A MATCHED-ERROR SENSITIVITY AUDIT OF GROUP-RELATIVE POLICY UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Marginal verifier error rates leave unspecified how errors perturb a group-relative policy update. We measure this sensitivity on fixed responses using mean-token log-probability gradients in specified LoRA coordinates. An exact decomposition compares the observed distortion with two references: one preserves error counts within each prompt, and the other preserves only the global false-positive and false-negative totals. This separates response-level error association from allocation across prompts. On a frozen 12,288-generation LiveCodeBench cohort, three-test observed-to-global-expectation ratios are 0.605 [0.504, 0.723] for Llama and 0.342 [0.225, 0.510] for Qwen Instruct, using validated labels and full-response FP32 Grams. Within-prompt association is negative for Llama and uncertain for Qwen; one-test ratio intervals include one for both policies. A retrospective APPS audit shows related allocation effects, with sparse, cap-limited Thinking evidence. Diagonal Grams preserve much of the distortion ranking, and tested full-matrix features do not establish a stable selection advantage. A restricted eight-action training experiment favors prompt-correlated noise, but free-form follow-ups and a frozen 48-run GSM8K study do not confirm that benefit. The audit identifies what a matched-error contrast measures while separating local distortion, audit-selection utility, and downstream training.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) replaces human preference labels with programmatic checks, such as final-answer matching or unit tests. Group Relative Policy Optimization (GRPO) and related methods compare several responses to the same prompt and center or standardize their rewards within that group (Shao et al., 2024). This creates an immediate mismatch between how verifier quality is commonly reported and how the optimizer consumes it. False-positive rate (FPR) and false-negative rate (FNR) are sample-level marginals; a group-relative update also depends on whether errors are isolated across prompts or concentrated within one prompt.

Recent analyses show that independent label noise can sometimes be corrected (El Mansouri et al., 2025; Cai et al., 2025), while content-dependent verifier errors can produce delay, plateaus, or collapse despite similar initial error rates (Egashira et al., 2026). We study a different and narrower question: if clean labels and the exact global FP/FN counts are held fixed, how much does the allocation of those errors across prompt groups change the update?

We distinguish two questions that require different evidence. The first asks whether an observed error allocation gives a different local update from other allocations with the same counts. The second asks whether changing a noise channel improves subsequent clean performance. Our APPS audit supports a local sensitivity result; our larger free-form experiments do not confirm the proposed training benefit. The broader claim that error rates alone are insufficient is already established by systematic-error studies. We therefore focus on a reproducible comparison that holds error counts fixed and evaluates the actual tangent vectors used in the audit.

Our contributions are:

- a matched-error audit with an explicit Gram-matrix estimand, fixed clean labels and FP/FN counts, and a clearly delimited permutation reference;

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.

- retrospective decomposition on three APPS policies and a frozen LiveCodeBench cohort, with numerical, evaluator and label-sensitivity checks; and
- a separation of homogeneous-group coefficient activation from training utility, including a restricted positive mechanism experiment and the negative free-form follow-ups and 48-run GSM8K confirmation.

The audit was motivated by an unsuccessful selective-auditing pilot. The inverse local-distortion hypothesis was specified before the two confirmatory policies; we retain the failed pilots and online follow-ups rather than treating the local result as evidence that correlation is beneficial in general.

2 Error geometry in group-relative updates

For prompt x, let K sampled responses have clean binary rewards $r^* \in \{0,1\}^K$ and verifier rewards $\widetilde{r}$. For a response of length L_i, the audit uses the length-normalized tangent vector

$$s_i = \nabla_\theta \left[\frac{1}{L_i} \sum_{t=1}^{L_i} \log \pi_\theta(y_{it} \mid x, y_{i,<t}) \right],$$

with sampled tokens and lengths held fixed and θ restricted to the specified LoRA coordinates, using the stated sampling distribution. This is not the unnormalized sequence score. With population-standard-deviation normalization, the corresponding local reward-surrogate group gradient is

$$g(r) = \frac{1}{K} \sum_{i=1}^{K} \frac{r_i - \overline{r}}{\mathrm{std}(r)} s_i, \tag{1}$$

where a zero-variance group contributes zero. We measure $\|g(r^*) - g(\widetilde{r})\|_2^2$ exactly within a fixed LoRA tangent space using the $K \times K$ score Gram matrix. Writing $A(r)$ for the normalized reward vector and $G_{ij} = s_i^\top s_j$, the per-group target is

$$D_x = \frac{1}{K^2} \Delta A^\top G \Delta A, \qquad \Delta A = A(r^*) - A(\widetilde{r}).$$

We average D_x over prompts, rather than take the squared norm of the batch-averaged gradient; cross-prompt cancellation is outside this estimand. This target does not model clipping, Adam state, or later policy shift.

Proposition 1 (Spurious activation of a homogeneous-negative group)**.** *Suppose $r_i^* = 0$ for all i and exactly m verifier labels are false positives. If $0 < m < K$, the normalized coefficients equal*

$$A_i = \begin{cases} \sqrt{(K-m)/m}, & \widetilde{r}_i = 1, \\ -\sqrt{m/(K-m)}, & \widetilde{r}_i = 0. \end{cases}$$

The coefficient vector is nonzero and has squared norm K. If $m = 0$ or $m = K$, the group remains degenerate and its update is zero.

Proof. The corrupted reward mean is m/K and its population standard deviation is $\sqrt{m(K-m)}/K$. Substitution gives the two displayed coefficient values, and their squared norm is $m(K-m)/m + (K-m)m/(K-m) = K$. At either endpoint the reward is constant, so our zero-variance convention assigns the group a zero update. □

Thus independent false positives activate an all-negative group with probability $1 - (1-p)^K - p^K$. A perfectly group-correlated channel that flips either zero or all K labels has the same marginal FPR p but activation probability zero. This is not an unconditional advantage for correlation: on a clean mixed group, clustered errors can erase useful contrasts. The net effect depends on clean group composition, the conditional distribution of the within-group error count, and score geometry. In particular, a nonzero coefficient vector need not yield a nonzero parameter gradient: if all s_i are equal, centering makes the weighted sum vanish. For mean-centered rewards without standardization, the analogous coefficient squared norm is $m(K-m)/K$, also zero at the endpoints. These identities explain coefficient activation, not a universal ordering of D_x.

2.1 SEPARATING ASSOCIATION FROM ALLOCATION

A global error permutation changes both the number of errors assigned to each prompt and their association with response gradients inside that prompt. Let R preserve global FP/FN counts and let U_c preserve the counts within each prompt and true-label class. For expectations, the exact decomposition is

$$D_{\text{obs}} - \mathbb{E}_R D = \underbrace{D_{\text{obs}} - \mathbb{E}_{U_c} D}_{\text{within-prompt association}} + \underbrace{\mathbb{E}_{U_c} D - \mathbb{E}_R D}_{\text{between-prompt allocation}} . \tag{2}$$

The first term holds prompt-level counts fixed; the second changes that allocation. Neither term is a causal training effect. We compute both expectations exactly on the cached eight-response groups, including constant-reward groups, rather than substituting permutation medians. Appendix B gives the calculation, exhaustive checks and unnormalized results.

3 MATCHED-ERROR AUDIT

Data and policies. We select 512 distinct problems from the APPS training partition (Hendrycks et al., 2021), each with 10–100 tests and an executable reference solution. For each prompt, each policy samples eight Python solutions at temperature 1.0, top-p 1.0, and a 4,096-token completion cap. The policies are Qwen3-4B-Instruct, Qwen3-4B-Thinking (Yang et al., 2025), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). Full-suite execution supplies the reference label. Executing one or three deterministically shuffled tests supplies nested proxy labels. Candidate programs run without network access in a bounded sandbox. In this dataset the proxies produce false positives but no false negatives. Length limits affect 448/4,096 Instruct responses, 3,433/4,096 Thinking responses (83.8%), and 5/4,096 Llama responses. These records remain in the analysis. In particular, the Thinking audit describes a severely cap-limited regime, not unrestricted reasoning performance.

Gradient geometry. For every response we differentiate its mean-token log probability with respect to a fixed rank-16 LoRA parameterization of the query and value projections. We retain each group's score Gram matrix, which suffices for the exact norm and cosine of every linear combination used here. This is a fixed-response, local surrogate quantity; the numerical backend and sampling distribution are specified separately. We define the spurious-signal fraction as the fraction of clean-homogeneous prompt groups whose proxy labels become nonconstant, and hence induce a nonzero normalized coefficient vector. The label-correction audit is documented in Appendix A.

Matched null. Within each clean reward class, we permute the observed error indicator over all 4,096 rollout positions. Every permutation preserves the exact FP and FN totals and the full clean reward array while destroying prompt-level clustering. The primary analysis uses exact expectations and a joint prompt bootstrap (Section 5). Qwen3-4B-Instruct was exploratory; the inverse hypothesis and gate were then frozen before generating either confirmatory policy. This reference changes both prompt allocation and the association of errors with individual response tangent vectors. The comparison characterizes the specified random-allocation reference; natural verifier errors are not assumed exchangeable, so these are not causal tests of changing correlation alone.

4 RESULTS

What the global contrast identifies. Figure 1 shows that the negative between-prompt allocation term can dominate a positive within-prompt association term. In the three-test Qwen Thinking data, 507/512 prompts have zero observed distortion; the retained zero groups and their clean-label composition are part of this estimand. The current analysis uses repaired labels and retains all prompt groups. A candidate-dependent single-test ambiguity is evaluated at both observed label states; it does not change the three-test endpoint. Coordinate sensitivity and training transfer remain separate questions.

APPS exact-expectation results. For the three-test proxy, observed/global-expectation ratios are 0.171 for Qwen Thinking (95% interval $[0.056, 0.434]$) and 0.637 for Llama ($[0.453, 0.893]$). These primary point estimates divide the observed prompt mean by the exact finite-population reference

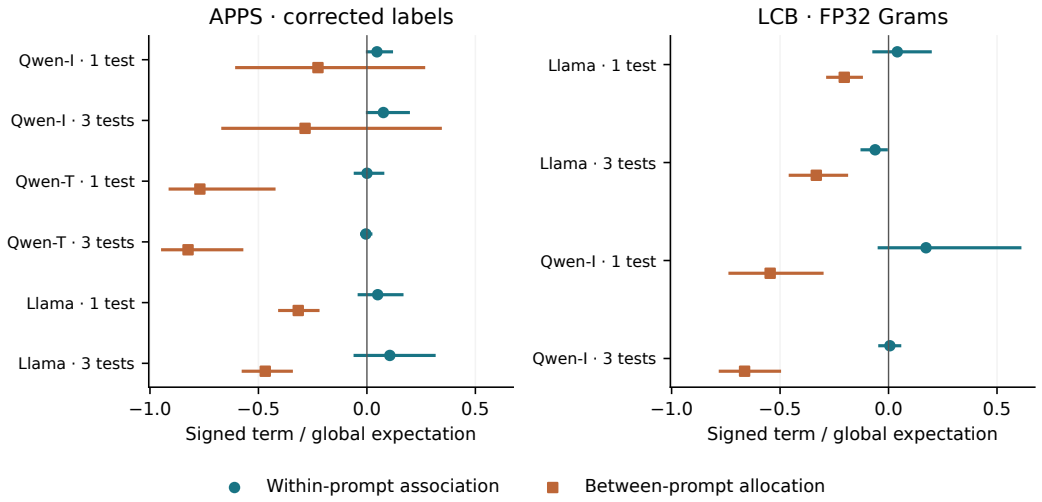


Figure 1: Retrospective exact-expectation decomposition of all 512 prompts per policy and both proxy strengths. Points show signed within and between terms divided by the cell's exact global-reference expectation; bars use shared-prompt bootstrap intervals, recomputing numerator and denominator in every resample and conditioning on a defined denominator. Qwen-I/T denote Instruct/Thinking. APPS uses corrected labels with archived Grams; its displayed one-test ambiguous label is zero (both states are in Appendix A), and Thinking is severely cap-limited; LCB uses repaired labels and FP32 Grams. Normalization does not imply parameterization invariance or training utility.

Table 1: Primary three-test audit: observed distortion divided by its exact global-reference expectation. APPS uses corrected labels with archived mixed-precision Grams; LCB uses full-response FP32 Grams. The cap-limited APPS Thinking cell has only five nonzero-distortion prompts. Ratio intervals condition on a defined denominator (27 undefined Thinking resamples); unscaled contrasts retain every resample. Intervals use 10,000 shared-prompt bootstrap samples with the global reference recomputed in every sample. The signed within term is in raw coordinate-dependent squared-gradient units.

Cohort	Policy	Ratio	Ratio 95% interval	Within-term 95% interval
APPS	Qwen Instruct	0.791	$[0.392, 1.425]$	$[1.80 \times 10^{-5}, 1.40 \times 10^{-3}]$
APPS	Qwen Thinking	0.171	$[0.056, 0.434]$	$[-1.00 \times 10^{-6}, 3.60 \times 10^{-7}]$
APPS	Llama Instruct	0.637	$[0.453, 0.893]$	$[-5.39 \times 10^{-4}, 3.19 \times 10^{-3}]$
LCB	Llama Instruct	0.605	$[0.504, 0.723]$	$[-9.47 \times 10^{-4}, -7.86 \times 10^{-5}]$
LCB	Qwen Instruct	0.342	$[0.225, 0.510]$	$[-1.93 \times 10^{-4}, 2.40 \times 10^{-4}]$

mean. They use the corrected labels throughout; superseded permutation summaries are not mixed with these estimates (Appendix A). Qwen Instruct's exploratory point ratio is 0.791, but its interval $[0.392, 1.425]$ includes one. All three three-test between-allocation point estimates are negative. The within-association term is positive for the exploratory policy and uncertain for the two originally reserved policies; the global contrast alone does not isolate favorable response-level association.

Frozen independent benchmark cohort. A frozen 512-question LiveCodeBench cohort (Jain et al., 2025) adds 12,288 responses. With repaired labels and full-response FP32 Grams, three-test observed/global-expectation ratios are 0.605 for Llama ($[0.504, 0.723]$) and 0.342 for Qwen Instruct ($[0.225, 0.510]$). Llama's within-association interval is below zero and remains negative in every leave-one-prompt-out calculation; Qwen's interval includes zero. The weaker one-test proxy has point ratios 0.836 and 0.627, but both intervals include one ($[0.695, 1.026]$ and $[0.269, 1.245]$). Thus the three-test contrast does not establish uniform robustness to verifier budget. All Qwen Thinking labels are retained, without a substituted approximate long-response Gram. The decomposition is a

retrospective endpoint on a frozen cohort; it is not the unexecuted preregistered audit-budget utility test (Appendix D.4).

Evaluator and label sensitivity. Rescoring all 12,288 archived APPS answers changes 18 records: 11 full-suite labels and 16 proxy-label records, with overlap. Independent three-repeat checks agree for 17 changed records. One candidate's unordered-set iteration changes its one-test label; its full-suite and three-test labels remain zero. We retain the candidate and report both one-test states rather than selecting a majority label. The primary three-test comparison is unaffected. Full-suite execution remains a reference under a declared runtime, not a proof of semantic correctness.

Controlled multi-step consequence. We construct 128-context, eight-action softmax policies with one clean correct action. Both noisy channels have exactly 25% FPR over incorrect context/action pairs. One accepts all wrong actions on a fixed quarter of contexts; the other disperses the same number of accepted wrong pairs. After 400 group-relative updates with $K = 8$, median clean correct-action probabilities are 0.995 (clean), 0.778 (correlated), and 0.443 (dispersed). Across 100 frozen seeds, the correlated-minus-dispersed difference is 0.334 with a 2.5/97.5-percentile interval of [0.280, 0.395]. $K = 4$ gives the same direction and magnitude.

Policy-matched online LLM mechanism. The fixed action-space masks above need not retain equal policy-weighted FPR after the policies diverge. We therefore froze a separate experiment in which Qwen3-0.6B is updated online through rank-8 query/value LoRA. At every update, every sampled wrong action has conditional false-positive probability 0.25 in both arms. The prompt-correlated arm draws one Bernoulli variable per prompt and applies it to every wrong response in that group; the sample-dispersed arm draws independently per wrong response. Correct responses are never flipped. The arms share initial adapters and separate paired action/noise random streams.

Across all 16 frozen seed pairs, the final clean correct-action probability is 0.133 higher on average under prompt-correlated noise (median 0.125; 12/16 positive; paired-seed bootstrap 95% interval [0.047, 0.211]; exact two-sided paired sign-flip permutation $p = 0.0113$ for the mean paired difference). Mean realized FPRs are 0.2520 and 0.2504, respectively. Figure 4 shows both the aggregate trajectory and every paired endpoint, including the negative seeds. This comparison controls policy-weighted marginals, unlike the earlier mask experiment, but still uses a restricted eight-action, one-token policy rather than free-form code RLVR.

Free-form follow-ups. We next trained Qwen2.5-7B-Instruct (Qwen Team, 2024) with the three online channels on 96-token worked arithmetic solutions. The first experiment's prespecified terminal endpoint saturated: correlated and dispersed held-out accuracies were 0.9676 and 0.9671, a difference of 0.0006 (eight seeds; paired-difference range $[-0.0179, 0.0089]$; five positive pairs; exact two-sided Fisher sign-flip permutation $p = 1.0$ for the mean difference). A post-hoc trajectory pattern motivated a second follow-up, for which trajectory AUC was specified before drawing new tasks and seeds. The pattern did not recur: mean AUC was 0.7339 versus 0.7378, a difference of -0.0039 (paired-difference range $[-0.0215, 0.0098]$; three positive pairs; exact paired sign-flip permutation $p = 0.4219$ for the mean difference). Terminal accuracies were 0.9600 and 0.9609. Thus these simple free-form tasks neither establish a benefit nor show a detectable penalty from correlation; they constrain extrapolation of the preceding 16-seed effect rather than extending it.

Frozen GSM8K training study. The study fixes 48 runs across two Qwen2.5-Instruct checkpoints (1.5B and 7B), eight seeds per model, and clean, clustered or dispersed reward channels. The official GSM8K (Cobbe et al., 2021) training partition supplies 1,024 training problems and separate validation and geometry-probe sets; 256 fixed official-test problems supply greedy evaluation. Each run uses rank-8 query/value LoRA ($\alpha = 8$, zero dropout), four prompts with eight completions each, two optimization epochs per rollout, AdamW learning rate 10^{-5}, clipping threshold 0.2, and KL coefficient 0.01. Decoding permits 1,024 new tokens. The primary endpoint is clean held-out accuracy AUC over the first one million *training completion tokens*, with linear interpolation at the boundary. This budget excludes evaluation and geometry-probe computation; it is not a matched total-compute comparison.

For a batch containing N incorrect completions, both noisy channels use the same stochastically rounded quota near $0.25N$. The dispersed channel chooses a uniform subset; the clustered channel

Table 2: Frozen GSM8K confirmation: accuracy AUC on $[0, 10^6]$ training completion tokens. Differences are clustered minus dispersed, in percentage points (pp). Intervals use 10,000 paired-seed bootstrap draws; p enumerates all 256 sign flips. The sign-flip reference assumes paired-arm exchangeability under its null. Neither model passes the frozen gate (mean at least 1 pp and interval lower bound above zero).

	Qwen2.5-1.5B	Qwen2.5-7B
Clean AUC	0.625879	0.919003
Clustered AUC	0.626094	0.919462
Dispersed AUC	0.626247	0.919819
Paired difference (pp)	-0.0153	-0.0357
95% interval (pp)	$[-0.5010, 0.5066]$	$[-0.1630, 0.0990]$
Exact two-sided p	0.96875	0.64844
Positive seed pairs	4/8	4/8

selects a random circular block of the incorrect completions ordered contiguously by prompt. Each incorrect completion has conditional inclusion probability K/N given the rounded quota K. Matched replay therefore fixes both the quota and per-item marginals, while online policy divergence can change N and realized FPR. This is a different noise construction from the earlier independent-Bernoulli experiment.

All 48 runs use the prespecified training budgets and paired initializations. Training task selection, numeric labels and corruption masks are retained for analysis alongside token costs, evaluation scores and midpoint Gram diagnostics. Both effects are close to zero (Table 2; Figure 2); this is a failure to confirm the proposed benefit, not proof of equivalence. Realized FPR averaged over seeds is 0.25037 versus 0.24983 for 1.5B and 0.25139 versus 0.25291 for 7B. The 1.5B sampled training completions hit the length cap in 9.3–10.1% of cases across arms; these records are retained. Greedy final-evaluation truncation is at most 0.049% averaged over seeds. The clean 1.5B arm improves from 60.55% to 63.48% final accuracy (positive in all eight seeds); clean 7B improves from 91.41% to 92.04%. Thus the experiment permits learning, although the 7B endpoint is near its ceiling. This diagnostic does not establish why the mechanism fails to translate into a useful training advantage.

5 Reliability and sensitivity of the audit

Exact references and joint uncertainty. For each eight-response group we enumerate the possible binary reward patterns. The within-prompt reference averages patterns with that group's observed error count; the global reference weights patterns by their finite-population probabilities under the observed cohort-wide FP/FN totals. The current analysis uses these expectations directly. It requires no sampled permutations and does not replace an expectation by a median. We draw 10,000 bootstrap samples of 512 prompt indices, sharing indices across policies and proxy strengths within each benchmark. In every sample we recompute the negative-class population, false-positive count and global-reference expectation. This retains the dependence induced by answering the same questions and propagates reference uncertainty. A zero global expectation makes a ratio undefined. For APPS Qwen Thinking, 6 one-test and 27 three-test resamples have this property; ratio and normalized-term intervals condition on a defined denominator. Raw signed contrasts retain all 10,000 resamples, without imputing an effect to undefined ratios. All intervals are retrospective, descriptive, and unadjusted for multiplicity. They condition on the retained responses, evaluator, models and tangent coordinates; they do not sample new policies or establish performance on unseen tasks.

Evaluator reliability changes the scientific result. The LiveCodeBench evaluator preserves candidate state across tests, streams one result per test, and compares outputs outside the jail without exposing expected outputs to candidate code. All 12,288 cached responses resolve under this contract. The driver passes 21 regression fixtures covering, among other cases, large output, shared object state, exceptions, and memory exhaustion. An earlier transport failure had incorrectly removed 309 full-suite-correct outcomes; the primary results use the complete rescoring. Appendix D.4 documents the failure and repair. The fixtures validate these execution cases, not all contest semantics.

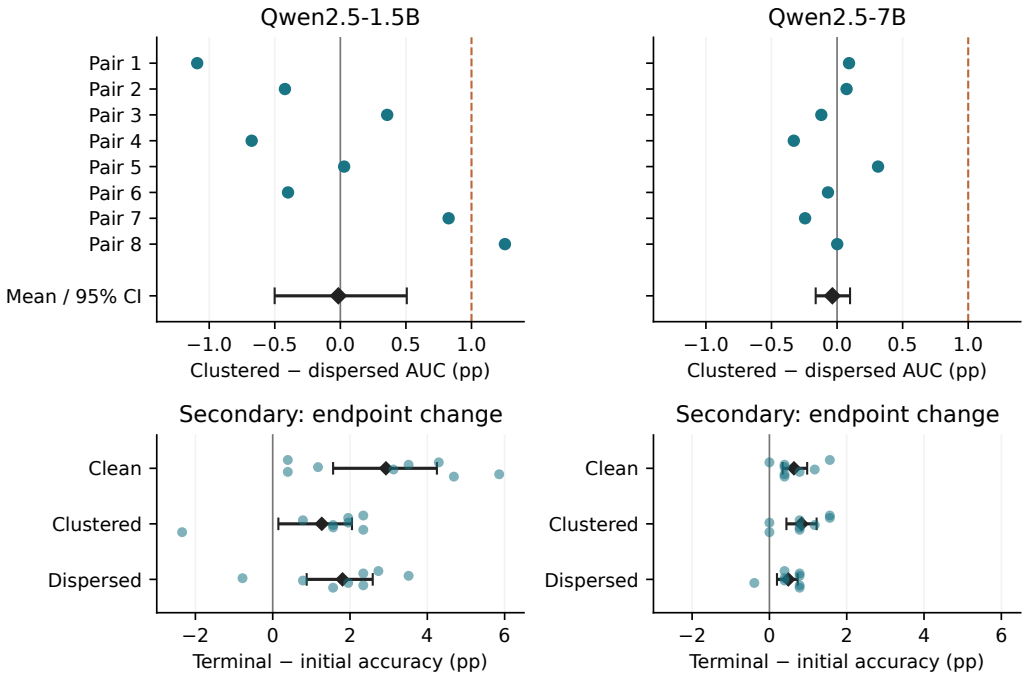


Figure 2: Frozen 48-run GSM8K study. Circles show all eight paired-seed clustered-minus-dispersed effects per model (pairs ordered by seeds 73000–73007) on a common percentage-point axis. Diamonds and horizontal bars show the means and 10,000-draw paired-seed bootstrap 95% intervals. Gray lines mark zero; dashed orange lines mark the +1 pp mean-effect gate. The prespecified mean-effect gate was $+1$ percentage point, and the interval lower bound also had to exceed zero. Neither gate is passed. Bottom: descriptive terminal-minus-initial accuracy for each reward channel, retaining all seeds; diamonds and bars show means and seed-bootstrap intervals. The clean arms show learning, but these secondary endpoints do not replace the frozen AUC test. Failure to confirm benefit is not an equivalence result.

Numerical validation is distinct from replay. Both primary LCB matrix sets use FP32 with TF32 disabled and zero dropout, preserving the realized LoRA tensors and all completion tokens for every prompt. Full-loss and chunked calculations agree on all 32 checked gradients (maximum relative error 6.4×10^{-5}; minimum cosine above 0.99999999). Exact replay of the archived BF16 implementation did not imply numerical agreement: three of four checked groups failed the full-loss/FP32 comparison. Archived BF16 matrices therefore serve only as a sensitivity check for LCB. The APPS matrices retain their archived backend and support a weaker numerical claim. These are local-score measurements on fixed cached responses, not new FP32 on-policy samples.

Sparse evidence and inexpensive controls. We retain all zero-distortion groups and report results stratified by the number of clean-correct responses. We also recompute the global reference after deleting each prompt, measure concentration of the signed association term, and isolate contributions from groups containing a length- or time-limited response. For APPS Qwen Thinking, 83.8% of responses reach the token cap and only five prompts have nonzero three-test distortion; all of the absolute within-association mass lies in at most six prompts. Its uncertainty must therefore be read as sparse conditional evidence, not a broad replication across 512 informative prompts. All of its observed three-test distortion lies in groups containing at least one capped response; this association does not identify a causal truncation effect. We compare full Grams with diagonal and same-trace isotropic matrices using both aggregate contrasts and per-prompt distortion rankings. Agreement in aggregate direction alone would not justify claiming that off-diagonal geometry is necessary. For an isotropic Gram, fixed within-prompt error counts fix the clean/noisy label contingency table and

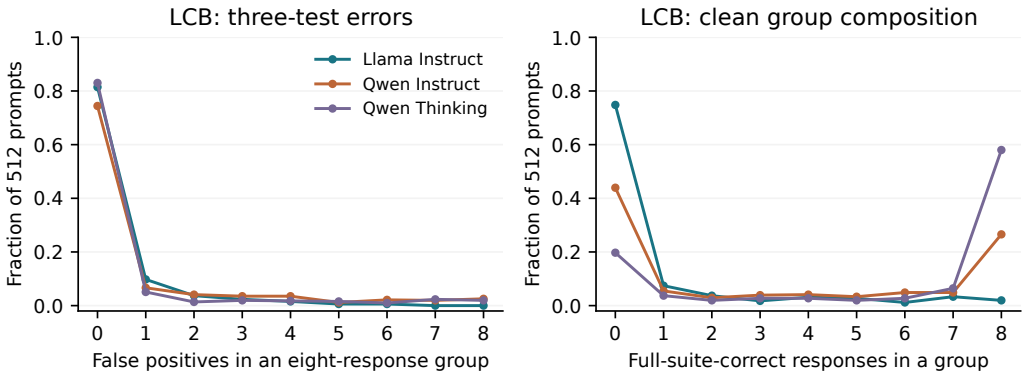


Figure 3: Repaired LiveCodeBench labels. Left: distribution of the three-test false-positive count within each eight-response group. Right: clean-correct response counts. Every curve includes all 512 prompts, including zero-count and homogeneous groups, on a linear probability scale. Thinking contributes labels only. Differences in decoding and token budgets preclude interpreting these curves as a controlled model ranking.

hence the squared advantage difference. Its within-association term is therefore identically zero; a nonzero global contrast in this control measures allocation rather than within-prompt association. The full matrix measures distortion magnitude and signed association, but need not be needed for ranking audit candidates. Appendix D.6 finds no stable selection advantage for the tested full-Gram features; diagonal Grams reproduce much of the observed ranking.

The error-count distribution and clean group composition in Figure 3 are complementary diagnostics. Concentrated errors can leave homogeneous groups degenerate, but the same count pattern can also erase contrasts in mixed groups. Neither marginal FPR nor the error-count histogram alone determines the signed association term: the correspondence between errors and response gradients still matters under the within-prompt reference. We therefore report both label-only summaries and the geometry audit, without treating one as a substitute for the other.

6 Related work

Noise-corrected GRPO derives corrections under Bernoulli reward corruption and argues that grouping can mitigate individual noise (El Mansouri et al., 2025). Other work derives backward and forward corrections under instance-independent class-conditional channels and explicitly notes that content-dependent errors violate the model (Cai et al., 2025). Our matched permutation leaves the marginal channel unchanged and jointly perturbs prompt allocation and error–tangent association across the groups used by the optimizer.

Systematic-verifier experiments show that error patterns with similar initial rates can lead to delay, plateau, or collapse (Egashira et al., 2026). We identify one algebraic regime in which positive prompt-level dependence reduces spurious within-group contrasts, not a general guarantee for systematic errors. Recent category-level verifier audits further show that aggregate errors can conceal sharply different implementation failure modes (Xin, 2026). We complement category-level decomposition by examining how errors interact with the optimizer's group-relative updates.

What the present study adds. Empirical robustness to imperfect verifiers has also been demonstrated across code and scientific reasoning settings (Plesner et al., 2026). Together with systematic-error studies, this means that "marginals are insufficient" alone is not a new general claim. Our narrower contribution is the exact FP/FN-preserving LoRA sensitivity audit and the algebra of within-group spurious activation. The null GSM8K result limits translation of that mechanism into training utility; we introduce neither a noise-correction algorithm nor evidence that our channel outperforms existing correction methods.

7 LIMITATIONS AND REQUIRED NEXT EVIDENCE

The real-data audit covers two code benchmarks, Python, nested partial-test proxies, three policies per benchmark, and only false positives; the archived APPS Thinking condition is severely length-limited and cannot support unrestricted-model generalization; full-suite execution is not proof of semantic correctness. LiveCodeBench exact Grams cover two policies, while Qwen3-4B-Thinking contributes labels only; complete long-response Grams have not passed the required numerical and completion checks. The restricted-LoRA tangent gradients describe a fixed-policy gradient surrogate, not an optimizer step. The matched permutation also changes which score vectors carry errors, making this a sensitivity audit rather than a causal estimate of clustering. The 16-seed online experiment isolates the mechanism but restricts an actual language model to eight one-token actions. The free-form follow-ups allow 96-token decoding only on simple arithmetic and both miss their prespecified criteria. The additional GSM8K study includes clipping and KL regularization, yet also misses its frozen criterion on both models. No online study includes code execution or a learned, evolving verifier. The GSM8K midpoint geometry uses only four fixed probe prompts per run and is a local mean-token LoRA diagnostic, not an explanation of Adam/clipping dynamics. The local comparisons leave the safety and usefulness of a correlated verifier unresolved.

A preregistered four-seed Qwen3-0.6B pilot had mean difference 0.183, but one pair was negative (-0.085); it failed the all-positive gate and is not confirmation. Its fixed masks matched action-space FPR, while policy-weighted FPR diverged as policy mass shifted. The subsequent 16-seed experiment matched conditional expected marginals at every update. The free-form studies do not confirm the restricted-action effect under their tested budgets. They support a limit on the demonstrated mechanism, rather than a general claim that it cannot transfer.

Reference sensitivity. A retrospective decomposition preserves lower observed distortion with diagonal and same-trace isotropic Grams. Holding the error counts fixed within each prompt changes the reference and can weaken or reverse that comparison. The evidence therefore concerns joint error allocation under a specified reference distribution; it does not demonstrate that off-diagonal geometry is necessary for the effect. Same-trace isotropic controls still retain each prompt's total score scale; they test the need for off-diagonal structure, not the absence of all gradient information. The corrected exact-expectation tables and accompanying arrays distinguish these references.

Reporting implication. A verifier report can pair marginal FP/FN counts with error-count histograms, clean-group composition, and homogeneous-group activation. If gradients are available, Eq. 2 adds the observed distortion and its within- and between-prompt contrasts in explicitly stated tangent coordinates. All groups belong in these summaries; infrastructure failures require resolution before scoring. Audit-selection cost and downstream training utility then require their own endpoints.

What an audit can decide. A large local distortion identifies a discrepancy at the audited policy and parameter coordinates. It does not identify which verifier would give the best terminal policy after many updates. That decision also depends on how the rollout distribution, error allocation and policy geometry change during training. A useful reporting distinction is therefore between diagnosing the current update and selecting a training procedure. The former is measured here; the latter requires a separate outcome-based comparison.

8 CONCLUSION

The matched-error audit separates two contributions that marginal FP/FN totals leave unspecified: allocation across prompts and association with response gradients within a prompt. All reported observed/global-reference point ratios are below one, but several intervals include one; the clearest interval-supported contrasts are the two LCB three-test results. Between-prompt allocation estimates are negative across the reported cells, while within-prompt association varies in sign and uncertainty. These measurements depend on the evaluator, tangent coordinates, and reassignment reference. Diagonal controls and retrospective selection results do not establish a practical advantage for full-Gram auditing, and the GSM8K study does not confirm a training benefit. The supported use is therefore diagnostic: report local distortion and its allocation/association terms alongside error counts, then test utility under a separate selection or training protocol.

Reproducibility Statement

The manuscript specifies the verifier-error decomposition, proxy construction, matched conditions, paired seeds, and formal training endpoints. Retained protocols, experiment code, and saved analysis records distinguish exact fixed-policy calculations, restricted-action online diagnostics, and free-form training confirmation. Reported seed differences and bootstrap intervals are computed from the saved formal runs; unsuccessful gates and negative results are retained. The development audit in Appendix D distinguishes output availability, reference validation, evaluator semantics, and sampling-score consistency. The supplementary material provides the evaluator code and constructed test cases used to distinguish these outcomes. The LiveCodeBench cohort is reported only after all 512 payloads, 12,288 responses, gold-isolated grading records, and 1,024 exact Gram matrices passed their recorded checks. Explicit candidate errors and specified fatal signals after acknowledged test entry are incorrect outcomes; generic missing driver records remain infrastructure failures until resolved. The original batch-output failure and BF16 numerical-check failures are retained alongside the repaired scoring and FP32 results.

AI Use Statement

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

References

APPS contributors. Apps evaluation implementation. https://github.com/hendrycks/apps/tree/b45c0ed78517a3a6492eb77b21cffbb79b1096f1, 2026. Pinned source used for evaluator-contract checks.

Xin-Qiang Cai, Wei Wang, Feng Liu, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Reinforcement learning with verifiable yet noisy rewards under imperfect verifiers. *arXiv preprint arXiv:2510.00915*, 2025.

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.

Kazuki Egashira, Mark Vero, Jasper Dekoninck, Florian E. Dorner, Robin Staab, and Martin Vechev. Delay, plateau, or collapse: Evaluating the impact of systematic verification error on rlvr. In *Conference on Language Modeling*, 2026. URL https://colm.eventhosts.cc/Conferences/2026/AcceptedPapers.

Omar El Mansouri, Fathinah Asma Izzati, Mohamed El Amine Seddik, and Salem Lahlou. Noise-corrected grpo: From noisy rewards to unbiased gradients. *arXiv preprint arXiv:2510.18924*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with apps. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/c24cd76e1ce41366a4bbe8a49b02a028-Abstract-round2.html.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free

evaluation of large language models for code. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/94074dd5a072d28ff75a76dabed43767-Abstract-Conference.html.

LiveCodeBench contributors. Livecodebench evaluation implementation. https://github.com/LiveCodeBench/LiveCodeBench/tree/16de2a1ef5c07342f6b9782a415cc29cee9ce6e4, 2026. Pinned source used for evaluator-contract checks.

Andreas Plesner, Francisco Guzman, and Anish Athalye. An imperfect verifier is good enough: Learning with noisy rewards, 2026. URL https://arxiv.org/abs/2604.07666.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

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.

Esther Xin. Where the verifier fails: A category-level audit of reward signals in rlvr. *arXiv preprint arXiv:2609.01354*, 2026.

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

A APPS EVALUATOR CORRECTION AND SINGLE-LABEL SENSITIVITY

The complete 12,288-answer replay preserves saved candidate extraction and the original deterministic test order. Repairs cover text/byte streams, builtin `pow`, function-return serialization, large-integer transport, and singleton replay of crashes after acknowledged test entry. Generic missing driver records are not assigned incorrect labels. All answer records resolve under this contract; this is not a claim of full official-runtime equivalence.

The 18 changed records include 11 full-label and 16 proxy-record changes. Three independent repeats agree for 17 records. Qwen Instruct's task 3099, answer 5, processes all game moves and chooses among final winners by iterating an unordered set. Its one-test outcomes are 1, 0, 0 across repeats, while its full and three-test outcomes are always 0. We preserve the original program, all questions and the repeated outcomes; no majority-vote replacement or favorable seed is used. Only the changed records received this three-repeat check, not the entire archive.

One-test state	Gram	Ratio	Conditional 95% interval
0	full	0.821	$[0.442, 1.313]$
0	diagonal	0.509	$[0.361, 0.682]$
0	isotropic	0.451	$[0.329, 0.594]$
1	full	0.821	$[0.443, 1.313]$
1	diagonal	0.509	$[0.362, 0.683]$
1	isotropic	0.452	$[0.330, 0.595]$

Both states use the same 10,000 shared-prompt bootstrap indices and exact global reference. Their interval envelope is a *label-sensitivity range*, not an additional 95% confidence interval. All main three-test results are unchanged between states. Historical permutation-median, pooled-bootstrap, group-size and additional-initialization summaries were computed under superseded labels. They are retained only as provenance in the artifact and are not quantitative evidence for this submission. Current conclusions use the corrected exact-expectation analysis rather than substituting those old summaries.

B EXACT SEPARATION OF WITHIN-PROMPT ASSOCIATION AND BETWEEN-PROMPT ALLOCATION

Let R be uniform error reassignment preserving total FP/FN counts and let U_c preserve those counts separately within each prompt and true-label class. With expectations (not medians), the exact identity is

$$D_{\text{obs}} - \mathbb{E}_R D = (D_{\text{obs}} - \mathbb{E}_{U_c} D) + (\mathbb{E}_{U_c} D - \mathbb{E}_R D).$$

The first term holds prompt allocation fixed; the second changes that allocation. Writing $\mu = \mathbb{E}_{U_c}[A(r^*) - A(\tilde{r})]$ and $\Sigma = \text{Cov}_{U_c}[A(r^*) - A(\tilde{r})]$ gives $\mathbb{E}_{U_c} D_x = (\mu^\top G_x \mu + \text{tr}(G_x \Sigma))/K^2$. Within a true-label class of size $n > 1$ and error fraction p, error-indicator covariance is $p(1-p)n/(n-1)(I - \mathbf{1}\mathbf{1}^\top/n)$. Constant-reward groups use zero advantage; singleton classes have zero covariance. We verified the expression against exhaustive binary assignments for groups up to eight and an independent global enumeration with both FP and FN errors. Global expectations for the historical data are computed exactly by summing all 2^8 per-prompt patterns with their finite-population probabilities.

Policy	Tests	Observed D	Within association	Between allocation
Qwen Instruct	1	1.06×10^{-2}	5.97×10^{-4}	-2.91×10^{-3}
Qwen Instruct	3	5.75×10^{-3}	5.52×10^{-4}	-2.07×10^{-3}
Qwen Thinking	1	1.08×10^{-5}	1.61×10^{-8}	-3.60×10^{-5}
Qwen Thinking	3	7.37×10^{-6}	-1.97×10^{-7}	-3.55×10^{-5}
Llama Instruct	1	1.39×10^{-2}	9.44×10^{-4}	-6.01×10^{-3}
Llama Instruct	3	6.51×10^{-3}	1.08×10^{-3}	-4.79×10^{-3}

These are retrospective recomputations of the original tangent matrices, not new independent policies. All 512 prompts per policy are retained, including zero distortion and negative contributions. A lower global-reference distortion can coexist with positive within-prompt association: it cannot by itself establish a beneficial error–tangent association. Diagonal and trace-isotropic controls are included in the accompanying raw decomposition files. The finite-policy identity does not imply a parameterization-invariant geometry or improved RL training.

C RESTRICTED-ACTION ONLINE MECHANISM

The eight-action experiment isolates a bounded mechanism; it is distinct from the free-form confirmation in Figure 2.

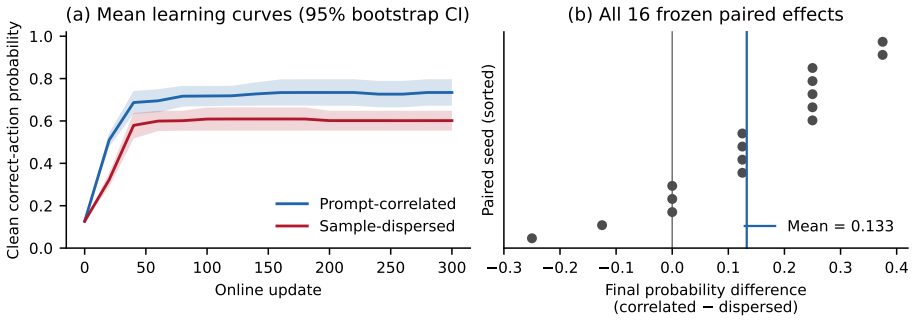


Figure 4: Frozen online Qwen3-0.6B LoRA experiment with equal conditional expected FPR at every update. Shading is a seed bootstrap interval, not within-run uncertainty. The right panel retains all 16 paired effects.

D DEVELOPMENT PROTOCOLS, EVALUATOR CHECKS, AND SCORE CONSISTENCY

These checks concern development and the frozen independent cohort implementation. They neither replace the archived APPS audit nor extend its confirmation claims. Task identities, decoding changes, execution completion, and scientific gates are recorded separately. A successful process exit is not evidence that the corresponding scientific gate passed.

D.1 OUTPUT AVAILABILITY UNDER SEPARATE DECODING PROTOCOLS

All rows below use the same eight APPS-training development questions and eight responses per question. They were not selected using generated-answer correctness. The original protocol uses temperature 1 and no top-k or top-p truncation. The later native-configuration study uses temperature 0.6, top-$k = 20$, and top-$p = 0.95$. The latter study pairs seeds and two-response microbatches across its two caps. Its results are not pooled with the original decoding study, whose batching and sampling conditions differ.

Table 3: Development output availability. "Final" means a nonempty final answer after the native reasoning delimiter; it is not a correctness label. For the two native-configuration rows, the last column requires both a final answer and a normal stop. Original-protocol stopping counts are reported separately below because their intersection with final availability was not reported in the original summary.

Protocol	Policy	Token cap	Final	Final + normal stop
Original	Llama-3.1-8B	8,192	64/64	—
Original	Qwen3-4B-Instruct	8,192	64/64	—
Original	Qwen3-4B-Thinking	8,192	15/64	—
Original	Qwen3-4B-Thinking	16,384	45/64	—
Native configuration	Qwen3-4B-Thinking	16,384	44/64	43/64
Native configuration	Qwen3-4B-Thinking	81,920	64/64	64/64

Normal stopping in the four original-protocol rows occurred 64, 63, 14, and 41 times, respectively. The native-configuration arms each produced all 64 records and exited normally. Their generated-token totals were 753,965 and 867,950; elapsed generation times were approximately 7,048 and 9,138 seconds. These are retained execution costs, not an exclusive-hardware throughput comparison. The prespecified 90% threshold requires at least 58 of 64 responses for each gate. Only the longer native-configuration arm passes both final availability and final-with-normal-stop gates. Neither gate measures correctness, verifier error, gradient distortion, or improvement during training.

D.2 A SAMPLING SCORE MUST USE ITS OWN DISTRIBUTION

Let $z_\theta(v \mid h)$ be the model logit at history h. For temperature T and the retained top-k/top-p set $S_\theta(h)$, the implemented sampling kernel has the form

$$q_\theta(v \mid h) = \frac{\mathbf{1}\{v \in S_\theta(h)\} \exp(z_\theta(v \mid h)/T)}{\sum_{u \in S_\theta(h)} \exp(z_\theta(u \mid h)/T)}. \tag{3}$$

On a region where the support is fixed and v belongs to it,

$$\nabla_\theta \log q_\theta(v \mid h) = \frac{1}{T}\left[\nabla_\theta z_\theta(v \mid h) - \sum_{u \in S_\theta(h)} q_\theta(u \mid h) \nabla_\theta z_\theta(u \mid h)\right]. \tag{4}$$

Both the temperature and restricted normalizer matter. This expression is not generally the temperature-1, full-vocabulary model score, and it does not assert differentiability across changes in the retained support. If a saved token falls outside the support recomputed during replay, its replay log probability is not finite; silently substituting a base-policy score changes the estimand.

We selected the median-length and longest completed responses from the 81,920-token development arm using response length alone. Each complete response was differentiated in a rank-16, alpha-16 LoRA tangent on query and value projections, with seed 83999 set before adapter creation, zero-initialized B, and no dropout. Gradient checkpointing and 256-token vocabulary-head chunks avoid materializing all full-vocabulary logits at once. The model conditions on the complete prompt and response prefix; chunks do not truncate the response or detach its history.

Table 4: Complete-response score engineering on two saved development responses. The gradient norm uses the base-policy mean-token log probability in the stated LoRA coordinates. These norms are not compared as task-utility effects. GiB refers to peak allocated GPU memory.

Case	Prompt	Response	Gradient norm	GiB	Native-score replay
Median length	243	13,001	0.05093	13.14	Off-support token(s)
Longest	398	35,015	0.04975	24.54	Off-support token(s)

Both base-policy gradients were finite and nonzero, while the A gradients were zero at $B = 0$, as expected. Both native-score checks failed the finite-support gate. This observation identifies a replay inconsistency in the tested numerical pipeline; it does not establish its unique cause or a general failure of truncated sampling. The checked generation configuration contains no additional sampling parameters beyond those recorded for the native arm. Possible effects of cached generation, teacher-forced execution, and numerical kernels require separate checks.

A further independently frozen development condition retains temperature 0.6 and the 81,920-token cap but removes top-k/top-p hard truncation. It targets sampling-score consistency rather than selecting a condition for higher measured correctness. This development condition completed: all 64 responses were graded, 25 passed the declared full test suite, and eight full-response Gram matrices were retained. It is a small engineering check and is excluded from benchmark confirmation and training-utility claims. Moreover, the mean-token score used here remains a local surrogate: replacing it by a trajectory-sum policy-gradient estimator multiplies each response's score by its length and requires a corresponding change to the interpretation of its Gram matrix.

D.3 Reference validation is not task-level correctness

The independent-cohort protocol freezes 512 APPS-test questions and 512 LiveCodeBench release-v6 questions independently of model outcomes and reference execution. Identities are retained when validation fails; failed questions are not replaced. Initially, 464 APPS questions had at least one passing reference among the first three published solutions under the existing sandbox convention. The 48 remaining questions were then investigated as follows.

Table 5: Sequential APPS reference investigations. Passing counts use the specific runtime and comparator of that row; the rows must not be combined into a claim that one official evaluator validated the entire cohort.

Investigation	Additional pass	Remaining
Up to 20 references in original order; original comparator	15	33
First 3 references; pinned upstream stdout comparator	9	24
Up to 20 references; same upstream stdout comparator	2	22
Pinned upstream runtime helpers; project jail	9	13
Expanded stdout and transport limits on three large-output cases	2	11

The upstream stdout-comparison block was extracted from a pinned APPS evaluator version; reference code continued to execute in the existing isolated runtime (APPS contributors, 2026). Thus comparator agreement is not full official-runtime equivalence. Differences can arise from output-order conventions, numerical comparison, language/runtime compatibility, resource limits, or multiple valid solutions to a constructive task. The retained diagnostics do not assign one of these causes without further evidence. In particular, an upstream comparator accepting a reference is not a proof that it implements the original contest's semantic checker. These investigations left 11 questions unresolved at that stage. Subsequent per-question runtime and semantic repairs validate reference execution for

all 512 fixed questions, without replacing questions. This validation is not a candidate-generation or new-cohort confirmation result. In the original cohort, question 1619 uses another published reference after the initial reference exceeded the unchanged time limit; the failed reference remains documented.

D.4 LIVECODEBENCH PAYLOAD, EVALUATOR CHECKS, AND CONFIRMATION

This cohort's task identities, decoding and budgets were frozen before generation. The frozen protocol retained an audit-budget utility criterion: at least 10% reduction at a 25% audit budget with a positive lower paired-confidence bound. The full cross-dataset, new-verifier utility experiment has not been executed. A retrospective held-out selection diagnostic is reported in Appendix D.6; it does not replace that endpoint. The association/allocation decomposition and its intervals are retrospective analyses of the frozen cohort, not successful tests of that preregistered endpoint. LiveCodeBench release-v6 also does not imply a temporal holdout relative to every model's pretraining.

Llama and Qwen Instruct use temperature 1.0 and an 8,192-token completion cap; Qwen Thinking uses temperature 0.6 and an 81,920-token cap. All use full-support sampling (top-$p = 1$, top-$k = 0$), eight responses per task and two-response microbatches. Length limits affect 2/4,096 Llama responses and 249/4,096 Qwen Instruct responses. Thinking has one length-limited and three time-limited responses; all four lack a final-answer segment. All remain in the denominator. These different conditions support within-policy comparisons, not a controlled ranking across models. The three-test proxy uses all available tests when fewer than three exist; four of the 512 LCB tasks have only two tests.

Initial isolated checks decoded 488 of 512 payloads; the remaining 24 passed when the decoder's memory limit alone was raised from 512 MiB to 4 GiB. Candidate execution limits were unchanged. All 512 payloads therefore passed structural checks, which do not by themselves establish correct model scoring.

A pinned upstream evaluator (LiveCodeBench contributors, 2026) and a gold-isolated adapter were then each tested on the same 13 constructed regression fixtures. Each passed 13/13 expected-outcome checks. The fixtures cover exact integers, preserved line boundaries, equivalent decimal representations, adjacent large integers, unequal nearby floats, whitespace, binary standard input, function arguments, tuple outputs, shared object state across tests, wrong answers, runtime errors, and timeouts. Negative fixtures are successful checks when the evaluator rejects them; 13/13 is not benchmark accuracy.

The upstream fixture harness removes an unused NumPy import after an AST check and disables fault-handler instrumentation inside the jail. The revised adapter retains upstream compilation and input-handling helpers, supplies inputs without expected outputs, and compares candidate outputs outside the jail. Candidate limits remain two seconds per test and 512 MiB; trusted result serialization is outside the candidate timer. Candidate code is compiled once and a function-style object persists across the full shuffled test sequence. The first definite failed test ends evaluation, while outcomes already sufficient for the one-/three-test proxies are retained. A missing result frame is infrastructure missingness, not automatically an incorrect answer. Segmentation, abort, arithmetic and CPU-limit signals after acknowledged test entry are explicitly recorded as candidate crashes; generic exits remain unresolved infrastructure events. The expanded driver passes 21/21 constructed fixtures, including a test suite whose serialized outputs exceed 1 MiB, a large single output, persistent object state across 29 tests, closed stdout and memory exhaustion. These checks cover the stated cases, not all possible evaluator behavior or contest-specific semantics.

The original batched driver did not satisfy this contract: output transport could lose successful results and candidate state could reset across batches. We preserve those records but supersede their labels. Rescoring all 12,288 unchanged responses and repairing five remaining transport records gives zero unresolved answers. Full-suite labels change from incorrect to correct for 10 Llama, 77 Qwen Instruct and 222 Qwen Thinking responses, with no correct-to-incorrect changes. Including proxy-label changes, 395 answer records change. Clean accuracies are 11.52%, 41.50% and 70.26%, respectively. These are diagnostics conditional on the declared tests, resource limits and decoding, not leaderboard or semantic-correctness scores.

For the three-test proxy, decompositions use all 512 Llama and Qwen Instruct FP32 matrices. Between-prompt allocation terms are -2.581×10^{-3} and -3.283×10^{-3}, respectively. Within-prompt terms are -4.788×10^{-4} and 3.013×10^{-5}; their intervals appear in Table 1. Qwen Thinking's 4,096 responses and error-count distribution are retained, but full long-response Gram computation did not complete under four memory-management implementations. We report no truncated, filtered or approximate replacement. This cohort supports a conditional local allocation contrast, not a downstream training benefit.

Table 6: Repaired frozen LiveCodeBench cohort. Accuracy and false positives (FP) are measured among 4,096 responses per policy; FN is zero for these nested proxies. Ratios use the three-test proxy and full-response FP32 Grams.

Policy	Correct	Accuracy	FP_1	FP_3	Ratio
Llama Instruct	472	.1152	621	189	.605
Qwen Instruct	1700	.4150	998	472	.342
Qwen Thinking	2878	.7026	649	340	—

D.5 Numerical, influence and inexpensive-geometry controls

All rows below retain all 512 prompts and use the same repaired labels. BF16 entries use the archived matrices; the other columns use newly computed FP32 matrices or their stated transformations. Ratios compare observed distortion with the exact global-reference expectation. Median relative Frobenius differences between archived BF16 and FP32 matrices are 4.6% for Llama and 9.6% for Qwen Instruct; their 95th percentiles are 29.7% and 74.7%. The modest change in an aggregate ratio therefore does not certify the original gradients' numerical fidelity.

Policy	Tests	Full FP32	Archived BF16	Diagonal FP32	Isotropic FP32
Llama Instruct	1	0.836	0.826	0.801	0.806
Llama Instruct	3	0.605	0.612	0.591	0.656
Qwen Instruct	1	0.627	0.627	0.370	0.500
Qwen Instruct	3	0.342	0.339	0.252	0.306

Policy	Tests	Cap-group D share	Top-six assoc. share	Within-term leave-one-out range
Llama Instruct	1	0.0%	44.4%	$[-4.01\times10^{-4}, 1.04\times10^{-3}]$
Llama Instruct	3	0.0%	33.7%	$[-5.21\times10^{-4}, -3.50\times10^{-4}]$
Qwen Instruct	1	4.6%	80.9%	$[-1.13\times10^{-4}, 1.48\times10^{-3}]$
Qwen Instruct	3	6.6%	55.1%	$[-1.16\times10^{-5}, 1.02\times10^{-4}]$

"Cap-group D share" is the fraction of all observed distortion contributed by groups containing at least one length- or time-limited response. There are two such Llama groups and 92 Qwen Instruct groups. "Top-six assoc. share" uses absolute within-association contributions, so opposite signs cannot cancel in its denominator. The leave-one-out calculation retains the remaining 511 groups and recomputes the global finite-population reference. Llama's three-test within term stays negative throughout. Qwen's one-test association is concentrated: six prompts account for 80.9% of its absolute contribution, and deleting a single prompt can reverse its sign. These diagnostics do not select a reduced cohort for the primary analysis.

Diagonal Grams preserve much of the three-test observed-distortion ranking: Spearman correlations with full Grams are 0.9998 (Llama) and 0.9999 (Qwen), with overlap of 49/52 and 51/52 among the 52 largest-distortion prompts. However, their relative ℓ_2 errors over per-prompt distortion values are 16.9% and 33.0%. Thus strong rank agreement need not imply accurate distortion magnitude, but these results also do not demonstrate a selection-utility advantage for full Grams. The isotropic control has exactly zero within-association term under fixed within-prompt error counts: the label contingency table fixes both advantage norms and their inner product. Floating-point residuals near zero are not substantive effects. Ranking overlap can be tie-dependent in the sparse archived APPS cells.

For the archived APPS cohorts, groups containing capped responses number 106 (Instruct), 478 (Thinking), and 5 (Llama). Their observed three-test distortion shares are reported from corrected labels in the artifact; the Thinking share is 100%. All groups remain included; these decompositions do not identify a causal effect of truncation. Clean-correct-count strata (0–8), unnormalized contributions, every shared bootstrap replicate, individual-prompt influence and all diagonal/isotropic comparisons accompany the artifact.

D.6 Retrospective selection and prefix-cap diagnostics

Status and estimand. After examining the repaired cohort, we froze an additional analysis of the two complete LCB FP32 policies and both proxy strengths. This is a retrospective comparison, not a new confirmation or an implementation of the original cross-dataset, new-verifier gate. Auditing one prompt replaces all eight proxy labels with their reference labels, removing that prompt's squared local update distortion at fixed parameters. This additive per-prompt target does not measure the squared sum of a minibatch's gradient errors, nor a change in training accuracy. In particular, cross-prompt cancellation is outside the estimand.

Features, selection and cost accounting. For each of five fixed split seeds (83001–83005), 128 prompts supply training labels and 384 supply evaluation labels. Basic features contain the proxy-positive fraction and variance and log-transformed mean, maximum and standard deviation of response length. Diagonal features add log-transformed mean, maximum and standard deviation of Gram diagonal entries. Full features additionally include the off-diagonal Frobenius norm divided by the full norm and the log-transformed nonnegative proxy-advantage quadratic form. Held-out gold labels are never features. Each standardized ridge predictor selects its penalty from $\{0.1, 1, 10\}$ on an inner 96/32 split, is refit on all 128 training prompts, and ranks the held-out prompts by nonnegative predicted distortion. Seeded random priorities break ties. The comparison also includes uniform expected selection, proxy disagreement and mean response length. We retain all methods, three audit fractions (10%, 25%, 50%), and all five splits: 60 conditions.

A 25% evaluation budget audits 96 of 384 held-out prompts; with the 128 training prompts, it consumes 224/512 (43.75%) fully labeled prompts. It is therefore *not* a 25% total-annotation-cost experiment. Prompt counts also omit variable test-runtime cost and matrix-construction cost. Uniform selection is its exact expectation, not a favorable random draw. For each condition, we compare against the smallest realized residual among the lower-feature-cost comparators; 10,000 paired prompt bootstrap resamples recompute this comparator minimum. These descriptive intervals condition on fitted scores and the selected audit set: resamples do not refit the learner or enforce a fresh fixed-size selection. The overlapping splits are sensitivity analyses, not independent replications. Here lower feature cost does not mean zero gradient cost: the diagonal baseline also requires score information. We do not establish an end-to-end compute saving.

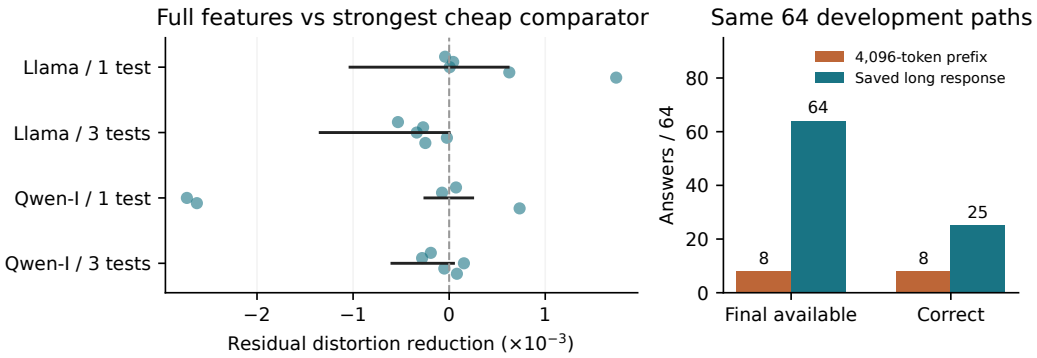


Figure 5: Retrospective checks, with their limited scopes. Left: at 25% held-out budget, each dot is one fixed split's absolute reduction in residual local distortion relative to its strongest lower-feature-cost comparator; positive favors full features. Bars show the first split's conditional 95% paired-bootstrap interval. Right: final-answer availability and correctness for the same 64 development paths before and after prefix censoring. These are neither independent confirmation nor training gains.

Results and interpretation. At 25% held-out budget, the first frozen split yields relative residual reductions of -0.8% and -57.7% for Llama with one and three proxy tests, and $+5.8\%$ and -34.6% for Qwen Instruct. None has a strictly positive lower confidence bound. Across five splits, one Llama one-test condition has a positive lower bound, but the pattern does not replicate across proxy strengths and policies. Negative values mean higher residual distortion than the strongest lower-feature-cost comparator. Thus these tested full-matrix summaries do not establish stable selection utility; this is not an impossibility result for every geometry-based policy. The artifact retains all split identities, feature arrays, targets and selection masks, including zero-distortion and unfavorable conditions.

What the long-response development study repairs. We separately apply a deterministic 4,096-token prefix cap to all 64 saved full-support development trajectories, decoding the retained prefix with the same tokenizer and requiring the same native final-answer contract. Reference tests and their order are unchanged. Only 8/64 prefixes contain a final answer, and those eight are correct; the saved long responses contain 64/64 final answers and 25/64 correct answers. The other 56 paths exceed 4,096 tokens. This paired prefix diagnostic shows that truncation removes usable answers in this development setting. It does not regenerate the historical temperature-1 APPS cohort: these paths use temperature 0.6, comprise only eight development tasks, and do not establish unrestricted Thinking geometry or a positive online RL effect. The original capped APPS observations and all null training results remain reported.