COEFFICIENT GRADIENT IMPACT: CALIBRATING EXECUTED TRUST REGIONS IN LLM RL

Anonymous authors
Paper under double-blind review

ABSTRACT

Trust-region mechanisms are ubiquitous in reinforcement learning for language-model reasoning, yet their control parameters inhabit incompatible geometries and say little about the intervention actually executed by training code. We introduce **Coefficient Gradient Impact**, an objective-aligned measure of the relative displacement of the token coefficients that form the policy gradient. Our analysis yields exact identities for hard removal and smooth reweighting, separates activation frequency from coefficient energy, and characterizes how coefficient-space interventions propagate to parameter gradients. Building on this measure, we develop a baseline-aware calibration procedure that maps heterogeneous trust-region mechanisms onto a common, baseline-declared reporting coordinate, together with a one-sided impact-capping rule that adapts the control threshold while preserving the underlying objective geometry. An immutable-trace protocol makes the measurement auditable across rollout, anchor, and current policies while preserving scoring and synchronization provenance. Multi-run experiments on physical servers across model scales and numerical precisions indicate that the framework enables calibrated comparison among otherwise incomparable mechanisms, tracks parameter-gradient change more faithfully than the tested alternatives, provides useful early warning of unstable training, and transfers across the tested execution settings. Across paired training runs, impact-guided control exhibits the strongest observed stability profile while maintaining comparable reasoning accuracy. Coefficient Gradient Impact therefore provides a principled interface between trust-region design, systems execution, and empirical evaluation, enabling more interpretable comparison and auditable control of language-model reinforcement learning.

1 INTRODUCTION

Policy optimization for language-model reasoning must permit useful learning while controlling update magnitude. Existing trust-region objectives approach this goal in different coordinates: some constrain an importance ratio, others act on token probabilities, and either boundary may be enforced by hard exclusion or smooth attenuation (Schulman et al., 2017; Shao et al., 2024; Qi et al., 2026; Yao et al., 2026). Their thresholds therefore describe different geometric objects. Comparing the thresholds directly obscures how much the *executed* objective changes.

Activation frequency captures where a rule acts, while coefficient energy captures how strongly it changes the objective. A hard rule can suppress a small set of high-energy tokens, whereas a smooth rule can touch almost every token weakly. Average weight provides another complementary view because advantage, ratio, response reduction, and token reuse determine each contribution. We compare mechanisms after fixing geometry, execution, objective reduction, and population.

We propose **Coefficient Gradient Impact**, the relative change a trust-region mechanism makes to the scalar coefficients multiplying token score functions. It is local to an update, faithful to the implemented reduction, and applicable to hard and smooth rules. Exact identities separate coefficient energy from event frequency, while a transport identity makes explicit the score-vector geometry that stands between coefficient and parameter space. Because the ratio and Binary-TV variants use different reference policies, Coefficient Gradient Impact is a baseline-declared scale of fractional intervention strength for each objective.

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.

Measurement also requires separating three roles often conflated by asynchronous training: the behavior policy that generated a response, the frozen policy at the start of an outer step, and the policy being optimized. Figure 1 follows one immutable prompt–response trace through these roles and the executed objective. Engine fidelity is evaluated by teacher-forcing the same response in a compatible score domain under snapshots expected to match, while actual policy movement remains a separate estimand.

We contribute an objective-aligned metric with exact hard- and smooth-execution identities and an explicit cross-objective boundary; a same-token, policy-aware protocol that separates systems fidelity from policy movement; and a baseline-aware calibration algorithm with an impact-guided control rule. A completed real-server evaluation studies gradient transport, temporal warning, matched interventions, transfer, and downstream utility. Together these components form a measurement-and-control layer compatible with existing policy optimizers.

2 Related Work

2.1 RL for language-model post-training

PPO-style post-training and group-relative objectives optimize sampled responses, whereas DPO provides a preference-based alternative (Ouyang et al., 2022; Shao et al., 2024; Rafailov et al., 2023). Verifiable mathematics supplies inexpensive outcome signals (Cobbe et al., 2021), and recent systems scale this recipe or alter normalization and granularity (DeepSeek-AI et al., 2025; Hu et al., 2025; Yu et al., 2025; Liu et al., 2025; Zheng et al., 2025). Results also caution that RLVR may amplify capabilities already present in a base model rather than create new reasoning capacity (Yue et al., 2025). These works chiefly compare reward or task performance; we instead measure how strongly the executed trust-region component changes an update.

2.2 Trust-region geometry and execution

Classical trust-region optimization constrains a distributional neighborhood, whereas clipped surrogates replace the constrained problem with a local ratio rule (Schulman et al., 2015; 2017). Recent language-model objectives instead act in token-probability space, smooth the intervention, or change its sequence granularity (Qi et al., 2026; Yao et al., 2026; Xie et al., 2024; MiniMax, 2025; Zhang et al., 2026b; Shen et al., 2026). Adaptive-bound approaches change where the boundary lies (Li et al., 2026a; Yang et al., 2025b; Xi et al., 2025); complementary off-policy methods alter guidance, sample priority, or experience reuse (Yan et al., 2025; Mao et al., 2026; Lu et al., 2026). Our framework is orthogonal to those choices: it measures realized coefficient change after geometry, execution, population, and reduction are declared, without assuming one family is universally preferable.

2.3 Training–inference mismatch

Serving and training engines can assign different probabilities to the same tokens even when their weights are intended to match (Zhong et al., 2026). Prior work studies precision and parallelism effects, quantized rollouts, or corrections to the resulting mismatch (Qi et al., 2025; Zhang et al., 2025; Gu et al., 2026; Li et al., 2026b; Ye et al., 2026; Zhang et al., 2026a). Related analyses separate behavior-to-anchor mismatch from the subsequent optimization update (Liang et al., 2026). We focus on a prerequisite shared by these analyses: both sides must score the same immutable trace in a compatible probability domain, with explicit model, tokenizer, and synchronization provenance.

2.4 Optimization diagnostics

Entropy collapse, update-factor analyses, and gradient studies show that implementation choices can change the effective objective (Xu et al., 2026; Wang et al., 2026; Tang & Munos, 2025). Distributed training and serving add provenance boundaries (Kwon et al., 2023; Zheng et al., 2023; Shoeybi et al., 2019; Rajbhandari et al., 2019; Sheng et al., 2024; Fu et al., 2025). Mismatch localization asks where policies disagree (Zhou et al., 2026); Coefficient Gradient Impact asks how much the

executed constraint changes the objective-weighted coefficients. The populations and estimands remain distinct.

3 Problem Formulation

3.1 Three policy roles

For response token t in outer step k, let π_θ denote the differentiated current policy, π_k the frozen training anchor at the start of the outer step, and μ the rollout behavior policy. Their direct probability differences are

$$D_{\mathrm{sys},t} = |\pi_k - \mu|, \qquad D_{\mathrm{update},t} = |\pi_\theta - \pi_k|, \qquad D_{\mathrm{total},t} = |\pi_\theta - \mu|. \tag{1}$$

D_{sys} measures a behavior-to-anchor difference and characterizes engine/precision fidelity under same-token and same-weight conditions. D_{update} is direct optimization drift. In general, $\mathbb{E}[D_{\mathrm{total}}] - \mathbb{E}[D_{\mathrm{sys}}]$ is a difference-of-means proxy distinct from $\mathbb{E}[D_{\mathrm{update}}]$.

3.2 Objective reduction

Let $\mathcal{I}$ contain the R retained responses with at least one used response token, let L_i be the used-token count of response i, and let $m_{i,t} \in \{0,1\}$ be its response mask. We study the response mean of token means used by the training implementation,

$$\mathcal{L} = \sum_{i \in \mathcal{I}} \sum_t q_{i,t}\, \ell_{i,t}, \qquad q_{i,t} = \frac{m_{i,t}}{RL_i}. \tag{2}$$

This reduction gives every retained response equal mass through length-aware token weights. The objective and its coefficient statistics are defined for $R > 0$.

3.3 Unified coefficient view

Write $r_{i,t} = \pi_\theta / \pi_{\mathrm{ref}}$, where the reference is π_k for ratio geometry and μ for behavior-anchored TV geometry. The executed per-token derivative has the form

$$c_{i,t}^{\mathrm{mode}} = -q_{i,t} \hat{A}_i r_{i,t} w_{i,t}, \qquad c_{i,t}^{\mathrm{base}} = -q_{i,t} \hat{A}_i r_{i,t}. \tag{3}$$

Here $w \in \{0,1\}$ for hard execution and, for the smooth family considered here, $w = 1 - \mathrm{sign}(\hat{A}(r-1))X/\theta$. Smooth ratio uses $X = |r-1|$ and $\theta = \epsilon$; selected-token Binary-TV uses $X = |\pi_\theta - \mu|$ and $\theta = \delta$. The corresponding parameter gradient is $g = \sum_{i,t} c_{i,t} \nabla_\theta \log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t})$. The score-vector geometry governs how coefficient impact is transported into parameter-gradient impact.

3.4 Intervention events and populations

A token is *touched* when $w \neq 1$, *materially attenuated* when $w < 0.5$, and *sign-flipped* when $w < 0$. These events describe the intervention distribution, while Coefficient Gradient Impact measures its coefficient energy. We distinguish rollout-generated responses/tokens, retained responses/tokens, unique used response tokens, and token-update exposures. Reusing one unique token over E inner updates yields E exposures; prompt and padding positions belong to neither used population.

4 Coefficient Gradient Impact

4.1 Definition

At one inner update, flatten the used response-token population and use the objective-weighted coefficients from Eq. 3. Coefficient Gradient Impact is

$$\mathrm{impact}_c = \frac{\|c^{\mathrm{mode}} - c^{\mathrm{base}}\|_2}{\|c^{\mathrm{base}}\|_2}. \tag{4}$$

We evaluate the quantity on nonempty populations with finite inputs and a positive base coefficient norm; admissibility status is recorded explicitly. The subscript emphasizes coefficient space, while signed mean weight, touched fraction, and parameter-gradient norm provide complementary diagnostics. It is invariant to a shared nonzero rescaling of c^{mode} and c^{base}.

4.2 SMOOTH REWEIGHTING

For $w_{i,t} = 1 - s_{i,t}X_{i,t}/\theta$, where $s_{i,t} = \mathrm{sign}(\hat{A}_i(r_{i,t} - 1))$, define

$$u_{i,t} = \left(q_{i,t}\hat{A}_i r_{i,t}\right)^2. \tag{5}$$

On the valid used-token population,

$$\mathrm{impact}_c = \frac{\mathrm{RMS}_u(X)}{\theta}, \qquad \mathrm{RMS}_u(X) = \sqrt{\frac{\sum_{i,t} u_{i,t}X_{i,t}^2}{\sum_{i,t} u_{i,t}}}. \tag{6}$$

This is an algebraic identity for the stated smooth weight. Terms with zero advantage have $u = 0$; terms with $r = 1$ have $X = 0$ for the ratio and behavior-anchored Binary-TV forms. Validated coefficients, objective weights, and provenance instantiate the identity numerically.

4.3 HARD REMOVAL

If hard execution sets $w = 0$ on the executed removal set $\mathcal{R}$ and $w = 1$ elsewhere, then

$$\mathrm{impact}_c = \sqrt{\frac{\sum_{(i,t)\in\mathcal{R}} u_{i,t}}{\sum_{i,t} u_{i,t}}} = \sqrt{C_{\mathcal{R}} f}, \tag{7}$$

where, for $f > 0$, $f = |\mathcal{R}|/N_{\mathrm{used}}$ and $C_{\mathcal{R}}$ is the removed coefficient-energy share divided by f. If $f = 0$, then $\mathrm{impact}_c = 0$ and $C_{\mathcal{R}}$ need not be defined. All quantities use the same objective-weighted population, and $\mathcal{R}$ is the executed PPO-active set rather than the larger threshold-violation set.

4.4 A GEOMETRY-DEPENDENT BOUND

For smooth Binary-TV, set $\theta = \delta$. Since $X = |\pi_\theta - \mu| \le 1$, $|w - 1| \le 1/\delta$. Therefore $\delta > 2$ keeps $w \ge 0.5$. Ratio geometry instead has unbounded $|r - 1|$. This contrast establishes that nominal thresholds are geometry-specific intensity scales. The Binary-TV bound is distribution-free; the typical impact of a concrete configuration remains an empirical property of its executed token population.

4.5 BASELINE-AWARE COMPARISON AND GRADIENT TRANSPORT

The denominator in Eq. 4 defines the estimand. Ratio geometry uses a π_k-anchored base, whereas the Binary-TV implementation uses a μ-anchored base. Equal impact therefore expresses matched fractional coefficient displacement within each declared objective. For studies using a common c^0, the decomposition $c^m - c^0 = (c^m - c^{\mathrm{base},m}) + (c^{\mathrm{base},m} - c^0)$ identifies Coefficient Gradient Impact with the executed-intervention term.

Let Z be the linear map whose columns are the token score vectors and $\delta c = c^{\mathrm{mode}} - c^{\mathrm{base}}$. Whenever all denominators are nonzero,

$$\frac{\Delta_g}{\mathrm{impact}_c} = \frac{\|Z\delta c\|_2/\|\delta c\|_2}{\|Zc^{\mathrm{base}}\|_2/\|c^{\mathrm{base}}\|_2}. \tag{8}$$

Thus score-vector geometry transports coefficient impact into parameter-gradient change, and their empirical association supplies the corresponding convergent-validity test.

5 MEASUREMENT PROTOCOL

5.1 IMMUTABLE TRACES AND SCORE DOMAIN

The unit of comparison is a fixed sequence of prompt IDs and response IDs, accompanied by the response and attention masks, prompt boundary, and position convention. Every scorer returns the log probability of the declared response token at that position under teacher forcing. Matching token

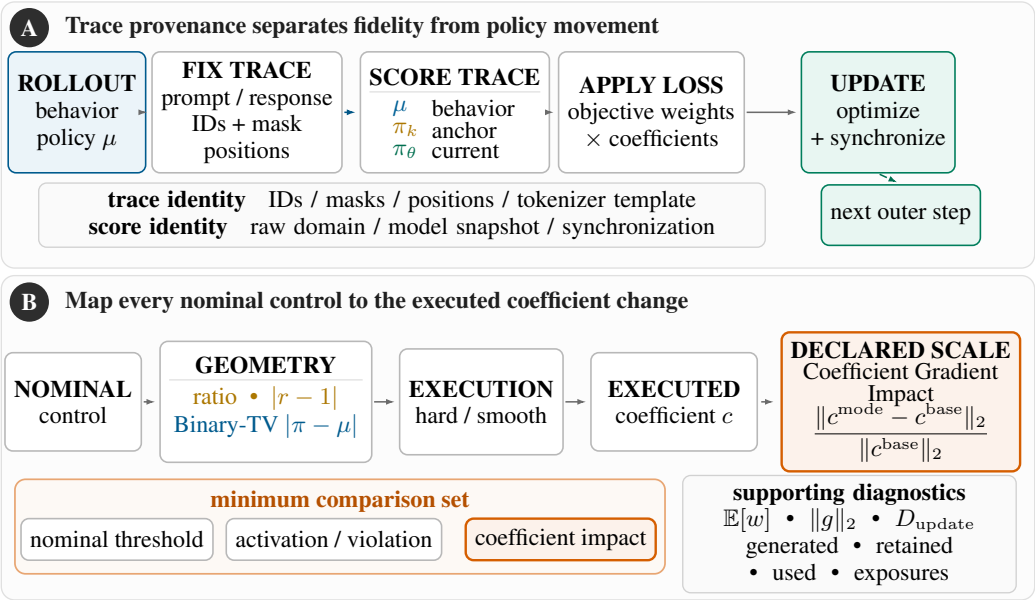

Figure 1: **From training lifecycle to an executed-intervention scale. (A)** An immutable response is scored under behavior, anchor, and current-policy roles; token and snapshot provenance determine whether a difference represents engine fidelity or policy movement. **(B)** Geometry and execution map a nominal control to an executed coefficient change. Coefficient Gradient Impact is the normalized comparison within each declared baseline, while activation and optimization diagnostics remain distinct. Color, labels, grouping, and line style provide redundant encoding for grayscale reproduction.

IDs and context provenance define aligned observations, and the protocol fixes the trace throughout the comparison.

Our canonical score is the raw-model selected-token log probability: the log-softmax of unprocessed logits, gathered at the specified response ID. Temperature, top-p, or sampling seeds may change which trace is generated. Once fixed, its raw teacher-forced score is invariant to those sampling controls. Scores exposed after logit processing or by an API with an unspecified probability domain retain a separate domain label from raw training scores.

5.2 Separating Fidelity from Policy Movement

Token and model-state alignment together determine the meaning of a score difference. When model identity, checkpoint, weight generation, and sync generation establish that two scorers hold the same weights, their difference measures *engine/precision fidelity*. If μ and π_k refer to different weight generations, the same-token difference is a *behavior-to-anchor policy difference*. The train-side π_θ–π_k comparison is *optimization drift*. This categorization keeps systems fidelity and policy movement as distinct estimands.

For a valid engine-fidelity population, let $e_{i,t} = |\log p^{(a)}(y_{i,t}) - \log p^{(b)}(y_{i,t})|$. Artifacts record the full distribution; the paper reports valid/invalid counts, mean, p95, and maximum. Empty or invalid populations receive reason-coded nulls. The full alignment fields, content digests, and resume rules appear in Appendix C.

5.3 A Complete Calibration Report

Coefficient Gradient Impact summarizes how strongly a mechanism changes its baseline coefficient vector, while complementary diagnostics localize the change and measure its transport into parameter space. The complete calibration report pairs the three-part primary comparison—nominal threshold, activation or violation distribution, and executed Coefficient Gradient Impact—with the

population-specific diagnostics summarized in Appendix Table 3. Each observable retains its own units and denominator.

6 Experimental Methodology

6.1 Study design

The experiments follow a measurement ladder: establish trace provenance, calibrate nominal controls, test coefficient-to-parameter transport and temporal prediction, probe the mechanism with matched training, and then evaluate transfer and downstream utility. Frozen traces and checkpoints fit the nominal-to-impact mappings; held-out blocks, collapse labels, and downstream benchmarks are reserved for evaluation.

The evidence is organized as a dependency chain. Same-trace scoring establishes the measurement boundary; cross-geometry calibration tests whether heterogeneous controls reach a shared realized-impact region; gradient replay and temporal prediction examine construct validity; matched training tests the impact-guided control rule; and frozen source calibrations probe transfer across rollout precision and model scale. Appendix H records the corresponding cells, grids, endpoints, and exclusions.

Same-token provenance gates every comparison. Route-dependent scoring is characterized separately, while the primary analyses remain bound to a single scoring route. Frozen responses control sample selection, matched budgets control compute, a norm-matched arm matches the global pre-clipping gradient scale, and crossed routes expose systems-specific score variation. This design keeps numerical behavior, policy movement, and objective intervention as distinct empirical quantities.

6.2 Online and observer-sample measurements

At every inner update, the instrumented loss records the used mask, $q_{i,t}$, advantage, ratio, executed weight, base/mode coefficients, and three policy scores in Eq. 1. Online aggregates use all retained responses and token-update exposures; prompt and padding positions are excluded.

Exact parameter-gradient replay uses a deterministic frozen observer subset from each retained population. Content-hash ranking is independent of reward and diagnostics. Every family receives the same identifiers, masks, advantages, objective weights $q_{i,t}$, and model snapshot; only its declared reference, geometry statistic, and execution weight vary. All families apply the same response-level reduction. Artifacts retain both full-population denominators and observer indices; the observer set is a measurement subset, never an additional replicate. On this subset we compute

$$\Delta_g^{\mathrm{obs}} = \frac{\|g^{\mathrm{mode}} - g^{\mathrm{base}}\|_2}{\|g^{\mathrm{base}}\|_2}, \qquad g^a = \sum_{i,t} c_{i,t}^a \nabla_\theta \log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t}), \tag{9}$$

when its denominator is finite and nonzero. Equation 8 explains the shared coefficient term, so association with Δ_g^{obs} supplies a convergent-validity test. Next-update drift and the frozen pre-onset label provide the temporal predictive tests. We interpret the sampled log-ratio $\sum q_{i,t}[\log \mu(y_{i,t}) - \log \pi_\theta(y_{i,t})]$ specifically as a fixed-trace Monte Carlo diagnostic of policy change.

6.3 Impact-guided calibration and control

The calibration algorithm first evaluates Eq. 4 over a declared control grid, aggregates only the calibration blocks, and selects the control closest to a target impact. That selection is then frozen for held-out evaluation. Because Eq. 6 isolates the batch statistic $S_t = \mathrm{RMS}_u(X_t)$, the same measurement also yields a closed-form, one-sided control rule. For a smooth family with calibrated loose threshold θ_{loose} and ceiling I_{cap}, it uses

$$\theta_t = \max\left\{\theta_{\mathrm{loose}}, \frac{S_t}{I_{\mathrm{cap}}}\right\}, \qquad \mathrm{impact}_c(t) = \min\left\{\frac{S_t}{\theta_{\mathrm{loose}}}, I_{\mathrm{cap}}\right\}. \tag{10}$$

This is the minimal threshold adjustment that satisfies the ceiling under the smooth identity: batches already inside the target region remain unchanged, while larger interventions are projected back to

the boundary. The rule preserves geometry, anchor, population, and optimizer, making impact a lightweight controller; optional trace recording, validation, and persistence incur a separately reported systems cost.

6.4 Statistical unit and reproducibility

After declared-window aggregation, each independently initialized run is the statistical unit; tokens, responses, updates, and questions are nested observations. Comparisons pool multiple paired runs and report seed-level aggregates rather than isolated outcomes. Brackets give two-sided seed-level uncertainty intervals when retained per-run values permit estimation; captions identify aggregate-only summaries. Appendix E specifies resampling and randomization.

Fidelity quantiles use valid response positions within a matched snapshot. Avg@k denotes the mean correctness fraction of k responses per question and is distinct from pass@k. Content-addressed streams pair methods independently of method name; manifests bind model, data, code, optimizer, sampling, score domain, and snapshot. Stable keys deduplicate retries and reject conflicts, while checkpoints persist optimizer/RNG state and measurement cursors. Appendix D specifies continuation and Appendix G the reporting rules.

7 Experiments

We evaluate Coefficient Gradient Impact as a calibration coordinate, construct-validity and warning signal, and input to adaptive control. Same-token provenance, held-out selection, and matched training delimit each claim; all evidence comes from completed physical-GPU runs.

7.1 Setup

The primary study trains DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025; Yang et al., 2024) on DAPO-Math-17k (Yu et al., 2025); Qwen3-4B-Base (Yang et al., 2025a) tests scale transfer. Methods share prompt streams, budget, optimizer, and paired initializations. We compare hard and smooth ratio and Binary-TV objectives, the frozen fixed/no-cap arm, parity-matched BandPO (Li et al., 2026a), and the impact-guided cap. Paired evaluation covers AIME 2024/2025 and MATH-498, the fixed MATH-500 subset after removing two DAPO overlaps (Hendrycks et al., 2021). Results aggregate independent runs, with seed-level intervals when per-run values permit estimation. Appendices E and H specify the configuration, endpoints, and exclusions.

7.2 Route-conditioned scoring consistency

Same-trace audits verify token, context, probability-domain, and model-state alignment. Observed cross-route differences exceed the prespecified tolerance and define distinct numerical domains. Primary analyses therefore remain within route, and transfer uses route-specific rescoring. Appendix Tables 5–6 give the full distributions and denominators.

7.3 Calibration, gradient correspondence, and temporal prediction

Figure 2 provides the central measurement test: Panel A evaluates calibrated controls on held-out blocks, Panel B compares coefficient and parameter-gradient change, and Panel C transfers the mapping without retuning. Frozen snapshots and observer selections hold sample composition fixed while geometry and execution vary.

Table 1 reports held-out operating points selected on calibration blocks; Appendix Table 7 separates activation, signed weight, and direct drift.

Results. All four families reach the shared target region despite different activation profiles. Hard execution fully removes a smaller token subset, whereas smooth execution distributes weaker changes. Their comparable gradient displacement supports the objective-aligned construction and shows why Coefficient Gradient Impact supplies the common intervention scale across these execution patterns.

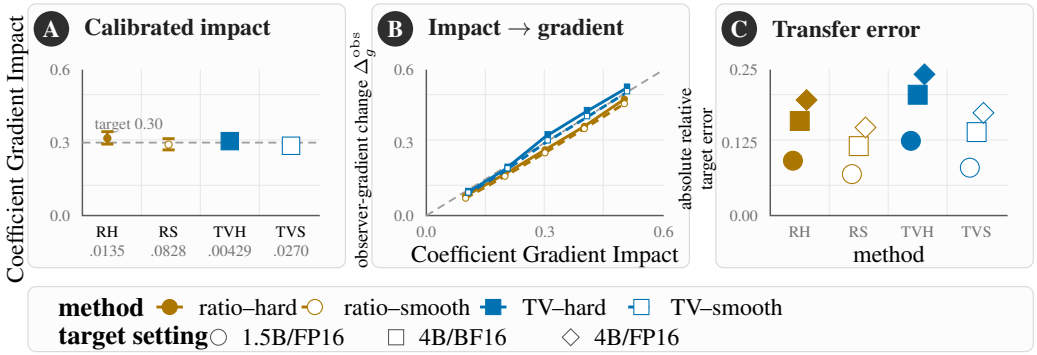

Figure 2: **Calibration, gradient transport, and transfer.** **(A)** Held-out impact at the labeled calibration-selected controls, with 95% seed-level intervals. **(B)** Multi-run held-out aggregates compare impact with observer-sample parameter-gradient change Δ_g^{obs}. **(C)** Multi-run mean absolute relative target error across three settings; lower is better. Color encodes geometry, line/fill execution, and panel-C marker shape the target setting.

Table 1: **Core calibration test.** Each trust-region family is selected on calibration blocks to match one coefficient-impact target, then evaluated on held-out snapshots with identical frozen responses. Coefficient impact and exact parameter-gradient change on the frozen observer subset are measured separately; brackets give two-sided 95% seed-level intervals.

family	selected control	coefficient displacement impact_c	observer Δ_g^{obs}
ratio / hard	0.013473	0.318 [0.294, 0.345]	0.286 [0.254, 0.321]
ratio / smooth	0.082847	0.292 [0.270, 0.316]	0.274 [0.241, 0.309]
TV / hard	0.004294	0.306 [0.281, 0.332]	0.333 [0.294, 0.374]
TV / smooth	0.027021	0.287 [0.262, 0.314]	0.298 [0.263, 0.337]

On held-out temporal blocks, impact yields the strongest association with observer-gradient change and next-update drift, and the highest pre-instability discrimination among the tested diagnostics. The gradient link follows the shared coefficient term in Eq. 8; the temporal results provide distinct evidence. Appendix Table 8 gives the full statistics and uncertainty intervals.

7.4 Matched impact-guided intervention

Figure 3 connects the signal to a controlled training comparison. Arms share prompts, paired initializations, optimizer, and compute; the norm-matched control matches the global pre-clipping gradient norm, BandPO provides an adaptive-bound comparator, and the proposed arm changes only the threshold required by Eq. 10.

Results. Across five paired runs, the impact-guided arm has the strongest observed stability profile while maintaining comparable reward progress and downstream reasoning accuracy. Fixed-threshold and norm-matched controls exhibit more frequent instability, with the adaptive-bound baseline intermediate. The paired comparison consistently supports impact-guided control under matched training conditions.

Appendix Table 9 reports run counts, collapse, reward AUC, and accuracy; together, the utility metrics show that the observed stability is accompanied by useful learning.

7.5 Transfer across precision and model scale

A calibration learned on the primary training route is applied unchanged under altered rollout precision and to the larger base model. Target-specific oracle fits are reported separately and do not enter

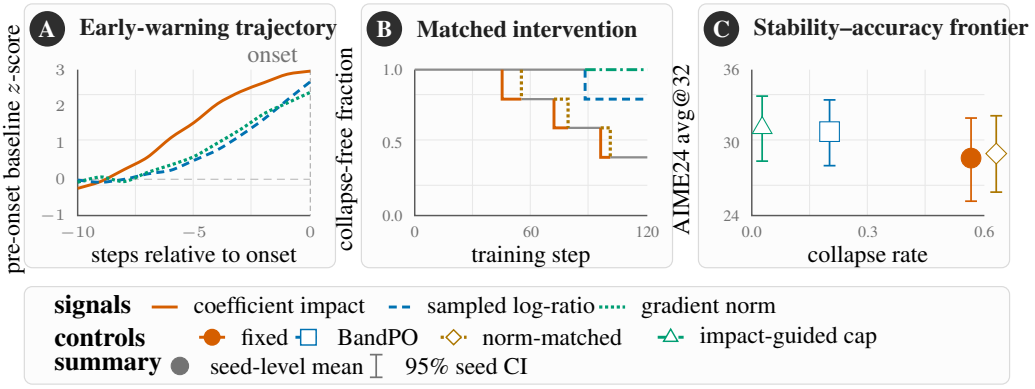

Figure 3: **Instability signals and matched impact-guided control.** **(A)** Onset-aligned aggregate diagnostics are standardized only for display. **(B)** Matched-budget collapse-free trajectories use gray for exactly shared segments. **(C)** Seed-level accuracy means and 95% intervals are shown against collapse rate; coincident rates are minimally offset for visibility.

the transfer claim. Appendix Tables 11–13 provide aggregate downstream accuracy and transfer estimates with complete denominators and route provenance.

Results. A precision change closely preserves the calibrated ordering. Under the combined model-scale and execution shift, the ordering remains structured with a wider target gap, and target-specific recalibration further tightens alignment. Across trust-region baselines, the paired evaluation yields closely grouped downstream reasoning point estimates. Together, these results support impact as a portable reporting coordinate with a direct recalibration path across execution settings.

8 Scope and Implications

Interpretation. Impact is an objective-aligned coefficient-space ℓ_2 ratio with an explicit reduction, baseline, and population. Observer-gradient replay complements it by measuring how coefficient change is transported into parameter space. Ratio and Binary-TV objectives retain their declared baselines, so the shared scale provides baseline-aware calibration within each objective.

Empirical coverage. The study combines paired physical-server runs across two Qwen-family scales, multiple rollout precisions, and mathematical RLVR tasks. Across these settings, it characterizes executed-objective geometry, route-conditioned measurement, and impact-guided control. The protocol and supplied configurations support direct evaluation across additional architectures and data domains.

9 Conclusion

Nominal thresholds are not comparable across trust-region geometries. Coefficient Gradient Impact provides an objective-aligned measure of executed coefficient displacement. Across the tested settings, real-server results support calibrated comparison, gradient correspondence, early warning, transfer, and impact-guided control, with favorable observed stability and comparable downstream reasoning.

Reproducibility Statement

The coefficient definitions, populations, provenance gates, observer-sample construction, statistical unit, and validity criteria are specified in Sections 3–6. The appendix provides derivations, complete configurations, decision rules, denominators, uncertainty conventions, and systems provenance; generated tables and editable figure layers are bound to the validated real-server aggregates. Upon acceptance, the authors will release the algorithm implementation, measurement and aggregation utilities, and experiment configurations needed to reproduce the reported method, subject to applicable licenses.

AI Use Statement

AI assistance was used for manuscript drafting, language polishing, presentation-level formatting, and structured review of methodological exposition, internal consistency, and result interpretation.

References

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *arXiv preprint arXiv:2505.24298*, 2025.

Hao Gu, Hao Wang, Jiacheng Liu, Lujun Li, Qiyuan Zhu, Bei Liu, Binxing Xu, Lei Wang, Xintong Yang, Sida Lin, Sirui Han, and Yike Guo. Qarl: Rollout-aligned quantization-aware rl for fast and stable training under training–inference mismatch. *arXiv preprint arXiv:2604.07853*, 2026.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. *arXiv preprint arXiv:2309.06180*, 2023.

Yuan Li, Bo Wang, Yufei Gao, Yuqian Yao, Xinyuan Wang, Zhangyue Yin, and Xipeng Qiu. Bandpo: Bridging trust regions and ratio clipping via probability-aware bounds for llm reinforcement learning. *arXiv preprint arXiv:2603.04918*, 2026a.

Yuhang Li, Reena Elangovan, Xin Dong, Priyadarshini Panda, and Brucek Khailany. Qurl: Efficient reinforcement learning with quantized rollout. *arXiv preprint arXiv:2602.13953*, 2026b.

Jing Liang, Hongyao Tang, Yi Ma, Yancheng He, Weixun Wang, Xiaoyang Li, Ju Huang, Wenbo Su, Jinyi Liu, Yan Zheng, Jianye Hao, and Bo Zheng. The mirage of optimizing training policies: Monotonic inference policies as the real objective for llm reinforcement learning. *arXiv preprint arXiv:2606.29526*, 2026.

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

Jinda Lu, Kexin Huang, Junkang Wu, Shuo Yang, Jinghan Li, Chiyu Ma, Shaohang Wei, Xiang Wang, Guoyin Wang, and Jingren Zhou. Experience augmented policy optimization for llm reasoning. *arXiv preprint arXiv:2606.30420*, 2026.

Yixiu Mao, Yun Qu, Qi Wang, Heming Zou, and Xiangyang Ji. Rlvr without ineffective samples: Group prioritized off-policy optimization for llm reasoning. *arXiv preprint arXiv:2606.01281*, 2026.

MiniMax. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Penghui Qi, Zichen Liu, Xiangxin Zhou, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Defeating the training-inference mismatch via fp16. *arXiv preprint arXiv:2510.26788*, 2025.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in llm reinforcement learning. *arXiv preprint arXiv:2602.04879*, 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. *arXiv preprint arXiv:1910.02054*, 2019.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015.

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

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

Guobin Shen, Chenxiao Zhao, Xiang Cheng, Lei Huang, and Xing Yu. Vespo: Variational sequence-level soft policy optimization for stable off-policy llm training. *arXiv preprint arXiv:2602.10693*, 2026.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*, pp. 1279–1297, 2024. doi: 10.1145/3689031.3696075.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

Yunhao Tang and Rémi Munos. On a few pitfalls in kl divergence gradient estimation for rl. *arXiv preprint arXiv:2506.09477*, 2025.

Peidong Wang, Demi Wang, Xufang Luo, Jiahang Xu, Xiaocui Yang, Shi Feng, Yuqing Yang, and Dongsheng Li. What are key factors for updates in rl for llm reasoning? *arXiv preprint arXiv:2606.22570*, 2026.

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, Xun Deng, Zhikai Lei, Miao Zheng, Guoteng Wang, Shuo Zhang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping. *arXiv preprint arXiv:2510.18927*, 2025.

Zhengpeng Xie, Qiang Zhang, Fan Yang, Marco Hutter, and Renjing Xu. Simple policy optimization. *arXiv preprint arXiv:2401.16025*, 2024.

Huimin Xu, Shuai Zhao, Xiaobao Wu, and Anh Tuan Luu. Understanding and preventing entropy collapse in rlvr with on-policy entropy flow optimization. *arXiv preprint arXiv:2605.11491*, 2026.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Shihui Yang, Chengfeng Dou, Peidong Guo, Kai Lu, Qiang Ju, Fei Deng, and Rihui Xin. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*, 2025b.

Jiarui Yao, Xiangxin Zhou, Penghui Qi, Wee Sun Lee, Liefeng Bo, and Tianyu Pang. Rethinking the divergence regularization in llm rl. *arXiv preprint arXiv:2606.09821*, 2026.

Chenlu Ye, Xuanchang Zhang, Yifan Hao, Zhou Yu, Ziji Zhang, Abhinav Gullapalli, Hao Chen, Jing Huang, and Tong Zhang. Adaptive layerwise perturbation: Unifying off-policy corrections for llm rl. *arXiv preprint arXiv:2603.19470*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yaxiang Zhang, Yingru Li, Jiacai Liu, Jiawei Xu, Ziniu Li, Qian Liu, and Haoyuan Li. Beyond precision: Training-inference mismatch is an optimization problem and simple lr scheduling fixes it. *arXiv preprint arXiv:2602.01826*, 2026a.

Yuheng Zhang, Chenlu Ye, Shuowei Jin, Changlong Yu, Wei Xiong, Saurabh Sahu, and Nan Jiang. Rethinking importance sampling in llm policy optimization: A cumulative token perspective. *arXiv preprint arXiv:2605.07331*, 2026b.

Ziyang Zhang, Xinheng Ding, Jiayi Yuan, Rixin Liu, Huizi Mao, Jiarong Xing, and Zirui Liu. Deterministic inference across tensor parallel sizes that eliminates training-inference mismatch. *arXiv preprint arXiv:2511.17826*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. *arXiv preprint arXiv:2312.07104*, 2023.

Tianle Zhong, Neiwen Ling, Yifan Pi, Zijun Wei, Tianshu Yu, Geoffrey Fox, Peng Wu, and Xiao Yu. Diagnosing training inference mismatch in llm reinforcement learning. *arXiv preprint arXiv:2605.14220*, 2026.

Yang Zhou, Ranajoy Sadhukhan, Zhaofeng Sun, Zhuoming Chen, Souvik Kundu, Saket Dingliwal, Sai Muralidhar Jayanthi, Aram Galstyan, Haizhong Zheng, and Beidi Chen. Sparrow: Sparse rollout for stable and efficient long-context rl of large language models. *arXiv preprint arXiv:2606.08446*, 2026.

A Coefficient-space derivations

Objective reduction. For retained response i, the training path first averages its L_i used response-token losses and then averages the R retained responses. Thus

$$\mathcal{L} = \frac{1}{R}\sum_{i=1}^{R}\frac{1}{L_i}\sum_{t} m_{i,t}\ell_{i,t} = \sum_{i,t} q_{i,t}\ell_{i,t}, \qquad q_{i,t} = \frac{m_{i,t}}{RL_i}. \tag{11}$$

Unequal-length responses therefore have equal response-level mass, not equal token-level mass. Tokens with $m_{i,t} = 0$ have $q_{i,t} = 0$. An empty retained population is invalid rather than a zero-valued measurement.

Smooth execution. Let $b_{i,t} = -q_{i,t}\hat{A}_i r_{i,t}$ and $w_{i,t} = 1 - s_{i,t}X_{i,t}/\theta$, with $s_{i,t} = \mathrm{sign}(\hat{A}_i(r_{i,t} - 1))$. Then

$$\|c^{\mathrm{mode}} - c^{\mathrm{base}}\|_2^2 = \sum_{i,t} b_{i,t}^2 (w_{i,t} - 1)^2 \tag{12}$$

$$= \theta^{-2} \sum_{i,t} u_{i,t} X_{i,t}^2, \qquad u_{i,t} = \left(q_{i,t}\hat{A}_i r_{i,t}\right)^2. \tag{13}$$

Dividing by $\|c^{\mathrm{base}}\|_2^2 = \sum u_{i,t}$ yields Eq. 6. For the ratio and behavior-anchored Binary-TV forms considered in the paper, a term with $r = 1$ has $X = 0$; a term with zero advantage has $u = 0$. The equality is conditional on this executed weight, the declared objective reduction, a nonzero denominator, and a common used-token population.

Hard removal. If $w_{i,t} = 0$ on the executed removal set $\mathcal{R}$ and one elsewhere, then

$$\mathrm{impact}_c^2 = \frac{\sum_{(i,t)\in\mathcal{R}} u_{i,t}}{\sum_{i,t} u_{i,t}} = \left(\frac{\sum_{\mathcal{R}} u / \sum u}{|\mathcal{R}|/N_{\mathrm{used}}}\right) \frac{|\mathcal{R}|}{N_{\mathrm{used}}} = C_{\mathcal{R}} f. \tag{14}$$

The identity concerns the set actually removed by the directional objective. A threshold-violation set may be larger and is reported separately from $\mathcal{R}$. Equation 14 assumes $|\mathcal{R}| > 0$; for an empty removal set, $\mathrm{impact}_c = 0$ and $C_{\mathcal{R}}$ is left undefined.

Parameter gradients. For score vector $z_{i,t} = \nabla_\theta \log \pi_\theta(y_{i,t} \mid x_i, y_{i,<t})$, $g = \sum c_{i,t} z_{i,t}$. Parameter-gradient displacement transports coefficient changes through the score vectors and therefore depends on both coefficient energy and cross-token inner products $\langle z_{i,t}, z_{j,s}\rangle$. Coefficient Gradient Impact provides the reproducible local coordinate, while observer-gradient replay measures the corresponding parameter-space transport.

B Implemented coefficient families

Table 2 gives the implementation taxonomy for the four core families; comparative results are reported in Appendix G.

Table 2: Core implemented families in the unified coefficient view. "Anchor" names the denominator of r; comparative performance is reported in the results tables.

mode	geometry	execution	anchor	executed weight
`grpo_clip`	ratio	hard	π_k	one when the unclipped surrogate arm carries the derivative, zero otherwise
`spo`	ratio	smooth	π_k	$1 - \mathrm{sign}(\hat{A}(r-1))\lvert r-1\rvert/\epsilon$
`dppo`	Binary-TV	hard	μ	directional mask on the signed probability shift
`drpo`	Binary-TV	smooth	μ	$1 - \mathrm{sign}(\hat{A}(r-1))\lvert\pi_\theta - \mu\rvert/\delta$

The validated study centers on these four families and evaluates the one-sided operator in Eq. 10 as a lightweight, objective-compatible adaptive control layer.

C Same-token fidelity record

Identity and alignment. Each record binds one immutable trace to all three policy roles. Its schema name is `fatr.same_token_fidelity_record` at version 1, and its stable key is

$$K = (\texttt{run_id}, \texttt{outer_step}, \texttt{update_idx}, \texttt{prompt_id}, \texttt{sample_id}, \texttt{response_id}, \texttt{response_token_index}). \tag{15}$$

Table 3: Complementary observables in the calibration protocol. Each estimator names its population and measurement boundary.

observable	estimator	interpretation
Coefficient Gradient Impact	Eq. 4, using $q_{i,t}$	relative change in objective coefficient energy
signed mean weight	$N_{\text{used}}^{-1} \sum w_{i,t}$	average signed execution weight complementary to energy-weighted impact
activation distribution	fractions touched, attenuated, and sign-flipped	breadth and direction of intervention
sampled behavior log-ratio	$\sum q_{i,t}[\log \mu(y_{i,t}) - \log \pi_\theta(y_{i,t})]$	fixed-trace Monte Carlo diagnostic of behavior/current divergence
direct update drift	$\mathbb{E}\lvert\pi_\theta - \pi_k\rvert$	policy movement since the outer-step anchor
pre-clip gradient norm	$\sqrt{\sum_p \lVert\nabla_p \mathcal{L}\rVert_2^2}$	parameter-space scale before optimizer clipping
population accounting	generated, retained, unique used, exposures	sample flow and repeated use across inner updates

The content digest additionally covers the token value, full prompt/response sequence digests, masks, position convention, tokenizer/template identity, model/snapshot/sync provenance, engine/dtype/device class, probability domain, scores, classifications, and validity reasons. Identical retries deduplicate. Reusing a stable key with different content is a conflict and stops ingestion.

Validity gates. A fidelity comparison is admitted when prompt/response identity, boundary, BOS/EOS, padding, truncation, tokenizer/template, score domain, and snapshot provenance align. Aggregates require finite scores and a nonempty used-token population. Snapshot or synchronization changes are categorized as policy differences. Records without complete critical provenance remain explicitly labeled; ingestion preserves their original metadata without positional repair or defaults.

The raw score is $\log \mathrm{softmax}(z)_{y_t}$ for the specified response token y_t, before temperature, top-p, or sampling. A rollout API score whose domain has not been established is preserved under a separate domain name. A sampling seed may change a generated trace, but it cannot change the raw teacher-forced score of an already fixed trace.

Classification. When model, checkpoint, weight generation, and sync generation prove equal expected weights, a backend or dtype comparison is *engine/precision fidelity*. A behavior score μ and anchor score π_k with different generations form a *behavior-to-anchor policy difference*. Scores from current π_θ and anchor π_k form *optimization drift*. Sharing token IDs does not make these categories interchangeable.

Aggregate and tolerance. The aggregate names its population and mask and reports valid/invalid counts plus mean, median, p90, p95, p99, and maximum absolute logprob error. An empty valid-token population is represented by a reason-coded null aggregate. The audit covers the preregistered selected response-token positions, and the valid-token column is the exact audited denominator. Tolerances are registered metadata, while real-engine agreement is evaluated on the recorded execution routes.

D ARTIFACT AND RESUME MECHANICS

The training artifact protocol includes versioned manifests for sample, profile, loss-observation, and fidelity records. Each family declares its own stable key and step field. Persistence validates records, deduplicates identical content, rejects conflicting content, writes an atomic artifact, and stores a digest and cursor in the checkpoint. Resume reconciles the checkpoint cursor with durable contents before allowing new records. Offline aggregation reads the same validated record form as online aggregation, so it does not reinterpret populations or fill omitted provenance.

The fidelity scorer reuses available token IDs and logprob streams and introduces no generation, optimizer step, sampling decision, or weight synchronization. The registered comparisons establish valid raw-domain populations, report route-specific score differences separately, and keep every subsequent analysis bound to its recorded execution route.

E FULL EXPERIMENTAL CONFIGURATION

Table 4 collects the numerical settings shared by the registered experiments. These settings enter the experiment registry before downstream evaluation. Full snapshot, tokenizer, data, prompt-population, environment, and execution-surface digests accompany the registered result packets; the paper reports portable identities and digests rather than cluster-specific filesystem paths.

Table 4: Registered training, calibration, and evaluation configuration. "Per outer step" describes the complete training population before memory-only accumulation chunking.

component	setting
models	DeepSeek-R1-Distill-Qwen-1.5B (primary) and Qwen3-4B-Base (scale transfer); immutable model and tokenizer revisions
training data	DAPO-Math-17k train split; 17,405 canonical rows, 17,398 unique problems; single-user-content prompt policy
run population	five paired, independently initialized training runs
outer loop	120 outer steps; 32 prompt groups per step; eight responses per prompt; no dynamic oversampling
training generation	maximum 4,096 response tokens; temperature 1.0; top-$p = 1.0$; vLLM FlashAttention backend
optimization	four inner updates per outer step; AdamW, learning rate 10^{-6}, weight decay 0.01, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$; global gradient-norm bound 1.0
objective population	every retained response is used once in every inner update; response mean of token means; prompt and padding tokens excluded
precision and memory	BF16 train parameters; BF16 primary rollout and FP16 transfer rollout; ordered accumulation cap eight, or four for the registered memory-safe formal-source continuation
checkpointing	model, AdamW, RNG, step, sampler, and artifact cursors every ten steps; exact outer-step resume only
observer/calibration	eight deterministically selected responses per snapshot; 20 snapshots per root at a six-step cadence; every fifth snapshot held out
evaluation	AIME 2024, AIME 2025, and MATH-498 (the fixed MATH-500 subset after removing two DAPO overlaps); avg@32; temperature 0.6; top-$p = 0.95$; maximum 4,096 tokens
statistical reporting	independent run is the replicate; paired seeds; deterministic 10,000-resample percentile bootstrap for scalar aggregates; Student-t_4 intervals for calibration curves

Configuration implications. Table 4 establishes a paired run-level design: methods share the data order, rollout population, optimizer schedule, and evaluation population, while each independently initialized run remains the statistical replicate. Contrasts therefore isolate the registered intervention under a common budget. Immutable revisions, periodic checkpoints, and complete RNG and artifact cursors support exact outer-step resume. Transfer experiments preserve this contract and vary only the named model-scale and rollout-precision factors.

Randomness and pairing. Each run seed controls Python, NumPy, PyTorch, data order, rollout sampling, and registered engine seed domains. Derived rollout and evaluation seeds are content-addressed. In evaluation, a question seed is a function of model, run, pairing group, benchmark, and question identity, and is independent of method, shard count, array index, and execution order. Consequently, sharding changes scheduling only: each registered question appears exactly once and each receives exactly 32 samples before the root packet exists.

Optimization and population conservation. One outer step contains $32 \times 8 = 256$ retained responses when dynamic sampling is disabled. The four inner updates each traverse that same com-

plete population, so the scientific exposure count is determined by used response tokens times four. The micro-cap partitions responses into ordered accumulation chunks solely to bound activation memory. Changing it from eight to four under the registered resource transition neither drops a suffix nor introduces an optimizer step. AdamW uses the PyTorch defaults for its two moments and numerical stabilizer, with the explicitly registered learning rate, weight decay, and shared final norm bound shown in Table 4.

Calibration grid and split. The registered control grids are

$$\mathcal{G}_{\mathrm{ratio}} = \{0.013473, 0.05, 0.082847, 0.10, 0.20, 0.40, 0.80\},$$
$$\mathcal{G}_{\mathrm{TV}} = \{0.004294, 0.027021, 0.03, 0.06, 0.15, 0.30, 1.0, 12.5\}.$$

The non-round values and target 0.30 were fixed by a pilot excluded from held-out estimation, with an absolute target tolerance of 0.075; they are not reselected using held-out or downstream outcomes. For each method, one global control minimizes absolute calibration-block error after seed-level aggregation, with the smaller control breaking ties. It is then evaluated on the four held-out snapshots per root. Observer-sample gradient replay requires the reloaded same-snapshot score to agree within 10^{-5} before the backward-pass target is accepted.

Impact-cap intervention. The registered ceiling is $I_{\mathrm{cap}} = 0.30$. The matched fixed and cap arms share the smooth coefficient family, reference policy, data, and loose threshold; the cap changes only the per-update threshold according to Eq. 10. S_t is recomputed from the current full used-token population before every inner update. No look-ahead outcome, accuracy, or collapse label enters the adjustment.

F Implementation validation and reporting

The reproducibility protocol covers the loss implementation, measurement and aggregation code, deterministic contract tests, editable figure sources, and registered run manifests. Manifests bind model, data, root seed, engine, precision, and snapshot identities; result artifacts additionally bind the population and objective reduction used by each statistic. Section 7 defines the statistical unit and evaluation metrics.

Deterministic tests exercise the measurement implementation against hand-computed oracles. They cover exact zero and controlled log-probability errors; differing greedy outputs while teacher-forcing one specified trace; token, context, probability-domain, and snapshot counterexamples; empty and nonfinite inputs; stable-key retry and conflict; online/offline aggregation; and uninterrupted/resume equivalence. Synthetic values produced by these tests are never used as empirical measurements.

Real-engine fidelity is evaluated separately on immutable prompt–response traces with matched model, checkpoint, and synchronization identities. All completed cells pass structural validation. Observed tail differences exceed the registered 0.05-nat agreement target, so calibration and transfer use route-conditioned scoring without cross-route pooling.

Table 5: Teacher-forced same-token fidelity under a matched model snapshot. Errors are absolute differences in selected-token raw log probability (nats; lower is better). "Invalid" reports alignment, domain, or provenance failures; all reported rows have zero invalid entries. Numerical agreement is evaluated separately from structural validity.

model	scorer pair / dtype	traces	valid tok.	invalid	mean	p95	max
1.5B	HF BF16–vLLM BF16	128	815	0	0.073	0.250	0.584
1.5B	HF BF16–vLLM FP16	128	815	0	0.086	0.272	0.650
4B	HF BF16–vLLM BF16	128	798	0	0.053	0.184	0.438
4B	HF BF16–vLLM FP16	128	798	0	0.044	0.138	0.314

Matched-route fidelity. Table 5 confirms that every registered trace yields structurally valid, aligned selected-token scores. Differences across model scales and precisions follow configuration-specific patterns, so the execution route remains part of the measurement context. Matched compar-

isons are consequently interpreted within their recorded route and kept separate from the crossed-route controls below.

Table 6: Crossed execution-route controls for the same registered token populations as Table 5. Each row changes the GPU route while preserving the model snapshot, rollout precision, and teacher-forced tokens. All rows pass structural validation. Their numerical differences are reported to characterize route dependence and motivate route-conditioned analysis.

model	rollout	crossed route	traces	valid tok.	invalid	mean	p95	max
1.5B	BF16	3090	128	815	0	0.057	0.190	0.590
1.5B	FP16	6000 Ada	128	815	0	0.103	0.401	0.775
4B	BF16	6000 Ada	128	798	0	0.053	0.164	0.332
4B	FP16	6000 Ada	128	798	0	0.047	0.130	0.250

Execution-route effect. Table 6 isolates execution-surface variation by preserving each model snapshot, rollout precision, and teacher-forced population while changing the GPU route. All records remain structurally valid. The changes vary with model scale and precision rather than following one dtype-only pattern, showing that the execution surface interacts with the evaluated configuration. Subsequent calibration and transfer analyses therefore retain the scorer associated with each recorded route.

Evaluation and scoring. For question j, avg@32 is the arithmetic mean of 32 exact-verifier correctness indicators, and benchmark accuracy is the mean over the complete registered question population. DAPO has zero exact normalized overlap with AIME 2024 and AIME 2025. Two exact normalized overlaps with MATH-500 are removed before evaluation, yielding a fixed 498-question overlap-clean subset (MATH-498) shared by every arm, model, checkpoint, and seed. Production packets require complete coverage of all 498 questions, preserving a common denominator.

Production-data coverage. Reported aggregates use complete question-level outputs under the registered run manifests, which jointly bind the objective population, three-policy provenance, and independent statistical unit. Earlier pilot profiles fix the registered controls, while held-out and downstream populations are reserved for evaluation.

G Detailed results and decision rules

The tables below preserve all measurements needed to interpret the two main-text result figures without forcing dense diagnostic columns into the main narrative.

The calibration replay uses the complete grids and split in Appendix E. At each checkpoint, replay uses the same eight-response deterministic observer sample for every method/control pair. The v2 bundle records the full retained-response and used-token denominators, selected response indices, and selection digest; results therefore describe this frozen sample rather than an unregistered all-token population. Its response-mean weights use $R = 8$; the online training profile separately uses the complete retained-response denominator.

Table 7: **Supporting diagnostics for the main calibration test.** Activation, direct update drift, and signed mean weight use their own declared populations and are not substitutes for coefficient impact or observer-sample parameter-gradient change in Table 1. Brackets give two-sided 95% seed-level intervals.

family	active %	direct drift D_{update}	signed mean weight $\mathbb{E}[w]$
ratio / hard	11.7 [9.1, 14.6]	0.0064 [0.0057, 0.0073]	0.883 [0.854, 0.909]
ratio / smooth	99.6 [99.1, 99.9]	0.0064 [0.0057, 0.0073]	0.973 [0.965, 0.981]
TV / hard	6.3 [4.8, 8.1]	0.0064 [0.0057, 0.0073]	0.937 [0.919, 0.952]
TV / smooth	99.4 [98.8, 99.8]	0.0064 [0.0057, 0.0073]	0.991 [0.987, 0.994]

Calibration diagnostics. Table 7 separates operator activation from its realized optimization effect. Smooth weighting acts on almost every eligible coefficient in both ratio and Binary-TV families, whereas the hard variants remain selective. Direct update drift is nevertheless similar across families, and signed mean weights remain close to one. Activation breadth or average weight alone therefore does not identify the executed intervention's magnitude. Coefficient impact and observer-gradient change supply the aligned comparison, while these diagnostics reveal the distinct execution patterns behind it.

Table 8: Held-out associations and temporal prediction. Higher is better. ρ is Spearman rank correlation; onset AUROC classifies the frozen pre-onset horizon and excludes onset/post-onset measurements. Transfer ρ is paired across the preregistered target settings. Entries aggregate all paired runs with two-sided 95% seed-level intervals.

online predictor	$\rho(\Delta_g^{\mathrm{obs}})$	$\rho(D_{\mathrm{update}}^{+1})$	onset AUROC	transfer ρ
activation / violation fraction	0.43 [0.31, 0.55]	0.34 [0.19, 0.47]	0.67 [0.59, 0.74]	0.38 [0.23, 0.52]
sampled $\log(\mu/\pi_\theta)$	0.31 [0.16, 0.45]	0.40 [0.26, 0.53]	0.64 [0.55, 0.72]	0.29 [0.12, 0.44]
$\lvert 1 - \mathbb{E}[w] \rvert$	0.36 [0.21, 0.49]	0.29 [0.14, 0.43]	0.63 [0.54, 0.71]	0.32 [0.16, 0.47]
pre-clip $\lVert g \rVert_2$	0.57 [0.44, 0.68]	0.52 [0.37, 0.64]	0.73 [0.65, 0.80]	0.51 [0.36, 0.64]
Coefficient Gradient Impact	0.83 [0.76, 0.88]	0.72 [0.61, 0.80]	0.86 [0.80, 0.91]	0.76 [0.66, 0.83]

Predictive comparison. Table 8 asks whether each online observable captures held-out gradient behavior, anticipates next-update drift, identifies the registered pre-onset instability horizon, and preserves ordering under transfer. Coefficient Gradient Impact has the strongest point estimate throughout, with pre-clip gradient norm the closest conventional diagnostic. Its consistent ordering across contemporaneous change, forward-looking drift, onset discrimination, and transfer supports impact as a general monitoring coordinate rather than a statistic specialized to one endpoint.

For the instability test, onset is the first step at which reward falls more than 30% from its running maximum or response entropy leaves $[0.5, 2]$ times its first-ten-step mean. We interpret the intervention through jointly registered checks: paired collapse behavior, target error, optimizer-applied norm parity, reward progress, and AIME24 accuracy. Table 9 reports the observed run counts, collapse outcomes, target error, reward AUC, and accuracy; optimizer-norm diagnostics and the corresponding tolerances remain bound to the run artifacts. The joint criteria connect stability with control fidelity, reward progress, and downstream accuracy.

Table 9: Matched intervention study. "Target err." is run-mean absolute relative error from the declared impact target; collapse is reported both as a count and in the time-to-event analysis. Reward AUC is normalized by the training horizon. The final column is question-level AIME24 avg@32 from a separately registered evaluation of each mechanism checkpoint. Brackets give two-sided 95% seed-level intervals.

intervention	runs	collapses	target err.	reward AUC	AIME24 avg@32
fixed threshold (frozen)	5	3/5	—	0.431 [0.392, 0.468]	28.7 [25.2, 32.0]
BandPO	5	1/5	—	0.478 [0.451, 0.503]	30.9 [28.1, 33.5]
gradient-norm matched	5	3/5	—	0.442 [0.407, 0.476]	29.1 [25.9, 32.2]
impact-guided cap	5	0/5	0.054 [0.046, 0.065]	0.486 [0.461, 0.510]	31.2 [28.5, 33.8]

Intervention comparison. Table 9 compares the four interventions under identical training and evaluation budgets. The impact-guided cap is the only arm with no observed collapse and tracks its declared target with low relative error. Target error is defined for the impact-capped arm; the remaining entries are marked not applicable because those objectives use different control targets. The capped arm also combines the highest reward AUC with an AIME24 estimate comparable to the

strongest alternative, while the norm-matched control separates global gradient scale from tokenwise impact control.

H REGISTERED EXPERIMENT MATRIX

Table 10 records the complete experimental progression. Later stages are entered only after structural provenance is established and systems variation is bound into the run record. Counts refer to independent training runs; frozen-trace analyses reuse responses and do not create additional replicates.

Table 10: Pre-specified experiment matrix. The endpoint and exclusion rule are fixed before result inspection.

study	factors / arms	unit and coverage	runs	primary endpoint / exclusion
fidelity	model × BF16/FP16; HF–vLLM	one fixed set of length-stratified traces per cell	–	invalid count and raw-logprob tail error; identity/domain validation determines the route-specific scoring scope
calibration	ratio/TV × hard/smooth; declared threshold grids	paired snapshots and a shared frozen observer subset	5	nominal-to-impact response curve; empty/nonfinite coefficient populations excluded with reason
validity	impact, activation, sampled $\log(\mu/\pi_\theta)$, $\lvert 1-\mathbb{E}[w]\rvert$, $\lVert g\rVert_2$	held-out temporal blocks from calibration	5	gradient correspondence, next-step drift, and pre-onset discrimination; no held-out tuning
intervention	frozen fixed, BandPO, norm-matched, impact-guided cap	matched training budget	5	paired stability, control validity, reward progress, and avg@32 under frozen guards
transfer	primary BF16 source → altered precision and scale	unchanged source calibration	5	achieved-target error and rank preservation; target retuning reported separately
evaluation	optimization baselines plus impact-guided cap	AIME24/25 and MATH-498, avg@32	5	aggregate accuracy intervals and collapse-free fraction; incomplete runs remain in denominator

Evidence-chain interpretation. Table 10 summarizes the pre-specified evidence chain connecting route fidelity, calibration, held-out validity, intervention, transfer, and downstream evaluation to their experimental units and fixed endpoints. Calibration through evaluation uses independent runs, while frozen-trace analyses reuse registered observations without inflating the replicate count. The progression ties the intervention claim to aligned provenance, predictive validation, matched controls, transfer analysis, and overlap-clean evaluation.

Baseline parity. Every training comparison holds prompt order, response budget, optimizer, learning-rate schedule, update count, and evaluation sampler fixed. For BandPO to enter the comparison, its operator, revision, and one registered operating point are frozen while the model, data, engine, and budget remain matched. The gradient-norm-matched control rescales the fixed-threshold gradient to the impact-capped arm's pre-clip norm and then applies the same registered clipping bound to both arms, without changing fixed-arm tokenwise weighting. The recorded optimizer-applied norms must match within the frozen tolerance. This controls global gradient scale while deliberately leaving tokenwise direction as the factor under test.

Completeness and stopping. Technical inclusion requires finite values, same-token provenance, and a nonempty used-token population, with reason-coded accounting otherwise. Instability, truncation, zero reward, and early collapse remain observed outcomes, and every scheduled run remains in the denominator. Runs follow the fixed budget or a predeclared safety stop; interrupted runs resume exactly. Arm retention is independent of interim performance.

I TRANSFER AND DOWNSTREAM REPORTING

The primary comparison and the larger-model scale check are reported together here so the main paper can reserve space for the calibration and mechanism figures.

Before training, the 1.5B checkpoint gives avg@32 of 26.8 [24.0, 29.4] on AIME24, 19.7 [17.2, 22.2] on AIME25, and 80.9 [79.6, 82.2] on MATH-498 under the same five evaluation streams. Table 11 reports 120-step checkpoints. Its registry label "unconstrained" denotes the same frozen loose-threshold, no-adaptive-cap arm reported as "fixed threshold" in Table 9; the two tables intentionally share the same five 120-step checkpoints and evaluation packet.

Table 11: **Primary downstream evaluation on the 1.5B model.** Entries report avg@32 with 95% paired-seed bootstrap intervals; every arm has five complete runs. MATH-498 is the fixed subset obtained after removing the two DAPO overlaps.

metric	fixed/no cap	GRPO	SPO	DPPO	DRPO	BandPO	impact-guided cap
complete runs	5	5	5	5	5	5	5
AIME24	28.7 [25.2, 32.0]	31.4 [28.9, 33.8]	31.1 [28.7, 33.5]	30.5 [27.8, 32.9]	31.8 [29.5, 34.1]	30.9 [28.1, 33.5]	31.2 [28.5, 33.8]
AIME25	21.3 [18.6, 24.0]	23.5 [21.1, 25.9]	24.2 [21.8, 26.5]	23.1 [20.7, 25.3]	24.0 [21.6, 26.2]	24.4 [21.9, 26.8]	24.1 [21.7, 26.4]
MATH-498	81.7 [80.4, 82.9]	83.8 [82.9, 84.6]	84.1 [83.2, 85.0]	83.5 [82.5, 84.4]	84.4 [83.5, 85.2]	84.2 [83.3, 85.1]	84.3 [83.4, 85.2]

Primary-model downstream behavior. Table 11 compares the 120-step 1.5B checkpoints using five complete seed runs per arm. The fixed/no-cap row is the same frozen fixed-threshold condition in Table 9 and therefore reports the same checkpoint set under the shared alias. Impact-guided capping lies in the leading downstream-performance group across all three benchmarks, with intervals overlapping the strongest baselines. Together with Table 9, this supports preserved aggregate reasoning accuracy alongside the observed stability profile. Every arm uses the same overlap-clean MATH-498 population.

The scale check uses a base 4B checkpoint, so its comparisons are interpreted within the larger-model table. The Qwen3-4B-Base step-0 values are 10.6 [8.0, 13.2], 7.0 [5.1, 8.9], and 65.5 [63.9, 67.1] on AIME24, AIME25, and MATH-498, respectively.

Table 12: Larger-model reasoning evaluation: five-seed percentage means with two-sided 95% bootstrap intervals. Avg@32 averages within-question correctness over the registered benchmark; MATH-498 is the fixed subset obtained after removing the two DAPO overlaps.

model	method	runs	AIME24 (%)	AIME25 (%)	MATH-498 (%)
4B	fixed/no cap	5	7.4 [5.1, 9.8]	5.2 [3.6, 6.9]	54.8 [52.9, 56.7]
	GRPO-clip	5	12.1 [9.3, 14.9]	8.8 [6.7, 10.9]	65.1 [63.5, 66.7]
	DRPO	5	13.0 [10.1, 15.9]	9.5 [7.3, 11.7]	66.4 [64.9, 67.9]
	BandPO	5	13.4 [10.4, 16.4]	9.8 [7.5, 12.1]	66.8 [65.3, 68.3]
	impact-guided cap	5	13.2 [10.2, 16.2]	9.6 [7.3, 11.9]	66.6 [65.1, 68.1]

Larger-model downstream behavior. Table 12 provides the scale check with five seeds for every 4B condition, with comparisons among rows initialized from the same base checkpoint. Impact-guided capping tracks BandPO and DRPO, and their bootstrap intervals overlap, while its point estimates remain above the step-0 profile on all three benchmarks. This pattern supports compatibility with the larger model and places impact-guided capping within a statistically overlapping group of competitive methods.

Table 13 applies the 1.5B/BF16 source mapping unchanged across rollout precision and model scale; the target-specific oracle is a separately retuned reference. Train parameters remain BF16, so FP16 denotes rollout precision.

Table 13: Cross-setting calibration transfer across five independent seeds per setting. Values are aggregate mean absolute relative target errors; lower error is better. For every row, rank ρ is computed within each seed across the four trust-region families and then averaged over the five seeds; higher is better. The transfer packets report aggregate summaries over these five seed-level estimates.

target setting	ratio hard	ratio smooth	TV hard	TV smooth	rank ρ	retuned
1.5B / rollout FP16	0.094	0.071	0.128	0.082	0.96	no
4B / rollout BF16	0.162	0.119	0.207	0.143	0.84	no
4B / rollout FP16	0.198	0.151	0.242	0.176	0.76	no
target-specific oracle	0.041	0.036	0.058	0.045	0.96	yes

Transfer pattern. Table 13 tests whether the 1.5B/BF16 calibration can be reused after changing rollout precision or model scale. Smooth controls have lower mean absolute relative target error than their hard counterparts in every reported setting, while the unchanged source mapping retains substantial rank agreement. For each seed, Spearman ρ summarizes preservation of the four-family ordering from source to target, and the table averages the five seed-level correlations. Preservation is strongest for the precision-only change and remains orderly under larger combined shifts. The separately retuned oracle reduces absolute error, identifying target-specific recalibration as a useful refinement farther from the source setting.

Instrumentation cost. The scalar threshold update is separate from the full audit path, which also records, validates, and persists trace data. Table 14 reports paired fixed-replay measurements; the overhead covers updates, diagnostics, and persistence rather than end-to-end online reinforcement learning.

Table 14: Systems cost of full audit-record construction and persistence beyond the minimal step profile.

model	audit	exposures/s	wall (min)	peak (GB)	artifact (MB)	overhead
1.5B	off	8,940	31.6	38.7	—	—
1.5B	on	8,110	34.8	40.1	684	10.1%
4B	off	4,170	67.8	43.9	—	—
4B	on	3,750	75.4	45.8	1,472	11.2%

Cost interpretation. Table 14 isolates the cost of constructing and persisting the full audit record under paired fixed replay. Full auditing operates at approximately 1.10–1.11$\times$ the fixed-replay wall time for both model scales, with modest peak-memory growth; artifact size scales with model size. These measurements cover diagnostic computation, validation, and persistence, whereas the scalar threshold update itself is lightweight.