RETHINKING SEQUENCE-LEVEL FEEDBACK FOR ON-POLICY DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence-level feedback can guide on-policy distillation beyond the current token, but estimating future teacher–student disagreement more accurately does not ensure better task decisions. We investigate this gap in mathematical reasoning by separating feedback estimation, gradient noise, and task utility. For candidate-based credit over a 16-token future window on Qwen3-4B, matched comparisons identify loss normalization as the source of an apparent gradient-variance increase under centering. Averaging more continuations reduces initial gradient variance by 80.3%, yet yields no measured training benefit over 40 updates across three seeds; reproducible candidate rankings likewise do not improve held-out correctness over the student's choices. These findings motivate a complementary use of feedback: selecting where to provide supervision, with reference solutions supplying the token targets. We formulate *Reference-supervised OPSD* (RS-OPSD), which retains distillation on student-generated prefixes and adds reference-token supervision on problems whose sampled answers fail verification. In single-seed evaluations, RS-OPSD exceeds OPSD's final four-benchmark macro accuracy by 1.76 and 0.84 percentage points for Qwen3-4B and Qwen3-8B, respectively. However, mean online differences are negative, and random reference selection reaches a similar 4B endpoint, leaving the advantage of failure selection unresolved. Together, these results show why feedback precision, task value, and supervision selection require distinct evidence.

1 INTRODUCTION

On-policy distillation (OPD) trains a student on its own generated responses, allowing a teacher to supervise the prefixes the student encounters at inference (Agarwal et al., 2024; Gu et al., 2024). A token choice, however, also determines which prefixes come next. Sequence-level feedback accounts for these later consequences, extending supervision beyond the mismatch at the current token.

Estimating those consequences is difficult because future responses must be sampled. Gradient decomposition and temporal credit assignment address this optimization problem (Gu et al., 2024; Liu et al., 2026), while trajectory guidance and token-importance methods refine the feedback itself (Jiang & Ferraro, 2026; Xu et al., 2026b). These approaches motivate a further question: when does a better estimate of future teacher–student disagreement yield a more useful update? Agreement proxies task performance, so estimation quality alone cannot answer this question (Agarwal et al., 2024; Ma, 2026).

We investigate this question through three links: scalar cost estimates to parameter gradients, initial gradient noise to training outcomes, and candidate rankings to task decisions. Each link requires its own control. Our starting scalar residual diagnostic arose in method development; we do not attribute it to all prior OPD methods. In a later, matched gradient audit, the apparent variance increase under sensitivity-weighted centering largely disappears when estimators share a normalization denominator. Averaging more continuations then supplies a direct denoising intervention for short training, while independent continuations test whether lower-cost actions improve correctness.

This distinction motivates a complementary role for outcome feedback: selecting where to provide supervision, while reference solutions supply what to predict. We formulate Reference-supervised OPSD (RS-OPSD) under on-policy self-distillation, where a frozen copy of the initial student receives reference context (Zhao et al., 2026). RS-OPSD retains distillation on every sampled response

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.

and adds reference-path cross-entropy (CE) on failed problems. Random selection tests whether the observed endpoint difference requires failure information or can arise from adding reference targets.

Our contribution is an account of what each controlled comparison identifies. Matched normalization identifies the source of a variance increase, short training bounds the observed benefit of an initialization-time denoising intervention, and held-out selection distinguishes score information from improvement over the student. The reference case extends the same attribution principle to problem selection. These conclusions concern the tested candidate estimator and training recipes, not sequence-level feedback in general.

2 Preliminaries: local matching and future-state credit

To distinguish immediate matching from later consequences, consider a student π_θ, a fixed teacher p_ϕ, and a problem x. For a sampled response y, the state $s_t = (x, y_{<t})$ consists of the problem and the tokens generated before the next-token decision; we call this conditioning context a *prefix*. The cost $d_\theta(s_t) = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s_t) \| p_\phi(\cdot \mid s_t))$ is the full-vocabulary reverse Kullback–Leibler (KL) divergence before sampling y_t. With horizon H and zero cost after termination, the cumulative objective is

$$J(\theta) = \mathbb{E}_{y \sim \pi_\theta} \left[\sum_{t=1}^{H} d_\theta(s_t) \right]. \tag{1}$$

This equals the sequence-distribution KL under the same termination convention. Differentiating both the divergence and the distribution from which the response is sampled gives (Gu et al., 2024; Schulman et al., 2015)

$$\nabla_\theta J = \underbrace{\mathbb{E}\left[\sum_{t=1}^{H} \nabla_\theta d_\theta(s_t) \right]}_{R_1} + \underbrace{\mathbb{E}\left[\sum_{i=1}^{H} \nabla_\theta \log \pi_\theta(y_i \mid s_i) \sum_{t=i+1}^{H} d_\theta(s_t) \right]}_{R_2}. \tag{2}$$

R_1 captures the change in teacher–student mismatch at visited prefixes; R_2 captures the change in which future prefixes are visited. Under gradient descent, the latter favors token choices that lead to smaller later divergence. The score function uses the same π_θ that samples the response. This decomposition explains why local matching leaves a role for future-state credit but does not establish that reducing future divergence improves task correctness.

The candidate studies use a finite-window auxiliary: at a student prefix, we force four candidate next tokens and sample M continuations per candidate. Variance and training studies sample eligible positions uniformly and use the highest-probability eligible tokens, excluding special/control tokens; held-out selection restricts positions to the response's first half.

Each candidate's $\widehat{Q}_s(a)$ averages discounted mean costs over at most 16 subsequent positions ($\gamma = 0.95$), excluding the forced token. Costs combine student-top-16 probability weights with teacher–student log-probability differences, rather than the full-vocabulary KL in Equation (1). The loss enumerates candidates with detached, candidate-normalized student weights; finite windows, restricted support, and clipping make it a surrogate for R_2, not an unbiased estimate.

Centering subtracts a shared prefix baseline $b(s)$ from candidate costs. Validity checks determine both which prefix contributions are averaged and their denominator. Appendix B.2 specifies eligibility, configuration, the loss, and the separate full-suffix geometry probe.

3 From estimated sequence credit to useful supervision

We connect the optimization concerns above to task performance through gradient geometry, repeated-gradient measurements, training, and held-out action selection. These tests distinguish signal scale from noise, scalar quality from gradient stability, and estimator improvements from task value. Protocols and replication units appear in Appendix C. Avg@K averages correctness over K sampled answers per problem; macro accuracy weights benchmarks equally.

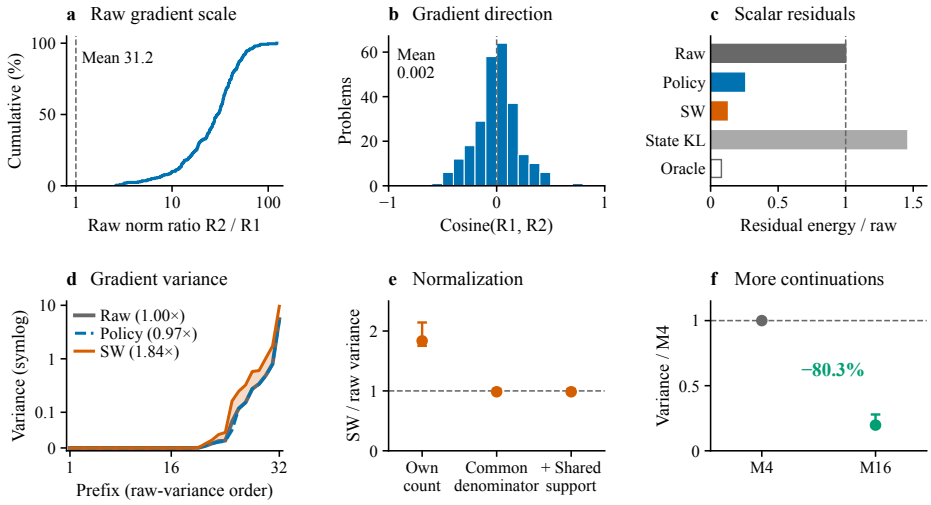

Figure 1: **Normalization explains the SW variance increase.** (a,b) Sequence gradients are larger than local gradients with little mean alignment (sampled 256 problems). (c) Centering reduces scalar residual energy; Oracle is an in-sample lower bound. (d,e) Sensitivity-weighted (SW) centering increases parameter variance, largely explained by normalization. (f) Averaging sixteen continuations per candidate (M16) reduces variance relative to four (M4). The three protocols for (a,b), (c), and (d–f) are distinct. Panels (d–f) use 16 repeats at 32 initial prefixes in two batches; (d) orders prefixes by raw variance and shades differences, not uncertainty. Bars in (e,f): descriptive paired-prefix 95% intervals. Detailed in Appendix D.

3.1 Establishing the scale of sequence credit

At a fixed checkpoint, a sampled-token, full-suffix probe measures how strongly sequence credit changes local matching and in which direction. Its loss and parameterization differ from the training auxiliary, so its scale does not measure applied training strength (Appendix B.2). The norm ratio $\kappa_i = \|R_2^{(i)}\|_2/\|R_1^{(i)}\|_2$ averaged 31.2; the cosine $c_i = \cos(R_1^{(i)}, R_2^{(i)})$ averaged 0.002, with both reinforcing and opposing directions (Figure 1a,b). The ratio reflects the loss reductions, including summed future returns versus local token averaging. Without a high-dimensional reference distribution, the near-zero cosine establishes neither unusual conflict nor independence. A positive coefficient λ scales κ_i but preserves c_i: **strength control alone does not align the two signals**.

3.2 Measuring and reducing gradient noise

We next ask how candidate gradients fluctuate across resampled continuations. Policy centering subtracts the candidate-probability-weighted mean cost; sensitivity-weighted (SW) centering weights additionally by a squared logit-gradient norm proxy. Both preserve candidate cost differences.

In our development audit, policy and SW centering reduced held-out scalar residual energy by 75.0% and 87.8% (Figure 1c). This weighted second moment uses the probability-times-sensitivity weights that define SW, minimizing it rather than parameter-gradient variance. Different prefix sets and sampling protocols separate this audit from the gradient audit below; they are not paired measurements of one intervention.

Holding the model, prefixes, and candidates fixed, we resample continuations and sum their gradient variance across parameters (the covariance trace). Under each estimator's own normalization, policy centering changed this variance by -2.5%, whereas SW increased it by 83.5% (Figure 1d).

SW's validity check retained 23 of 32 prefixes, whose contributions were amplified by the smaller averaging denominator. On the same continuations, a common denominator reduced the SW/raw variance ratio from 1.835 to 0.986; intersecting active supports also gave 0.986 after rounding (Fig-

ure 1e). Normalization thus explains the increase without revealing substantial denoising: policy and SW remain within 3% of raw variance. Their per-prefix mean-gradient cosines to raw, 0.72 and 0.84, also indicate changed update directions (Appendix D.4).

The implemented candidate sum helps explain why scalar energy is insufficient. At the collection parameters, with candidate support and normalization fixed, the unnormalized prefix gradient satisfies

$$h_b(s) = h_0(s) - b(s)\nabla_\theta \log Z_s, \qquad Z_s = \sum_{a \in \mathcal{A}_s} \pi_\theta(a \mid s). \tag{3}$$

Centering changes only the candidate-mass gradient direction, leaving orthogonal continuation noise intact. Enumerating a fixed candidate set changes the usual sampled-action baseline argument (Greensmith et al., 2004). This identity supplies a possible mechanism, without quantifying the observed residual/variance difference (Appendix B.3).

Averaging more continuations provides a direct denoising intervention: M16 reduced initial parameter variance by 80.3% relative to M4, with similar mean-gradient norms and directions (cosine 0.989; Figure 1f). The variance ratio, 0.197 (descriptive interval $[0.185, 0.279]$), is consistent with the 0.25 expected from four times as many independent samples. Does this measured initial denoising improve training?

3.3 Does measured denoising improve training?

We compare local-only OPD, M4, and M16 over 40 updates and three seeds, fixing the future window and evaluation. At $\lambda = 0.0015$, the mean auxiliary gradient was only 0.06–0.09% of the local norm, below the later calibration target; this inconclusive comparison tests a weak intervention.

The larger coefficient $\lambda = 0.05$ passed the initialization calibration: weighted gradient ratios were 1.84–3.13%, and the relative change in the mathematical AdamW update was 12.63–18.94%. M16 then trailed M4 in every seed by -0.66 pp on average (family-wise 95% interval $[-1.53, +0.20]$; Figure 2a). The upper bound excludes the prespecified $+0.5$ pp improvement under the paired seed-t model, using Bonferroni adjustment over three contrasts. Three seeds cannot validate that model's distributional assumption, and a failed zero-coefficient monitor leaves sustained update strength unverified (Appendix E.2).

The centering matrix asks whether observed gains require lower initial variance. At $\lambda = 0.0015$, SW exceeded raw M4 credit in all three seeds (mean $+0.88$ pp; family-wise interval $[+0.35, +1.41]$), but comparisons with local-only and shuffled credit remained unresolved (Figure 2b). Shuffling preserves the centered values while breaking their assignment to actions. This contrast does not isolate denoising: SW changed mean-gradient direction, and shuffled credit also exceeded raw in two of three seeds (post hoc; Appendix E.3). Together, these comparisons separate measured initial variance from observed short-training differences without identifying a unique mechanism for the SW result.

3.4 Do the feedback scores prefer better actions?

Limited training gains may reflect either optimization or the signal's task value. To examine task value independently, we test action selection without parameter updates. Centering preserves candidate rankings, $\arg\min_a(\widehat{Q}_s(a) - b(s)) = \arg\min_a \widehat{Q}_s(a)$; more continuations can improve ranking precision, but even population costs need not favor correct answers.

On new problems, we select among four next-token candidates at a fixed prefix using either the lowest future distillation cost $\widehat{Q}_s(a)$ (Section 2) or the highest sampled correctness. Independent continuation sets select and evaluate actions, separating selection noise from measured value. Random score reassignment tests the value of score–action correspondence; the student's candidate probabilities test improvement over its existing choices. Appendix F.1 specifies sampling and labels future-cost selection *Future-KL* (KL in Figure 2).

The 16-token future-cost rankings were reproducible across continuation sets (mean Spearman correlation 0.86), yet selection differed from the student by -0.92 pp (family-wise 95% interval $[-2.71, +0.41]$) and from reassignment by $+0.43$ pp ($[-0.87, +1.69]$). Correctness-based selection

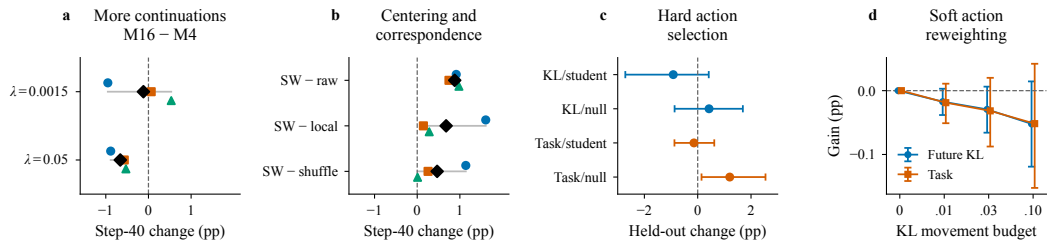

Figure 2: **Estimator improvements require separate tests of task utility. blue/orange/green: different seeds, diamonds: means, gray segments: ranges.** (a) M16-minus-M4 training effects. (b) SW centering versus raw, local-only, and shuffled credit. Both use five-benchmark macro Avg@12; (c,d) Held-out hard selection and soft reweighting (sampled problems): KL uses estimated future distillation cost; Task uses sampled final-answer correctness. Bars: family-wise 95% intervals over four comparisons in (c), pointwise 95% source-bootstrap intervals in (d). All effects are in pp, detailed in Appendices E, F.

beat reassignment by 1.20 pp but differed from the student by -0.14 pp (Figure 2c): information relative to reassignment need not improve the student's choices.

To retain more of the student's preference, we also shifted probability toward better-scoring candidates within a KL budget. This soft reweighting likewise established no gain (Figure 2d). An optimistic same-sample calculation reaches 0.40 pp at the largest budget; because evaluation outcomes choose the policy, this establishes neither achievable held-out gain nor test power (Appendix F.1).

The outcome control also lacks a demonstrated student-relative advantage, limiting a KL-specific interpretation. Positive covariance of candidate correctness across independent continuation halves supports persistent task-value differences, yet correctness rankings were poorly reproduced (mean Spearman correlation 0.047). Thus, the outcome control detects information relative to reassignment without establishing sensitivity to student-relative gains. These results bound the tested selection effects without ruling out task information in future disagreement.

3.5 From reference-conditioned scores to reference targets

The preceding studies use future costs to assign credit to sampled tokens. Reference-conditioned self-distillation offers a different teacher: a frozen copy of the initial student receives the reference solution (Zhao et al., 2026), motivated by the role of teacher–student compatibility (Li et al., 2026a). Development probes under this setting remained mixed; a single-seed matched pair moved from $+1.15$ pp at step 40 to -0.75 pp at its endpoint (Appendix G.1). These different-recipe probes motivate further study rather than independently confirming the candidate-credit findings.

A reference can also supply target tokens directly: answer failure selects a problem for additional training, and its stored solution specifies what to predict. This motivates RS-OPSD, which preserves supervision on the student's own prefixes while adding reference targets on selected problems. Reference exposure and failure-based selection then become separately testable design choices.

4 Reference-supervised OPSD

RS-OPSD combines distillation at student-visited prefixes with reference CE on failed problems (Figure 3). Each example provides a problem x_i, a stored solution z_i, and answer metadata a_i^*; the sampled answer determines whether reference CE is activated. Our OpenThoughts references (Guha et al., 2026a) are neither newly generated student solutions nor assumed to be human-written or error-free.

Distillation on sampled prefixes. The student generates y_i from x_i alone; a frozen copy of its initial model scores the same prefix using the privileged prompt $P(x_i, z_i)$. The sampled-prefix loss $\mathcal{L}_{\text{sampled}}$ applies to every response, preserving on-policy supervision regardless of answer cor-

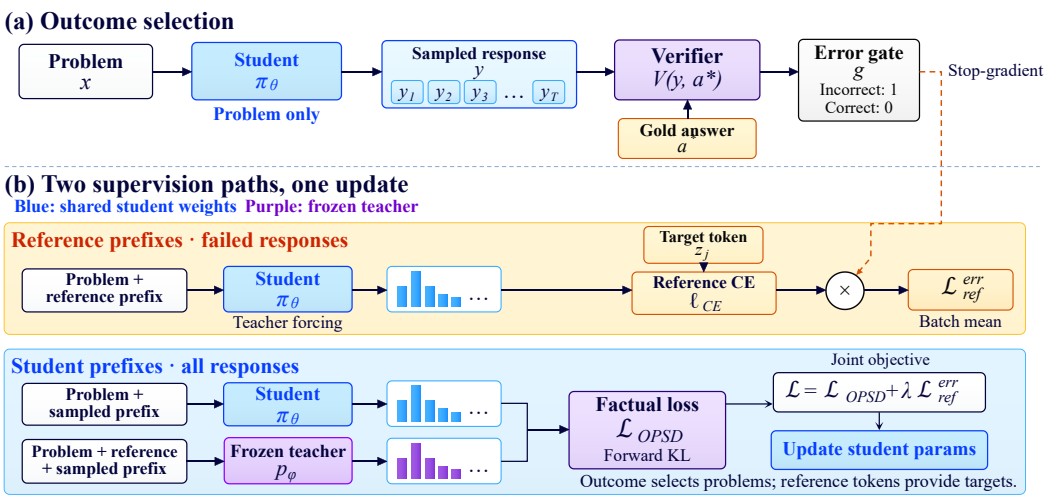

Figure 3: **Error-conditioned reference supervision.** (a) A student response supplies the verification gate. (b) Sampled-prefix distillation applies to every response; reference CE additionally supervises reference prefixes on failed problems. Only the frozen teacher receives privileged reference context; both losses update the student. The diagram's Factual/FKL label refers to the sampled-prefix OPSD loss $\mathcal{L}_{\text{sampled}}$; implementation details are given in Appendix C.2.

rectness. Our OPSD implementation and experimental procedure are specified in Appendices C.2 and C.

Supervision on selected reference prefixes. Answer verification gives $u_i = \text{Verify}(y_i, a_i^*) \in \{0, 1\}$ (Appendix C.6). A failure activates supervision on the stored solution, without assigning token-level blame within the sampled response. The detached gate defines

$$g_i = \text{sg}(\mathbf{1}[\text{clip}(u_i, 0, 1) < 0.5]),$$
$$\mathcal{L}^{\text{err}}_{\text{ref}} = -\frac{1}{B} \sum_{i=1}^{B} \frac{g_i}{\max(1, L'_i)} \sum_{j=1}^{L'_i} \log \pi_\theta(z_{i,j} \mid x_i, z_{i,<j}). \tag{4}$$

Here, sg denotes stop-gradient, $\mathbf{1}$ is the indicator, and L'_i counts valid reference tokens after truncation and masking. The bounded denominator and empty-sum convention give zero loss when $L'_i = 0$. The student predicts each reference token from $x_i, z_{i,<j}$. Thus, outcome feedback determines which problems receive additional targets, while the reference determines the update direction. Dividing by the full batch size B makes the CE contribution scale with the fraction of selected problems.

Joint update. The two paths form the RS-OPSD objective:

$$\mathcal{L} = \mathcal{L}_{\text{sampled}} + \lambda \mathcal{L}^{\text{err}}_{\text{ref}}, \qquad \lambda = 0.001. \tag{5}$$

Only the student is updated. Reusing the distillation rollout for verification adds teacher-forced reference computation but no sampled response. This auxiliary supervises off-policy reference prefixes; it does not estimate R_2. Appendix C details target processing and optimization.

5 Evaluating Reference Supervision and Error Selection

The joint loss changes two things at once: it exposes the student to reference-token targets and uses failures to select the problems that receive them. A comparison with OPSD measures their combined effect. Random selection is needed to examine whether the outcome gate adds value, while always-on supervision probes the broader choice of exposing every problem to its reference. We organized the evaluation around these attribution questions.

Table 1: **Final reference-supervision results.** Seed 2026, step-229 Avg@12 (%), 38,912-token limit. RS-OPSD adds error-conditioned CE to OPSD; differences use unrounded scores.

	Qwen3-4B			Qwen3-8B		
Dataset	OPSD	RS-OPSD	Δ	OPSD	RS-OPSD	Δ
AIME24	70.00	**72.78**	+2.78	**76.94**	76.11	−0.83
AIME25	61.11	**64.17**	+3.06	63.89	**67.22**	+3.33
AMC23	89.86	**90.56**	+0.70	89.26	**89.96**	+0.70
Olympiad-Bench	67.25	**67.77**	+0.52	68.62	**68.80**	+0.17
Average	72.05	**73.82**	+1.76	74.68	**75.52**	+0.84

5.1 Setting and evaluation protocol

We trained Qwen3-4B and Qwen3-8B on a prepared OpenThoughts mathematics subset (Guha et al., 2026b). The baseline / RS-OPSD pairs share initialization and settings, except for CE; their later trajectories can diverge. The 4B random-reference run uses the same nominal recipe. The Appendices C and G.2 specify the training configuration and selection control. Training uses non-thinking mode and a 1,024-token limit; evaluation uses thinking mode under the archived reasoning-evaluation recipe. This is a train–evaluation mode difference, not evidence that leakage is prevented. Online checkpoints 50/100/150/200 use 8,192 tokens and Avg@4, while the final checkpoint uses 38,912 tokens and Avg@12 (Appendix C, Table 4). Avg@$K = (N_d K)^{-1} \sum_{i,k} r_{i,k}$ measures the average correctness. Macro accuracy equally weights AIME24, AIME25, AMC23, and Olympiad-Bench. Minerva was excluded from this reporting scope at the authors' request after results were available; this was not a prespecified exclusion. The preceding credit studies retain their original five-benchmark scope.

These reference comparisons use one training seed per model size. Same-protocol step-0 scores and an independent check of effective adapter loading are unavailable for the original pairs, which retain aggregate logs but not question-level outputs. The positive endpoint differences cannot distinguish absolute improvement from reduced degradation. Training logs record 49–52% response truncation and 57–58% failure; without their joint counts, the fraction of selected failures caused by truncation cannot be identified. The gate includes incomplete or unparseable answers as well as completed incorrect answers (Appendix C.6).

5.2 Positive endpoints and mixed online differences

RS-OPSD exceeded OPSD by 1.76 pp for 4B and 0.84 pp for 8B at the final checkpoint (Table 1). AIME24/25 contribute 83% of the 4B macro difference, while 8B declines on AIME24. Reweighting benchmark means by counts 30/30/83/673 reduces the differences to $+0.71/+0.31$ pp. These counts are verified for the later random run, so the historical weighting calculation is a sensitivity analysis, not recovered pooled-question accuracy.

Online evaluations give a less consistent picture: 4B RS-OPSD-minus-OPSD differences range from -1.87 to $+0.44$ pp and average -0.54 pp (Figure 4), opposite in mean sign to the endpoint difference. Because checkpoint, generation budget, and K change together, these single-seed comparisons cannot separate evaluation effects from checkpoint variation.

We also examine whether the reference-supervision difference extends to another distillation recipe. In the ordinary-OPD bridge, reference CE changes macro accuracy from 77.320% to 77.036% (-0.284 pp). Because this comparison changes both teacher and objective, it does not isolate which component accounts for the different result (Appendix G.5). The observed reference-CE benefit therefore remains specific to the tested setting.

5.3 Does failure selection account for the difference?

To separate reference exposure from selection, the random control selects as many problems as its own current failure mask within each reference-length bin, retaining each selected problem's

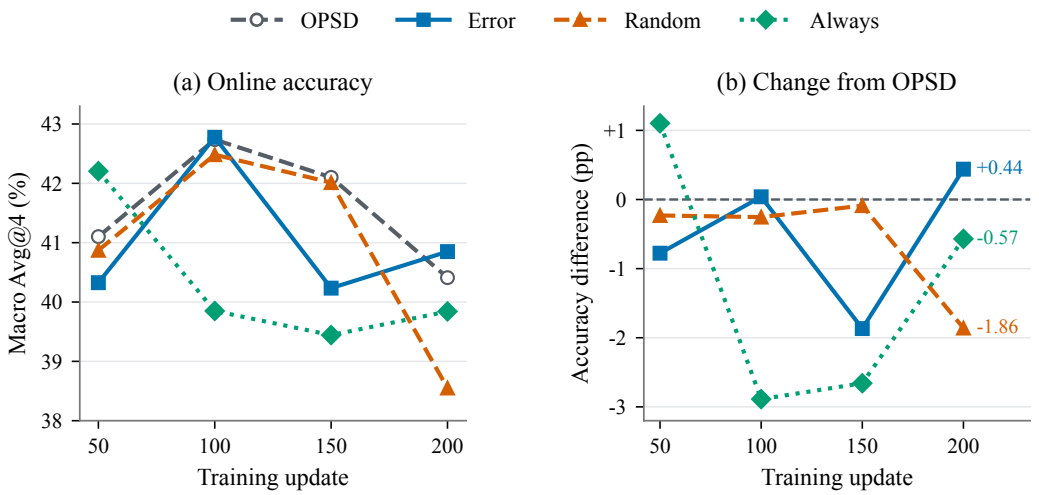

Figure 4: **Qwen3-4B online accuracy varies across checkpoints.** (a) Macro Avg@4. (b) Each CE arm minus OPSD at the same checkpoint; the dashed zero line marks equal accuracy. Colors, markers, and line styles identify methods in both panels. Error denotes RS-OPSD; Random uses count/length-bin-matched reference selection; Always uses all references. All runs use thinking mode, an 8,192-token limit, and seed 2026. Lines connect observed values without uncertainty bands. Step 229 uses a different evaluation budget and is excluded. Exact values: Appendix G.3.

reference. Random selection reached 73.86% final macro accuracy, compared with 73.82% for error selection and 72.05% for OPSD (Figure 5). Both CE arms had positive final and negative mean online differences from baseline. Thus, the positive endpoint does not isolate a failure-selection benefit, although one seed cannot establish gate equivalence. Full-training random and always-on controls are unavailable for 8B.

The similar error/random endpoints raise the possibility that reference exposure matters more than which problems are selected. Applying CE to every reference at the same nominal $\lambda = 0.001$ gives 40.33% mean online accuracy, versus 41.05% for error selection. With the completed OlympiadBench score of 64.83%, always-on reaches 72.77% final macro accuracy: $+0.72$ pp over OPSD, but 1.04 pp below error selection and 1.08 pp below random selection (Figure 5). Always-on is higher than both selection arms on AMC23 and lower on the other three benchmarks. This pattern favors selective exposure in the observed endpoints, but the different total CE exposure prevents attribution to selection alone (Appendix H.4).

Since full-training trajectories diverge, single-update tests provide a more local comparison of the gates. These tests compare error selection with count-and-length-matched and raw-gradient-norm-matched random selection. Error-reference changes accuracy by $+0.34$ pp at initialization and -0.34 pp at step 200 relative to OPSD; neither starting point establishes an advantage over random selection (Appendix G.4). Together, the full-training and local controls leave added reference exposure as an explanation for the endpoint difference. Similar selected-problem counts do not verify equal selected-token totals or optimizer dose (Appendix G.2).

6 Related Work

Distillation signals and task value. Sequence-level KD uses teacher outputs (Kim & Rush, 2016), whereas OPD supervises student prefixes; DistiLLM-2 contrasts these sources (Ko et al., 2025). TIP, FiRe-OPD, and REOPOLD refine token importance or teacher supervision (Xu et al., 2026b; Li et al., 2026c; Ko et al., 2026); TOPD and R2-OPD use near-future trajectories or reasoning progress (Jiang & Ferraro, 2026; Yang et al., 2026). GKD emphasizes student-generated sequences, while MiniLLM develops reverse-KL optimization (Agarwal et al., 2024; Gu et al., 2024). Complementing work on compatibility, outcome confounding, and state coverage (Li et al., 2026a; Ma,

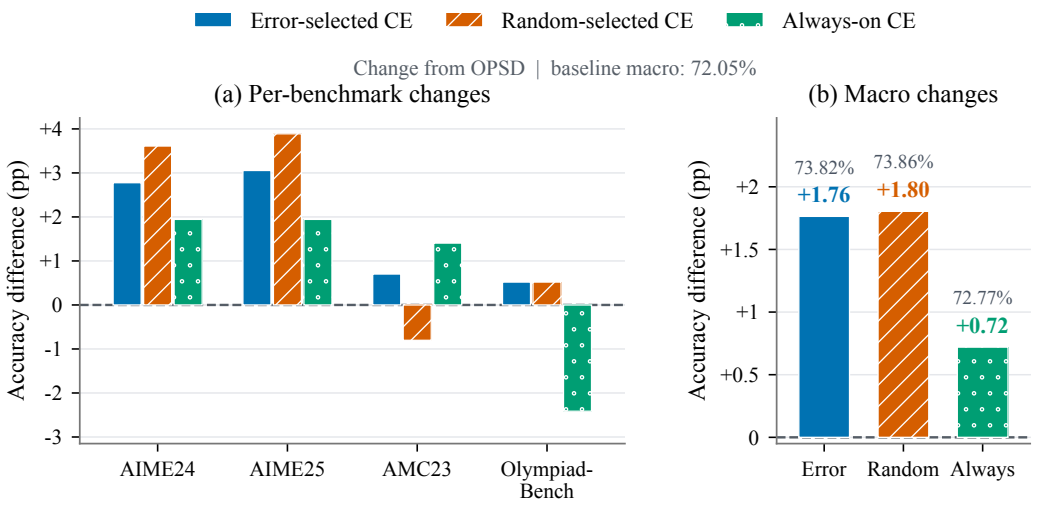

Figure 5: **Qwen3-4B reference-selection endpoints.** Error denotes RS-OPSD. Bars show accuracy differences from OPSD in percentage points: (a) each benchmark and (b) their equal-weight macro. Zero denotes the OPSD score for the corresponding comparison. Labels above the macro bars give the difference and absolute accuracy. Evaluation uses step-229 Avg@12, thinking mode, a 38,912-token limit, and seed 2026. Always-on includes the separately reported Olympiad-Bench completion (Appendix H.4). These single-seed runs do not match trajectories or total CE exposure. Absolute scores: Appendix G.3.

2026; Fu et al., 2026), our diagnostics connect feedback to task improvement by distinguishing scalar quality, gradient variance, and student-relative value.

Baselines and temporal credit. Variance-optimal score-function baselines depend on gradient sensitivity (Greensmith et al., 2004); centering our enumerated candidate sum can also change the mean update (Equation (3)). β-OPSD and OPPO accumulate future log-ratio returns and oracle-conditioned likelihood ratios, respectively (Xu et al., 2026a; Li et al., 2026b). SC-GRPO weights policy gradients by self-conditioned divergence; TurnSight adjusts turn-level advantages using hindsight (Shan et al., 2026; Qu et al., 2026). These mechanisms assign credit, whereas RS-OPSD selects problems for reference targets. The corresponding controls are action reassignment and matched random problem selection.

Reference exposure and selection. SCOPE routes supervision by correctness (Zheng et al., 2026). OPSD and SDPO use privileged context or textual feedback (Zhao et al., 2026; Hübotter et al., 2026), whose value can depend on reference content (Kaur et al., 2026; Shrestha & Tessier, 2026). STaR learns from newly generated correct rationales (Zelikman et al., 2022); RS-OPSD uses existing references on failed problems. Error/random selection tests this gate; separating teacher-conditioning and CE-target roles requires further controls.

7 Conclusion

We studied when sequence-level feedback improves on-policy distillation. For the tested finite-window candidate estimator, normalization explains SW's variance increase; lower initial gradient noise and reproducible rankings do not establish better short-training outcomes or held-out decisions. RS-OPSD retains sampled-prefix distillation and uses failures to select reference targets. Its single-seed endpoint gains over OPSD on Qwen3-4B/8B coexist with negative mean online differences and a similar random-selection endpoint, leaving the gate's benefit unresolved. These findings distinguish estimation quality, effective updates, and task value, while separating reference exposure from feedback-based selection.

AI USE STATEMENT

During the preparation of this draft, the authors used an LLM assistant solely to polish the language and research the literature. Its assistance was limited to improving the clarity, grammar, and fluency of the written text of the authors and helping to identify relevant publications. *Author verification pending:* Human authors must verify the final text, references, and this disclosure before submission.

REPRODUCIBILITY STATEMENT

The appendices specify the implemented losses, training and evaluation protocols, and gradient derivations. An anonymized source package with experiment configurations, analysis scripts, figure inputs, and a file-checksum manifest is available at https://anonymous.4open.science/r/OPSD_with_Reference-518D. Available artifacts and exclusions are documented in Appendix H. Model checkpoints and response-level outputs are not included in that package, and the documented overlap checks do not establish the absence of data leakage.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Zixuan Fu, Bingxiang He, Yuxin Zuo, Haohuan Huang, Jinqian Zhang, Ruhang Xiao, Cheng Qian, Qinyu Luo, Huan-ang Gao, Yudong Wang, Zhiyuan Liu, Ning Ding, and Chaojun Xiao. Rethinking On-Policy Distillation of Large Language Models II: One Training Example, 2026. URL https://arxiv.org/abs/2609.04172.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://www.jmlr.org/papers/v5/greensmith04a.html.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. OpenThoughts: Data Recipes for Reasoning Models. In *International Conference on Learning Representations*, pp. 108059–108130, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b010241b9f1cdfc7d4c392db899cef86-Abstract-Conference.html.

Etash Guha, Ryan Marten, Sedrick Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Benjamin Feuer, Leon Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Stutee Acharya, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya

Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alex Dimakis, and Ludwig Schmidt. OpenThoughts: Data Recipes for Reasoning Models. In *International Conference on Learning Representations*, pp. 108059–108130, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b010241b9f1cdfc7d4c392db899cef86-Abstract-Conference.html.

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement Learning via Self-Distillation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64121.

Yuxuan Jiang and Francis Ferraro. Bridging Reasoning Trajectories in On-Policy Distillation via Near-Future Guidance, 2026. URL https://arxiv.org/abs/2606.00305v2.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking On-Policy Self-Distillation for Thinking Models, 2026. URL https://arxiv.org/abs/2607.05184.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. DistiLLM-2: A Contrastive Approach Boosts the Distillation of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 31044–31062. PMLR, 2025. URL https://proceedings.mlr.press/v267/ko25a.html.

Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. Scaling Reasoning Efficiently via Relaxed On-Policy Distillation, 2026. URL https://arxiv.org/abs/2603.11137.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huanang Gao, Wenkai Yang, Zhiyuan Liu, and Ning Ding. Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe, 2026a. URL https://arxiv.org/abs/2604.13016v2.

Yu Li, Rui Miao, Tian Lan, and Zhengling Qi. OPPO: Bayesian Value Recursion for Token-Level Credit Assignment in LLM Reasoning, 2026b. URL https://arxiv.org/abs/2605.21851v2.

Yuying Li, Leqi Zheng, Yongzi Yu, Wenrui Zhou, Xuchang Zhong, Xing Hu, Jing Jin, Hangjie Yuan, and Tao Feng. Filter, Then Reweight: Rethinking Optimization Granularity in On-Policy Distillation, 2026c. URL https://arxiv.org/abs/2606.02684.

Shiqi Liu, Zeyu He, Letian Tao, Guojian Zhan, Jiaxin Gao, Feihong Zhang, Jingliang Duan, Wei Xiong, Kehua Sheng, Bo Zhang, Yang Guan, and Shengbo Eben Li. Beyond Token-Local Imitation: Reward-Compatible Temporal Credit Assignment for On-Policy Distillation, 2026. URL https://arxiv.org/abs/2609.16937.

Guoqing Ma. Outcome-Confounded Local Supervision in On-Policy Distillation, 2026. URL https://arxiv.org/abs/2607.23731.

Changle Qu, Sunhao Dai, Hengyi Cai, Yuqi Zhou, Xinran Chen, Simon, and Jun Xu. TurnSight: Turn-Level Hindsight Self-Distillation for Tool-Integrated Reasoning, 2026. URL https://arxiv.org/abs/2608.04007.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient Estimation Using Stochastic Computation Graphs. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/hash/de03beffeed9da5f3639a621bcab5dd4-Abstract.html.

Yingyu Shan, Yuhang Guo, Zihao Cheng, Zeming Liu, Xiangrong Zhu, Xinyi Wang, Jiashu Yao, Wei Lin, Hongru Wang, and Heyan Huang. Learning from Own Solutions: Self-Conditioned Credit Assignment for Reinforcement Learning with Verifiable Rewards, 2026. URL https://arxiv.org/abs/2606.18810.

Samyak Shrestha and Alexander Tessier. Rethinking Privileged Information in On-Policy Self-Distillation, 2026. URL https://arxiv.org/abs/2608.18271.

Jiawei Xu, Minghui Liu, Juzheng Zhang, Tom Goldstein, and Furong Huang. β-OPSD: Deriving with Policy Optimization, Training with Self-Distillation, 2026a. URL https://arxiv.org/abs/2607.28582.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token Importance in On-Policy Distillation, 2026b. URL https://arxiv.org/abs/2604.14084.

Chen Yang, Haiyuan Wan, Rengrong Xiong, Yize Chen, and Danny H. K. Tsang. Beyond Imitation: Filtering On-Policy Distillation by Reasoning Progress, 2026. URL https://arxiv.org/abs/2608.19408.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STaR: Bootstrapping Reasoning With Reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488. Curran Associates, Inc., 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64784.

Binbin Zheng, Xing Ma, Yiheng Liang, Jingqing Ruan, Xiaoliang Fu, Kepeng Lin, Benchang Zhu, Ke Zeng, and Xunliang Cai. SCOPE: Signal-Calibrated On-Policy Distillation Enhancement with Dual-Path Adaptive Weighting, 2026. URL https://arxiv.org/abs/2604.10688.

A Discussion and limitations

Our results separate three properties of feedback: estimation quality, its effect on the update, and its value for the task. For the tested candidate estimator, matching the normalization denominator removes SW's apparent variance increase, leaving centered and raw gradients with similar variance. The candidate-sum identity also shows that centering can change the mean update through candidate probability mass. Together, these findings explain why reducing scalar residuals alone does not establish parameter-gradient denoising; the scalar and gradient audits use separate protocols.

Increasing continuation count supplies a direct denoising intervention, but its measured initial benefit does not translate into a positive mean training difference over 40 updates. The three-seed interval bounds that difference under the stated model; sustained auxiliary strength remains unverified. Frozen selection examines task value independently of training, yet neither future-cost selection nor its outcome control demonstrates an advantage over the student. These results concern single-token choices and leave open the value of longer reasoning segments and other return-to-go objectives (Xu et al., 2026a).

RS-OPSD makes a complementary design choice: outcomes select problems, and references supply token targets. Its positive final differences from OPSD coexist with negative mean online differences and a similar random-selection endpoint. Reference exposure therefore remains a plausible explanation, while a specific advantage from selecting failures is unresolved. The ordinary-OPD bridge further limits transfer of this result across distillation recipes.

A.1 Scope and unresolved questions

For sequence credit, the next tests are later-checkpoint gradient measurements, a validated auxiliary-update monitor, and held-out selection over longer decisions. Optimistic same-sample ceilings cannot substitute for those tests. For reference supervision, stronger attribution requires multiple seeds, matched token exposure and update magnitude, standard-divergence objectives, and separate controls for teacher context and CE targets. Missing step-0 evaluations, effective-adapter verification, and joint failure/truncation counts currently limit interpretation. Verified initialization-to-endpoint comparisons also require fixed decoding, complete outputs, scorer/version records, and overlap audits.

The data and compute scope also constrain interpretation. MATH-500 informed development and is not untouched confirmation data. M16 uses four times as many auxiliary continuations, and reference CE adds teacher-forced computation. Neither matched updates nor one sampled rollout establishes compute efficiency.

Guide to the appendices

The supporting material follows the main argument: objective definitions and shared settings precede gradient diagnostics, training comparisons, and held-out selection. Reference-supervision controls then examine the two roles of outcome feedback and reference targets. Table 2 provides direct links; the final appendix records artifact availability and exclusions. Legacy E1–E3, N1–N3, and P0 labels identify archived experiments, not additional methods.

B Gradient decomposition and the implemented objectives

The ideal sequence objective provides a starting point for understanding future-state credit. We derive its gradient, specify the implemented surrogates, and then show how candidate centering and detached reference selection alter the update. This order separates properties of the ideal objective from those of the losses used in the experiments.

B.1 Derivation for a pre-action state cost

Fix a problem and a finite horizon. Assume differentiable probabilities on a fixed support and sufficient integrability to interchange differentiation and expectation. Both trajectory sampling and

Table 2: Main-text questions and their supporting appendices. Each experimental appendix specifies the replication unit, calculations, and complete comparisons.

Appendix	Question and supporting material	Main-text link
A	How far do the findings extend? Discussion, evidence gaps, and limitations	Section 7
B	What does sequence credit differentiate? Gradient identity and detached reference gate	Sections 2, 4
C	What is optimized? Sampled-prefix loss, reference CE, and shared settings	Sections 4, 5
D	Is the signal large or noisy? Geometry, residuals, and repeated gradients	Sections 3.1, 3.2
E	Does estimator refinement improve learning? All training seeds and controls	Section 3.3
F	Do preferred tokens help? Held-out selection and soft reweighting	Section 3.4
G	Does reference selection help? OPSD probes, full-training random selection, and local controls	Sections 3.5, 5
H	Which outputs are usable? Recovery records and excluded evaluations	Section 5

the score function use π_θ; temperature or truncation must therefore be included in that distribution when applying the identity. Let $C_\theta(y) = \sum_t d_\theta(s_t)$, with any terminating trajectory extended by absorbing states of zero cost. The score identity yields

$$\nabla \mathbb{E}[C_\theta(y)] = \mathbb{E}\left[\nabla C_\theta(y) + C_\theta(y) \sum_i \nabla \log \pi_\theta(y_i \mid s_i)\right]. \tag{6}$$

For $t \le i$, $d_\theta(s_t)$ is measurable before sampling y_i, while $\mathbb{E}[\nabla \log \pi_\theta(y_i \mid s_i) \mid s_i] = 0$. Its contribution to that score term is therefore zero. Removing such terms gives Equation (2). A state baseline $b(s_i)$ can be subtracted from the future return for the same reason. It may reduce score-estimator variance, but $\arg\min_a (Q(s,a) - b(s)) = \arg\min_a Q(s,a)$, so it does not improve the underlying within-state ranking.

The chain-rule KL identity requires full conditional reverse KL and the same termination convention. It does not apply unchanged to top-K surrogates, sampled-action costs, or masked means. For $J = N(\theta)/D(\theta)$, $\nabla J = (\nabla N - J\nabla D)/D$; omitting a policy-dependent denominator drops a gradient term. Likewise, $\mathbb{E}[C/T] \neq \mathbb{E}[C]/\mathbb{E}[T]$ in general.

B.2 Implemented future-state auxiliaries

The empirical labels R_1, R_2 refer to implemented gradients, denoted $\widehat{R}_1, \widehat{R}_2$ here to distinguish them from Equation (2). The geometry and candidate implementations below differ in loss, reduction, and parameterization; neither gives an unbiased population sequence-credit estimate.

Sampled-token geometry probe. The early probe in Section 3.1 computes gradients in a rank-8 low-rank adaptation (LoRA) parameterization (Hu et al., 2022), differentiating

$$\mathcal{L}_{\text{geom}} = \frac{1}{T} \sum_{i=1}^{T} \operatorname{sg}\left[G_i - \bar{G}_{\text{bin}(i)}\right] \log \pi_\theta(y_i \mid s_i), \qquad G_i = \sum_{t=i+1}^{T} d_{\text{old}}(s_t). \tag{7}$$

Here d_{old} is the stored full-vocabulary reverse KL, T counts the retained response tokens, and $\bar{G}_{\text{bin}(i)}$ is the empirical mean future return within one of 20 relative-position bins in the diagnostic records. The recorded configuration uses discount 1, no return standardization, and no PPO ratio or clipping in this probe. The corresponding local loss is the mean full-vocabulary reverse KL on the same tokens. Gradients are computed without an optimizer update. The empirical position baseline is not asserted to be independent of the current sampled action; the probe measures the geometry of its stated loss, not an unbiased population gradient.

Candidate-based variance and training studies. For these candidate-based studies, the implemented auxiliary is a cost-minimizing clipped policy surrogate:

$$\mathcal{L}_{\text{aux}}(\theta) = \frac{1}{D_b} \sum_{s \in \mathcal{B}} \sum_{a \in \mathcal{A}_s} m_{s,a} w_{s,a} \max\{\rho_\theta(s,a) A_{s,a},\ \text{clip}(\rho_\theta(s,a), 1-\epsilon, 1+\epsilon) A_{s,a}\}. \tag{8}$$

Here $A_{s,a} = \mathrm{sg}[\widehat{Q}_s(a) - b(s)]$ is a centered future cost, $\rho_\theta = \pi_\theta(a \mid s)/\pi_{\mathrm{old}}(a \mid s)$, and $w_{s,a}$ is the detached old-policy probability renormalized over valid candidates. The mask $m_{s,a}$ retains valid active candidates; $D_b = \max(1, \sum_{s\in\mathcal{B}} \mathbf{1}[\sum_a m_{s,a} > 0])$ counts active prefixes, and $\epsilon = 0.2$. Only the current log probabilities are differentiated, giving $\widehat{R}_2 = \nabla_\theta \mathcal{L}_{\mathrm{aux}}$ before auxiliary weighting.

At each pivot s, four candidate tokens are forced separately and followed by M sampled continuations. Write $c_{s,a,j,h}$ for the detached local distillation cost at future position h of continuation j, $v_{s,a,j,h}$ for its valid-token mask, and $\mathcal{J}_{s,a}$ for the retained continuations with nonempty, finite windows. For a valid candidate and $j \in \mathcal{J}_{s,a}$, the H16 masked-mean return is

$$q_{s,a,j} = \frac{\sum_{h=1}^{16} \gamma^{h-1} v_{s,a,j,h} c_{s,a,j,h}}{\sum_{h=1}^{16} \gamma^{h-1} v_{s,a,j,h}}, \qquad \widehat{Q}_s(a) = \frac{1}{|\mathcal{J}_{s,a}|} \sum_{j\in\mathcal{J}_{s,a}} q_{s,a,j}, \quad \gamma = 0.95. \tag{9}$$

The forced token's own pre-action cost is excluded. At each future state s', the ordinary-OPD cost is $c(s') = \sum_{v\in S_{16}(s')} \widetilde{\pi}_{\mathrm{old}}(v \mid s')[\log \pi_{\mathrm{old}}(v \mid s') - \log p_\phi(v \mid s')]$, where S_{16} is student-top-16 and $\widetilde{\pi}_{\mathrm{old}}$ renormalizes its weights on that support. The log probabilities inside the difference remain full-model probabilities. This restricted-support cost is not the full-vocabulary divergence in Equation (1). Padding and masked positions do not contribute; a retained partial window is normalized by its observed discount-weight sum. For the primary variance and training matrices, a candidate requires a finite log probability and at least two valid continuations; a pivot requires all four valid candidates. Empty or nonfinite windows are invalid.

Pivot and candidate selection. The variance/training launcher selects one uniform pivot per response, up to 16 per batch. Eligible positions satisfy the response-length and response-fraction limits and leave context space for the forced token and future window. The trainer takes the four highest-probability eligible student tokens, excluding special, EOS/padding, and think/tool control tokens, and forces each to start a branch. The held-out selection study further restricts pivots to the response's first half. These rules follow the inspected code and launch records; historical source hashes are incomplete.

The raw baseline is zero. Policy and SW baselines are defined in Equation (17); training uses all retained continuations for its baseline, whereas the held-out scalar audit separates estimation and evaluation halves. For SW centering, a pivot also requires a finite probability-norm estimate and $\sum_a w_{s,a} u_a > \varepsilon_{\mathrm{fp32}}$. This check creates the extra inactive prefixes in the variance audit. The sensitivity u_a influences the baseline and this mask; it does not multiply the candidate's contribution to Equation (8).

The maximum in Equation (8) implements cost minimization with ratio bounds $[0.8, 1.2]$ and no additional advantage clipping. A global active-prefix denominator combines microbatches and workers. Appendix D.4 varies this denominator and the active masks while retaining candidate weights and costs; the ideal score identity need not hold for every configured decoder.

Future-divergence feedback ($\widehat{R}_2^{\mathrm{KL}}$) uses these later distillation costs. Historical verifier feedback ($\widehat{R}_2^{\mathrm{ver}}$) instead uses final-answer costs; its separate settings appear with those experiments.

B.3 What centering changes in an enumerated candidate loss

Fix a prefix, candidate support $\mathcal{A}_s$, and common validity mask. Evaluate at the collection parameters, where $\rho = 1$ lies inside the clipping interval. Detached weights satisfy $w_a = \pi(a \mid s)/Z_s$, so

$$\sum_{a\in\mathcal{A}_s} w_a \nabla \log \pi(a \mid s) = \frac{\sum_a \nabla \pi(a \mid s)}{Z_s} = \nabla \log Z_s. \tag{10}$$

Subtracting the same detached baseline from all candidate costs gives Equation (3). Any fixed projection orthogonal to $\nabla \log Z_s$ has the same centered and raw gradient for each continuation sample, hence the same conditional variance. The baseline can change both mean and variance in the remaining direction. The identity does not compare different masks or denominators, and need not apply once PPO clipping activates away from the collection parameters.

Classical score-function baselines minimize variance under an action-sampling distribution (Greensmith et al., 2004). Here all retained candidates are summed, and the SW logit-score proxy mini-

mizes a scalar weighted second moment; neither property guarantees minimum parameter-gradient covariance. The archived repeated-gradient audit has mean candidate mass 98.69% across 32 prefixes (minimum 87.73%). Mass alone does not determine $\|\nabla_\theta \log Z_s\|$, so this identity is not a quantitative attribution of the observed cosine or variance changes.

B.4 The reference auxiliary is an adaptive supervised semi-gradient

The reference loss $\mathcal{L}_{\text{ref}}^{\text{err}}$ differs from both action-credit auxiliaries: failed responses select problems for direct supervision on reference tokens, rather than assign scalar credit to the sampled actions. For a fixed (x, z), let $C_\theta(z \mid x)$ be the per-example reference CE in Equation (4), before gating and batch averaging, using the same $\max(1, L')$ denominator and empty-sum convention. Let $g(y)$ be the binary failure indicator and write $p_{\text{err},\theta}(x) = \mathbb{E}_{y\sim\pi_\theta}[g(y)]$. The hypothetical expected objective is

$$F(\theta) = \mathbb{E}_y[g(y)C_\theta(z \mid x)] = p_{\text{err},\theta}(x)C_\theta(z \mid x). \tag{11}$$

Its exact derivative contains two components:

$$\nabla F = p_{\text{err},\theta}(x)\nabla C_\theta(z \mid x) + C_\theta(z \mid x)\nabla p_{\text{err},\theta}(x). \tag{12}$$

The detached gate estimates the first component under the rollout policy and decoder, omitting differentiation through failure probability. The implemented update is therefore an adaptively selected supervised semi-gradient, rather than the full gradient of F or J. Teacher probabilities, sampled token IDs, masks, and reference targets are held fixed during the actor update.

The derivations identify what each empirical comparison must measure: cost estimates describe the feedback, parameter gradients describe its immediate direction and variability, and training or held-out selection tests its task value. The following appendices connect these quantities without treating one as a substitute for another.

C Experimental settings and reporting conventions

Table 3 maps each study to its model, protocol, and replication unit. We then specify the reference-supervision recipe and common reporting conventions; study-specific details accompany the results in Appendices D–G. Benchmark scopes differ: credit training and OPSD credit probes retain five benchmarks including Minerva, terminal-verifier training uses MATH-500/Minerva/Olympiad-Bench, and reference CE uses the post-hoc four-benchmark scope. These studies must be compared within their respective protocols.

Table 3: Protocol map. A repeated gradient, a new continuation, a data replicate, and a training seed measure different sources of variation. Frozen experiments take no optimizer steps. Model names abbreviate Qwen3; the 4B-Instruct teacher is the Instruct-2507 checkpoint.

Study	Model and parameter setting	Unit and scope
Gradient geometry	Sampled-token full suffix; 4B student / 8B teacher; rank-8 LoRA; frozen	256 problems; one gradient record per problem
Scalar residual audit	Four candidates; 4B / 4B-Instruct; frozen initial model, full parameters	112 retained pivots in 20 batches; cross-fitted scalar costs
Parameter-gradient variance	H16 candidate credit; 4B / 4B-Instruct; full parameters; frozen	32 fixed prefixes in two batches; 16 repeated estimates
Credit-training matrices	H16 candidate credit; 4B / 4B-Instruct; full-parameter updates on DeepMath	Three training seeds per matrix; 40 updates per run
Frozen candidate policies	H16 candidate cost; 4B / 4B-Instruct; no parameter updates	256 new problems in the M64 study; independent evaluation halves
Historical OPSD credit training	4B with same-checkpoint self-teacher; full-parameter updates	Seed 2026; two 40-update probes and one 93-update matched pair
Main reference study	4B or 8B with same-checkpoint self-teacher; rank-64 LoRA; OpenThoughts	One training seed per model size; 229 updates; four benchmarks
Single-update gate tests	Reference recipe at initialization and step 200	16 data replicates per starting point; one update per arm

Table 4: Protocols for the reference-supervision study. Both matched arms use the same settings within each column. Response limits are in tokens; K is the number of evaluated responses per question.

	Training	Online evaluation	Final evaluation
Mode	Non-thinking	Thinking	Thinking
Update / checkpoint	1–229	50/100/150/200	229
Response limit	1,024	8,192	38,912
Responses per question	One per visit	$K = 4$	$K = 12$

C.1 Main OPSD training and evaluation recipe

Tables 4 and 5 specify evaluation and training for the OPSD/RS-OPSD pairs. The random-reference control uses the same nominal recipe, with its selection rule defined in Appendix G.2.

Table 5: Training and evaluation settings for the 229-update OPSD/RS-OPSD pairs. Earlier diagnostic studies use the separate protocols specified with their results.

Component	Setting
Student / frozen teacher	Qwen3-4B or Qwen3-8B; same initial checkpoint
Trainable parameters	LoRA rank 64, alpha 128, all-linear
Precision / workers	BF16 / four GPUs; exact GPU model requires manifest audit
Data	29,427 prepared examples; 29,312 consumed, 115 dropped
Optimizer	AdamW, LR 5×10^{-6}, weight decay 0.01
Gradient clipping / batch	0.1 / 128
Training seed / steps	2026 / 229
Student generation	One response; non-thinking; max 1,024 tokens
Training sampling	Temperature 1.1; top-p 0.95; top-k 20
Sampled-prefix student prompt	Problem only; max 1,024 tokens
Teacher context	Problem + reference; reference max 6,144; total max 9,216
Distillation	Teacher top-64; support-renormalized clipped forward KL
Pointwise clipping	Upper bound 10^{-6} before summing contributions
Reference auxiliary	Uniform target-token CE; error threshold 0.5
Reference target	Max 1,024 tokens; EOS appended only when it fits
Auxiliary scale	$\lambda = 0.001$ for all steps; full-batch denominator
Online evaluation	Steps 50/100/150/200; Avg@4; max 8,192
Final evaluation	Step 229; Avg@12; max 38,912
Evaluation decoding	Thinking; temperature 1; top-p 0.95; top-k disabled
Other evaluation settings	min-p 0; presence penalty 0; normal EOS stopping

Observed training behavior. For 4B, mean training reward rose from 0.416 to 0.433, while sampled-response truncation fell from 51.7% to 49.2%. Approximately 9.8% of reference targets were truncated, so the auxiliary did not always expose a complete solution. At fixed decoding, increasing evaluation K improves sampling precision without changing expected accuracy; changing the length limit can change the response distribution itself.

C.2 Implemented sampled-prefix distillation loss

The reference-supervision comparisons hold the sampled-prefix loss fixed when adding CE. We call this implementation OPSD in the main text and *OPSD-c64* here to identify its teacher-top-64 support and pointwise cap. At each sampled response position, teacher and student probabilities are renormalized on the teacher's top-64 support $\mathcal{S}_{i,t}$:

$$\bar{p}_{i,t}(v) = \frac{p_{i,t}(v)}{\sum_{w \in \mathcal{S}_{i,t}} p_{i,t}(w)}, \qquad \bar{\pi}_{i,t}(v) = \frac{\pi_{i,t}(v)}{\sum_{w \in \mathcal{S}_{i,t}} \pi_{i,t}(w)}. \tag{13}$$

Pointwise upper clipping precedes summation:

$$\widetilde{d}_{i,t} = \sum_{v \in \mathcal{S}_{i,t}} \min\left\{ \bar{p}_{i,t}(v) \log \frac{\bar{p}_{i,t}(v)}{\bar{\pi}_{i,t}(v)}, c \right\}, \quad c = 10^{-6}, \qquad \mathcal{L}_{\text{sampled}} = \frac{\sum_{i,t} m_{i,t} \widetilde{d}_{i,t}}{\sum_{i,t} m_{i,t}}. \tag{14}$$

Here $m_{i,t}$ selects valid sampled-prefix tokens; teacher probabilities and support indices are detached. The displayed ratio assumes at least one valid token; an all-masked batch contributes zero. Renormalization discards the student's total mass on the retained support, matching only relative probabilities within it. A contribution above c has zero derivative through the cap; negative contributions remain uncapped, so the sum can be negative. This is a pointwise-capped, conditional-support forward-KL surrogate, not a standard KL divergence or the original full-vocabulary OPSD objective. Holding it fixed identifies the effect of adding CE within this recipe, not transfer to other sampled-prefix objectives.

Gradient of the capped surrogate. For support logits ℓ_j, let $U = \{v : \bar{p}_v \log(\bar{p}_v/\bar{\pi}_v) < c\}$. Away from cap boundaries,

$$\frac{\partial \widetilde{d}}{\partial \ell_j} = \bar{\pi}_j \sum_{v \in U} \bar{p}_v - \mathbf{1}[j \in U] \bar{p}_j. \tag{15}$$

A capped term has no direct derivative, but uncapped terms still affect all support logits through normalization. For teacher $(0.9, 0.1)$ and student $(0.5, 0.5)$, only the second term remains uncapped and the logit gradient is $(0.05, -0.05)$; descent increases the teacher-disfavored second probability. This example demonstrates a possible direction, not its frequency in training. The inspected implementation in `compute_opsd_topk_token_divergence` renormalizes log probabilities and clamps individual terms before summing, matching Equation (14); it does not independently verify every historical run's source version.

Scope of the reference comparison. The ordinary-OPD bridge reaches 77.320% macro accuracy, 5.26 pp above this OPSD implementation, and 77.036% with reference CE (-0.284 pp; Appendix G.5). Because it changes both teacher and objective to student-top-16 weighted supervision, the bridge cannot isolate clipping. Reference CE may compensate for the capped objective; a benefit across distillation objectives remains unestablished.

C.3 Reference processing and optimizer update

Reference targets are limited to 1,024 tokens, with EOS appended only when it fits. CE averages valid tokens within each example and then over the full batch, assigning zero to empty targets. Its contribution therefore depends on both the selected fraction and the active-example mean loss. Student and teacher tokenization must be compatible.

Each update caches one student response and its verification gate, evaluates the frozen teacher and student on sampled prefixes, and computes student CE on the selected reference prefixes. Teacher scores, tokens, and gates are detached. The two losses are combined before microbatch accumulation, gradient clipping, and one AdamW step; correct responses remain in the reference loss's batch denominator. Legacy activity fields measure target availability or padded-token activity, so they cannot recover failure coverage or selected-token totals.

C.4 Experimental units and uncertainty

Paired training seeds are the replication unit for the short-training matrices. Seed-t intervals use two degrees of freedom and Bonferroni adjustment over three contrasts. Continuations measure conditional gradient or policy variability; one-update data replicates are not training seeds. Captions specify the relevant units. Macro averages weight benchmarks equally; aggregate historical reference logs cannot supply prompt-paired intervals.

C.5 Sampling labels and loss reductions

Table 6 distinguishes continuation counts per estimate from total cross-fitting pools.

Table 6: Diagnostic vocabulary. Sampling counts are per candidate token, not training batches or optimizer updates.

Term or label	Meaning
Pivot	A next-token decision whose preceding student prefix is held fixed.
Candidate	A possible token forced at the pivot before sampling a suffix.
Continuation	One sampled suffix after that candidate token.
H16	Up to sixteen future-cost positions after the forced token; distinct from the full response limit.
M4 / M16, gradient studies	Four / sixteen continuations per candidate in one gradient estimate.
M16, historical selection	A complete pool of sixteen continuations per candidate, split into two halves of eight.
M64	The frozen selection pool per candidate, split into 32 estimation and 32 evaluation continuations.
Soft-reweighting M	Estimation continuations per candidate; the evaluation half stays fixed at 32.

A *masked mean* divides the discounted cost sum by the observed token-weight sum. *Active-count normalization* instead divides the batch auxiliary by retained prefixes. The common-denominator controls isolate the latter.

C.6 Answer verification and scoring provenance

The 4B/8B reference pairs and random control use `ttrl_math.reward_func`, with format reward disabled. Binary correctness supplies both the failure gate and evaluation score.

The available scorer extracts the final `\boxed{}` answer; missing or incomplete boxes score zero. It checks normalized strings, then SymPy equivalence (including tuples and fractions), then Math-Verify when `fast=False`. SymPy and Math-Verify limits are five and one seconds; final fallback parsing failures/timeouts are non-matches, while exceptions reaching the wrapper are raised. No separate universal numerical tolerance is configured.

Historical scorer hashes and dependency versions are incomplete, so the current implementation does not certify identical historical parsing or a scorer-error rate. Truncation does not by itself prove failure: a truncated response could already contain a complete answer. Without joint truncation/failure records, ratios of marginal rates cannot identify gate composition. The random gate matches counts and reference-length bins, not truncation status.

D Gradient geometry and variance diagnostics

We first characterize gradient scale, then examine how centering affects scalar residuals and repeated parameter gradients. These are separate frozen protocols (Figure 1a,b, c, and d–f); only the last estimates conditional parameter variance by resampling futures at fixed prefixes. Its matched controls isolate normalization before testing continuation averaging as a direct denoising intervention.

D.1 Across-problem gradient geometry

The early non-thinking diagnostic used Qwen3-4B as student and Qwen3-8B as teacher. It computed one gradient record per problem in a rank-8, alpha-16 LoRA space, with a 4,096-token response limit. All 256 problem records and 213,635 tokens were retained in the data audit. Figure 6a shows the exact empirical distributions without smoothing or subsampling. Across-problem SD uses denominator n, matching the original summary; it is not a training-seed confidence interval. For problem i, the unweighted norm ratio and directional cosine are

$$\kappa_i = \frac{\|\widehat{R}_2^{(i)}\|_2}{\|\widehat{R}_1^{(i)}\|_2}, \qquad c_i = \frac{\langle \widehat{R}_1^{(i)}, \widehat{R}_2^{(i)} \rangle}{\|\widehat{R}_1^{(i)}\|_2 \, \|\widehat{R}_2^{(i)}\|_2}. \tag{16}$$

The unweighted sum $\widehat{R}_1 + \widehat{R}_2$ had mean cosine 0.997 to $\widehat{R}_2$ and 0.051 to $\widehat{R}_1$. These per-problem comparisons quantify how the larger sequence gradient dominated the combined direction.

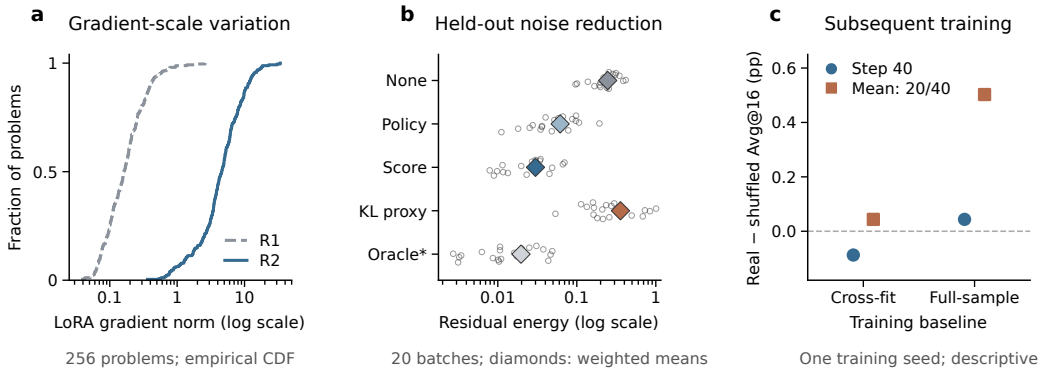

Figure 6: Gradient geometry, scalar residuals, and earlier training comparisons. (a) All 256 per-problem LoRA gradient norms. (b) All 20 frozen-batch residual-energy records and valid-pivot-weighted means; the oracle uses the evaluation half. (c) Real minus shuffled question-weighted Avg@16 in single-seed training probes. The panels use separate development protocols; scalar energy in (b) is distinct from parameter-gradient variance in Figure 7.

Figure 1a retains all norm ratios on a log axis, while panel (b) uses 20 bins on $[-1, 1]$. These displays describe across-problem geometry; the following studies ask how estimates vary when continuations are resampled.

Table 7: Raw per-problem LoRA gradient diagnostics. Norms are in the recorded loss scale, before an auxiliary training coefficient. These observations mix problem heterogeneity with stochastic variation.

Metric	Mean	Across-problem SD
$\widehat{R}_1$ gradient norm	0.224770	0.247597
$\widehat{R}_2$ gradient norm	5.720515	4.521992
Per-problem norm ratio $\widehat{R}_2/\widehat{R}_1$	31.168585	18.021884
$\widehat{R}_1$–$\widehat{R}_2$ gradient cosine	0.001894	0.195487
Top-1% absolute surrogate contribution	0.338706	0.121448

The token-tail statistic is the largest 1% of absolute $A_t \log \pi(y_t \mid s_t)$ contributions divided by their total, averaged across problems; it is not a token-gradient norm. Across-problem norm ratios likewise do not estimate conditional signal-to-noise.

D.2 Cross-fitted scalar residual energy

To measure the scalar effect of centering, the residual audit uses Qwen3-4B and Qwen3-4B-Instruct-2507 with LoRA rank zero and learning rate zero, keeping the initial model parameters fixed. Each frozen batch requested eight pivots, with four candidate actions and four continuations per candidate. The report retained 112 valid pivots out of 160 requested (mean valid fraction 0.70), over 20 batches. Candidate and pivot validity were determined by finite costs, available samples in both halves, at least two valid candidates, and a non-degenerate score measure. All recorded batches and all five baseline comparators are shown; no outcome-based filtering was introduced for the figure.

For a fixed prefix, abbreviate $p_a = \pi_\theta(a \mid s)$ and write $w_a = p_a / \sum_{a' \in \mathcal{A}_s} p_{a'}$ for its candidate-normalized weight. Using continuation costs $\widehat{Q}_a^{\text{est}}$ estimated from one half, the two centering baselines are

$$b_{\text{policy}}(s) = \sum_a w_a \widehat{Q}_a^{\text{est}}, \qquad b_{\text{SW}}(s) = \frac{\sum_a w_a u_a \widehat{Q}_a^{\text{est}}}{\sum_a w_a u_a}. \tag{17}$$

The independent half evaluates the weighted squared residual

$$\mathcal{E}_s(b) = \frac{\sum_a w_a u_a \big(\widehat{Q}_a^{\mathrm{eval}} - b(s)\big)^2}{\sum_a w_a u_a}. \tag{18}$$

Both sums run over valid candidates $\mathcal{A}_s$. The implementation uses the nonnegative logit-score proxy

$$u_a = \max\left\{0, 1 - 2p_a + \widehat{\sum_v p_v^2}\right\}. \tag{19}$$

With the exact probability squared norm, the expression inside the maximum equals $\|\nabla_{\ell_s} \log \pi_\theta(a \mid s)\|_2^2$, where ℓ_s is the vector of next-token logits. The stored support approximation does not include the model's parameter Jacobian. Each $\widehat{Q}$ is the mean of valid continuation costs within its half. The two cross-fit directions are averaged within prefix, then aggregated with valid-prefix weights across batches. Writing the resulting energy as $\overline{\mathcal{E}}(b)$, its reduction relative to raw credit is $1 - \overline{\mathcal{E}}(b)/\overline{\mathcal{E}}(0)$. Baselines are zero, candidate-policy mean, sensitivity-weighted mean, current-state KL, and an evaluation-half score optimum. The last is an in-sample diagnostic lower bound, not a deployable baseline.

Table 8: Held-out baseline audit. Energy and P99 are valid-pivot-weighted summaries of batch records; P99 is not a newly pooled residual quantile. Reduction is one minus the ratio of aggregate energies, not the average of per-pivot reduction ratios.

Baseline	Residual energy	Reduction (%)	Residual P99
Zero	0.24609	0.00000	1.12347
Policy mean	0.06157	74.98238	0.71024
Sensitivity-weighted	0.03000	87.80939	0.57485
Current-state KL	0.35694	−45.04242	1.26812
Evaluation-half oracle	0.01967	—	—

SW reduces the scalar second moment it targets by 87.81%, but this metric excludes the model's parameter Jacobian. The following audit measures parameter covariance directly. Its different prefixes and sampling protocol preclude treating the residual reduction and the 83.5% variance increase as paired effects of one intervention.

D.3 Repeated parameter gradients and applied strength

We now hold prefixes fixed and resample continuations to measure noise in the implemented parameter gradient. With $R = 16$ resamples at each of 32 prefixes, the mean gradient and covariance-trace estimator are

$$\bar{g}_{s,b} = \frac{1}{R}\sum_{r=1}^{R} g_{s,b}^{(r)}, \qquad \widehat{V}_{s,b} = \frac{1}{R-1}\sum_{r=1}^{R} \|g_{s,b}^{(r)} - \bar{g}_{s,b}\|_2^2. \tag{20}$$

The trace sums variance across parameters; $\bar{V}_b$ averages it over all 32 prefixes, including inactive zero contributions. For a given baseline, we abbreviate $g_{s,b}^{(r)}$ as $g_s^{(r)}$. Each vector includes the mask and batch denominator in Equation (21); it is a prefix's contribution to the batch gradient, not a separately normalized one-prefix update.

D.3.1 Fixed-prefix protocol and estimator

This frozen audit uses a Qwen3-4B student and a Qwen3-4B-Instruct-2507 teacher. It measures full-parameter gradients with fully sharded data parallel (FSDP) backpropagation and takes no optimizer steps. The design contains two batches of 16 prefixes, four candidate actions per prefix, H16 future costs, and 64 continuations per candidate. Consecutive groups of four produce 16 repeated estimates shared by the raw, policy-centered, and SW-centered estimators. The stored outputs contain 1,536 prefix-level and 96 full-batch repeated gradient norms, totaling 1,632 scalars.

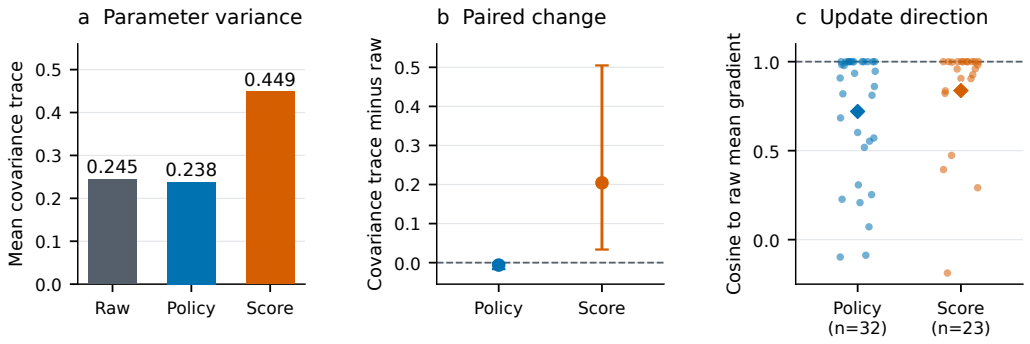

Figure 7: Direct gradient measurements distinguish centering from variance reduction. (a) Mean conditional covariance trace over 32 fixed prefixes. (b) Paired change from raw, with descriptive 95% prefix-bootstrap intervals conditional on two fixed batches. (c) Per-prefix mean-gradient cosine to raw; diamonds denote means. SW centering has nine zero-gradient prefixes, retained in (a,b); their undefined cosines are omitted in (c). All methods use the same continuation groups at the initial checkpoint and the implemented baseline-specific active-pivot normalization.

Baseline definitions follow Equation (17). Each prefix contribution includes its validity mask and batch denominator. Raw and policy retain all 32 prefixes; SW retains 23. All-prefix variance includes the nine zero contributions, whereas undefined cosines are omitted. Figure 1d orders these prefixes by raw variance and shades observed differences, not uncertainty; its symlog axis is linear below 0.1 to retain zero SW gradients.

Table 9: Repeated parameter-gradient measurements (N1). Variance averages all 32 prefixes; changes are paired against raw. Descriptive 95% intervals use 4,000 prefix-bootstrap resamples (seed 7181), conditional on the two fixed batches. Ratios divide summed variances rather than averaging per-prefix ratios.

Baseline	Mean variance	Ratio to raw	Change	95% interval
Raw	0.244548	1.000000	0.000000	—
Policy	0.238362	0.974703	−0.006186	$[-0.016473, -0.000520]$
SW	0.448858	1.835456	+0.204309	$[0.033594, 0.504802]$

The average mean-gradient cosines to raw are 0.720274 for policy ($n = 32$) and 0.837559 for SW ($n = 23$). These are arithmetic averages of prefix cosines, not the cosine between two globally pooled gradients. The full-batch measurements are computed separately, preserving cross-prefix gradient covariance.

D.3.2 Gradient variability versus the applied coefficient

For each complete batch, let $G^{(r)}$ be the repeated auxiliary gradient, with mean $\bar{G}$ and covariance trace V_G. The corresponding full-batch sampled-prefix gradient is denoted G_{local} in both ratios. We report relative mean strength $\lambda\|\bar{G}\|/\|G_{\text{local}}\|$ and root-mean-square (RMS) noise $\lambda\sqrt{V_G}/\|G_{\text{local}}\|$, with $\lambda = 0.0015$. Table 10 expresses these pre-optimizer gradient ratios as percentages, separately from AdamW parameter increments.

These measurements reveal a weak initial auxiliary at the applied coefficient. Before testing training transfer, we determine whether SW's variance increase comes from normalization and whether additional continuations reduce variance directly. Later-checkpoint variance remains unmeasured.

Table 10: Actual complete-batch auxiliary strength, expressed as percent of the sampled-prefix $\widehat{R}_1$ gradient norm. Each row is one measured batch, with no inferred between-seed uncertainty.

Batch	Baseline	Weighted mean (%)	Weighted noise RMS (%)
1	Raw	0.057510	0.032776
1	Policy	0.057424	0.032642
1	SW	0.084215	0.047855
2	Raw	0.094099	0.080764
2	Policy	0.067927	0.079498
2	SW	0.093696	0.106584

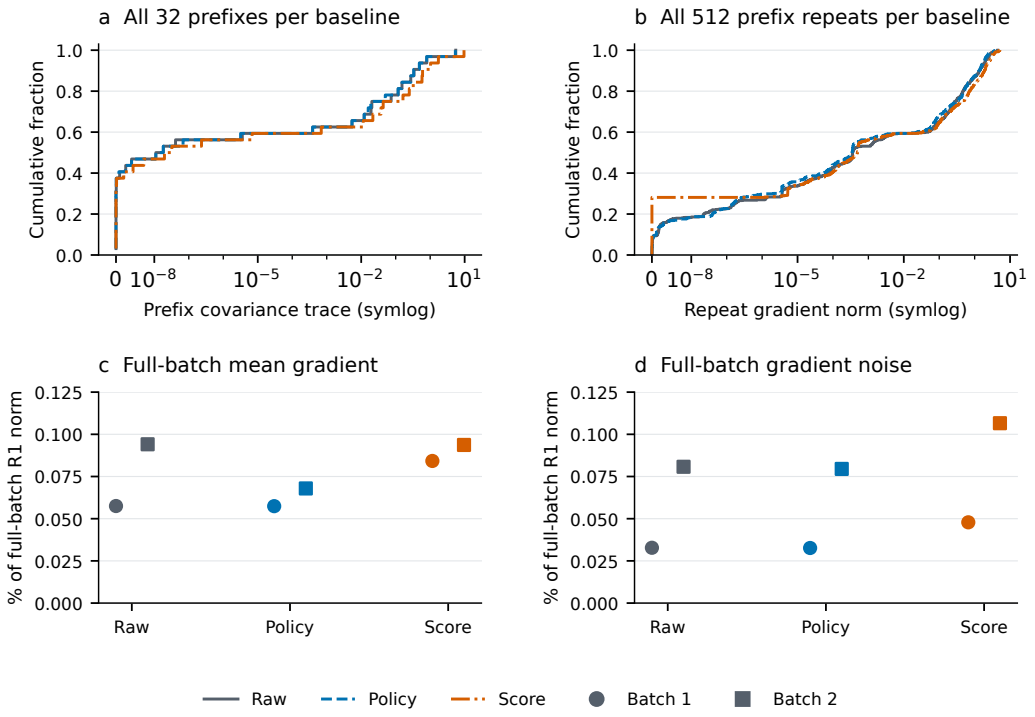

Figure 8: Complete N1 distributions and applied gradient strength. (a) Covariance-trace ECDF for all 32 prefixes per baseline. (b) ECDF of all 512 prefix-level repeated gradient norms per baseline. Symmetric-log axes preserve zero contributions and the full range without tail removal; the linear region extends to 10^{-8}. (c,d) Separately measured complete-batch mean and noise contributions, weighted by $\lambda = 0.0015$ and divided by the same batch's sampled-prefix gradient norm. Circle/square markers identify the two batches.

D.4 Common-denominator and common-support controls

Because SW changes which prefixes remain active, its apparent variance increase could arise from normalization rather than noisier costs. The follow-up tests this explanation using 32 fixed prefixes in two batches, four candidate actions per prefix, and 16 disjoint groups of four continuations per candidate. Each prefix and full batch was evaluated under nine conditions: raw, policy-centered, and SW-centered gradients, each with (i) its own active-count denominator, (ii) the full-batch denominator of 16 while retaining its own mask, or (iii) the common active support and the same full-batch denominator. Common support intersected the three baseline masks over all 16 repeats. It retained 23 prefixes; the other nine contributed zero and remained in all-prefix averages. Candidate weights and advantages were not re-estimated after intersection.

Let $h_{s,b}^{(r)}$ be the prefix gradient before batch normalization and $m_{s,b}^{(r)} \in \{0,1\}$ its active mask. The contribution measured in Equation (20) is

$$g_{s,b}^{(r)} = \frac{m_{s,b}^{(r)} h_{s,b}^{(r)}}{D_b^{(r)}}, \qquad D_b^{(r)} = \max\left(1, \sum_{s \in \mathcal{B}} m_{s,b}^{(r)}\right), \tag{21}$$

where $\mathcal{B}$ is the fixed batch of prefixes. The common-denominator control replaces $D_b^{(r)}$ with $|\mathcal{B}| = 16$; common support additionally replaces each mask with their intersection across baselines and repeats. Reported variance ratios divide the mean variance of each condition by its matched raw mean, rather than averaging per-prefix ratios.

Table 11: Normalization/support controls at initialization. Covariance traces and mean-gradient norms average all 32 prefixes; each variance ratio uses raw under the same normalization/support regime. Cosine is the mean per-prefix cosine to the original raw gradient, computed only where defined.

Regime	Baseline	Variance	Ratio	Mean norm	Cosine
Own denominator	Raw	0.244539	1.0000	0.298669	1.0000
	Policy	0.238347	0.9747	0.261100	0.7203
	Score	0.448832	1.8354	0.376028	0.8375
Common denominator	Raw	0.244515	1.0000	0.298675	0.9997
	Policy	0.238333	0.9747	0.261095	0.7203
	Score	0.241007	0.9857	0.269060	0.8385
Common support	Raw	0.244539	1.0000	0.298682	0.9997
	Policy	0.238337	0.9746	0.261092	0.8341
	Score	0.240993	0.9855	0.269070	0.8385

SW/raw variance ratios were 1.8354, 0.9857, and 0.9855 under own-count, common-denominator, and common-support conditions. Full-batch ratios fell from 2.1320/1.7415 to 1.0074/0.9782 with the common denominator. Intersecting support reduced raw/policy coverage from 32 to 23 prefixes but added little variance change.

Corresponding pointwise 95% paired-prefix bootstrap intervals for the three prefix-level ratios were $[1.7525, 2.1425]$, $[0.9762, 1.0328]$, and $[0.9761, 1.0325]$. These intervals use 10,000 resamples with seed 91601 and are descriptive conditional on the two batches, not training-seed inference. All 12 planned diagnostic contrasts are retained in the source summary; intervals are not multiplicity-adjusted.

D.5 Continuation count and conditional gradient variance

The matched-normalization results do not establish substantial denoising from centering. To obtain a direct denoising intervention, the frozen M-scaling audit instead increases continuation count in two batches of 16 prefixes at initialization. Each of four candidate actions received a pool of 256 continuations, split into 16 disjoint blocks of 16. In each repeat, raw M4 used the first four continuations of the block and raw M16 used all 16. Both estimators retained the same coefficient, candidate weights, clipping, and active-pivot normalization. Here M denotes continuations per gradient estimate, unlike the total-pool labels in the earlier cross-fitted E1 experiments. The audit stored all 1,088 repeated gradient norms across 32 prefixes, two full batches, two conditions, and 16 repeats, with no optimizer updates. Each estimate is $\widehat{Q}_{s,M}(a) = M^{-1} \sum_{m=1}^{M} C_s^{(m)}(a)$, where $C_s^{(m)}(a)$ is the recorded future cost after candidate a. The reported variance reduction is $1 - \bar{V}_{\mathrm{M16}}/\bar{V}_{\mathrm{M4}} = 0.803$ after rounding.

Full-batch M16/M4 variance ratios were 0.2046 and 0.1938. With $\lambda = 0.0015$, M4/M16 mean-gradient strengths were 0.05628/0.05532% of the sampled-prefix gradient norm in batch 1 and 0.09094/0.09387% in batch 2. Their corresponding noise RMS ratios were 0.04279/0.01936% and 0.08209/0.03613%. Thus, the observed reduction is not explained by a smaller auxiliary coefficient

Table 12: Raw M4/M16 gradient measurements over 32 fixed prefixes. All prefixes are active. The M16/M4 variance ratio is 0.197 (descriptive 95% interval $[0.185, 0.279]$); mean cosine is relative to the M4 estimate at the same prefix.

Estimator	Mean variance	Mean norm	Cosine	Active prefixes
M4	0.291544	0.293774	1.000000	32/32
M16	0.057552	0.291523	0.989436	32/32

Table 13: Continuation averaging and centering at the original coefficient. Entries are paired step-40 differences in five-benchmark macro Avg@12 (pp); positive values favor the first arm. Rows summarize separate three-seed matrices. Training seeds, not sampled answers, are the replication units. Full endpoints and intervals appear in Appendices E.1 and E.3.

Comparison	Seed 2026	Seed 2027	Seed 2028	Mean
More sampling: M16 minus M4	-0.95	$+0.06$	$+0.53$	-0.12
Centering: SW minus raw	$+0.91$	$+0.74$	$+0.98$	$+0.88$

or a comparable collapse of mean-gradient magnitude. These pre-optimizer measurements do not establish equal AdamW increments or the same variance reduction at later checkpoints.

E TRAINING TRANSFER OF FUTURE-STATE CREDIT

The frozen audits establish how the estimators differ at initialization; this appendix tests whether those differences improve learning. We compare continuation counts at the original and a calibrated larger coefficient, then examine centering with an action-shuffling control. Each matrix uses three paired training seeds and is analyzed separately. Earlier refinement and verifier studies provide development context after the primary comparisons.

E.1 CONTINUATION COUNT: MATCHED THREE-SEED TRAINING

We compared local-only OPD, raw M4 credit, and raw M16 credit using Qwen3-4B, a frozen Qwen3-4B-Instruct-2507 teacher, full-parameter FSDP training on four H200 GPUs, and seeds 2026/2027/2028. Each arm ran for 40 updates on the same prepared DeepMath stream: batch size 16, actor microbatch size 1, learning rate 10^{-6}, weight decay 0.01, and gradient clipping 1.0. The auxiliary coefficient was 0.0015 for M4/M16 and zero for local-only, which retained M16 branching overhead. Future costs used H16, discount 0.95, four candidates, and a masked mean. This ordinary-OPD recipe is separate from the LoRA OPSD reference recipe in Table 5.

Training responses allow 7,168 tokens. Step-20/40 evaluation uses non-thinking Avg@12, temperature 1, top-$p = 0.95$, and 8,192 tokens. A 1,024-token prompt filter retains 1,088/1,090 questions: AIME24/AIME25/AMC23/Minerva/Olympiad-Bench counts 30/30/83/272/673. Two overlength Olympiad-Bench questions are excluded, not scored as incorrect. Output identity, hashes, and sample-coverage checks are summarized in Appendix H.2.

The completed matrix contains nine runs, 18 validated endpoint files, and 235,008 responses, including one replacement run documented in Appendix H.2.

Final M16-minus-M4 differences are $-0.952, +0.064, +0.534$ pp for seeds 2026–2028 (Table 15). The intervals still allow the prespecified $+0.5$ pp benefit, and three seeds cannot reliably test the normality assumption. Step 20 and the two-checkpoint mean are secondary endpoints. Because M16 uses four times as many auxiliary requests, the comparison matches updates rather than compute. The small initial auxiliary contribution motivates the higher-coefficient test below.

Table 14: Complete matched-training endpoints: five-dataset macro Avg@12 (%) with a non-thinking 8,192-token response budget. Step 40 is the primary endpoint; step 20 is secondary. Each row pairs arms by training seed.

Seed	Step	local-only	Raw M4	Raw M16
2026	20	40.502	40.285	40.517
2026	40	41.052	41.567	40.614
2027	20	40.645	40.343	40.353
2027	40	40.641	41.347	41.411
2028	20	39.980	40.632	40.237
2028	40	41.454	40.397	40.931

Table 15: Primary step-40 differences for the continuation-count comparison (pp). Intervals use paired training-seed differences ($n = 3$, t distribution with two degrees of freedom). The last column controls the three-comparison family by Bonferroni adjustment.

Comparison	Mean difference	Seed-t 95% CI	Family-wise 95% CI
M16 − M4	-0.118	$[-2.006, 1.770]$	$[-3.473, 3.237]$
M16 − local-only	-0.064	$[-1.860, 1.732]$	$[-3.256, 3.128]$
M4 − local-only	$+0.054$	$[-2.349, 2.457]$	$[-4.218, 4.326]$

E.2 Higher-coefficient training: calibration and results

We next calibrate a larger coefficient to increase the auxiliary's initial contribution, then repeat the M4/M16 comparison. Calibration uses gradients and optimizer increments, independently of validation accuracy.

E.2.1 Snapshot and coefficient selection

The larger-coefficient matrix uses the same full-parameter OPD recipe, three seeds, and 40-update endpoint. All nine runs and 18 endpoint files are complete (235,008 responses), without substituting results from the earlier coefficient matrix. Appendix H.2 describes the common output checks.

Calibration used the first 32 problems of a fixed development bank, two batches of 16, and sixteen repeated gradients per estimator. M4 and M16 shared candidate pools of 256 continuations. Let G_{local} be the complete-batch local gradient and $\bar{G}_{\text{sequence}}$ the mean repeated auxiliary gradient. At a fixed model and AdamW state, the calibration measures

$$\alpha(\lambda) = \frac{\lambda \|\bar{G}_{\text{sequence}}\|_2}{\|G_{\text{local}}\|_2}, \qquad \rho(\lambda) = \frac{\|U(G_{\text{local}} + \lambda \bar{G}_{\text{sequence}}) - U(G_{\text{local}})\|_2}{\|U(G_{\text{local}})\|_2}. \tag{22}$$

The ratio α measures pre-optimizer auxiliary strength; ρ measures the relative change in the parameter increment. The map U applies global gradient clipping and the FP32 mathematical AdamW update, including moments, bias correction, and weight decay. The quantity uses the mean estimator gradient, not the mean of independently sampled optimizer updates or BF16 stored-parameter differences.

The fixed coefficient grid was $\{0.0015, 0.005, 0.015, 0.05, 0.15, 0.5\}$. Selection chose the smallest coefficient meeting both targets in both batches and both M conditions: $\rho \in [0.05, 0.30]$ and weighted gradient ratio in $[0.01, 0.25]$. These are operational design thresholds, not statistical tests. Only $\lambda = 0.05$ passed; validation accuracy was not used. Its calibrated update ratios ranged from 12.63% to 18.94%, with weighted gradient ratios from 1.84% to 3.13%. Full-batch M16/M4 variance ratios were 0.2046 and 0.1937. These measurements support the calibration decision at initialization, not sustained effective strength during training.

Table 16: Completed higher-coefficient matrix. Five-benchmark macro Avg@12 (%) uses non-thinking decoding and 8,192 tokens. M4 and M16 use $\lambda = 0.05$; local-only uses zero. All three seeds and both planned evaluation checkpoints are shown. The final row averages the three step-40 endpoints.

Seed	Step	local-only	M4	M16
2026	20	40.448	40.219	40.550
2026	40	41.720	41.593	40.708
2027	20	40.431	40.196	40.426
2027	40	41.351	42.479	41.912
2028	20	39.952	41.036	40.216
2028	40	41.410	41.259	40.730
Mean	40	41.494	41.777	41.117

E.2.2 Training outcomes

Step-40 mean macro Avg@12 is 41.494% (local-only), 41.777% (M4), and 41.117% (M16). M16 trails M4 in every seed; other contrasts vary in sign. The M16-minus-M4 interval includes zero but its upper bound $+0.20$ pp excludes the prespecified $+0.5$ pp benefit under the seed-t model. Per-dataset and question-weighted results remain in the source data.

Table 17: Paired step-40 contrasts for the coefficient-0.05 matrix (pp). Seeds, not sampled answers, are the replication units. Intervals use the same seed-t procedure and Bonferroni adjustment over three contrasts as the earlier matrices. Figure 2 (a) displays every M16-minus-M4 seed difference at both coefficients.

Contrast	2026	2027	2028	Mean	Family-wise 95% CI
M16 minus M4	-0.885	-0.567	-0.529	-0.660	$[-1.525, +0.204]$
M16 minus local-only	-1.012	$+0.561$	-0.680	-0.377	$[-4.039, +3.284]$
M4 minus local-only	-0.127	$+1.128$	-0.152	$+0.283$	$[-2.948, +3.514]$

E.2.3 Zero-coefficient validation of the update monitor

To assess whether the calibrated contribution persists, the monitor compares hypothetical AdamW updates from recomputed local-only and production gradients at the same parameter/optimizer state. Its zero-coefficient control nevertheless reads 5.08–5.29% at step 1, versus 7.35–16.26% in auxiliary arms (Table 18), and both readouts later fall. This failed control prevents interpreting the monitor as auxiliary-only strength.

Table 18: Logged production-versus-recomputed-local AdamW discrepancy, expressed as a percentage of the recomputed-local update norm. These are monitor outputs, not validated auxiliary-only contributions. local-only is the zero-coefficient control; all completed runs are shown.

Seed	Arm	Step 1	Step 20	Step 40
2026	local-only	5.176	0.194	0.176
2026	M4	8.781	0.736	0.941
2026	M16	10.259	0.230	0.807
2027	local-only	5.287	0.201	0.163
2027	M4	8.060	1.141	0.333
2027	M16	7.355	0.614	0.872
2028	local-only	5.077	0.196	0.325
2028	M4	16.255	0.721	0.500
2028	M16	12.994	1.057	0.873

Resolving this discrepancy requires matching production and recomputed gradients at zero coefficient, then separating both losses within the same forward pass. Subtracting norms across diverged trajectories cannot fix it. Training endpoints remain valid observations, but sustained auxiliary update strength is unverified.

E.3 SW CENTERING AND ACTION CORRESPONDENCE

Unlike continuation averaging, centering also changes the mean-gradient direction. We therefore include shuffled credit to test whether its action correspondence contributes to training. The matrix contains 12 complete runs, including one replacement documented in Appendix H.2.

This experiment compared local-only, raw sequence credit, SW-centered sequence credit, and a shuffled-advantage control at seeds 2026/2027/2028. All arms computed four continuations for each of four candidate actions. The local-only coefficient was zero; the other arms used $\lambda = 0.0015$. The shuffled control permuted the final SW-centered advantages among valid candidates without recomputing the baseline, preserving their values while disrupting action correspondence. This does not guarantee equal weighted parameter-gradient norms. The experiment used the same student, teacher, DeepMath stream, optimizer settings, and 40-update evaluation protocol described in Appendix E.1, but the earlier inference-batching profile. Comparisons are made within each matrix rather than pooling runs across profiles.

Table 19: Three-seed centering and action-shuffling matrix (N2). Values are five-dataset macro Avg@12 (%) with non-thinking decoding and an 8,192-token budget. Raw, SW-centered, and shuffled arms share the nominal auxiliary coefficient.

Seed	Step	local-only	Raw	SW	Shuffled
2026	20	40.361	40.593	40.100	40.455
2026	40	40.438	41.129	42.044	40.905
2027	20	40.108	39.897	40.283	40.331
2027	40	41.216	40.616	41.361	41.106
2028	20	40.006	40.286	40.786	40.492
2028	40	41.033	40.338	41.316	41.307

Table 20: N2 step-40 contrasts (pp). The first three were planned; their intervals use paired training seeds ($n = 3$; t, two degrees of freedom), with Bonferroni adjustment across those three comparisons. The last row is a post-hoc descriptive comparison; no additional test is claimed.

Comparison	Mean difference	Seed-t 95% CI	Family-wise 95% CI
SW − Raw	$+0.879$	$[0.579, 1.179]$	$[0.346, 1.412]$
SW − local-only	$+0.678$	$[-1.327, 2.682]$	$[-2.885, 4.241]$
SW − Shuffled	$+0.467$	$[-1.009, 1.944]$	$[-2.158, 3.092]$
Raw − local-only (post-hoc)	-0.201	—	—

SW-minus-raw differences are $+0.915, +0.745, +0.978$ pp, with an adjusted interval above zero; SW-minus-local-only and SW-minus-shuffled intervals include zero (Table 20). Post-hoc raw-minus-local-only differences are $+0.691, -0.600, -0.696$ pp (mean -0.201). Shuffled-minus-raw is approximately $-0.22, +0.49, +0.97$ pp (mean $+0.41$), consistent with explanations involving centered values rather than their action assignment. The initial variance measurements cannot explain the SW gain as denoising; mean-gradient changes remain an alternative. All seed-level inferences rest on three replications.

E.4 EARLIER BASELINE-REFINEMENT STUDIES

These development studies preceded the primary matrices and used different reductions and evaluation scopes. Two 40-step, seed-2026 comparisons used four candidates, four continuations, H16,

and coefficient 0.01. Each compared real and within-pivot shuffled targets against the same reused OPD control. Cross-fitted training formed targets from both cross-fit directions; full-sample training pooled all valid continuations and used cross-fitting only for diagnostics. The shared control is one baseline replication.

Validation used Avg@16 on AIME24/AIME25/AMC23 with counts 30/30/83, weighted by question count. The mean-checkpoint statistic averages steps 20 and 40; it is not the equal-dataset macro used in the main OPSD table. No seed-level or prompt-paired confidence intervals were supplied, and the two checkpoints are not independent training repeats.

Table 21: Complete training contrasts following SW-baseline refinement (pp). Both positive and negative effects are retained. Point differences alone do not establish significance or equivalence.

Training baseline	Comparator	Step 40	Mean of steps 20/40
Cross-fitted	Real minus OPD	$+0.087$	$+0.000$
Cross-fitted	Real minus shuffled	-0.087	$+0.044$
Full-sample	Real minus OPD	$+0.087$	$+0.044$
Full-sample	Real minus shuffled	$+0.044$	$+0.503$

The raw score-space auxiliary/direct ratios averaged 0.8475/1.0173 for cross-fit real/shuffled and 0.2483/0.3479 for full-sample real/shuffled. These are not matched AdamW update norms. Full-sample real retained a valid-pivot fraction of 0.5687, below its predeclared 0.80 coverage gate. Training divergence changes realized coverage and dose, limiting a variance-only causal interpretation. A same-recipe raw future-credit training arm is also absent from these comparisons.

For context, an earlier full-suffix training comparison yielded combined AIME24/AIME25 Avg@16 of 58.4375% for OPD and 57.8125% for both real and shuffled auxiliaries. Its real-minus-OPD 95% interval included zero. This earlier, different protocol motivates the subsequent diagnostics; it is not a pre-refinement arm of the SW-baseline probes.

E.5 Earlier terminal-verifier studies

A separate development study replaced future distillation costs with terminal-verifier feedback. Its frozen sequence audit permuted complete signed leave-one-out advantages, preserving their multiset and active support while breaking response-to-advantage assignment. It supported assignment-specific effects in raw and AdamW-aware comparisons, without taking an optimizer step. The subsequent training contrasts are listed below. A graded verifier expanded informative coverage from 46.64% to 82.42%, but reduced the two-checkpoint validation macro relative to its dose-matched binary counterpart by 0.346 pp.

Table 22: Historical terminal-verifier sequence auxiliaries. Differences are auxiliary minus matched OPD (pp). The 40-step validation statistic averages steps 20 and 40 over MATH-500, Minerva, and Olympiad-Bench; intervals are reported four-seed bootstrap intervals. The 93-step rows are separate longer-run configurations.

Setting	Train reward Δ	Validation Δ	Uncertainty / endpoint
40 steps, $\lambda = 0.000625$	$+0.959$	$+0.103$	95% CI $[-0.088, 0.298]$
40 steps, $\lambda = 0.00125$	$+2.174$	$+0.017$	95% CI $[-0.275, 0.525]$
93 steps, active normalization	$+1.639$	-0.698	Final checkpoint
93 steps, active + cap 1.20	$+1.894$	-0.258	Final checkpoint
93 steps, total-response norm.	$+1.253$	-0.308	Final checkpoint

F Task utility of frozen candidate policies

The training comparisons leave open whether optimization fails to exploit useful feedback or the feedback itself favors unhelpful actions. Frozen selection tests action value without parameter updates. Two controls distinguish improvement over the student's candidate policy from information

Table 23: Held-out hard-selection effects in the independent M64 study (E1), in percentage points. Intervals use 20,000 source-paired bootstrap resamples with Bonferroni adjustment over four primary comparisons (family-wise 95%; each interval 98.75%). AB/BA estimates are averaged within each of the 256 sources.

Selection comparison	Gain (pp)	Adjusted interval
Future-KL minus student	-0.9171	$[-2.7147, 0.4144]$
Future-KL minus assignment null	$+0.4257$	$[-0.8667, 1.6937]$
Outcome minus student	-0.1404	$[-0.8720, 0.6216]$
Outcome minus assignment null	$+1.2024$	$[0.1439, 2.5431]$

in the score–action assignment. We first define hard selection and soft reweighting on the shared M64 pool, then report the earlier alignment and M16 studies separately.

F.1 Hard action selection and the outcome control

The M64 experiment selects a next token using one continuation half and evaluates it on the other. The student control renormalizes its probabilities over the same candidates; the assignment null permutes scores among them. Outcome selection improves on reassignment but has no demonstrated advantage over the student (Table 23). The definitions below formalize this distinction.

F.1.1 Policy value and ranking reliability

Let A and B denote independent continuation halves, with estimated future costs $\widehat{Q}_s^A(a)$ and correctness $\widehat{v}_s^B(a)$. The selector q^A uses only half A: it minimizes estimated future cost or maximizes estimated correctness. Its held-out value weights candidate correctness by the selector's probabilities:

$$\widehat{\mathrm{Acc}}_s^{A\to B}(q) = \sum_{a\in\mathcal{A}_s} q^A(a \mid s)\widehat{v}_s^B(a). \tag{23}$$

The paired student-relative gain is

$$\begin{aligned} \widehat{\Delta}_s^{A\to B}(q) &= \widehat{\mathrm{Acc}}_s^{A\to B}(q) - \widehat{\mathrm{Acc}}_s^{A\to B}(p_{\mathcal{A}}), \\ \Delta_{\mathrm{student}}(q) &= \frac{100}{2N}\sum_{s=1}^{N}\left[\widehat{\Delta}_s^{A\to B}(q) + \widehat{\Delta}_s^{B\to A}(q)\right], \end{aligned} \tag{24}$$

where $N = 256$ and the factor 100 converts accuracy differences to percentage points. Swapping halves prevents the continuation noise used for selection from also determining its measured value. Ranking reliability is measured separately by the Spearman correlation

$$r_s^{\mathrm{rank}} = \mathrm{Corr}_{a\in\mathcal{A}_s}\left(\mathrm{rank}\,\widehat{Q}_s^A(a),\ \mathrm{rank}\,\widehat{Q}_s^B(a)\right). \tag{25}$$

We use average ranks for ties and average correlations over problems with nonconstant rankings in both halves. This statistic measures reproducibility of the candidate order, independently of whether the preferred token improves correctness.

F.1.2 Sampling protocol and inference

The M64 experiment used 256 preselected new problems, four candidates per problem, and 65,536 continuations, with no missing or excluded problems. It used Qwen3-4B with a Qwen3-4B-Instruct-2507 teacher in frozen ordinary OPD, not OPSD reference-CE training. Sampled-prefix generation allowed 7,168 tokens in non-thinking mode; terminal continuations allowed 4,096. Future cost used H16, discount 0.95, and a masked mean. The prompt limit was 1,280 tokens, checked before sampling. The 16 sampling batches are not optimizer updates. Selection, rollout, and analysis seeds were 6301, 6311, and 6391, respectively.

The source manifest excluded identified overlaps with old E1, pilot, E2, training, and benchmark sets using normalized source IDs and word-trigram SimHash64 at Hamming distance at most three.

This improves the documented exclusion scope but does not certify removal of every semantic near-duplicate. The fixed 256-source stopping rule was not changed according to interim significance.

Even/odd continuation indices formed independent selection/evaluation halves of 32, exchanged in AB/BA cross-fitting. Ties used uniform weights over the best candidates. Averaging all candidate-assignment permutations gives the uniform four-candidate evaluation value exactly. Continuations were retained regardless of EOS or answer-parsing failure. Flat-half correlations were undefined and excluded only from that diagnostic, with effective sample counts reported below; they were not replaced by zeros.

Held-out correctness is 68.0496% (student), 67.9092% (outcome), 67.1326% (future cost), and 66.7068% (null). Outcome/student effects change sign across cross-fit directions ($+0.0826/-0.3634$ pp); outcome/null remains positive ($+1.3509/+1.0539$ pp). Neither rule beats both comparators with adjusted lower bounds above zero and positive student-relative effects in both directions. The future-cost/student upper bound $+0.4144$ pp excludes the prespecified $+2$ pp effect for this protocol.

F.1.3 Action variation, reliability, and generation diagnostics

Let $\hat{p}_a$ be the mean correctness of candidate a, and let $\hat{v}_a$ estimate its single-continuation Bernoulli variance. The signed diagnostic is

$$\widehat{V}_{\text{signal}} = \text{Var}_a(\hat{p}_a) - \frac{1}{4}\sum_{a=1}^{4}\frac{\hat{v}_a}{64}.$$

Negative source-level estimates are retained before averaging. The split-half candidate covariance supplies an alternative estimate, not an independent experiment.

Table 24: M64 descriptive diagnostics. Intervals are unadjusted source-bootstrap 95% intervals, unlike the primary family in Table 23. Variances concern correctness estimates, not full-parameter gradient variance.

Metric	Estimate	95% CI	Valid n
Observed candidate correctness range (pp)	9.0271	$[7.4463, 10.8093]$	256
Observed between-candidate variance	0.006231	$[0.003233, 0.009827]$	256
Single-continuation Bernoulli variance	0.097913	$[0.086623, 0.109386]$	256
Mean-estimation noise (within/64)	0.001530	$[0.001353, 0.001709]$	256
Signed signal variance	0.004702	$[0.001702, 0.008297]$	256
Split-half candidate covariance	0.004594	$[0.001575, 0.008203]$	256
Task split-half Spearman	0.046734	$[-0.039894, 0.135863]$	184
Task weighted cosine	0.037826	$[-0.072852, 0.150017]$	176
KL split-half Spearman	0.860156	$[0.825781, 0.891406]$	256
KL weighted cosine	0.839506	$[0.777990, 0.895303]$	236

Positive signed variance and split-half covariance indicate candidate task-value variation, not reliable per-problem rankings. All continuations were correct for 43 sources and incorrect for six; 72 sources (28.125%) had a flat half. Positive KL rank reliability with non-positive/undefined task rank occurred in 164/256 sources (64.0625%), a descriptive statistic rather than a significance test or selection rule.

Mean candidate mass was 99.5232%, all Q16 source records were complete, EOS frequency was 99.0204%, and parsing failure was 0.8286%. Non-EOS frequency of 0.9796% is not synonymous with verified token-limit truncation. Higher continuation count also does not create an oracle: selection uses only 32 outcomes per candidate in either direction.

Optimistic empirical selection ceiling. A post-hoc calculation chooses and scores a policy on the same evaluation half, then averages both directions over all 256 sources. Hard selection gains 4.24 pp over the student; KL-constrained tilting gains 0.124, 0.217, and 0.400 pp at budgets 0.01, 0.03, and 0.10. These optimistic same-sample maxima describe the existing candidate pool, not achievable held-out gains, population bounds, or test power. They require no new continuations and leave out-of-sample ranking precision unresolved.

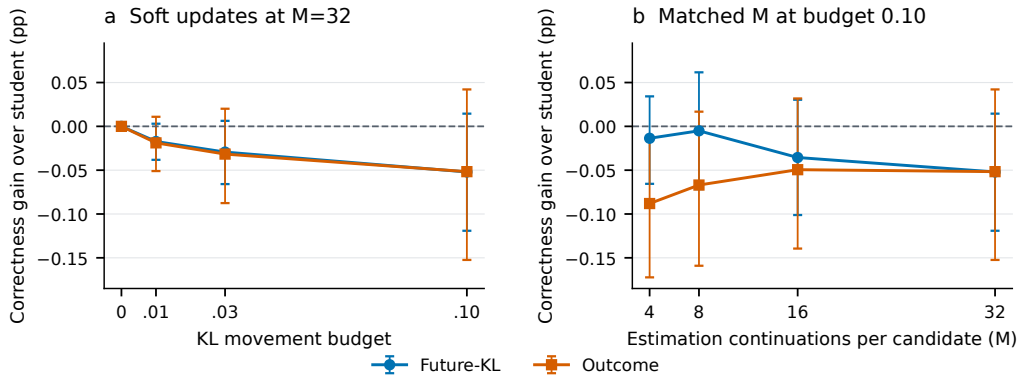

Figure 9: Gentler policy changes do not reveal a positive task increment. (a) Held-out correctness gain over the student candidate policy at $M = 32$. (b) Matched estimation counts at budget 0.10, with evaluation fixed at 32 continuations per candidate. Points average AB/BA within 256 sources; error bars are pointwise 95% source-bootstrap intervals. This analysis reuses the frozen M64 sample and measures candidate-policy value, not validation after training. Complete multiplicity-adjusted results appear in Appendix F.2.

F.2 Policy-preserving soft reweighting

Hard selection discards the student's relative preferences among nonselected candidates. To test whether a gentler change preserves useful preferences, the soft intervention retains the candidate prior $p_{\mathcal{A}}$ and tilts it using estimation-half negative future KL or terminal correctness, denoted $\widehat{u}_s(a)$:

$$q_\eta(a \mid s) = \frac{p_{\mathcal{A}}(a \mid s)\exp\{\eta \widehat{u}_s(a)\}}{\sum_{a' \in \mathcal{A}_s} p_{\mathcal{A}}(a' \mid s)\exp\{\eta \widehat{u}_s(a')\}}, \qquad D_{\mathrm{KL}}(q_\eta \| p_{\mathcal{A}}) \le \delta. \tag{26}$$

Here $\eta \ge 0$ is chosen from estimation scores alone, and δ limits movement from the student prior. Held-out values and paired gains follow Equations (23) and (24).

F.2.1 Selection, evaluation, and assignment controls

Soft reweighting reuses all 256 M64 sources and the even/odd halves of 32. In each direction, the first $M \in \{4, 8, 16, 32\}$ estimation continuations determine scores, while all 32 continuations in the other half evaluate correctness. Values are averaged across directions within source. Varying M therefore changes estimation effort while holding prefixes, candidates, and evaluation size fixed.

For each source and signal, Equation (26) tilts the conditional student prior toward negative future KL or terminal correctness. The temperature parameter is chosen using estimation scores alone to meet a KL budget of 0, 0.01, 0.03, or 0.10. A flat score vector retains the student prior. At zero budget, the real and null policies reduce to that prior. The realized KL can remain below the requested budget, so equal requested budgets do not imply identical realized movement.

The assignment null averages all 24 permutations of the four estimated scores. Each permutation is recalibrated against the same student prior under the same budget before held-out evaluation. Unlike the hard-selection null, this soft null is generally not a uniform-action policy. Table 25 reports the actual mean KL and both comparator differences for every nonzero configuration.

F.2.2 Uncertainty and complete results

The analysis uses 20,000 paired source-bootstrap resamples with seed 6392. Pointwise intervals use the 2.5th and 97.5th percentiles. The complete family contains $4 \times 3 \times 2 \times 2 = 48$ contrasts: estimation counts, nonzero budgets, signals, and comparators. The adjusted intervals use percentiles $0.025/48$ and $1 - 0.025/48$. AB/BA directions and nested estimation counts remain within the same source; they are not independent replications. These are exploratory results from the existing M64 sample.

All 24 nonzero student-relative point estimates are negative, with adjusted intervals spanning zero (Figure 10); the pointwise $M = 32$ intervals also span zero (Figure 9). Some $M = 4$ outcome intervals exclude zero on the negative side before adjustment. Thus, preserving more of the student's prior does not reveal a positive held-out increment in these data. Generation and coverage diagnostics are shared with the M64 study above.

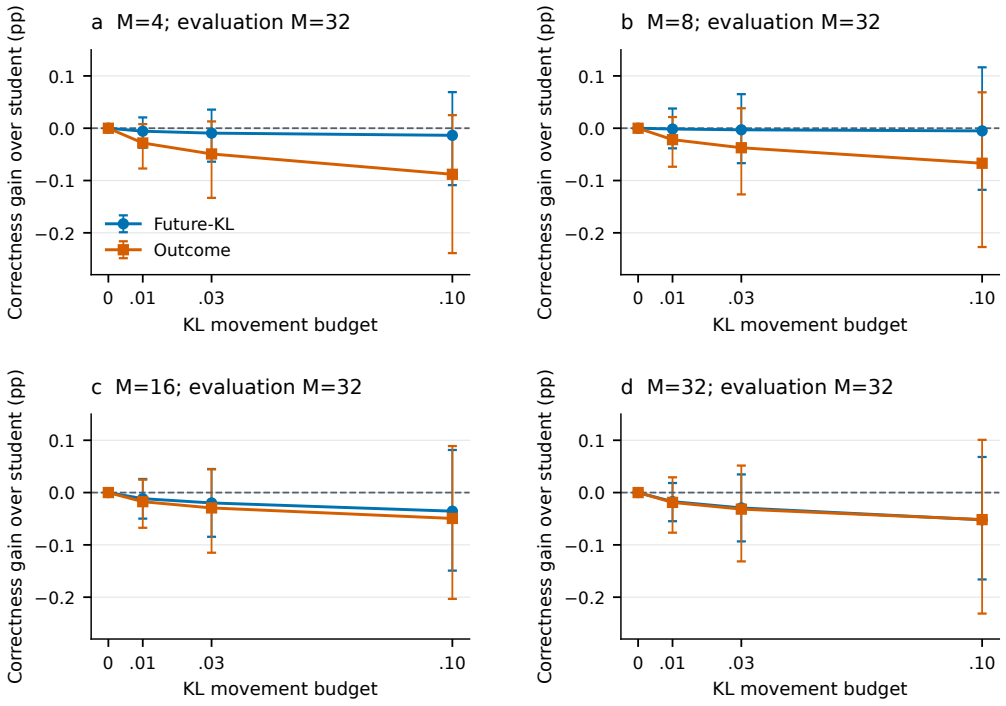

Figure 10: All soft-reweighting configurations relative to the student candidate policy. Panels vary estimation count while fixing evaluation at 32 continuations per candidate. Intervals use the 48-contrast Bonferroni adjustment; the zero-budget policy is the student itself. All nonzero student-relative point estimates are negative, with adjusted intervals spanning zero. The complete values against the separately calibrated assignment null are in Table 25.

F.3 Earlier future-cost and terminal-cost alignment studies

These earlier audits motivated the subsequent held-out selection tests. Their horizon-adjusted alignment intervals include zero (Table 26); the independent M64 study provides a stronger outcome control while still distinguishing reassignment-relative information from improvement over the student.

F.4 Earlier M16 selection study and tie-handling correction

The earlier H16 audit completed 256 sources and 16,384 continuations with eight per cross-fit half. Its original analysis reported future-KL/student gain -0.625 pp (95% CI $[-1.598, 0.301]$), gain over a sampled assignment null $+0.484$ pp ($[-0.496, 1.566]$), and task-positive-control gain -0.015 pp ($[-1.028, 1.042]$). Its weighted KL reliability was 0.691; this cosine is not the split-half rank correlation.

Reanalysis with uniform ties and the exact assignment null gives outcome/student $+0.1663$ pp (adjusted interval $[-1.1485, 1.5558]$) and outcome/null $+1.2431$ pp ($[0.0509, 2.6449]$). These revise the same M16 data; differing sources, pivots, and seeds prevent interpreting M16 versus M64 as a matched sampling-count experiment.

Table 25: Complete nonzero-budget N3 results. Gains and adjusted intervals are in percentage points; movement is the mean realized $D_{\mathrm{KL}}(q\|p_{\mathcal{A}})$ in nats. All 256 sources contribute to each row.

M	Budget	Signal	Movement	vs. student	Adjusted interval	vs. null	Adjusted interval
4	0.01	KL	0.00784	-0.006	$[-0.037, +0.021]$	-0.005	$[-0.037, +0.021]$
4	0.01	Outcome	0.00380	-0.028	$[-0.077, +0.008]$	-0.027	$[-0.076, +0.009]$
4	0.03	KL	0.02324	-0.009	$[-0.064, +0.036]$	-0.007	$[-0.060, +0.039]$
4	0.03	Outcome	0.01098	-0.049	$[-0.133, +0.013]$	-0.047	$[-0.130, +0.016]$
4	0.10	KL	0.07606	-0.014	$[-0.109, +0.069]$	-0.006	$[-0.100, +0.076]$
4	0.10	Outcome	0.03543	-0.088	$[-0.239, +0.025]$	-0.082	$[-0.231, +0.030]$
8	0.01	KL	0.00767	-0.002	$[-0.038, +0.038]$	-0.001	$[-0.038, +0.038]$
8	0.01	Outcome	0.00427	-0.022	$[-0.074, +0.021]$	-0.020	$[-0.072, +0.023]$
8	0.03	KL	0.02280	-0.003	$[-0.067, +0.065]$	-0.000	$[-0.064, +0.067]$
8	0.03	Outcome	0.01253	-0.037	$[-0.127, +0.038]$	-0.034	$[-0.123, +0.041]$
8	0.10	KL	0.07503	-0.005	$[-0.118, +0.116]$	$+0.003$	$[-0.109, +0.122]$
8	0.10	Outcome	0.04056	-0.067	$[-0.227, +0.069]$	-0.059	$[-0.218, +0.074]$
16	0.01	KL	0.00748	-0.012	$[-0.050, +0.026]$	-0.011	$[-0.050, +0.027]$
16	0.01	Outcome	0.00481	-0.017	$[-0.067, +0.025]$	-0.016	$[-0.065, +0.026]$
16	0.03	KL	0.02214	-0.020	$[-0.085, +0.045]$	-0.017	$[-0.085, +0.047]$
16	0.03	Outcome	0.01408	-0.029	$[-0.115, +0.045]$	-0.025	$[-0.110, +0.048]$
16	0.10	KL	0.07276	-0.036	$[-0.149, +0.081]$	-0.028	$[-0.146, +0.086]$
16	0.10	Outcome	0.04532	-0.049	$[-0.203, +0.089]$	-0.039	$[-0.193, +0.094]$
32	0.01	KL	0.00745	-0.017	$[-0.055, +0.018]$	-0.016	$[-0.054, +0.020]$
32	0.01	Outcome	0.00539	-0.019	$[-0.077, +0.029]$	-0.018	$[-0.075, +0.030]$
32	0.03	KL	0.02206	-0.029	$[-0.093, +0.035]$	-0.026	$[-0.091, +0.038]$
32	0.03	Outcome	0.01577	-0.032	$[-0.132, +0.052]$	-0.029	$[-0.128, +0.054]$
32	0.10	KL	0.07255	-0.052	$[-0.166, +0.068]$	-0.043	$[-0.159, +0.075]$
32	0.10	Outcome	0.05109	-0.052	$[-0.231, +0.101]$	-0.044	$[-0.225, +0.107]$

Table 26: Historical frozen terminal-alignment audits. Entries are cost–cost Spearman correlations with reported 95% step-cluster bootstrap intervals. Each audit completed 40 frozen steps and reported 93 eligible alignment observations per horizon; these are not 93 independent training runs. No horizon passed the one-sided family-wise rule across H8/H16/H32.

Horizon	4B teacher	30B teacher
8	$0.0368\ [-0.0710, 0.1437]$	$0.0269\ [-0.0709, 0.1192]$
16	$0.0457\ [-0.0702, 0.1669]$	$0.0287\ [-0.0696, 0.1248]$
32	$0.0133\ [-0.1166, 0.1452]$	$0.0401\ [-0.0689, 0.1475]$

A separate historical SW audit used a top-K probability-norm approximation and averaged source-wise reductions (87.18%). That statistic differs from the aggregate energy ratio in Appendix D.2 and from parameter variance.

G REFERENCE-SUPERVISED OPSD AND SELECTION CONTROLS

RS-OPSD uses outcomes to select problems and references to supply token targets. We first describe the earlier OPSD credit studies motivating this design (Section 3.5), then test reference exposure and failure selection through full-training and single-update controls. A final comparison examines dependence on the distillation recipe. Throughout, the self-teacher is a frozen copy of the initial student with privileged reference context; the earlier credit studies use a different training recipe and are not RS-OPSD replications.

G.1 FUTURE-STATE CREDIT UNDER A REFERENCE-CONDITIONED SELF-TEACHER

Reference-conditioned feedback. These historical probes use Qwen3-4B for both the student and its frozen teacher, with reference access restricted to the teacher. For shared initialization θ_0, their next-token distributions are

$$p_t^{\text{self}} = p_{\theta_0}(\cdot \mid P(x,z), y_{<t}), \qquad \pi_t = \pi_\theta(\cdot \mid x, y_{<t}), \qquad y \sim \pi_\theta(\cdot \mid x). \tag{27}$$

Here $P(x,z)$ contains the problem and reference (Section 4). The reference changes the conditioning context, and the student changes during training, so shared initialization does not order self-teacher and external-teacher KL distances. These within-protocol comparisons neither measure that distance gap nor identify cross-model mismatch as a cause of the ordinary-OPD results.

Local alignment and residual reduction. An initial frozen audit used four student candidates and four continuations each over 40 batches. H32 future-cost/task-cost rank correlation was 0.150 (95% step-cluster interval $[0.033, 0.267]$; 66 observations), with a positive lower bound after correction over H8/H16/H32. A follow-up window gave 0.086 ($[-0.012, 0.186]$; 165 observations), with no horizon meeting the alignment criterion. In that follow-up, cross-fitted SW centering reduced residual energy from 22.317 to 4.426 (80.2%). This is the scalar proxy in Appendix D.2, not parameter-gradient covariance; centering does not change within-state action rankings.

Training the centered future-cost signal. The three probes use full-parameter Qwen3-4B updates, seed 2026, batch size 16, learning rate 10^{-6}, and teacher-top-64 sampled-prefix supervision, unlike the later rank-64 LoRA reference recipe. Each main response supplies lexical/number pivots for four candidate tokens and four continuations per candidate. The H16 auxiliary uses discount 0.95, a detached full-sample SW baseline, and a differentiable score loss with nominal coefficient 0.0014929497. The capped variant attenuates this coefficient above a score-space auxiliary/local norm ratio of 0.075, not a parameter-gradient or AdamW-increment ratio. Validation averages eight responses per question at a 7,168-token limit, then equally weights AIME24, AIME25, AMC23, Minerva, and Olympiad-Bench.

Table 27: Historical OPSD future-credit training. Residual reduction is the mean logged scalar score-energy reduction, not parameter-gradient denoising. Accuracy differences are auxiliary minus the corresponding OPSD control in five-benchmark macro Avg@8 (pp). Dashes denote checkpoints outside that run. Each row uses one training seed.

Variant	Updates	Residual ↓ (%)	Δ_{20}	Δ_{40}	Δ_{60}	Δ_{80}	Δ_{93}
Centered H16	40	80.0	$+0.39$	-0.35	—	—	—
Capped H16	40	74.8	-0.86	$+0.08$	—	—	—
Capped H16, matched	93	81.4	$+0.11$	$+1.15$	-0.87	-1.15	-0.75

The 40-update arms reuse one earlier OPSD control; the 93-update pair uses a new control configured with identical auxiliary branching but zero coefficient. These are development comparisons, not independent replications: positive intermediate differences did not yield a consistent endpoint benefit (Table 27).

From privileged scores to explicit reference targets. Single-rollout frozen audits of sampled privileged residuals and terminal-verifier credit gave task-gradient cosines of 0.0090 ($[-0.0120, 0.0282]$; 20 valid batch clusters) and -0.0087 ($[-0.0207, 0.0037]$; 19), respectively.

Neither established positive alignment. RS-OPSD gives the reference an additional role as a prediction target, beyond teacher conditioning. The controls below test this design directly; the frozen diagnostics alone establish neither its benefit nor that of failure selection.

G.2 Full-training reference-selection control

The random control asks whether failure-based selection is needed to obtain the observed RS-OPSD endpoint. It preserves selected-problem counts within reference-length bins while randomizing problem identity. This matches one property of the selection rule; exposure and trajectory differences remain, as described below.

Selection rule. The completed seed-2026 run uses the 4B recipe, reference coefficient 0.001, and full-batch normalization. Within reference-length bins 1–256, 257–512, 513–768, and 769–1024, it selects the same number of problems as its current failure mask. Selection is randomized within each bin, using gate seed 6222; each selected problem keeps its own reference. This preserves per-bin counts, not exact token totals, another arm's evolving failure rate, or AdamW increments.

Evaluation and results. Online and final evaluation follow Table 4, matching the nominal settings of the OPSD/RS-OPSD pair. Final four-benchmark macro Avg@12 was 73.8590%, or $+1.8042$ pp relative to baseline and $+0.0402$ pp relative to error-reference. Mean online accuracy was 40.9791%, or -0.6086 pp relative to baseline and -0.0672 pp relative to error-reference. Tables 28 and 29 show all checkpoints and benchmark endpoints, respectively.

Output checks and comparison scope. The random run has 229 steps and five complete validation files: each online file contains 3,264 responses; the final file contains 9,792, covering the same 816 questions (30/30/83/673). Hashes, sample coverage, metadata, and benchmark means match the saved certificate. The original pair lacks comparable question-level records, so cross-run prompt pairing and uncertainty cannot be recovered.

Recoverable supervision exposure. The random run logs 16,608 selected problem visits over 229 updates. For error-reference, binary reward means and the fixed batch size of 128 imply 16,623 failed visits in 4B and 16,758 in 8B; these are reconstructed failure counts, not independently saved gate masks. Both 4B CE runs log 18,762,245 reference tokens before gating and mean post-gate padded-token fractions 0.41196/0.40611 (error/random). Dynamic microbatch averaging and missing denominators prevent conversion to verified selected-token totals. The legacy active-response field records reference availability, not gate coverage.

G.3 Online and final reference-selection results

Tables 28 and 29 provide the values underlying Figures 4 and 5. Positive selected-CE endpoints coexist with negative mean online differences from OPSD. The 8B pair likewise averages 41.740/41.477% online accuracy for OPSD/RS-OPSD (-0.26 pp); it is omitted from the 4B curves. Count-weighted rows are sensitivity calculations with the coverage limits in Appendix H.1.

G.4 Single-update controls for failure selection

Changing the gate changes subsequent training trajectories and reference exposure. Single-update interventions address this difficulty locally by fixing the starting state and update data. They test immediate gate effects and complement the full-training comparison.

At initialization and a step-200 OPSD snapshot, 16 data replicates compared four one-step interventions. Arm A used OPSD; B added error-conditioned reference CE with $\lambda = 0.001$. Arm C0 used random-gated CE at the same coefficient; C1 calibrated random-gated CE to B's raw auxiliary gradient norm. Random gates preserved active counts within reference-length bins, and every active problem retained its own reference. These controls changed gate assignment while preserving each problem's reference target.

Each comparison restored parameters, AdamW state, scheduler, and random state, then used one cached student response per update problem. A replicate contained 128 update and 64 independent

Table 28: **Online macro Avg@4 (%).** Thinking mode, 8,192-token limit. Base: OPSD; Error: RS-OPSD; Random: count-and-length-bin selection; Always: all-reference CE at $\lambda = 0.001$ without dose calibration. Dashes indicate unavailable results. Final evaluations change both checkpoint and budget (Appendix H.3).

	Qwen3-4B			
Step	Base	Error	Random	Always
50	41.102	40.326	40.872	42.205
100	42.737	42.776	42.482	39.849
150	42.101	40.234	42.012	39.443
200	40.410	40.849	38.551	39.841
Mean	41.588	41.046	40.979	40.335

Table 29: **4B selection controls.** Step-229 Avg@12 (%), 38,912-token limit; $E - R$: Error minus Random (pp). These single-seed runs do not match trajectories or total CE dose. Always-on incorporates the separately reported Olympiad-Bench completion (Appendix H.4).

Dataset	OPSD	Error	Random	Always-on	$E - R$
AIME24	70.00	72.78	73.61	71.94	-0.83
AIME25	61.11	64.17	65.00	63.06	-0.83
AMC23	89.86	90.56	89.06	91.27	$+1.51$
Olympiad-Bench	67.25	67.77	67.77	64.83	$+0.00$
Macro	72.05	73.82	73.86	72.77	-0.04
Count-weighted*	69.42	70.14	70.05	67.72	$+0.09$

evaluation problems; the same data groups were reused across starting states. Thus, the 32 cells are not 32 independent full-training seeds. Each updated model generated four fresh thinking-mode responses per evaluation problem with an 8,192-token limit, giving 128 evaluated arm states and 32,768 answers overall. The late snapshot intervention did not resume the old training-loader cursor.

Table 30: Completed E2 paired effects in held-out Avg@4, in pp. Intervals use the reported replicate/source two-level paired bootstrap (95%, not multiplicity-adjusted). All include zero.

Starting point	Comparison	Effect (pp)	95% CI
Initialization	B minus A	$+0.3418$	$[-1.1963, 1.8799]$
Initialization	B minus C0	$+0.2197$	$[-1.2451, 1.6846]$
Initialization	B minus C1	$+0.2930$	$[-1.2207, 1.7822]$
Step 200	B minus A	-0.3418	$[-2.0264, 1.2207]$
Step 200	B minus C0	$+0.6592$	$[-0.9521, 2.2461]$
Step 200	B minus C1	$+0.0488$	$[-1.4404, 1.5137]$

For A/B/C0/C1, mean correctness was 70.264/70.605/70.386/70.312% at initialization and 66.968/66.626/65.967/66.577% at step 200. Error-gate coverage was 51.8555/50.1465%, sampled-response truncation was 44.5313/42.7246%, and reference-target truncation was 9.6680% at both starting points. These observations limit a complete-reference interpretation of the intervention.

Neither starting point resolves an error-gate advantage or equivalence. Matching raw gradient norms does not match AdamW increments; these local tests do not replace full-training gate comparisons.

G.5 Sensitivity to the distillation recipe

The selection controls above hold the sampled-prefix loss fixed. To assess the scope of their conclusions, Table 31 compares the reference-CE effect across OPSD and an ordinary-OPD bridge. Other development recipes differ in teacher, objective, support, generation budget, and parameterization:

an older 459-step run and a second-seed intervention window gave final differences of -0.487 and -1.065 pp, respectively. Neither replicates the current recipe.

Table 31: Protocol sensitivity and the ordinary-OPD bridge. Online averages four checkpoint macros (Avg@4, 8,192 tokens); final is step-229 Avg@12 at 38,912 tokens. Differences are auxiliary minus baseline. The bridge changes teacher and objective, not teacher context alone.

Setting	Online Δ (pp)	Final without / with CE (%)	Final Δ (pp)
OPSD, 4B	-0.541	$72.055 \rightarrow 73.819$	$+1.764$
OPSD, 8B	-0.263	$74.678 \rightarrow 75.522$	$+0.844$
Ordinary OPD, 4B	$+1.234$	$77.320 \rightarrow 77.036$	-0.284

The bridge used a frozen Qwen3-4B-Instruct-2507 teacher and student-top-16, student-probability-weighted supervision. Its higher absolute final score and mixed reference-CE difference do not establish an isolated benefit from privileged context.

H Evaluation coverage and artifact provenance

This appendix specifies which records support the reported results and which comparisons are excluded. Aggregate scores, complete response files, and verified model identity provide different levels of evidence. We document the main runs first, followed by recovered diagnostics, replacement runs, withheld evaluations, and the separately completed always-on endpoint.

H.1 Availability of the main-run artifacts

For the original OPSD/RS-OPSD pairs, final validation dictionaries and all 229/229 progress entries are retained. Some 8B/bridge logs report cleanup exceptions after the endpoints were written. Checkpoint saving and validation JSONL were disabled, leaving aggregate scores available but preventing independent verification of prompt-paired intervals, effective adapter loading, and retained question identities. Count-weighted rows use $\sum_d n_d A_d / \sum_d n_d$, with $n_d = (30, 30, 83, 673)$ verified for the later random run only; they are sensitivity calculations, not recovered pooled accuracy.

If the two AIME sets each contained 30 questions, their 4B differences would correspond to 10 and 11 net correct responses at $K = 12$; paired question improvements cannot be reconstructed from aggregate means.

H.2 Recovered diagnostic records and replacement training runs

Completed diagnostic records were checked after a launcher failure; incomplete training arms without checkpoints were rerun from their registered initializations. The following records identify the outputs used in each comparison. Failed attempts are archived and do not count as additional replications.

Normalization/support audit. All 32 prefix and two batch records were saved (4,896 repeated gradient norms) before the launcher exited with code 127 after an argument was parsed as a shell command. Recovery verified condition definitions, finite statistics, normalizers, ownership, and 8,192 unique continuation seeds. The failed process status and source hashes remain archived.

Continuation-count matrix. Seed-2028 M16 failed during final validation without a checkpoint. That arm alone was repeated from its registered initialization, seed, settings, and inference profile. Both replacement endpoints are used; the failed attempt is archived, not counted as another replication. The selected matrix has 18 files and 235,008 responses.

SW-centering matrix. Eleven centering runs completed; the seed-2028 shuffled arm failed after saving step-20 responses but before logging its metrics, without a checkpoint. That arm was repeated, and both replacement endpoints are used. The 24 selected files contain 313,344 responses

validated against the same 1,088-question set. Historical configuration/data hashes and current replacement source/model metadata are retained without claiming identical historical code or tensor hashes.

H.3 Checkpoint-by-budget evaluation: withheld comparisons

A separate evaluation was designed to distinguish checkpoint changes from generation-budget changes, but model-identity checks prevent its use as performance evidence. This 4B pair (training/evaluation seeds 6230/6231) saved steps 200/229 and generated eight method/checkpoint/budget cells. Evaluation fixes Avg@12, thinking mode, temperature 1, and top-$p = 0.95$, varying only length budget within each checkpoint. It does not recover the original seed-2026 checkpoints.

A tokenizer audit explains two excluded overlength Olympiad-Bench prompts, leaving 816 questions in the four-benchmark scope. This resolves coverage, not model identity; paired intervals and checkpoint-by-budget effects remain withheld.

Table 32: E3 evaluation availability. Scores are withheld (dashes) pending pipeline validation and are not included in the four-benchmark results. At step 229, arm outputs are identical within each length budget.

Step	Token budget	OPSD	RS-OPSD	Comparison status
200	8,192	—	—	Withheld: pipeline validation
200	38,912	—	—	Withheld: pipeline validation
229	8,192	—	—	Withheld: integrity audit
229	38,912	—	—	Withheld: integrity audit

OPSD/RS-OPSD step-229 JSONL files were byte-identical within each budget, although all 504 corresponding saved adapter tensors differed. Saved-weight differences do not verify effective adapter application during generation. The cause remains unresolved, and step-200 outputs also require pipeline validation. We therefore retain Table 32 as a provenance record, excluding the matrix from replicated performance and checkpoint-by-length claims.

H.4 Always-on reference CE: recovered outputs and completed endpoint

The always-on endpoint combines local records with an author-reported aggregate from another machine. Local online files are complete: each of steps 50, 100, 150, and 200 contains 3,264 readable responses for the same 816 problems. The step-229 file contains 2,248 readable responses across 188 problems, with 187 complete groups of 12, before ending in a truncated record (Table 33). The separate aggregate completes the reported accuracy, not the local response archive.

Table 33: Locally archived always-on output integrity within the four-benchmark reporting subset. The truncated step-229 file remains unchanged; the separately reported Olympiad-Bench aggregate below completes the final accuracy table, not the local response-level archive.

Step	Readable responses	Sources	Complete groups	Status
50	3,264	816	816 ($k = 4$)	Complete
100	3,264	816	816 ($k = 4$)	Complete
150	3,264	816	816 ($k = 4$)	Complete
200	3,264	816	816 ($k = 4$)	Complete
229	2,248	188	187 ($k = 12$)	Truncated

The retained final records contain 360 AIME24, 360 AIME25, 996 AMC23, and 532 Olympiad-Bench responses. The first three benchmarks have complete 12-response groups for the same problem identities as the online files. Recomputing accuracy from their saved binary scores gives 71.94%, 63.06%, and 91.27%, respectively (Table 29). The local Olympiad-Bench subset is incomplete and is not used to estimate its full-dataset accuracy. A subsequent evaluation on another machine, reported by the authors under the same settings, yielded 5,236 correct responses out of

$673 \times 12 = 8,076$, giving 64.8341% Avg@12. Combining this aggregate with the three complete benchmark scores gives 72.7748% macro accuracy and 67.7185% count-weighted accuracy. The completion is recorded as an author-supplied aggregate; its response-level outputs are not present in the local archive.

Recovered online means match the logged macro of 40.335%. A disk-quota write failure explains the truncated endpoint file; question-level pairing and effective adapter loading remain unverified for the separate completion. Finally, the common nominal coefficient 0.001 does not equalize total CE exposure between always-on and selective supervision.