from a small pool. We therefore select 16 of the 48 prompts per model by a second, outcome-independent hash ordering and generate 128 new responses per prompt. Generation settings and LoRA score coordinates are unchanged; generation seeds differ from those of the 32-response calibration pool. This adds 4,096 responses across the two models. No calibration-unanimous prompt is excluded.

For each prompt and $G \in \{4, 8\}$, the calibration pool supplies the reward frequency and class-conditional score moments in the exact covariance formula. The new responses are divided into disjoint groups in generation order, giving 32 or 16 group-gradient vectors. We estimate their covariance trace directly with the unbiased sample covariance. If M groups have Gram matrix K, this is

$$\widehat{\operatorname{tr}\operatorname{Cov}} = \frac{\operatorname{tr} K - M^{-1}\mathbf{1}^\top K \mathbf{1}}{M - 1}.$$

Thus no resampling of the calibration responses enters the observed covariance. Table 2 reports the ratio of summed calibration predictions to summed held-out estimates, retaining all 16 prompts. The intervals resample whole prompts 20,000 times and are descriptive for the selected prompt set; they do not certify performance on other tasks or coordinate spaces.

The 1.5B calibration estimates underpredict aggregate variability by about 16%; agreement within the original empirical pool does not remove this estimation error. For 7B the aggregate ratios bracket one across the two group sizes, with wider intervals. Two calibration prompts for 1.5B and nine for 7B have unanimous rewards and therefore zero empirical predictions. For each model, one of these prompts exhibits nonzero covariance in the independent sample. A finite pool can miss a reward class entirely. Exact finite-G statistics consequently address estimator algebra, while independent draws are needed to assess how accurately the required moments have been estimated.

Table 2: Independent-draw validation of sequence-score covariance traces. A ratio of one indicates agreement of aggregate magnitudes. Each row uses 16 prompts, 32 calibration responses and 128 fresh responses per prompt.

Policy	G	Predicted / observed	Prompt-bootstrap 95% interval
Qwen2.5-1.5B	4	0.838	[0.724, 0.958]
Qwen2.5-1.5B	8	0.836	[0.730, 0.936]
Qwen2.5-7B	4	1.113	[0.855, 1.383]
Qwen2.5-7B	8	0.951	[0.785, 1.162]

7 Sampling allocation and training outcomes

7.1 A fixed-policy allocation check

We select 32 GSM8K training questions by an outcome-independent hash order. For each model, 16 responses per question calibrate the success rate $\hat{p}_x$ and the group-size-four covariance trace $\hat{V}_x$. Another 16 fresh responses per question form independent validation pools, adding 2,048 responses across both models. None of these questions is selected using test accuracy.

For $n = 32$ prompts and m_x independent groups at each, the equal-prompt average has conditional covariance trace

$$\mathrm{TrCov}\left(\frac{1}{n}\sum_x \frac{1}{m_x}\sum_{j=1}^{m_x}\hat{g}_{xj}\right) = \frac{1}{n^2}\sum_x \frac{V_x}{m_x}. \tag{3}$$

The continuous fixed-budget allocation is proportional to $\sqrt{V_x}$ when all $V_x > 0$, by a Lagrange multiplier calculation. We implement a conservative integer heuristic: assign one of 64 groups to every prompt, then distribute the remaining 32 by largest remainders. The three weights are uniform, $\sqrt{\hat{p}_x(1-\hat{p}_x)}$, and $\sqrt{\hat{V}_x}$; ties use the fixed prompt order, and an all-zero weight vector falls back to uniform. This rounding rule is not claimed to solve the constrained integer optimization exactly.

Keeping the allocations fixed, we evaluate Equation (3) using the independent validation pools' moments. Relative to success-only allocation, covariance-informed allocation lowers this conditional variance by 2.47% for 1.5B but raises it by 0.46% for 7B. These are finite-pool comparisons, not estimates of training accuracy or population-optimal budgets.

7.2 An online training intervention

We next compare the three frozen allocations on Qwen2.5-1.5B-Instruct, using three sampling/order seeds per condition and the same adapter initialization. Each run makes 256 updates from fresh groups of four responses on the same 32 training prompts. Four shuffled cycles repeat each prompt according to its allocated group count. A group at prompt x receives weight $1/(32q_x)$, where $q_x = m_x/64$, to offset static prompt reweighting. The last-layer query/value LoRA adapters now train with AdamW, learning rate 10^{-5}, zero weight decay, and gradient-norm clipping at one. We use sequence-sum scores and standardized binary rewards, with no reference KL or PPO ratio clipping. Generation uses the audit's settings. Appendix D specifies scheduling and cost.

Table 3: Held-out GSM8K accuracy (%) after 256 updates. All runs start at 54.33% on the same 300 questions. Seeds change sampling and prompt order; adapter initialization is shared. Tokens are mean generated training tokens, excluding calibration and evaluation.

Allocation	Seed 41	Seed 42	Seed 43	Mean	Tokens (k)
Uniform	57.67	56.00	57.67	57.11	256.4
Success-only	55.00	57.67	56.00	56.22	256.7
Covariance-informed	57.33	58.33	57.33	57.67	263.4

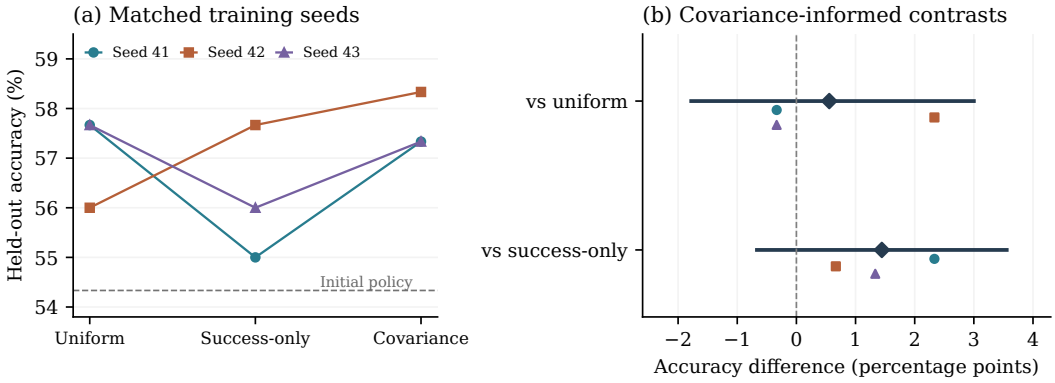


Figure 3: From allocation to training outcomes. Left: final accuracy for each sampling/order seed; lines connect matched seeds and the dashed line is the common starting accuracy. Right: covariance-informed minus each comparator, with individual seed differences (small points), their mean (diamond), and post hoc paired seed/question bootstrap 95% intervals. Neither interval excludes zero. The nine runs share one model and a 32-prompt training set.

The covariance-informed mean exceeds uniform by 0.56 percentage points, but the paired seed differences are -0.33, $+2.33$, and -0.33 points. A post hoc paired bootstrap over seeds and shared test questions gives a 95% interval of $[-1.78, 3.00]$ points. The difference from success-only allocation is $+1.44$ points, with interval $[-0.67, 3.56]$. These descriptive intervals use 20,000 draws and only three seeds; they do not establish a reliable accuracy advantage (Figure 3).

Equal response counts do not imply equal computation. Covariance-informed runs generate 2.73% more training tokens than uniform runs on average. Their calibration additionally requires 512 responses (104,143 output tokens) and 512 score backward passes, shared across seeds. Success-only calibration uses the responses but needs no score backward passes; uniform sampling needs neither. The comparison therefore provides no evidence of end-to-end compute savings. Gradient clipping is active in 67.2–70.7% of updates. Adam, clipping, changing policies, and sequential rather than fixed-policy stratified updates all separate this intervention from Equation (3). These differences limit the connection between the fixed-policy variance criterion and the observed training outcomes.

8 Limitations and Conclusion

At one prompt, a/σ is a positive rescaling of a, but averaging across prompts reweights prompt-specific gradients. This field need not be the gradient of the original expected-reward objective. In particular, a bias bound of order G^{-1} relative to the unnormalized gradient is generally false. Convergence and group-size selection therefore require an additional analysis of the optimization objective, prompt sampling, and successive policy updates.

The bounded-score assumptions are not globally certified for these neural policies. The Bernoulli calculation is a mathematical check, not evidence of LLM performance. The fixed-policy audit uses only 48 prompts per model and one untrained last-layer LoRA coordinate space. The independent validation covers only 16 of those prompts per model, and answer extraction affects the reward classes. Calibration-pool unanimity can miss nonzero covariance in fresh samples. The online intervention adds one model, three sampling/order seeds, and only 32 training prompts. Frozen allocation does not track changing score geometry, and the descriptive bootstrap cannot compensate for this narrow scope.

Success counts determine group degeneracy, but leave out variation among equally rewarded responses. Our exact decomposition quantifies this missing term, which dominates most mixed pools in the measured LoRA coordinates. Fresh responses expose errors in estimating its magnitude, and the nine training runs do not establish an accuracy benefit from the tested allocation. Improving this application requires more reliable covariance estimates and an allocation rule that accounts for policy changes and optimizer dynamics.

Reproducibility Statement

The supplement provides the data and code needed to reproduce the reported analyses, following Pineau et al. (2021). The estimator, zero-variance convention, and assumptions are specified in the mathematical sections, which include proofs of the covariance results. The fixed-policy audit describes question selection, response generation, score coordinates, precision, and empirical-pool resampling. The numerical checks separate exhaustive verification of finite examples from conditional analyses of language-model responses. The supplement also preserves the 4,096 independent validation responses and their Gram matrices, so calibration predictions and held-out covariance estimates can be recomputed without another model run. The additional allocation pools, all nine training trajectories, final per-question predictions, and paired analysis code reproduce the sampling intervention.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29304–29320. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

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, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

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.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Aleksandar Fontana, Marco Simoni, Giulio Rossolini, Paolo Mori, and Andrea Saracino. On the hidden objective biases of group-based reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 109–121, 2026. doi: 10.18653/v1/2026.acl-short.11. URL https://aclanthology.org/2026.acl-short.11/.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://www.jmlr.org/papers/v5/greensmith04a.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning That Matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11694. URL https://ojs.aaai.org/index.php/AAAI/article/view/11694.

Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

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. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Lariviere, Alina Beygelzimer, Florence d'Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL https://jmlr.org/papers/v22/20-303.html.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

John Schulman, Sergey Levine, Philipp Moritz, Michael Jordan, and Pieter Abbeel. Trust Region Policy Optimization. In *International Conference on Machine Learning*, volume 37, pp. 1889–1897, 2015. URL https://proceedings.mlr.press/v37/schulman15.html.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. URL https://arxiv.org/abs/1707.06347.

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. URL https://arxiv.org/abs/2402.03300.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/464d828b85b0bed98e80ade0a5c43b0f-Abstract.html.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256. URL https://doi.org/10.1017/CBO9780511802256.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

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/2603.01162v3.

A PROOF OF THE INFLUENCE-FUNCTION RESULT

Proof. Write $Y_i = r_i - \mu$, $X_i = Y_i s_i$, $\bar{Y} = G^{-1}\sum_i Y_i$ and $\bar{s} = G^{-1}\sum_i s_i$. The numerator is $\bar{X} - \bar{Y}\bar{s}$, and $\hat{\sigma}^2 = G^{-1}\sum_i Y_i^2 - \bar{Y}^2$. Both products of centered sample means are $O_p(G^{-1})$. Taylor expansion of $u/v^{1/2}$ at (a, σ^2) gives

$$\hat{g}_G - \frac{a}{\sigma} = \frac{1}{G}\sum_i \left[\frac{X_i - a}{\sigma} - \frac{a(Y_i^2 - \sigma^2)}{2\sigma^3}\right] + O_p(G^{-1}).$$

The multivariate central limit theorem yields the distributional result. For moment convergence, restrict first to $\hat{\sigma}^2 \geq \sigma^2/2$; there the second derivatives are bounded in a fixed neighborhood, and bounded sample moments give an L^2 remainder of order G^{-1}. The complementary event has exponentially small probability by concentration of bounded Y_i and Y_i^2. The bound $\|\hat{g}_G\| \leq S$ proved below controls that event, so its contribution to the scaled second moment vanishes. The expansion then gives covariance convergence. Boundedness and convergence in probability also give convergence in mean. □

B SCALAR POLICIES AND PROMPT HETEROGENEITY

A constant baseline is a different operation. For $g_b = (r - b)s$, the mean is a for every fixed b. When $\mathbb{E}\|s\|^2 > 0$, minimizing its covariance trace gives

$$b^* = \frac{\mathbb{E}[r\|s\|^2]}{\mathbb{E}[\|s\|^2]},$$

by differentiating $\mathbb{E}[(r-b)^2\|s\|^2] - \|a\|^2$. This need not equal μ (Greensmith et al., 2004).

B.1 RANDOM-DENOMINATOR EFFECTS IN A BERNOULLI POLICY

Take a Bernoulli-logit policy, with reward $r \in \{0, 1\}$, success probability p, and score $s = r - p$. If $K \sim \text{Binomial}(G, p)$ and $u = K/G$, then

$$\hat{g}_G = \sqrt{u(1-u)}, \qquad a = p(1-p), \quad a/\sigma = \sqrt{p(1-p)}.$$

The expression includes $K = 0, G$ under the zero-variance convention. Thus its finite-G moments can be computed by summing over $K = 0, \ldots, G$, without sampling. The delta method gives

$$\lim_{G\to\infty} G\text{Var}(\hat{g}_G) = \frac{(1-2p)^2}{4}.$$

At $p = 0.2$, the true gradient is 0.16, the estimator mean tends to 0.4, and the scaled variance tends to 0.09. Omitting denominator fluctuations would instead predict 0.36. At $p = 1/2$, the first-order variance vanishes; this is a degenerate first-order limit, not identically zero finite-G variance. Our accompanying deterministic check enumerates $p \in \{0.2, 0.5, 0.8\}$ and $G \in \{2, 8, 32, 1000, 10000\}$.

B.2 GROUP DEGENERACY AND PROMPT HETEROGENEITY

Proposition 5 (Binary reward groups)**.** *For i.i.d. binary rewards at a fixed prompt with success probability p, the probability of a group with no relative reward signal is*

$$D(G, p) = p^G + (1-p)^G.$$

For $0 < p < 1$, the smallest integer G satisfying $D(G, p) \leq 1/2$ is a majority-informative threshold. As $p \to 1$, this threshold is asymptotic to $\log 2/(1-p)$; the symmetric statement holds as $p \to 0$.

Proof. The disjoint events of all-success and all-failure have probabilities p^G and $(1-p)^G$. Near $p = 1$, solving $p^G \approx 1/2$ yields $G \approx \log 2/(-\log p) \sim \log 2/(1-p)$; the other term is negligible at this scale. Integer rounding does not change the asymptotic equivalence. □

At $p = 0$ or $p = 1$, no finite group size is informative. For a prompt mixture with conditional success probability p_x, the aggregate degeneracy is $\mathbb{E}_x[D(G, p_x)]$, not generally $D(G, \mathbb{E}_x p_x)$. For $G \geq 2$, convexity and Jensen's inequality give

$$\mathbb{E}_x[D(G, p_x)] \geq D(G, \mathbb{E}_x p_x).$$

Using the overall accuracy in the fixed-prompt formula can therefore underestimate the fraction of uninformative groups. This limits claims based on aggregate training curves or schedules estimated from a small sample.

C Using the covariance diagnostic

C.1 From an exact identity to a calibrated diagnostic

The formula has two distinct uses. Conditional on a saved response pool, it is exact for resampling that pool. Applied to unseen model draws, it is a prediction based on estimated reward frequencies, class means, and covariances. Agreement in the first use does not validate the second. The 1.5B validation shortfall illustrates this distinction without contradicting the identity.

Unanimous pools are an extreme example. Their empirical covariance prediction is zero at every group size, and resampling cannot reveal a missing reward class. Fresh responses can, as observed for one prompt per model. Mixed pools also leave uncertainty when one class contains few responses. Aggregate agreement therefore need not imply accurate prompt-wise predictions.

The practical diagnostic has three parts: record the reward counts, estimate class-conditional score moments in declared coordinates, and compare predicted covariance with independent response groups. These distinguish signal availability, within-class variability, and calibration error. A silent-group rate supplies only the first part.

Score reduction remains part of the estimator definition. Dividing by realized response length changes conditional means and covariances rather than applying one common scale. The sequence-sum and token-mean sensitivities must therefore be read separately. Neither raw trace predicts the effects of adaptive preconditioning, clipping, or subsequent policy changes. Evaluating a training intervention requires measurements beyond this fixed-policy audit.

D Allocation and training protocol

Data and calibration. Training questions are the first 32 in a SHA-256 ordering of GSM8K training indices with prefix `allocation-train-v1:`. The 300 evaluation questions use the test split and prefix `allocation-heldout-v1:`; all conditions share their order. The two splits have separate roles. The test set belongs to the same benchmark used by the earlier fixed-policy audit; this is not a new-task replication. Calibration and independent-validation generation seeds are 20260923 and 20261023, offset by question index. Each uses 16 responses per prompt at temperature one, top-$p = 1$, top-$k = 0$, and a 512-token ceiling. The selected questions, texts, token IDs, rewards, and score Grams are provided. All prompts, including unanimous pools, enter the allocation.

Integer allocation and update schedule. For nonnegative weights w_x, put $a_x = 32 w_x / \sum_z w_z$ and initially $m_x = 1 + \lfloor a_x \rfloor$. Assign the remaining groups to the largest fractional parts of a_x, with stable prompt-order tie breaking. Uniform weights give two groups per prompt. For each training cycle, form a list with m_x copies of each question and shuffle it using the run seed plus cycle number. Four cycles give 256 updates and 1,024 fresh training responses per run. Generation uses seed $10000 \times$ run seed + update index. Seeds 41, 42, and 43 vary sampling and scheduling, while adapter initialization always uses 20260922. These are not three independently initialized neural models.

Optimization. The adapter has rank eight, scale 16, and zero dropout, and targets the last layer's query and value projections. Both adapter factors train. The optimizer is AdamW with learning rate 10^{-5} and zero weight decay; other optimizer settings are the framework defaults. The four response losses are accumulated before one optimizer step. Binary advantages use population-divisor sample standard deviation plus 10^{-8}, making all-equal groups contribute zero. Completion log probabilities are summed, not averaged by response length. The static weight $1/(32 q_x)$ offsets unequal prompt

frequencies for a fixed-policy average. It does not make the adaptive sequence of clipped Adam updates identical across allocations or certify unbiased conditional updates under sampling without replacement. The conditional covariance in Equation (3) applies to independent groups at a fixed policy; it is not an optimizer-level covariance identity.

Evaluation and uncertainty. Evaluation is greedy, with the same numerical parser and 512-token ceiling, on all 300 selected questions. We retain every seed. The reported intervals are descriptive post hoc percentile intervals from 20,000 paired draws with random seed 20260922. Each draw samples three seed indices with replacement and 300 question indices with replacement, using the same question draw for both methods and every sampled seed. Question-only intervals, conditional on the three runs, are $[-0.89, 2.11]$ points versus uniform and $[-0.11, 3.11]$ versus success-only. None of these intervals provides a multiplicity-adjusted or broad cross-model claim. Per-run outcomes are in Table 3.

Resource accounting. All training conditions have equal generation-call counts and output ceilings; actual response lengths differ. Mean output-token counts are 256,391, 256,660, and 263,388 for uniform, success-only, and covariance-informed allocation. The calibration overhead is not included in those training totals. For 1.5B, calibration costs 104,143 output tokens and the independent validation costs 103,819; for 7B these counts are 98,458 and 96,992. Each pool contains 512 responses and 512 backward score evaluations. Only the 1.5B calibration is used in training; the independent pools evaluate the fixed-policy diagnostic. Teacher-forced score evaluation, prompt processing, model loading, and final test generation add costs not captured by output-token counts. The study therefore establishes neither equal total FLOPs nor an end-to-end speedup.