RELIABLE LOCAL CREDIT FROM STRATIFIED CONTINUATION EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Sample-based local credit uses auxiliary continuations to evaluate intermediate reasoning decisions. With a finite evidence budget, accurate value estimates are only part of the problem: their errors interact as local credits are converted into a policy update. We characterize this evidence-to-update relationship and show how reliability depends on both evidence covariance and update geometry. We develop Stratified Segment Policy Optimization (S-SPO), which changes the joint design of auxiliary continuations while preserving the operational value target, local-credit rule, and continuation count per query. Under an explicit independence and reuse contract, equal-mass stratification guarantees non-increase of every fixed linear quadratic update risk; a first-order guarantee with a controlled remainder extends the analysis through normalization. At nine continuations per query, S-SPO improves full-parameter frozen-gradient reliability. Across three matched RhoMath-1.1B training seeds, it raises fixed-endpoint validation accuracy by 2.75 percentage points on average and official GSM8K test accuracy by 2.26 points, with every seed pair positive on both evaluations. A second-model comparison on Llama-3.2-1B-Instruct yields a 2.85-point validation gain. These results connect the design of finite continuation evidence to reliable local credit and improved learning.

1 INTRODUCTION

Local credit gives reasoning policies a training signal at a finer granularity than the final outcome of a complete response. Methods such as VinePPO and Segment Policy Optimization (SPO) estimate continuation values at intermediate prefixes and use their differences to evaluate progress along a generated solution (Kazemnejad et al., 2025; Guo et al., 2025). This approach avoids a learned critic while retaining feedback about intermediate decisions. Its practical resource is auxiliary evidence: each queried prefix requires additional continuations whose terminal rewards estimate the value used for credit assignment.

A finite continuation budget introduces uncertainty at precisely the point where these local signals are constructed. The relevant question is therefore not only how accurately a prefix value can be estimated, but how reliably its evidence is converted into the policy update. The same value error can enter several credit differences, be shared across repeated prefix occurrences, be replicated over different sets of tokens, and participate in normalization with other estimated credits. Measuring value error alone leaves this structured conversion unresolved.

We study *conditional evidence-to-update reliability*: the discrepancy caused by auxiliary-evidence randomness when the policy, primary trajectories, query identities, and downstream execution context are held fixed. In a chain, an intermediate boundary error enters its two neighboring local contrasts with opposite signs. Its raw-gradient contribution is therefore weighted by the difference between adjacent segment score vectors: an error may cancel or be amplified depending on the update geometry. This makes evidence covariance, rather than an isolated value variance, the relevant object for a downstream reliability guarantee.

Our intervention improves that covariance at its source. Stratified Segment Policy Optimization (S-SPO) draws one auxiliary continuation from each equal-mass stratum of an arithmetic decoding coordinate, then uses the same unweighted value estimate and downstream SPO machinery. Building on arithmetic sampling and stratification (Vilnis et al., 2023; Li et al., 2026a; Owen, 2013), the

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

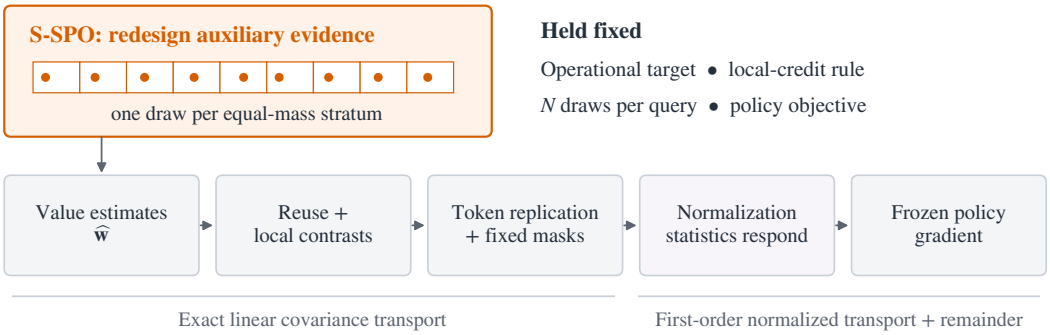


Figure 1: **Changing the evidence, retaining the credit pipeline.** S-SPO changes only the joint sampling of auxiliary continuations. The same value estimator, reuse, contrasts, masks, normalization, and policy objective carry evidence uncertainty into the update. Fixed linear stages admit exact covariance transport; normalization introduces a controlled nonlinear remainder.

method preserves the operational continuation target and each query's evidence budget. It requires neither a learned critic nor an estimate of which queries most strongly influence the gradient. Under independent primitive-query randomization and unchanged reuse, its variance reduction yields a simultaneous non-increase guarantee for all admissible fixed linear quadratic risks.

The theoretical and empirical questions remain distinct at nonlinear stages. We derive the Jacobian of regularized normalization and an explicit second-order remainder, establishing non-increase of the leading normalized-update risk. Full-parameter gradient measurements test the intervention through the executed nonlinear pipeline. Matched online experiments then assess whether the same evidence design improves learning: the RhoMath study spans three training seeds and the official GSM8K test, complemented by a second-model comparison on Llama.

Our contributions are threefold. **(1)** We characterize how reuse and adjacent segment-score differences convert value errors into update errors, explaining why value MSE alone cannot determine update reliability. **(2)** Target-preserving S-SPO guarantees simultaneous non-increase of all admissible fixed linear quadratic risks without estimating downstream influence; normalization admits a first-order guarantee with an explicit remainder. **(3)** With the arithmetic decoder, $N = 9$ budget, and downstream SPO settings matched, we measure improved full-parameter frozen-gradient reliability. Three matched RhoMath seed pairs yield mean gains of 2.75 validation and 2.26 official-test points; Llama adds a 2.85-point validation gain.

2 Related Work

Local credit and reliability. VinePPO uses Monte Carlo continuation estimates for reasoning-step credit (Kazemnejad et al., 2025); SPO develops segment-level estimation and probability-masked optimization (Guo et al., 2025). TEMPO aggregates descendant outcomes in prefix trees, while LEAF constructs sparse retrospective trees for speech-aware models (Tran et al., 2025; Gerogiannis et al., 2026). S-SPO retains SPO's credit rule and objective, and characterizes how finite auxiliary-evidence uncertainty propagates through reuse, contrasts, and normalization into policy updates.

Arithmetic, stratified, and QMC sampling. Arithmetic Sampling maps uniform codes to language-model sequences and studies expectation fidelity and variance (Vilnis et al., 2023). QuasiMoTTo (Li et al., 2026a) includes stratified arithmetic sampling, inference scaling, dependence-aware estimator analysis, and training with GRPO (Shao et al., 2024). Randomized QMC also reduces variance in policy-gradient and actor–critic estimation (Arnold et al., 2022). Building on these constructions and classical stratification (Owen, 2013), S-SPO independently stratifies continuations within each auxiliary query, conditional on fixed primary trajectories, and connects source covariance to local-credit update risk through reuse and normalization.

Estimator-aware interventions. PAIR optimizes rollout continuation probabilities and corrects pairwise inclusion for a scoped group-relative gradient target (Nomand et al., 2026). VIP predicts per-prompt success probabilities and allocates rollout counts across prompts to minimize a modeled gradient-variance objective under a fixed total budget (Nguyen et al., 2026). SKPO addresses unreliable early-token Monte Carlo advantages through an upstream/downstream formulation (Teng et al., 2026); ECPO calibrates agent credit using action-level shrinkage and reliability gating (Li et al., 2026b). S-SPO preserves the unweighted value estimator, operational target, credit rule, and sample count per auxiliary query; it changes the joint design of that query's continuations.

3 Setup and Evidence-to-Update Reliability

Operational continuation values. Let π_θ be the actor used in the policy loss and $\mu = \mu_{\theta_0,\lambda}$ the auxiliary continuation law at the generating checkpoint θ_0. The configuration λ includes temperature, top-p, EOS handling, generation limits, and other decoding transformations. A query state s contains the prefix and protocol state determining this law and the terminal reward R_s. For a continuation τ, define

$$V^\mu(s) = \mathbb{E}_{\tau\sim\mu(\cdot|s)} R_s(\tau). \tag{1}$$

For consecutive boundaries s_k, s_{k+1}, SPO-chain constructs $c_k = V^\mu(s_{k+1}) - V^\mu(s_k)$ (Guo et al., 2025). With undiscounted terminal payoff and suffix-consistent continuation semantics, this equals the advantage of forcing the realized segment and then resuming the operational process. Our comparison also covers boundary-specific operational contrasts when generation limits reset at each query. Completed terminal states use their deterministic observed payoff; Appendix A details these conventions.

The frozen object. Condition on a context $\mathcal{F}$ containing the actor checkpoint, primary trajectories, boundaries, primitive-query identities, reuse, token masks, normalization-pool membership, and score features. Auxiliary evidence supplies fresh randomness conditional on $\mathcal{F}$. Pool membership is fixed, while normalization statistics respond to the estimated credits. The analyzed update is the *full-parameter frozen policy-gradient component before optimizer transformations*. At the first PPO step (Schulman et al., 2017), detached advantages and ratios inside the clipping interval make the coefficient-to-gradient map linear before normalization; the executed loss reductions are included in that map.

Let $\mathbf{w} \in \mathbb{R}^q$ contain the q primitive query values, with error $\boldsymbol{\epsilon} = \widehat{\mathbf{w}} - \mathbf{w}$. The reuse matrix $R \in \{0,1\}^{b\times q}$ selects the primitive value for each of b boundary occurrences. With occurrence-level contrast operator D, set $B = DR$, $\mathbf{c} = B\mathbf{w}$, and $\widehat{\mathbf{c}} = B\widehat{\mathbf{w}}$. Known payoffs are zero-error coordinates; zero rows represent fixed zero-credit conventions. All expectations and covariances below condition on $\mathcal{F}$.

Proposition 1 (Structured transport)**.** *For fixed B, define $\Sigma = \mathrm{Cov}(\boldsymbol{\epsilon})$ and $\Sigma_c = \mathrm{Cov}(\widehat{\mathbf{c}} - \mathbf{c})$ under finite second moments. Then*

$$\widehat{\mathbf{c}} - \mathbf{c} = B\boldsymbol{\epsilon}, \qquad \Sigma_c = B\Sigma B^\top. \tag{2}$$

On a chain of m segments, let ϵ_k denote the occurrence-level error at boundary k and g_k the fixed score contribution of segment k, $0 \le k < m$. The raw policy-gradient error is

$$\delta G_{\mathrm{raw}} = -\epsilon_0 g_0 + \sum_{j=1}^{m-1} \epsilon_j (g_{j-1} - g_j) + \epsilon_m g_{m-1}. \tag{3}$$

Equation (3) weights each interior boundary error by adjacent segment-score differences. Cancellation or amplification depends on these scores and evidence reuse, so value MSE alone cannot characterize update reliability. The identities follow by linearity (Appendix B). Uncorrelated boundary errors induce adjacent credit covariance $-\mathrm{Var}(\epsilon_{k+1})$, while the underlying chain differences telescope to $\epsilon_m - \epsilon_0$.

For unbiased errors, any fixed linear map H and positive-semidefinite metric Q define the risk

$$\mathcal{R}_{H,Q} = \mathbb{E}\|H\boldsymbol{\epsilon}\|_Q^2 = \mathrm{tr}(QH\Sigma H^\top), \qquad \|v\|_Q^2 = v^\top Q v. \tag{4}$$

Here H can incorporate reuse, contrasts, token replication, masks, and the frozen score map. Reducing Σ in positive-semidefinite order guarantees non-increase across this family without estimating H during sampling. We use Euclidean vector norms and spectral matrix norms.

4 STRATIFIED EVIDENCE AND RELIABILITY GUARANTEES

4.1 S-SPO AT THE AUXILIARY-EVIDENCE LAYER

Let $\mathcal{T}_s$ be the continuation space for query s. An arithmetic autoregressive map $\Phi_s : [0,1) \to \mathcal{T}_s$ transforms one scalar input into a continuation (Vilnis et al., 2023). At each token, it selects an interval in the cumulative operational probabilities and rescales the residual coordinate within it. Using those probabilities, a fixed token ordering, and the operational stopping and reward semantics, $\Phi_s(U)$ follows $\mu(\cdot \mid s)$ for uniform U. Write $f_s = R_s \circ \Phi_s$.

Following arithmetic equal-mass stratification (Vilnis et al., 2023; Li et al., 2026a; Owen, 2013), S-SPO partitions the input interval into N cells and draws

$$U_i = \frac{i + \xi_i}{N}, \quad \xi_i \overset{\text{iid}}{\sim} \text{Unif}[0,1), \quad \widehat{V}_N^{\text{S}}(s) = \frac{1}{N} \sum_{i=0}^{N-1} f_s(U_i). \tag{5}$$

We independently permute the draw labels uniformly and use independent randomization across primitive queries. The permutation leaves the mean unchanged and makes each labeled draw marginally follow μ, while draws within a query remain jointly dependent. The same pipeline then reuses values, takes local differences, replicates and masks token credits, normalizes them, and applies the SPO objective. Stratification removes variation due to random stratum counts; within-stratum variation remains (Owen, 2013).

4.2 EXACT VALUE, COVARIANCE, AND FROZEN-RISK GUARANTEES

Theorem 1 (Target preservation and covariance-to-risk dominance)**.** *Condition on the same frozen context for arithmetic-iid and S-SPO. At primitive query j, use the same map Φ_j, reward integrand $f_j = R_j \circ \Phi_j$, budget N_j, and reuse. Assume finite second moments and independent primitive-query randomization within each design. For uniform U, let I_j be its equal-mass stratum index, $\sigma_j^2 = \text{Var}(f_j(U))$, and $b_{N_j,j}^2 = \text{Var}(\mathbb{E}[f_j(U) \mid I_j])$. Then*

$$\mathbb{E}\widehat{w}_j^{\text{S}} = w_j, \qquad \text{Var}(\widehat{w}_j^{\text{S}}) = \frac{\sigma_j^2 - b_{N_j,j}^2}{N_j} \le \frac{\sigma_j^2}{N_j} = \text{Var}(\widehat{w}_j^{\text{iid}}). \tag{6}$$

Putting $d_j = b_{N_j,j}^2 / N_j$ gives

$$\Sigma_{\text{iid}} - \Sigma_{\text{S}} = \text{diag}(d_j) \succeq 0, \qquad B\Sigma_{\text{S}}B^\top \preceq B\Sigma_{\text{iid}}B^\top. \tag{7}$$

For every fixed H and $Q \succeq 0$,

$$\mathcal{R}_{H,Q}^{\text{iid}} - \mathcal{R}_{H,Q}^{\text{S}} = \sum_j d_j \|Q^{1/2} H_{:j}\|_2^2 \ge 0. \tag{8}$$

The value-variance reduction is strict exactly when stratum reward means differ. The risk reduction is strict exactly when the displayed sum is positive.

Proof and interpretation. The estimator averages all stratum means equally; total variance gives Equation (6). Independent primitive queries give the diagonal covariance gap, and linear covariance transport gives Equations (7)–(8); Appendix B supplies full proofs. In the gain decomposition, d_j is the variance removed at query j, and $H_{:j}$ its downstream influence column after reuse; $Q^{1/2}$ is the positive-semidefinite square root. Each removed variance is weighted by squared influence in the chosen metric, without S-SPO estimating H or these weights. The guarantee holds simultaneously for every admissible fixed linear quadratic risk, including nonbinary rewards; equality remains possible.

4.3 FIRST-ORDER TRANSFER THROUGH NORMALIZATION

Let C implement token replication and fixed zeroing, M select the actual normalization-pool entries, and $K = MCB$. The score matrix S includes the loss reduction and has zero columns for any normalization-only positions. For a pool of $n \ge 2$ entries define

$$P = I - \mathbf{1}\mathbf{1}^\top / n, \quad \kappa = (n-1)^{-1}, \quad L = PK, \quad x = L\mathbf{w}, \quad d^2 = \kappa \|x\|_2^2 + \eta,$$

where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector, I is the identity matrix, $\eta > 0$, and $d > 0$. The regularized frozen-gradient map and its Jacobian with respect to $\mathbf{w}$ are

$$G(\mathbf{w}) = S\frac{L\mathbf{w}}{\sqrt{\kappa\|L\mathbf{w}\|_2^2 + \eta}}, \qquad J_G = S(d^{-1}P - \kappa d^{-3}xx^\top)K. \tag{9}$$

Theorem 2 (First-order normalized-risk guarantee)**.** *Under Theorem 1's contract, assume also* $m_{4,A} = \mathbb{E}\|L\boldsymbol{\epsilon}_A\|_2^4 < \infty$ *for* $A \in \{\mathrm{iid}, \mathrm{S}\}$. *Set* $q_A = \mathrm{tr}(J_G\Sigma_A J_G^\top)$ *and* $c = 5\|S\|_2\sqrt{\kappa}/(2d^2)$. *Then*

$$G(\mathbf{w} + \boldsymbol{\epsilon}_A) - G(\mathbf{w}) = J_G\boldsymbol{\epsilon}_A + r_A, \qquad \|r_A\|_2 \le c\|L\boldsymbol{\epsilon}_A\|_2^2, \tag{10}$$

$q_{\mathrm{S}} \le q_{\mathrm{iid}}$, *and the exact normalized-update MSE* $\mathcal{R}_{G,A} := \mathbb{E}\|G(\mathbf{w} + \boldsymbol{\epsilon}_A) - G(\mathbf{w})\|_2^2$ *obeys*

$$|\mathcal{R}_{G,A} - q_A| \le 2c\sqrt{q_A m_{4,A}} + c^2 m_{4,A} =: \mathcal{E}_A. \tag{11}$$

Consequently $\mathcal{R}_{G,\mathrm{S}} - \mathcal{R}_{G,\mathrm{iid}} \le -(q_{\mathrm{iid}} - q_{\mathrm{S}}) + \mathcal{E}_{\mathrm{S}} + \mathcal{E}_{\mathrm{iid}}$.

Appendix C proves the global remainder using augmented unit normalization. A gain $q_{\mathrm{iid}} - q_{\mathrm{S}} > \mathcal{E}_{\mathrm{S}} + \mathcal{E}_{\mathrm{iid}}$ is a sufficient finite-sample certificate; its verification is not assumed in the $N = 9$ experiments. Raw-gradient quadratic error measures the quantity aligned with the exact frozen-risk analysis, whereas normalized-gradient geometry tests transfer through the executed nonlinear pipeline. Online training separately tests whether the intervention improves learning.

5 Experiments

5.1 Experimental setup

We organize the experiments around three questions: whether finite continuation evidence materially affects the policy update, whether stratification improves reliability at the same evidence budget, and whether that improvement transfers to online learning.

Frozen-gradient measurements. Offline experiments use the RhoMath-1.1B SFT checkpoint and GSM8K with SPO-chain. We hold primary trajectories and the executed gradient context fixed, and compare estimated-value gradients against *finite high-budget reference estimates*: 128 independent continuations per query (Full128), with two disjoint 64-draw halves for reference sensitivity. Measurements use the full-parameter first-step policy-gradient component before optimizer transformations, with cutpoint interval 5, the strict probability mask $p < 0.9$, microbatches of eight trajectories, Bessel-corrected masked variance, and $\eta = 10^{-8}$.

To isolate stratification, the fixed-budget comparison holds the arithmetic decoder fixed: arithmetic-iid (M1) versus S-SPO (M2). Both use $N = 9$ continuations per query. We average 16 randomization repeats within each of 24 pools, then conduct paired inference across pools. Value MSE, local-credit MSE, and gradient direction are primary outcomes; raw-gradient quadratic error is supporting. A separate marginal control of the official ancestral-iid decoder (M0) against M1 is reported in Appendix A.

Online training and evaluation. For all online matched comparisons, the SPO control uses arithmetic-iid auxiliary sampling (M1), while S-SPO uses the same arithmetic decoder with equal-mass stratification (M2). The operational target, $N = 9$ budget, local-credit construction, and downstream SPO objective and training recipe are matched within each seed pair; the intervention is the joint design of auxiliary continuation evidence.

RhoMath-1.1B uses three matched SPO/S-SPO training-seed pairs (41, 42, 43). Llama-3.2-1B-Instruct supplies one matched second-model comparison. All runs use GSM8K and a fixed endpoint of 256 policy iterations. Training schedules match within each model setting: Rho executes 4096 optimizer updates and Llama 768. Appendix F gives the Rho protocol and available Llama execution records.

For J questions and K draws per question, accuracy is

$$\widehat{a}_K = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\mathbf{1}\{\text{draw } k \text{ answers question } j \text{ correctly}\}. \tag{12}$$

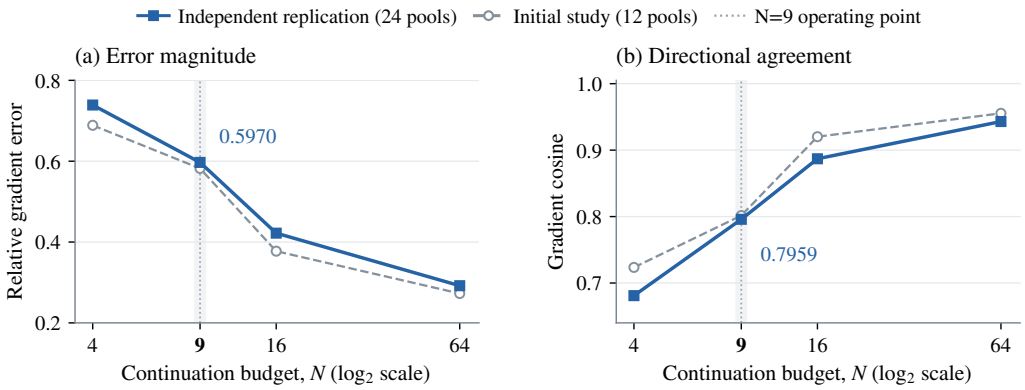


Figure 2: **Finite evidence changes the actual frozen policy gradient.** Increasing the continuation budget improves gradient reliability in two independent studies. The continuation-budget axis uses a base-2 logarithmic scale; the vertical marker identifies $N = 9$. The 24-pool estimates at $N = 4, 16, 64$ have three-decimal source precision; lines connect observed budgets without smoothing or inferred uncertainty bands.

This is a sampled estimate of mean single-draw correctness, not the probability of at least one correct answer among K draws. Dense validation curves use four draws on 373 questions every eight iterations; fixed endpoints use 16 draws. For Rho, these questions form an RL-held-out validation split from GSM8K training data; the initial checkpoint is `rho-1b-sft-GSM8K`. The official GSM8K test provides the separately held-out final evaluation: all six final Rho checkpoints use 1319 questions with 16 draws each. Test decoding uses temperature 0.35, top-$p = 0.9$, and a 1024-token generation limit, with identical prompt, parser, scoring, and evaluation-seed protocols. Across Rho training seeds, we report the mean and *sample standard deviation of paired gains*; these describe three training-seed pairs rather than confidence intervals.

5.2 Does finite evidence materially affect the policy update?

Finite evidence produces substantial gradient deviation through the executed SPO-chain pipeline. In an independent 24-pool problem study, increasing N from 4 to 16 to 64 reduces mean relative gradient error from approximately 0.739 to 0.422 to 0.292, and raises mean cosine from approximately 0.681 to 0.887 to 0.943 (Figure 2). The recovery between the smallest and largest budgets is consistent across all 24 pools. At the $N = 9$ operating point, mean relative error is 0.5970 and mean cosine is 0.7959: finite-evidence effects remain substantial at the budget used for online training.

An initial 12-pool study shows the same budget-dependent recovery and supplies a direct check of structured transport. Adjacent credit-error covariance is negative at all three budgets; divided by the shared-boundary error variance, its magnitude is 0.966, 1.029, and 0.986, as predicted by Equation (2). The association between coefficient error and gradient-direction error is strong (Spearman 0.8417). Its $N = 9$ relative error of 0.5821 and cosine of 0.8015 are also close to the independent study. Appendix D reports reference sensitivity and compact normalization, cross-fitting, and truncation checks. Together, these observations identify a measurable update-reliability problem at finite evidence budgets.

5.3 Does stratification improve reliability at the same budget?

At $N = 9$, S-SPO improves value-level and downstream gradient reliability (Table 1). Value MSE decreases by 9.7%, while mean normalized-gradient cosine increases from 0.8215 to 0.8466. Point estimates improve throughout the measured value-to-gradient pipeline.

The gradient-direction gain is 0.0251, with a paired 95% pool-bootstrap interval $[0.0091, 0.0431]$; 20 of 24 pools improve. The initial 12-pool intervention study gives a similar mean gain of 0.0254. These two intervention studies are distinct from the problem studies above. Cosine and its corresponding

Table 1: **Stratified evidence improves fixed-budget reliability.** At $N = 9$, S-SPO and arithmetic-iid sampling share the decoder and frozen context. Means average 16 repeats within each of 24 pools.

Metric	Arithmetic-iid	S-SPO	Effect
Value MSE ↓	0.0112	0.0101	−9.7%
Local-credit MSE ↓	0.0170	0.0158	−7.2%
Gradient cosine ↑	0.8215	0.8466	+0.0251
Raw-gradient quadratic error ↓	75.77	66.30	−12.5%

The first three outcomes are primary (cosine presents the equivalent $1 - \cos$ine outcome); raw-gradient quadratic error is supporting. The local-credit paired 95% CI includes zero. Complete paired intervals and reference sensitivity appear in Appendix E.

risk $1 - \cos$ describe the same directional outcome. All three primary point estimates favor S-SPO against Full128 and either half-reference.

Raw-gradient quadratic error decreases by 12.5%, with paired difference −9.47 and 95% interval $[-16.11, -2.66]$. This supporting measurement targets the frozen quadratic-risk object in Theorem 1. Normalized relative gradient error also decreases, from 0.5394 to 0.5013. The raw and normalized metrics use the same frozen auxiliary draws: together they show how an evidence intervention reaches the full-parameter gradient through the executed downstream operations.

5.4 Does the improvement transfer to online learning?

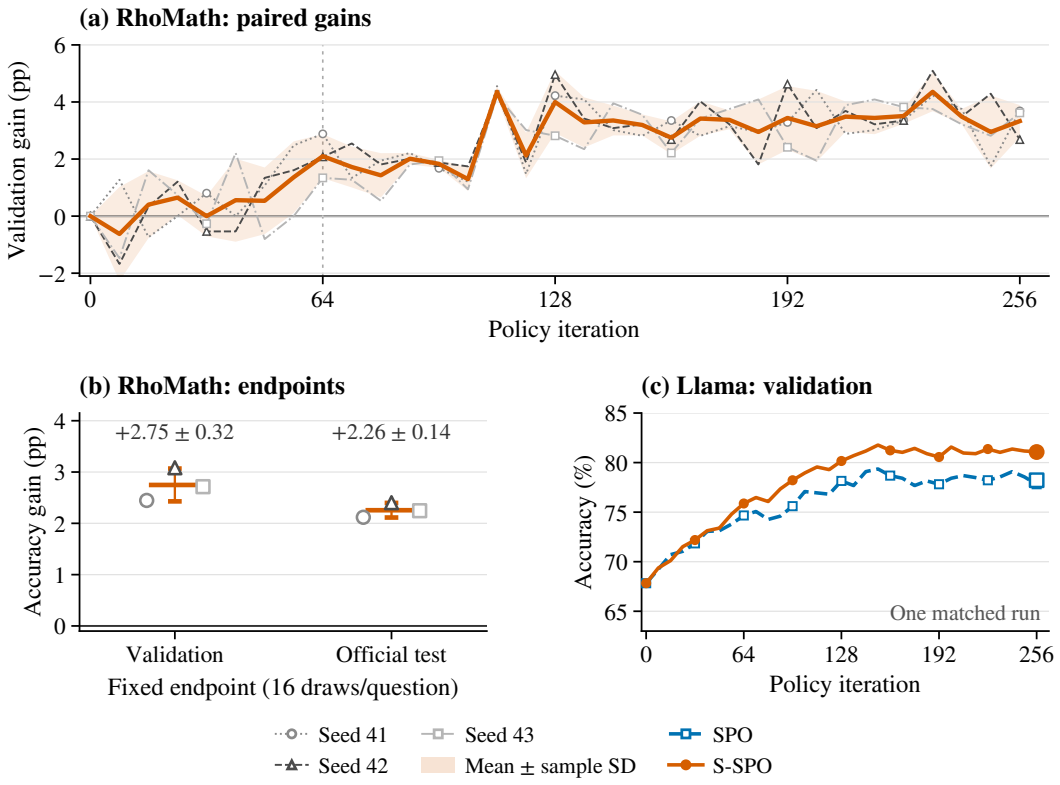


Figure 3: **Online gains on GSM8K.** (a) Rho paired four-draw validation gains: three seed trajectories and their mean ± sample SD. (b) Markers show the three 16-draw validation and official-test gains at iteration 256; horizontal mean marks and error bars show paired mean ± sample SD. (c) Llama's single matched run: four-draw curves with larger 16-draw final markers.

Table 2: **Fixed-endpoint accuracy after 256 policy iterations.** All values are percentages; gains are percentage points. Rho accuracy columns are means over three matched training seeds, and ± is the *sample SD of their paired gains* (not a CI or SEM). Llama has one matched run. Every endpoint uses 16 draws per question.

Model	Evaluation	Seeds	SPO	S-SPO	Paired gain
RhoMath-1.1B	Validation	3	65.59	68.34	$+\mathbf{2.75 \pm 0.32}$
	Official test	3	49.71	51.97	$+\mathbf{2.26 \pm 0.14}$
Llama-3.2-1B-Instruct	Validation	1	78.22	81.07	$+\mathbf{2.85}$

Consistency across RhoMath training seeds. S-SPO improves fixed-endpoint validation accuracy in all three Rho seed pairs (Table 2). Mean accuracy rises from 65.59% to 68.34%, with paired gains of 2.45, 3.08, and 2.71 percentage points. The mean paired gain is 2.75 ± 0.32 points, where the spread is the sample standard deviation across the three training seeds.

The improvement extends over the learning trajectory (Figure 3). From iteration 64 through 256, every recorded dense validation checkpoint favors S-SPO in every seed pair. The area under the curve (AUC), defined as the trapezoidal average of four-draw accuracy over iterations 0–256, also improves in all three pairs, by 2.36 percentage points on average. Thus the observed advantage is present across training and at the prespecified final horizon.

Official GSM8K test. The improvement persists on the official test split. SPO reaches 49.71% mean accuracy and S-SPO 51.97%, with positive paired gains of 2.12, 2.40, and 2.25 points. The mean gain is 2.26 ± 0.14 points across training seeds. This evaluation uses the final iteration-256 checkpoints and the common test protocol defined above. It therefore provides a held-out assessment of the same matched training intervention, separate from the validation trajectories.

Second-model corroboration. On Llama-3.2-1B-Instruct, S-SPO raises validation accuracy from 78.22% to 81.07%, a 2.85-point gain in one matched run. Its AUC improves by 1.95 points; every recorded checkpoint from iteration 24 onward favors S-SPO. This provides a second model setting under its own executed training schedule. Appendix G reports per-run endpoints, AUC, and active-training cost for both models. Equal N fixes the continuation count per query; complete-run runtime also depends on the evolving queries and generated lengths.

6 Discussion and Scope

The theory and measurements identify a useful design principle: evidence can be made more reliable before it enters an unchanged local-credit rule, without estimating gradient influence during sampling. Rho shows consistent gains across three training seeds and the official test, while Llama supplies a second-model comparison. Both online settings use GSM8K; cross-task generalization remains an empirical question. The exact guarantee concerns frozen linear update risks. Normalization is controlled to first order with an explicit remainder, and finite-budget gradient direction and multi-step learning are evaluated experimentally.

7 Conclusion

The reliability of finite continuation evidence depends on the update it supports. By characterizing its structured transport and improving its joint design, S-SPO connects target-preserving variance reduction to exact frozen-risk guarantees, better full-parameter gradient reliability, and higher online accuracy. Three matched RhoMath seeds show gains on both validation and the official GSM8K test, complemented by a second-model improvement. The results establish auxiliary-evidence design as a practical component of reliable sample-based local credit.

Reproducibility Statement

The operational assumptions and implementation contract are specified in Appendix A; complete proofs and normalization bounds appear in Appendices B and C. Appendices D–G describe the offline inference units, matched online comparison, recorded protocol details, evaluation rules, and per-run results, with Llama limited to available execution records. The accompanying anonymous supplementary archive provides the available analysis data and scripts for reproducing the reported tables and figures; Appendix H summarizes these materials.

AI Use Statement

Generative AI tools were used to assist with code implementation, grammar editing, manuscript structuring, and the preparation of figures and tables. These tools were also employed to aid in literature search and synthesis, and to support the review and refinement of mathematical derivations and experimental procedures. All AI-assisted outputs, including mathematical claims, experimental analyses, and manuscript content, were independently reviewed and validated by the authors. The authors made the final scientific and methodological decisions and take full responsibility for the integrity and final content of this work.

References

Sébastien M. R. Arnold, Pierre L'Ecuyer, Liyu Chen, Yi-Fan Chen, and Fei Sha. Policy learning and evaluation with randomized quasi-monte carlo. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 1041–1061. PMLR, 2022. URL https://proceedings.mlr.press/v151/arnold22a.html.

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.

Argyrios Gerogiannis, Yekaterina Yegorova, Mark Hasegawa-Johnson, and Venugopal V. Veeravalli. LEAF: Growing trees without branching for speech-aware large language model post-training. *arXiv preprint arXiv:2606.07610*, 2026. URL https://arxiv.org/abs/2606.07610v1.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in RL for large language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2505.23564v2.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590. PMLR, 2025. URL https://proceedings.mlr.press/v267/kazemnejad25a.html.

Michael Y. Li, Anthony Zhan, Kanishk Gandhi, Noah D. Goodman, and Emily B. Fox. QuasiMoTTo: Quasi-monte carlo test-time scaling. *arXiv preprint arXiv:2607.01179*, 2026a. URL https://arxiv.org/abs/2607.01179v1.

Yuanfan Li, Qi Zhou, Wenjing Duan, and Lu Chen. When denser credit is not enough: Evidence-calibrated policy optimization for long-horizon LLM agent training. *arXiv preprint arXiv:2606.05885*, 2026b. URL https://arxiv.org/abs/2606.05885v1.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*, 2024. URL https://arxiv.org/abs/2404.07965.

Meta. Llama 3.2 model card. Model documentation, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct.

Hieu Trung Nguyen, Bao Nguyen, Wenao Ma, Yuzhi Zhao, Ruifeng She, and Viet Anh Nguyen. Adaptive rollout allocation for online reinforcement learning with verifiable rewards. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.01601v3.

Pixel Nomand, Elena Voss, Marcus Hale, and Sofia Reyes. PAIR: Pairwise-aware inclusion reweighting for adaptive rollout allocation in RLVR. *arXiv preprint arXiv:2608.11368*, 2026. URL https://arxiv.org/abs/2608.11368v2.

Art B. Owen. *Monte Carlo Theory, Methods and Examples*. Online manuscript, 2013. URL https://artowen.su.domains/mc/.

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.

Fengwei Teng, Jinyi Bai, Xinhao Yao, Demi Ruohan Wang, Jiahao Zhao, and Zhijiang Guo. Skip-connected policy optimization for implicit advantage. *arXiv preprint arXiv:2604.08690*, 2026. URL https://arxiv.org/abs/2604.08690v1.

Hieu Tran, Zonghai Yao, and Hong Yu. Exploiting tree structure for credit assignment in RL training of LLMs. *arXiv preprint arXiv:2509.18314*, 2025. URL https://arxiv.org/abs/2509.18314v2.

Luke Vilnis, Yury Zemlyanskiy, Patrick Murray, Alexandre Tachard Passos, and Sumit Sanghai. Arithmetic sampling: Parallel diverse decoding for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35120–35136. PMLR, 2023. URL https://proceedings.mlr.press/v202/vilnis23a.html.

A Operational contract and implementation details

A.1 Operational continuation values and boundary semantics

The continuation law in the main paper includes the transformations applied at generation time. In particular, write

$$q_{\theta_0,\lambda}(a \mid h) = \mathcal{T}_\lambda\big(\pi_{\theta_0}(\cdot \mid h)\big)(a),$$

where $\mathcal{T}_\lambda$ includes temperature, support truncation, and renormalization. The state s contains the prompt, prefix, task and verifier identity, and any remaining generation budget or stopping state needed to determine the continuation experiment. Its payoff $R_s(\tau)$ includes the implemented answer parsing and finish-reason conventions. Two equal text prefixes with different remaining horizons need not define the same value. In the problem-validation records, the reward takes values in $\{-2, 0, 1\}$; -2 denotes a malformed response with multiple #### delimiters, rather than a truncation penalty. The general results require a square-integrable payoff, without a Bernoulli assumption.

For a realized segment z_k from s_k to s_{k+1}, define $Q^\mu(s_k, z_k)$ by forcing that segment and then resuming the continuation process. Under deterministic prefix transitions, an undiscounted terminal payoff, and *suffix consistency*,

$$Q^\mu(s_k, z_k) - V^\mu(s_k) = V^\mu(s_{k+1}) - V^\mu(s_k). \tag{13}$$

Suffix consistency means that the experiment after forcing z_k is exactly the experiment defining $V^\mu(s_{k+1})$. Restarting a fresh generation cap at each boundary need not satisfy this condition; the comparison still holds for the resulting boundary-specific operational value contrasts. A separate actor KL regularizer is not part of the terminal outcome return.

Under the terminal-payoff convention $V^\mu(s_T) = R(s_T)$, a deterministic terminal payoff has zero estimation error after the primary trajectory is fixed. A stochastic verifier's uncertainty would instead remain in the primitive error model. Fresh auxiliary sampling after segmentation preserves conditional unbiasedness; an on-policy macro-action gradient interpretation additionally requires alignment of segmentation and score laws.

A.2 Primitive queries, sharing, and frozen operators

The common context $\mathcal{F}$ fixes the actor, primary trajectories, boundaries, operational queries, sharing, masks, pool membership, and score map. Auxiliary draws are fresh conditional on $\mathcal{F}$ and do not select the compared operators or batch. Primary rollouts already conditioned on are not counted as fresh auxiliary evidence.

Each coordinate of $\mathbf{w}$ is one independently estimated query group, determined by actual cache/sharing scope. The same estimate may occur several times through R. Distinct groups use independent jitters, permutations, and any verifier randomness. The comparison preserves R and can be averaged over a shared distribution of contexts; evolving training runs generally have different context distributions.

A.3 Arithmetic sampling and the decoder control

M0 denotes official ancestral-iid auxiliary sampling. M1 applies an arithmetic decoder Φ_s to iid uniform inputs. M2 applies the same decoder to one independent uniform input per equal-mass cell, followed by an independent uniform permutation. M2 versus M1 isolates stratification in both the offline and online matched comparisons. M0 versus M1 is a marginal decoder-replacement control connecting the arithmetic-iid SPO control to the canonical ancestral-iid implementation. A query's fixed map includes a measurable ordering at every prefix, which can depend on the prefix and logits. Ordering affects the stratum reward means and gain while preserving the ideal marginal law.

Arithmetic pushforward. Let q_a be the operational token probabilities in the chosen order, with cumulative masses F_a. Select the positive-mass token whose half-open interval contains the current coordinate u, and update

$$u \longleftarrow \frac{u - F_{a^-}}{q_a}, \qquad u \in [F_{a^-}, F_a). \tag{14}$$

For a uniform input, the first token has probability q_a, and the residual conditional on selecting it is uniform. Induction gives each prefix cylinder probability $\prod_t q(a_t \mid s, a_{<t})$. With a finite

Table 3: Marginal decoder control: arithmetic-iid minus ancestral-iid. Each paired 95% interval is contained in $[-\text{margin}, \text{margin}]$.

Statistic	Difference	Paired 95% interval	Margin
Absolute reference deviation	-0.004987	$[-0.012799, 0.002493]$	0.08
EOS rate	0.000332	$[-0.002992, 0.003989]$	0.04
First-token TV diagnostic	0.001401	$[-0.011450, 0.014105]$	0.05
Reward-zero frequency	0.002327	$[-0.008976, 0.013630]$	0.05
Reward-one frequency	-0.002660	$[-0.013298, 0.008311]$	0.05
Reward-minus-two frequency	0.000332	$[-0.002327, 0.003657]$	0.05
Reward mean	-0.003324	$[-0.015293, 0.008976]$	0.08
Standardized per-token logprob	-0.008241	$[-0.045535, 0.028294]$	0.20
Standardized mean length	-0.030669	$[-0.062667, 0.000352]$	0.20
Truncation rate	-0.000332	$[-0.003989, 0.002992]$	0.02

cap or almost-sure termination, terminal cylinders partition a set of measure one, establishing the pushforward to $\mu(\cdot \mid s)$. Null endpoints do not affect the result; positive-probability nontermination must be included in the output/return definition if no cap is used. The arithmetic construction (Vilnis et al., 2023) retains this residual across tokens.

This is an exact-distribution statement. A fixed finite-precision map $\widetilde{\Phi}_s$ induces a law $\widetilde{\mu}$; stratification reduces variance relative to iid inputs through that same map when cell input laws are correct. For reward range width W, total-variation discrepancy at most δ gives $|V^{\widetilde{\mu}} - V^{\mu}| \leq W\delta$. Equivalence to a separate production decoder requires an implementation-level comparison.

Marginal decoder control. The frozen M0/M1 control compares Transformers ancestral sampling with arithmetic-iid sampling at temperature 0.6, top-$p = 0.9$, and a 512-token generation cap. The arithmetic implementation uses ascending token-ID order, float64 coordinates/CDFs, and iid initial uniforms. Boundary-level sufficient statistics cover 94 nonterminal boundaries in pools 0, 1, 2, and 5, with $N = 32$, accounting for 3,008 continuations per method. The KS distance between the two sets of 94 boundary reward means is 0.053191, below the frozen margin 0.2; all ten paired 95% intervals lie inside their frozen equivalence margins (Table 3). Six synthetic-test records report passes, and the final diagnostic records zero first-token support violations. The final analysis corrects the exact-batch first-token reference distribution; this changes the reference-support diagnostic, not the recorded reward/value statistics. Individual continuation records are unavailable, limiting independent record-level completeness and support checks. The evidence supports marginal agreement within the stated criteria, rather than gradient or training-trajectory equivalence.

Intervals use 10,000 paired boundary-cluster bootstrap samples, with criterion-specific deterministic seeds derived from 20260904 and $N = 32$, and percentile endpoints. Length and logprob use pooled standard deviations retained in the original result, 65.320825 and 0.058893, respectively. Length excludes EOS; per-token logprob is each sequence's operational logprob divided by its emitted-token count, including EOS, before boundary averaging. Reference deviation uses the fixed 128-draw boundary reference; the first-token diagnostic uses the corrected exact-batch operational reference. The synthetic records cover two- and three-category distributions, small probabilities, top-p filtering, 32-step recursive coordinates, and half-open interval ties.

A.4 Executed replication, masks, and normalization pools

SPO-chain uses successive value differences and a probability mask (Guo et al., 2025). The offline implementation obtains old token probabilities from teacher-forced full-vocabulary logits at temperature 0.6. Tokens with probability below 0.9 define the cutpoint sequence; an interval of five is a stride through this sequence, with boundaries immediately before the selected tokens. Credits $V_{\text{next}} - V_{\text{previous}}$ are broadcast backward to segment tokens. Generation-stage zeroing uses $p > 0.9$; the statistics/loss mask uses strict $p < 0.9$.

The $N = 9$ frozen-gradient experiment uses eight-trajectory microbatches, Bessel-corrected masked variance, and $\eta = 10^{-8}$. Whitening does not cross gradient-accumulation microbatches in the verified single-device implementation. More generally, the executed statistics call determines the

pool, including distributed aggregation and any empty-mask fallback. The operator C records replication and fixed zeroing, M selects statistics positions, and $K = MCB$; normalization-only positions have zero columns in S. Bessel normalization requires $n \geq 2$, independently of the positive stabilizer.

A.5 Independent references and uncertainty in a stratified batch

For any estimator $\widehat{c}$ and an independent unbiased reference $\widehat{c}_{\mathrm{ref}}$, finite second moments give

$$\mathbb{E}\|\widehat{c} - \widehat{c}_{\mathrm{ref}}\|_Q^2 = \mathbb{E}\|\widehat{c} - c\|_Q^2 + \mathrm{tr}(Q\Sigma_{\mathrm{ref}}), \tag{15}$$

$$\mathbb{E}[(\widehat{c} - \widehat{c}_{\mathrm{ref},1})^\top Q(\widehat{c} - \widehat{c}_{\mathrm{ref},2})] = \mathbb{E}\|\widehat{c} - c\|_Q^2, \tag{16}$$

where the second identity requires two unbiased references independent of each other and of $\widehat{c}$. Both follow by expanding the products: every cross term containing an independent centered reference error vanishes. Shared evidence need not satisfy these identities. After nonlinear normalization, an unbiased value reference need not produce an unbiased update reference, so this subtraction does not directly calibrate normalized-gradient error.

The within-batch scatter of stratified rewards is also different from the variance of their mean. With $s_Y^2 = (N-1)^{-1}\sum_i (Y_i - \bar{Y})^2$, $N \geq 2$, and between-cell variance b_N^2 as in Theorem 1,

$$\mathbb{E}s_Y^2 = \sigma^2 + \frac{b_N^2}{N-1}, \qquad \mathbb{E}(s_Y^2/N) - \mathrm{Var}(\bar{Y}) = \frac{b_N^2}{N-1}. \tag{17}$$

Use $\sum_i (Y_i - \bar{Y})^2 = \sum_i Y_i^2 - N\bar{Y}^2$ and the variance identity of Theorem 1 to obtain both formulas. Independent complete stratified randomizations, as used for the repeated offline measurements, provide independent sample means; the individual outputs of one permuted batch are not iid repetitions of that estimator.

B Proofs of the main evidence-to-update results

B.1 Structured transport and boundary influence

All quantities below are conditional on $\mathcal{F}$; expectations and covariance omit that conditioning for readability. The primitive error is $\epsilon = \widehat{\mathbf{w}} - \mathbf{w}$, the occurrence-level values are $R\mathbf{w}$, and $B = DR$ forms the fixed local contrasts. For a positive-semidefinite matrix Q, write $\|v\|_Q^2 = v^\top Q v$. Matrix norms without a subscript are operator norms.

Proof of Proposition 1. Linearity gives $\widehat{\mathbf{c}} - \mathbf{c} = B\epsilon$. Subtracting its expectation in the covariance definition gives $\mathrm{Cov}(B\epsilon) = B\Sigma B^\top$, without requiring independent primitive errors. On a chain, using occurrence-level boundary errors as in Proposition 1, $e_k = \epsilon_{k+1} - \epsilon_k$ and

$$\begin{aligned}\mathrm{Cov}(e_k, e_{k+1}) = \mathrm{Cov}(\epsilon_{k+1}, \epsilon_{k+2}) - \mathrm{Var}(\epsilon_{k+1}) \\ - \mathrm{Cov}(\epsilon_k, \epsilon_{k+2}) + \mathrm{Cov}(\epsilon_k, \epsilon_{k+1}).\end{aligned}$$

Uncorrelated boundary errors therefore yield the stated negative adjacent covariance. To derive the raw-gradient identity, expand

$$\delta G_{\mathrm{raw}} = \sum_{k=0}^{m-1} (\epsilon_{k+1} - \epsilon_k) g_k.$$

The coefficients of the first and final boundary errors are $-g_0$ and g_{m-1}; the coefficient of an interior error ϵ_j is $g_{j-1} - g_j$. This proves Equation (3). For a fixed H and zero-mean error,

$$\mathbb{E}\|H\epsilon\|_Q^2 = \mathrm{tr}\big(QH\mathbb{E}[\epsilon\epsilon^\top]H^\top\big) = \mathrm{tr}(QH\Sigma H^\top).$$

With bias $b = \mathbb{E}\epsilon$, the right side additionally contains $\|Hb\|_Q^2$. □

For the underlying chain differences, $\sum_k e_k = \epsilon_m - \epsilon_0$, so exact endpoints give zero total contrast error even when individual credits are noisy. On a fixed graph, $D^\top v = 0$ implies $v^\top DR\epsilon = 0$ pathwise. The score-weighted operator determines the update effect of this incidence structure. With replication and extraction the raw map is $H = SMCB$; retaining centering while removing scaling gives $H = SPMCB$.

B.2 TARGET PRESERVATION AND THE VALUE-VARIANCE IDENTITY

Proof of the value statement in Theorem 1. Fix a query and let $Y = f(U)$, where U is uniform on $[0, 1)$. For the equal-mass cell index $I = \lfloor NU \rfloor$, define $m_i = \mathbb{E}[Y \mid I = i]$ and $s_i^2 = \mathrm{Var}(Y \mid I = i)$, $i = 0, \ldots, N-1$. The pushforward in Appendix A gives $\mathbb{E}Y = V^\mu(s)$. Integrability and equal cell masses imply

$$\mathbb{E}\widehat{V}_N^{\mathrm{S}} = \frac{1}{N}\sum_i m_i = \mathbb{E}Y.$$

Under finite second moments, the independent pre-permutation draws satisfy

$$\mathrm{Var}(\widehat{V}_N^{\mathrm{S}}) = \frac{1}{N^2}\sum_i s_i^2.$$

The decomposition $Y - \mathbb{E}Y = (Y - m_I) + (m_I - \mathbb{E}Y)$ has zero cross term because $\mathbb{E}[Y - m_I \mid I] = 0$. Thus

$$\sigma^2 = \frac{1}{N}\sum_i s_i^2 + b_N^2, \qquad b_N^2 = \frac{1}{N}\sum_i (m_i - \mathbb{E}Y)^2,$$

which proves $\mathrm{Var}(\widehat{V}_N^{\mathrm{S}}) = (\sigma^2 - b_N^2)/N \le \sigma^2/N$. The iid sample mean has variance σ^2/N. Equality holds exactly when all m_i agree; strict gain holds exactly when $b_N^2 > 0$. Zero stratified variance is the separate condition $s_i^2 = 0$ for every cell. □

For the cell-conditional projection $\Pi_N f = \mathbb{E}[f(U) \mid I]$, the same decomposition has the compact form

$$N\,\mathrm{Var}(\widehat{V}_N^{\mathrm{S}}) = \|f - \Pi_N f\|_{L^2}^2, \qquad b_N^2 = \|\Pi_N f - \mathbb{E}f\|_{L^2}^2. \tag{18}$$

Stratification removes between-cell variation while retaining within-cell variation.

An independent permutation leaves the sample mean unchanged pathwise and makes any labeled input a uniform mixture of the cell laws, hence uniform on $[0, 1)$. For distinct output labels r, t and $N \ge 2$,

$$\mathbb{E}[Y_r Y_t] = \frac{\sum_{i \ne j} m_i m_j}{N(N-1)} = V^2 - \frac{b_N^2}{N-1}.$$

Consequently the reward covariance is $-b_N^2/(N-1)$. Correct marginal outputs after permutation are dependent. Before permutation the draws are independent with different cell-conditional laws.

B.3 COVARIANCE DOMINANCE AND THE UPDATE-WEIGHTED GAIN

Proof of the covariance and risk statements in Theorem 1. Conditional independence across primitive queries makes both primitive covariance matrices diagonal. The preceding value proof applied at budget N_j gives

$$\Sigma_{\mathrm{iid}} - \Sigma_{\mathrm{S}} = \mathrm{diag}(d_1, \ldots, d_q), \qquad d_j = \frac{b_{N_j,j}^2}{N_j} \ge 0.$$

For any vector v,

$$v^\top B(\Sigma_{\mathrm{iid}} - \Sigma_{\mathrm{S}})B^\top v = \|\,\mathrm{diag}(\sqrt{d_j})B^\top v\|^2 \ge 0,$$

proving contrast covariance dominance. With the same fixed H, Q in the two designs,

$$\begin{aligned}\mathcal{R}_{H,Q}^{\mathrm{iid}} - \mathcal{R}_{H,Q}^{\mathrm{S}} &= \mathrm{tr}\big(QH\,\mathrm{diag}(d_j)H^\top\big) \\ &= \sum_j d_j \|Q^{1/2}H_{:j}\|^2.\end{aligned}$$

All terms are nonnegative, proving Equation (8). Equality holds when every reduced-variance query has zero influence in the chosen metric; strict gain holds exactly when the sum is positive. □

Reuse is already included in each column of H. If query j contributes occurrence-specific vectors $h_{j,r}$, then $H_{:j} = \sum_r h_{j,r}$ and its primitive risk contribution is $\mathrm{Var}(\epsilon_j)\|\sum_r h_{j,r}\|_Q^2$. Independent estimates per occurrence would define a different sharing and sampling design. Keeping reuse fixed permits both constructive and canceling influence without altering the proof. S-SPO therefore requires

no estimate of these influence vectors to obtain the simultaneous non-increase for every admissible fixed linear quadratic risk.

Independence is a sufficient condition for the PSD covariance gap. More generally, a verified PSD gap $\Delta\Sigma$ suffices, with strict gain exactly when $Q^{1/2}H(\Delta\Sigma)^{1/2} \neq 0$. Coordinate-wise variance reductions alone do not: I_2 and $\left(\begin{smallmatrix} .9 & .9 \\ .9 & .9 \end{smallmatrix}\right)$ have quadratic variances 2 and 3.6 in direction $(1,1)$ despite the latter's smaller diagonal. Likewise, unequal estimator biases require comparison of $\Sigma + bb^\top$, rather than covariance alone. Under a common distribution of frozen contexts, conditional risk ordering can be averaged directly; it is not a comparison of the different context distributions produced by online optimization.

C Normalization, direction, and nonlinear scope

C.1 The executed map and its differential

Use the main-text notation P, $\kappa = (n-1)^{-1}$, $L = PK$, $x = L\mathbf{w}$, and $d^2 = \kappa\|x\|^2 + \eta$, with $n \geq 2$ and $\eta > 0$. At $a = K\mathbf{w}$, the coefficient map and frozen policy-gradient component are

$$\mathcal{W}_\eta(a) = \frac{Pa}{\sqrt{\kappa\|Pa\|^2 + \eta}}, \qquad G(\mathbf{w}) = S\mathcal{W}_\eta(K\mathbf{w}).$$

Writing $F(x) = x/(\kappa\|x\|^2 + \eta)^{1/2}$ and differentiating d gives

$$DF(x) = d^{-1}I - \kappa d^{-3}xx^\top, \quad J_{\mathcal{W}}(a) = d^{-1}P - \kappa d^{-3}xx^\top, \quad J_G = SJ_{\mathcal{W}}(a)K. \tag{19}$$

Here $Px = x$ and $P^2 = P$. Under $J_{\mathcal{W}}$, the constant direction has eigenvalue zero; for $x \neq 0$, centered tangential and radial directions have eigenvalues $1/d$ and η/d^3, respectively. At $x = 0$, every centered direction has eigenvalue $\eta^{-1/2}$. Globally, $\|J_{\mathcal{W}}\| \leq \eta^{-1/2}$.

C.2 A global quadratic remainder and proof of Theorem 2

The oracle-scale constant in Equation (10) follows from an elementary normalization identity that is valid for arbitrarily large perturbations.

Lemma 1 (Unit-normalization remainder)**.** *For $u \neq 0$, $u + h \neq 0$, and $\mathcal{N}(u) = u/\|u\|$,*

$$\|\mathcal{N}(u+h) - \mathcal{N}(u) - D\mathcal{N}(u)h\| \leq \frac{5\|h\|^2}{2\|u\|^2}.$$

Proof. Set $v = u/\|u\|$, $e = h/\|u\|$, $t = \|e\|$, $r = \|v + e\| > 0$, $a = v^\top e$, and $v' = (v+e)/r$. The reverse triangle inequality yields $|r-1| \leq t$, and $\|v' - v\| \leq \|v' - (v+e)\| + \|e\| = |1-r| + t \leq 2t$. Since $r^2 = 1 + 2a + t^2$ and the supporting-hyperplane inequality gives $r \geq 1 + a$,

$$0 \leq r - 1 - a = \frac{t^2 - (r-1)^2}{2} \leq \frac{t^2}{2}.$$

Using $e = rv' - v$ gives the exact decomposition

$$v' - v - (I - vv^\top)e = -(r - 1 - a)v + (1 - r)(v' - v).$$

The two terms have norms at most $t^2/2$ and $2t^2$, respectively. Since $D\mathcal{N}(u)h = (I - vv^\top)e$, their sum proves the claim. □

Proof of Theorem 2. The derivative is Equation (19). For $z = L\epsilon$, apply Lemma 1 to

$$u = (\sqrt{\kappa}\,x, \sqrt{\eta}), \qquad h = (\sqrt{\kappa}\,z, 0), \qquad \|u\| = d.$$

Both augmented vectors are nonzero because $\eta > 0$. The first coordinates of $\mathcal{N}(u)$, divided by $\sqrt{\kappa}$, equal $F(x)$. Projection is a contraction, so

$$\|F(x+z) - F(x) - DF(x)z\| \leq \frac{5\sqrt{\kappa}}{2d^2}\|z\|^2.$$

Multiplication by S gives $\|r\| \le c\|z\|^2$, where $c = 5\|S\|\sqrt{\kappa}/(2d^2)$. This bound is global; it does not assume that z remains in a neighborhood of zero.

For either design $A \in \{\mathrm{iid}, \mathrm{S}\}$, define

$$q_A = \mathrm{tr}(J_G \Sigma_A J_G^\top), \quad m_{4,A} = \mathbb{E}\|L\epsilon_A\|^4, \quad \mathcal{R}_{G,A} = \mathbb{E}\|G(\mathbf{w}+\epsilon_A) - G(\mathbf{w})\|^2.$$

Mean-zero primitive errors give $\mathbb{E}\|J_G\epsilon_A\|^2 = q_A$. If $m_{4,A} < \infty$, then $\mathbb{E}\|r_A\|^2 \le c^2 m_{4,A}$. Expanding the squared norm and applying Cauchy–Schwarz yields

$$|\mathcal{R}_{G,A} - q_A| \le 2c\sqrt{q_A m_{4,A}} + c^2 m_{4,A} =: \mathcal{E}_A. \tag{20}$$

Theorem 1 with $H = J_G$, $Q = I$ gives

$$q_{\mathrm{iid}} - q_{\mathrm{S}} = \sum_j d_j \|(J_G)_{:j}\|^2 \ge 0.$$

Subtracting the two error bounds proves

$$\mathcal{R}_{G,\mathrm{S}} - \mathcal{R}_{G,\mathrm{iid}} \le -(q_{\mathrm{iid}} - q_{\mathrm{S}}) + \mathcal{E}_{\mathrm{S}} + \mathcal{E}_{\mathrm{iid}}. \tag{21}$$

Hence $q_{\mathrm{iid}} - q_{\mathrm{S}} > \mathcal{E}_{\mathrm{S}} + \mathcal{E}_{\mathrm{iid}}$ is sufficient for strict finite-sample actual-update MSE improvement. The first-order ordering holds without this extra inequality. □

The proof also bounds normalized-update bias by $\|\mathbb{E}G(\mathbf{w}+\epsilon) - G(\mathbf{w})\| \le c\mathbb{E}\|L\epsilon\|^2$, despite unbiased primitive values. The $N = 9$ experiment measures actual nonlinear gradients; it does not assume that the sufficient moment-based certificate holds.

Hessian and local form. Direct differentiation also gives

$$\begin{aligned} D^2F(x)[u,v] = &-\frac{\kappa}{d^3}\left[(x^\top v)u + (x^\top u)v + (u^\top v)x\right] \\ &+ \frac{3\kappa^2}{d^5}(x^\top u)(x^\top v)x. \end{aligned} \tag{22}$$

Using $\sqrt{\kappa}\|x\| \le d$ bounds its norm by $6\sqrt{\kappa}\|u\|\|v\|/d^2$. Taylor's integral remainder is $S\int_0^1 (1-t)D^2F(x+tz)[z,z]\,dt$. On $\sqrt{\kappa}\|z\| \le \alpha d$, $0 < \alpha < 1$, the denominator along this segment is at least $(1-\alpha)d$, yielding

$$\|r\| \le \frac{3\|S\|\sqrt{\kappa}}{(1-\alpha)^2 d^2}\|z\|^2.$$

This local bound requires its stated event or complementary-event control in expectations; the global augmented-vector bound used in Theorem 2 does not.

C.3 Full-pool noise and consistency

With $z = L\epsilon$, the estimated squared denominator obeys

$$\widehat{d^2} = d^2 + 2\kappa x^\top z + \kappa\|z\|^2, \qquad \mathbb{E}\widehat{d^2} = d^2(1+\beta), \qquad \beta = \frac{\kappa\,\mathrm{tr}(L\Sigma L^\top)}{d^2}.$$

Covariance dominance weakly decreases this full-pool noise-to-signal quantity. It describes the expected squared denominator. For $\alpha > 0$, Markov's inequality gives $\Pr(\sqrt{\kappa}\|z\| > \alpha d) \le \beta/\alpha^2$; small β alone does not control the fourth moment needed in Equation (20).

The global Jacobian bound yields $\|F(x+z) - F(x)\| \le \eta^{-1/2}\|z\|$, preserving mean-square consistency for fixed $\eta > 0$ when $\mathbb{E}\|L\epsilon\|^2 \to 0$. Weak-signal pools can therefore have substantial finite-budget sensitivity while remaining consistent.

C.4 Direction and aggregation across pools

For one pool, $G(\mathbf{w}+\epsilon) = S(x+z)/\widehat{d}$ has a positive scalar denominator. Whenever both numerators are nonzero,

$$\cos(G(\mathbf{w}+\epsilon), G(\mathbf{w})) = \cos(S(x+z), Sx). \tag{23}$$

Set $y = Sx \neq 0$, $u = y/\|y\|$, and $P_u^{\perp} = I - uu^{\top}$. The direction map $\mathcal{D}(\mathbf{w}) = G(\mathbf{w})/\|G(\mathbf{w})\|$ has Jacobian

$$J_{\mathcal{D}} = \frac{P_u^{\perp} SL}{\|Sx\|}.$$

Its leading cosine loss is $\frac{1}{2}\operatorname{tr}(J_{\mathcal{D}}\Sigma J_{\mathcal{D}}^{\top})$, since the squared distance between unit vectors is twice their cosine loss. Covariance dominance orders this term. Lemma 1 bounds the directional remainder by $5\|SL\epsilon\|^2/(2\|Sx\|^2)$ provided the perturbed numerator is also nonzero. A finite fourth moment then gives an error allowance by the same expansion as Equation (20), divided by two. This retains the nonzero-direction condition.

For several actual normalization pools, with an evidence-independent regularizer,

$$G(\mathbf{w}) = \sum_{\ell} S_{\ell}\mathcal{W}_{\eta_{\ell}}(K_{\ell}\mathbf{w}) + G_{\mathrm{reg}}, \qquad J_G = \sum_{\ell} S_{\ell} J_{\mathcal{W}_{\ell}} K_{\ell}.$$

Theorem 1 applies to this combined Jacobian, retaining cross-pool terms from reused queries. Distinct random denominators change relative pool weights; an unscaled regularizer can also affect the total gradient's direction.

At the first PPO step, ratios equal one, inside the clipping interval; detached advantages and fixed masks give the stated linear coefficient-to-gradient map. At a later fixed checkpoint, this map is locally linear only while active clipping branches stay fixed. A checkpoint optimized using the same evidence is not independent of it merely because its numerical weights are subsequently frozen. A fixed preconditioner T is covered by replacing H with TH. Adam's current second-moment estimate generally depends on the noisy gradient and requires a separate optimizer map and remainder analysis. These results concern the frozen policy-gradient component before such optimizer transformations.

C.5 A counterexample within the stratified sampler class

Take $N = 2$ and $f(u) = \mathbf{1}_{[1/2,5/8)}(u)$. Its value is $V = 1/8$, with cell success probabilities $(0, 1/4)$. Hence

$$\widehat{V}^{\mathrm{S}} \sim \tfrac{1}{2}\,\mathrm{Bernoulli}(1/4), \qquad \widehat{V}^{\mathrm{iid}} \sim \tfrac{1}{2}\,\mathrm{Binomial}(2, 1/8).$$

The variances are $3/64 < 7/128$, but the probabilities of $|\widehat{V} - 1/8| > 1/4$ are $1/4 > 15/64$. Thus variance ordering does not imply ordering of every tail probability.

Now use $a(v) = (v - 1/4,\, 1/4 - v)$, $S = I_2$, and $\kappa = 1$. This is a fixed linear contrast after adjoining the known value $1/4$ as a zero-error primitive coordinate. The oracle numerator at $v = 1/8$ is nonzero. As $\eta \downarrow 0$, the normalized vectors have unit norm, with opposite oracle direction exactly when $\widehat{V} > 1/4$. Their limiting squared errors are four times these probabilities: 1 for stratification and $15/16$ for iid. Every possible numerator is nonzero, so continuity preserves the strict ordering for all sufficiently small positive η. The same estimator has strictly lower value variance and larger normalized-update MSE. This establishes the need for the first-order qualifier or an explicit finite-sample certificate; it is not an empirical model of the observed SPO pools.

D Finite-Evidence Problem Validation

D.1 Frozen-gradient protocol and independent replication

We evaluate the RhoMath-1.1B GSM8K SFT checkpoint with SPO-chain, fixing the actor, primary trajectories, segment boundaries, reuse identities, masks, and whitening pools while varying auxiliary evidence. An initial study uses 12 official pools; an independent replication uses 24 pools with new trajectories and stochastic evidence. Each pool contains eight trajectories. The cutpoint interval is five, the probability threshold is 0.9, and whitening uses the masked mean, unbiased variance, and

regularization 10^{-8}. Full-parameter gradients are measured at the first PPO step, with ratio one and clipping threshold 0.2, before optimizer transformations.

The initial study (sampling seed 20260830) contains 304 complete banks of 64 low-budget and 128 independent reference continuations, giving 58,368 records, 303 local-credit pairs, 203 adjacent pairs, and 101 complete chains. One incomplete bank is excluded. Its 12 pools contain 1,230 effective loss-mask tokens. The probe reproduces the official log probabilities, masks, and first-step clipping derivative; the maximum whitening residual is 2.20×10^{-7}, and the duplicate-reference gradient discrepancy is zero. All approximately 1.100 billion parameter gradients are finite. Reference sensitivity uses the two disjoint 64-continuation halves of the Full128 bank.

D.2 Budget-dependent recovery and the operating point

Let G_N and G_{ref} denote the estimated and reference full-parameter gradients in a pool. Table 4 reports relative error $\|G_N - G_{\mathrm{ref}}\|_2 / \|G_{\mathrm{ref}}\|_2$ and gradient cosine. Relative error decreases monotonically across $N = 4, 16, 64$ in all 12 initial pools; all 12 improve endpoint cosine, and 11 have monotone cosine. The independent 24-pool replication has a consistent endpoint recovery direction. At $N = 9$, substantial discrepancy remains in both studies. The $N = 9$ values are separately observed, not interpolated from other budgets. The independent replication's $N = 4, 16, 64$ means are available at three-decimal summary precision; Figure 2 plots these summaries without inferred uncertainty bars.

Table 4: Frozen-gradient recovery with continuation budget. Brackets give available 95% intervals; $\approx$ denotes reported summary precision.

Study	N	Relative gradient error	Gradient cosine
12 pools	4	$0.6888\ [0.5764, 0.7932]$	$0.7235\ [0.6318, 0.8091]$
	9	0.5821	0.8015
	16	$0.3774\ [0.2991, 0.4473]$	$0.9202\ [0.8932, 0.9464]$
	64	$0.2727\ [0.2039, 0.3408]$	$0.9554\ [0.9336, 0.9740]$
Independent 24 pools	4	≈ 0.739	≈ 0.681
	9	0.5970	0.7959
	16	≈ 0.422	≈ 0.887
	64	≈ 0.292	≈ 0.943

In the 12-pool study, the value RMSE at $N = 4, 16, 64$ is $0.1478, 0.07539, 0.04241$, the credit RMSE is $0.1810, 0.09242, 0.05308$, and the post-whitening coefficient RMSE is $0.7973, 0.5091, 0.3155$. Coefficient cosines are $0.6790, 0.8691, 0.9497$ and sign-disagreement rates are $0.2251, 0.1343, 0.0773$. Mean gradient norm ratios are $0.9392, 0.9827, 0.9933$, so the residual error includes directional discrepancy even when the gradient norm is close to its reference. Coefficient RMSE correlates with relative gradient error (Spearman 0.8533, 95% CI $[0.7411, 0.9344]$) and with gradient cosine risk (0.8417, $[0.6934, 0.9324]$).

At $N = 9$, value/credit/coefficient RMSEs are $0.0980/0.1211/0.6306$ in the initial study and $0.1090/0.1374/0.6232$ in the independent replication; coefficient cosines are 0.7992 and 0.8040. Adjacent credit-error covariances are -0.00723 and -0.00976, with covariance-magnitude/shared-variance ratios 1.021 and 0.978, respectively.

D.3 Robustness and normalization-regime diagnostics

Table 5 consolidates structural, normalization, reference, and reward checks. BNSR measures centered evidence noise relative to regularized signal in the full normalization pool. Its association with error attenuation is stronger in the independent 24-pool study; it is a diagnostic of finite-budget behavior, not a universal failure threshold. All three BNSR tertiles in the initial study improve with increasing evidence budget. Centering affects sign disagreement, while scaling carries most of the BNSR-dependent coefficient attenuation difference. This coefficient result is consistent with the cancellation of a positive scalar denominator from single-pool gradient cosine.

The -2 reward denotes malformed outputs with multiple #### markers, not truncation. Consequently the relevant variance is σ^2/N, rather than its binary-reward special case. True truncation affects

Table 5: Robustness checks. Pairs separated by semicolons denote the initial 12-pool study and independent 24-pool replication, respectively. All intervals are reported 95% CIs.

Check	Observed result
Contrast structure	Initial-study adjacent covariance at $N = 4, 16, 64$: -0.01585 $[-0.02129, -0.01145]$, -0.004196 $[-0.005628, -0.002883]$, -0.001274 $[-0.002248, -0.000595]$. Magnitude/shared-variance ratios: $0.966, 1.029, 0.986$. Contrast-identity residual: zero.
Continuous BNSR	Spearman: 0.427; 0.614. Theil–Sen slope: 0.0287 $[-0.0820, 0.1449]$; 0.0833 $[0.0489, 0.1419]$. High-minus-low attenuation difference: 0.0579; 0.1215.
Reference cross-fit	Initial folds A/B: slopes -0.0128 $[-0.1043, 0.0497]$ and 0.0481 $[-0.0528, 0.1380]$, Spearman $0.238/0.385$. Independent folds A/B: 0.0408 $[-0.0448, 0.1339]$ and 0.1720 $[0.0795, 0.2607]$, Spearman $0.280/0.572$.
Leave-one-pool-out	In the 24-pool study, the high-minus-low difference remains positive in all 24 removals, range $[0.09429, 0.15498]$; robust slopes range $[0.07555, 0.09896]$.
Transformation	High-minus-low differences for raw / centering-only / scaling-only / full whitening: $(-0.0506, -0.0522, 0.0587, 0.0579)$; $(0.0137, 0.0170, 0.1101, 0.1215)$.
Rewards and truncation	Reward counts for $(-2, 0, 1)$: $(208, 32423, 25737)$; $(1044, 69356, 47104)$. True continuation truncation: $600/58368$ (1.028%); $1089/117504$ (0.927%). Primary trajectories reaching the 512-token cap: $1/200$; $2/192$.
Truncation sensitivity	Exclusion coefficient RMSE: 0.6025; 0.6369, retaining $11/12$; $23/24$ pools. One boundary per study has all 192 continuations truncated.
Finite-reference floor	Initial Ref-A/B: value RMSE 0.04947, correlation 0.99337, credit RMSE 0.06242; coefficient RMSE 0.3888, cosine 0.9237; gradient relative discrepancy mean 0.3538, median 0.3004.

approximately 1% of continuations; exclusion sensitivity is supplementary because it cannot retain every pool. The finite reference also contributes uncertainty: the initial study's $N = 64$ relative gradient error is below the half-reference discrepancy. We therefore interpret budget-dependent recovery and independent replication directly, without mechanically subtracting a nonlinear reference floor.

E Fixed-Budget S-SPO Reliability

E.1 Matched evidence design and paired inference

The intervention comparison holds the arithmetic decoder fixed and compares arithmetic-iid sampling (M1) with S-SPO (M2), using $N = 9$ continuations per primitive query. The independent 24-pool comparison contains 192 trajectories, 612 sampled queries, and 192 terminal records. Sixteen repetitions yield 88,128 continuations per method. All raw and post-whitening gradient measurements use these same frozen draws. The common Full128 reference comprises 128 independent continuations per query; sensitivity references use its disjoint first and second 64 draws.

Each metric is computed per pool and repetition, averaged across the 16 repetitions within each pool, then equally averaged across 24 pools. Value MSE averages query-level squared errors before averaging repetitions. Credit MSE uses the complete chains' 612 credit pairs, including terminal-related pairs and the recorded forced-zero first-segment convention. Pairing is by pool, with difference M2 minus M1. The percentile bootstrap resamples 24 pools 20,000 times with seed 20260905 and common bootstrap indices across metrics, methods, and references. Intervals are pointwise 95% intervals; repetitions, queries, and tokens are not independent inferential units.

The frozen BF16 actor is differentiated through all 201 parameter tensors (1,100,048,384 parameters), using float32 streamed norm and inner-product accumulation. Raw and normalized measurements share the estimated values, contrast, replication, probability mask, loss reduction, and first-step PPO derivative. The normalized branch additionally applies official eight-trajectory whitening with strict $p < 0.9$, unbiased variance, and 10^{-8} regularization. These full-parameter measurements precede optimizer transformations.

E.2 Independent confirmation and complete outcomes

An initial 12-pool method comparison increases gradient cosine from 0.82306 to 0.84844, with 11 pools improving; its paired interval includes zero. Value MSE decreases from 0.010348 to 0.007706, credit MSE from 0.014860 to 0.011285, and coefficient cosine increases from 0.80113 to 0.82904. The independent 24-pool comparison gives a closely aligned gradient-cosine gain of 0.025081, from 0.821496 to 0.846577, with paired 95% CI $[0.0090507, 0.0430989]$, exact sign-flip $p = 0.002615$, and 20 of 24 improving pools.

The primary outcomes are value MSE, local-credit MSE, and post-whitening gradient cosine risk. Cosine in the main table is the equivalent presentation of $1 - \text{cosine}$; the two are one directional outcome. Raw-gradient quadratic error and the remaining diagnostics are supporting. At Full128, value MSE decreases by 9.72%, credit MSE by 7.20%, gradient cosine risk by 14.05%, and raw-gradient quadratic error by 12.50%. The credit-MSE paired interval includes zero, while the value, normalized-direction, and raw-gradient quadratic error intervals exclude it.

For gradient error $e = \widehat{G} - G_{\text{ref}}$ and reference direction $u = G_{\text{ref}}/\|G_{\text{ref}}\|_2$, quadratic error is $\|e\|_2^2$, relative error is $\|e\|_2/\|G_{\text{ref}}\|_2$, tangent error is $\|(I - uu^\top)e\|_2$, and radial squared error is $(u^\top e)^2$. Relative tangent error divides the tangent norm by the reference norm. All are computed before repeat averaging; a mean squared error is not obtained by squaring an aggregated norm.

E.3 Reference sensitivity

Table 6 gives all reported metrics and paired intervals under each finite reference. All three primary point-estimate directions favor S-SPO under Full128 and either half-reference. Credit-MSE intervals include zero under all three references; gradient-direction intervals exclude zero throughout. The value interval excludes zero under Full128 and Ref-A but narrowly includes zero under Ref-B. Reference-dependent supporting intervals are reported in full. No nonlinear reference floor is subtracted, and this fixed-budget comparison does not identify an equivalent iid continuation count.

Table 6: Complete $N = 9$ reliability results. Δ =M2−M1; intervals are paired 95% pool-bootstrap CIs. † marks primary metrics; other rows are supporting. The Pools column counts repeat-averaged improvements out of 24. Cosine and cosine risk are equivalent presentations.

Metric	M1	S-SPO	Δ	95% CI for Δ	Pools
Full128					
Values and local credit					
Value MSE ↓†	0.0112055	0.0101167	-0.00108881	$[-0.00178690, -0.000343738]$	18/24
Credit MSE ↓†	0.0170203	0.0157953	-0.00122498	$[-0.00267891, 0.000364242]$	17/24
Before whitening					
Coefficient RMSE ↓	0.127353	0.120389	-0.00696410	$[-0.0119585, -0.00183859]$	16/24
Coefficient cosine ↑	0.814650	0.836302	0.0216526	$[0.00468572, 0.0416599]$	19/24
Raw-gradient quadratic error ↓	75.7691	66.2955	-9.47364	$[-16.1104, -2.66419]$	18/24
Relative gradient error ↓	0.630468	0.598860	-0.0316077	$[-0.0572019, -0.00172698]$	17/24
Gradient cosine ↑	0.829657	0.851675	0.0220180	$[0.00737735, 0.0388497]$	19/24
Gradient cosine risk ↓	0.170343	0.148325	-0.0220180	$[-0.0388497, -0.00737735]$	19/24
Relative tangent error ↓	0.589357	0.562176	-0.0271816	$[-0.0494868, -0.00351867]$	15/24
Tangent error ↓	7.25639	6.79628	-0.460118	$[-0.776666, -0.141924]$	15/24
Radial squared error ↓	11.3806	8.44550	-2.93506	$[-5.92019, -0.532578]$	18/24
After whitening					
Coefficient RMSE ↓	0.567584	0.534198	-0.0333866	$[-0.0587348, -0.00868146]$	18/24
Coefficient cosine ↑	0.806300	0.829931	0.0236303	$[0.00496043, 0.0449341]$	19/24
Gradient quadratic error ↓	1459.26	1238.44	-220.817	$[-355.773, -97.2983]$	19/24
Relative gradient error ↓	0.539351	0.501264	-0.0380866	$[-0.0578347, -0.0176443]$	20/24
Gradient cosine ↑	0.821496	0.846577	0.0250811	$[0.00905071, 0.0430989]$	20/24
Gradient cosine risk ↓†	0.178504	0.153423	-0.0250811	$[-0.0430989, -0.00905071]$	20/24
Relative tangent error ↓	0.489119	0.461665	-0.0274544	$[-0.0424926, -0.0111547]$	19/24
Tangent error ↓	31.1099	29.2526	-1.85729	$[-2.89271, -0.752785]$	19/24
Radial squared error ↓	322.009	233.641	-88.3680	$[-155.946, -31.0213]$	19/24

Table 6 continued

Metric	M1	S-SPO	Δ	95% CI for Δ	Pools
Ref-A					
Values and local credit					
Value MSE ↓†	0.0122643	0.0109314	-0.00133291	$[-0.00206746, -0.000607475]$	20/24
Credit MSE ↓†	0.0180902	0.0167722	-0.00131796	$[-0.00262546, 0.0000764007]$	17/24
Before whitening					
Coefficient RMSE ↓	0.131095	0.124163	-0.00693197	$[-0.0116472, -0.00215530]$	16/24
Coefficient cosine ↑	0.805498	0.825255	0.0197569	$[0.00353846, 0.0381201]$	18/24
Raw-gradient quadratic error ↓	79.9549	71.9950	-7.95987	$[-14.5432, -1.54089]$	17/24
Relative gradient error ↓	0.632486	0.602340	-0.0301456	$[-0.0550395, -0.00215667]$	18/24
Gradient cosine ↑	0.819876	0.839996	0.0201199	$[0.00569306, 0.0364998]$	18/24
Gradient cosine risk ↓	0.180124	0.160004	-0.0201199	$[-0.0364998, -0.00569306]$	18/24
Relative tangent error ↓	0.588135	0.564179	-0.0239560	$[-0.0471952, 0.000697287]$	15/24
Tangent error ↓	7.45795	7.03598	-0.421967	$[-0.733357, -0.119082]$	15/24
Radial squared error ↓	12.1068	9.65353	-2.45329	$[-5.19536, -0.141144]$	16/24
After whitening					
Coefficient RMSE ↓	0.583432	0.552127	-0.0313051	$[-0.0552763, -0.00767722]$	18/24
Coefficient cosine ↑	0.796528	0.818187	0.0216594	$[0.00389374, 0.0416488]$	18/24
Gradient quadratic error ↓	1553.95	1356.19	-197.765	$[-342.396, -72.1451]$	19/24
Relative gradient error ↓	0.550282	0.516763	-0.0335186	$[-0.0529378, -0.0133722]$	19/24
Gradient cosine ↑	0.810950	0.833621	0.0226705	$[0.00693790, 0.0407397]$	18/24
Gradient cosine risk ↓†	0.189050	0.166379	-0.0226705	$[-0.0407397, -0.00693790]$	18/24
Relative tangent error ↓	0.495485	0.472066	-0.0234188	$[-0.0386301, -0.00661481]$	19/24
Tangent error ↓	31.8163	30.1907	-1.62554	$[-2.67503, -0.518673]$	19/24
Radial squared error ↓	373.534	285.694	-87.8397	$[-166.145, -21.4446]$	18/24
Ref-B					
Values and local credit					
Value MSE ↓†	0.0118315	0.0109867	-0.000844718	$[-0.00163692, 0.0000243693]$	17/24
Credit MSE ↓†	0.0179998	0.0168678	-0.00113200	$[-0.00303534, 0.000794795]$	16/24
Before whitening					
Coefficient RMSE ↓	0.131192	0.124762	-0.00643063	$[-0.0120731, -0.000763296]$	15/24
Coefficient cosine ↑	0.803780	0.826009	0.0222293	$[0.00396656, 0.0437376]$	18/24
Raw-gradient quadratic error ↓	81.1263	70.2301	-10.8962	$[-19.3554, -2.78035]$	18/24
Relative gradient error ↓	0.641605	0.610713	-0.0308919	$[-0.0579231, 0.000337739]$	18/24
Gradient cosine ↑	0.820408	0.843465	0.0230567	$[0.00722426, 0.0409275]$	20/24
Gradient cosine risk ↓	0.179592	0.156535	-0.0230567	$[-0.0409275, -0.00722426]$	20/24
Relative tangent error ↓	0.598810	0.572043	-0.0267666	$[-0.0505948, -0.00186863]$	15/24
Tangent error ↓	7.45659	6.99407	-0.462517	$[-0.783059, -0.141804]$	15/24
Radial squared error ↓	12.7664	9.57570	-3.19068	$[-6.48297, -0.576179]$	15/24
After whitening					
Coefficient RMSE ↓	0.584268	0.552316	-0.0319519	$[-0.0578034, -0.00698851]$	17/24
Coefficient cosine ↑	0.795410	0.819214	0.0238038	$[0.00424587, 0.0463628]$	18/24
Gradient quadratic error ↓	1528.42	1300.47	-227.951	$[-366.633, -100.599]$	20/24
Relative gradient error ↓	0.557390	0.518809	-0.0385816	$[-0.0589061, -0.0179138]$	19/24
Gradient cosine ↑	0.812059	0.837896	0.0258368	$[0.00857398, 0.0449415]$	20/24
Gradient cosine risk ↓†	0.187941	0.162104	-0.0258368	$[-0.0449415, -0.00857398]$	20/24
Relative tangent error ↓	0.505162	0.477094	-0.0280680	$[-0.0424389, -0.0130894]$	18/24
Tangent error ↓	31.9407	30.0813	-1.85946	$[-2.83845, -0.820938]$	18/24
Radial squared error ↓	332.311	245.214	-87.0971	$[-152.893, -31.2661]$	18/24

F ONLINE PROTOCOLS

F.1 MATCHED TRAINING DESIGN

For all online matched comparisons, SPO uses arithmetic-iid auxiliary sampling (M1), while S-SPO uses the same arithmetic decoder with equal-mass stratification (M2). The operational target, nine continuations per query, local-credit construction, and downstream SPO objective and training recipe are matched within each run pair; the intervention is the joint design of auxiliary continuation evidence.

Table 7: RhoMath training recipe. Within each training seed, the comparison retains the SPO-chain recipe and changes auxiliary-evidence sampling.

Component	Setting
Policy iterations; prompts per iteration	256; 64
Main trajectories per prompt; per iteration	8; 512
Auxiliary continuations per query	$N = 9$
Carrier; cutpoint interval	SPO-chain; 5
Probability mask	$p < 0.9$
Training temperature; top-p	0.6; 0.9
Maximum generated tokens; effective context	1,024; 2,047
Maximum step for value estimation	25
PPO epochs; effective batch	2; 64
Microbatch; gradient accumulation	8; 8
Optimizer updates per iteration; total	16; 4,096
Optimizer; weight decay	AdamW; 0
Adam (β_1, β_2); numerical ϵ	$(0.9, 0.999)$; 10^{-8}
PPO clip; gradient-norm clip	0.2; 1
KL coefficient; adaptive KL controller	10^{-4}; disabled
Advantage whitening; reward whitening	enabled; disabled
Peak learning rate; warmup updates	10^{-6}; 128
Decay; final learning rate	single cosine; 10^{-7}

The primary online comparison uses RhoMath-1.1B SFT on GSM8K with three paired training seeds, $s \in \{41, 42, 43\}$. For each seed, SPO and S-SPO use a fixed 256-iteration budget and 4,096 optimizer updates. The endpoint is selected by this fixed budget. Every completed iteration in all six Rho runs satisfies continuations $= 9 \times$ queries. A separate Llama-3.2-1B-Instruct comparison provides one matched second-model run per method, also at 256 iterations and $N = 9$, with its own 768-update schedule.

F.2 RhoMath training recipe

The initial model is `realtreetune/rho-1b-sft-GSM8K`, with 1,100,048,384 parameters and BF16 computation. The GSM8K training split contains 7,473 examples. A NumPy `default_rng(42)` permutation assigns the first 373 indices to RL-held-out validation and the remaining 7,100 to RL training, preserving their order; the split seed is distinct from the three training seeds. RL-held-out denotes exclusion from RL updates; the initial checkpoint has already undergone GSM8K supervised fine-tuning. The 1,319 official test examples provide the separately held-out final evaluation. Exact, Unicode-normalized/case-folded/whitespace-normalized, and alphanumeric-only comparisons find no cross-split duplicates. Table 7 summarizes the Rho training recipe.

The learning rate is indexed by successful optimizer updates. For update u, the schedule is

$$\eta_u = \begin{cases} 10^{-6}u/128, & 1 \le u \le 128, \\ 10^{-7} + \dfrac{9 \times 10^{-7}}{2}\left[1 + \cos\left(\dfrac{\pi(u-128)}{3968}\right)\right], & 129 \le u \le 4096. \end{cases} \tag{24}$$

The scheduler advances once per successful update. The prompt is `[MATH_TASK] Problem:`, a newline, the question, two newlines, and `Solution:`. Training appends BOS to the query and EOS to the response. The backend's 2,048-token limit accommodates the tokenizer-added BOS within the effective 2,047-token budget. Training uses the native GSM8K reward function with calculator expressions removed and the unfinished-response penalty set to zero.

F.3 Validation and dense learning curves

Each of the eight runs is evaluated on 373 validation questions at the 33 observed checkpoints $0, 8, \ldots, 256$, using four draws per question. The Rho runs additionally have 16-draw estimates at iterations $0, 64, 128, 192, 224, 256$; the Llama runs have 16-draw estimates at iterations 0 and 256. The primary endpoint comparisons use the 16-draw estimates at iteration 256. Each Rho seed pair

shares its initial estimates. Accuracy is the proportion of correct responses across questions and draws, so both evaluation budgets estimate mean single-draw correctness.

Rho validation uses temperature 0.35, top-$p = 0.9$, the training prompt, a 1,024-token generation limit, and the 2,047-token effective context. Generation stops on EOS or three newline characters followed by `Problem:`. The parser extracts the suffix after the last `####`, with a missing marker treated as failure. It strips surrounding whitespace, removes commas from the prediction, lower-cases the prediction and gold answer, and compares the resulting strings. No additional numeric extraction or answer repair is applied. The dense four-draw series and the 16-draw anchor estimates are kept distinct in the analysis; their complete values are provided in the supplementary CSV data.

F.4 Official GSM8K test

All six final Rho checkpoints are evaluated on the same 1,319 official test questions, with 16 draws per question and 21,104 evaluated responses per checkpoint. The evaluation temperature is 0.35, top-p is 0.9, and the maximum number of generated tokens is 1,024. The prompt, answer parser, and correctness rule are identical across the six checkpoints. The prompt is the template specified above. The prediction is the trimmed suffix after the last `####`; an absent marker is incorrect. The gold answer is the official answer's final-answer suffix. Comparison strips surrounding whitespace, removes commas from the prediction, and lower-cases both strings, with no numeric rescue. Length-truncated responses remain in the denominator and are scored by the same rule.

For each question, one ordinary iid generation request returns $n = 16$ responses. Every request explicitly uses seed 42, and the inference server seed is also 42. Seeds are not derived from question or draw identifiers, and evaluation does not consist of 16 separately reseeded rounds. The aggregate statistic is total correct/total draws, an estimate of single-draw correctness rather than pass@16, best-of-16, or majority-vote accuracy. These common evaluation seeds are distinct from training seeds 41, 42, and 43.

F.5 Llama second-model setting

Available execution records for Llama-3.2-1B-Instruct cover GSM8K and one matched run per method. Both records contain 256 completed iterations with three successful optimizer updates per iteration, ending at update 768. Their last-update learning-rate sequences agree exactly: the recorded learning rate reaches 10^{-6} at iteration 32 (update 96) and is 8.9144×10^{-7} at iteration 256. This schedule is specific to Llama. Each iteration uses nine auxiliary continuations per query. Validation uses 373 questions, 1,492 responses per dense checkpoint, and 5,968 responses at the initial and final 16-draw checkpoints. Appendix G reports its endpoint, AUC, and active-training cost alongside the three-seed Rho comparison.

F.6 Recorded computation

Active-training time is obtained from the recorded per-iteration durations. The timing interval includes generation, PPO, and checkpoint verification, and excludes validation, restart/setup, and post-checkpoint cleanup. Across all eight runs, summed iteration durations agree with final cumulative-time fields, and recorded GPU-hours equal these sums divided by 3,600. This measures the active cost of the complete training run. The evolving policies can produce different query counts and output lengths despite matching N, so complete-run time is reported separately from the per-query evidence budget.

F.7 Relation to the published SPO result

The SPO paper reports 56.7% GSM8K test accuracy for SPO-chain and 40.3% for its initial SFT model (Guo et al., 2025). Several core settings align with ours: cutpoint interval 5, $N = 9$, training temperature/top-p of $0.6/0.9$, learning rate 10^{-6}, two policy epochs, and evaluation temperature/top-p of $0.35/0.9$ with 16 samples. The paper does not identify its 56.7% endpoint as a 256-iteration checkpoint, and its training plots cover longer horizons. Our scientific comparison is arithmetic-iid SPO (M1) versus stratified S-SPO (M2) under the same executed 256-iteration protocol and common evaluation protocol; the published score is not a fixed-budget reproduction target.

Table 8: Per-seed online results at the fixed training budget. Endpoints use 16 draws per question. SPO and S-SPO entries are percentages; gains are percentage points. In Rho mean rows, ± is the sample standard deviation of the three paired gains. Llama supplies one matched second-model comparison.

Setting and measure	Seed	SPO	S-SPO	Paired gain
Rho validation	41	65.7172	68.1635	+2.4464
	42	65.4658	68.5489	+3.0831
	43	65.5999	68.3143	+2.7145
	Mean	65.5943	68.3423	$+2.7480 \pm 0.3197$
Rho official test	41	49.8815	51.9996	+2.1181
	42	49.5498	51.9522	+2.4024
	43	49.7015	51.9475	+2.2460
	Mean	49.7110	51.9665	$+2.2555 \pm 0.1424$
Rho validation AUC	41	62.6623	65.0721	+2.4097
	42	61.8486	64.3034	+2.4548
	43	62.4466	64.6647	+2.2181
	Mean	62.3192	64.6800	$+2.3609 \pm 0.1257$
Llama validation	Single pair	78.2172	81.0657	+2.8485
Llama validation AUC	Single pair	76.1153	78.0643	+1.9489

G Online results and statistical summaries

G.1 Per-seed validation and official-test results

Table 8 reports the final validation and official-test accuracies for each Rho training seed, together with the dense-curve AUC. Every Rho seed pair favors S-SPO on all three measures. For an accuracy or AUC expressed as a fraction, define the paired gain in percentage points as

$$d_s = 100(a_{\mathrm{S},s} - a_{\mathrm{SPO},s}), \qquad \bar{d} = \frac{1}{3}\sum_{s\in\{41,42,43\}} d_s, \qquad s_d = \sqrt{\frac{1}{3-1}\sum_{s\in\{41,42,43\}} (d_s - \bar{d})^2}. \tag{25}$$

Reported $\bar{d} \pm s_d$ values are the mean and sample standard deviation of the three paired training-seed gains, using denominator $3-1$ (`ddof=1`). They are neither confidence intervals nor standard errors. The mean validation gain is 2.7480 ± 0.3197 points, and the mean official-test gain is 2.2555 ± 0.1424 points. Differences and summary statistics are computed before rounding.

Llama's final validation counts are 4,668 correct responses for SPO and 4,838 for S-SPO, each out of 5,968. Its initial 16-draw accuracy is 65.5161% for both methods. This comparison supplies an observed second-model gain; no across-seed dispersion estimate is assigned to the single pair.

G.2 Learning-curve summary

Let $a_{m,s}(t_j)$ denote four-draw validation accuracy for method m and seed s at $t_j = 8j$, $j = 0, \ldots, 32$. The iteration-normalized trapezoidal area is

$$\mathrm{AUC}_{m,s} = \frac{1}{256}\sum_{j=1}^{32} \frac{a_{m,s}(t_{j-1}) + a_{m,s}(t_j)}{2}(t_j - t_{j-1}). \tag{26}$$

Sixteen-draw anchors do not replace points in this calculation. The three Rho AUC gains are all positive, with mean 2.3609 percentage points. From iteration 64 through 256, every recorded dense checkpoint in every Rho seed pair favors S-SPO. Llama's dense curve favors S-SPO at every recorded checkpoint from iteration 24 onward. Figure 3 displays the Rho paired-gain curves and the separate Llama comparison. The supplementary CSV data contain all 33 observed dense checkpoints for each of the eight runs, including correct-response counts, total draws, and cumulative active-training time. They also include each available 16-draw anchor and the six final official-test counts. The plotted paired differences and AUC values use these observed checkpoints; no additional checkpoints are imputed.

Table 9: Recorded active-training cost for all matched runs. GPU-hours sum completed iteration durations; the final column is $100(T_{\mathrm{S}}/T_{\mathrm{SPO}} - 1)$. Each Rho run executes 4,096 optimizer updates; each Llama run executes 768.

Model	Training seed	SPO GPU-h	S-SPO GPU-h	Increase (%)
RhoMath-1.1B	41	27.3745	28.0170	2.3474
	42	23.1852	23.4325	1.0665
	43	25.9636	26.8949	3.5872
Llama-3.2-1B-Instruct	Single pair	19.8352	20.2407	2.0442

G.3 Active-training cost

Table 9 reports the active GPU-hours for all eight completed runs. Rho's within-seed relative increases are 2.3474%, 1.0665%, and 3.5872% for seeds 41, 42, and 43, respectively. Their arithmetic mean is 2.3337%; this averages the three within-seed percentages. Llama's increase is 2.0442%. These figures accompany the accuracy gains at the common per-query budget.

The supplementary per-iteration data retain auxiliary-query counts, continuation counts, generated-token counts, and generation times. They verify the $N = 9$ continuation budget throughout all eight runs while allowing total query counts and response lengths to vary with the evolving policies. Accordingly, the measured cost differences characterize complete training runs rather than an isolated or universal sampler overhead.

H Reproducibility Summary

Contracts and proofs. Appendix A specifies the operational continuation law, primitive-query identity and reuse, arithmetic decoding, mask and normalization pool, and decoder-replacement control. Appendices B and C provide the linear covariance-to-risk proofs and the normalized-update expansion, including its remainder and scope. Appendices D and E give the frozen-gradient protocols, finite-reference definitions, inference units, and paired offline results. Appendices F and G report the Rho training and evaluation protocols, available Llama execution records, per-run measurements, and resource accounting.

Data and checkpoint identity. Both model settings use GSM8K (Cobbe et al., 2021). The RhoMath setting uses the GSM8K-supervised checkpoint `realtreetune/rho-1b-sft-GSM8K` from the Rho-1 family (Lin et al., 2024). The second setting starts from `meta-llama/Llama-3.2-1B-Instruct` (Meta, 2024). The Rho recipe and the verified recorded protocol details for Llama are reported separately.

Analysis materials. The accompanying analysis materials contain manuscript sources, bibliography, display data, vector figures, and figure-generation and summary-analysis scripts. The scripts reproduce the conceptual pipeline, budget-dependent gradient summaries, complete offline tables, and online curves and aggregates from their corresponding inputs. Offline paired intervals use pools as the inference unit; the tabulated summaries preserve the archived intervals rather than reconstructing the original pool bootstrap. For the three Rho training seeds, the reported sample standard deviation summarizes paired gains across training runs.