Block-Level Credit Assignment in On-Policy Distillation

Anonymous authors
Paper under double-blind review

Abstract

On-policy distillation (OPD) transfers reasoning behavior through teacher feedback on student-generated trajectories. Common sampled-token methods assign this feedback independently to each token, although a local reasoning decision may span several tokens. We introduce Block3 Mean, which averages teacher–student log-probability differences within three-token blocks and optimizes a clipped objective with a joint block importance ratio and valid-length weighting. This gives adjacent tokens a shared learning signal without requiring an additional reward model or process annotations. We analyze how the update couples neighboring credit, changes gradient scale and alters clipping, and evaluate mathematical reasoning across teacher–student pairs and saved checkpoints. In the historical Qwen3-1.7B comparison at Step200, Block3 improves Avg@8 over sampled-token OPD by 8.90 percentage points on MATH500 and 5.42 on AMC23. The official Qwen3-8B-to-1.7B Instruct pair also has positive Step200 Avg@8 differences on all four evaluated benchmarks. Trajectory diagnostics connect these results to entropy, teacher–student overlap and local credit redistribution, while distinguishing numerical stability from generation quality. Our study makes feedback granularity an explicit design dimension in reasoning distillation.

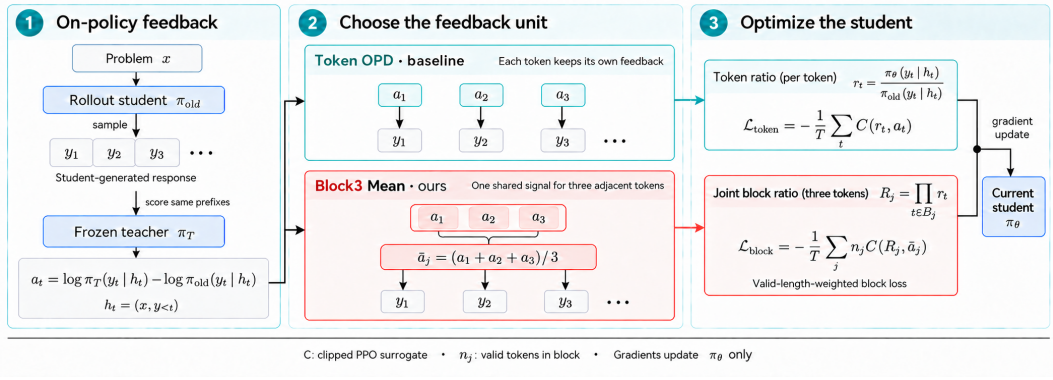

Figure 1: From token-level to block-level updates. The student generates a response and the frozen teacher scores the same prefixes. Token OPD keeps each token's feedback separate; Block3 Mean averages feedback within three adjacent tokens and shares it across the block. The implemented method also uses a joint block ratio and valid-length-weighted block loss. The two rows are alternative training runs, not summed objectives. Teacher feedback and old-policy probabilities are detached; only the current student receives gradients. A full three-token block is illustrated; masked tails use their valid length n_j. Both methods sample complete trajectories before updating.

1 Introduction

Transferring reasoning behavior from a capable teacher is an important route to improving language models. On-policy distillation (OPD) trains a student on its own responses, with

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.

teacher feedback at the prefixes the student actually visits (Agarwal et al., 2023; Gu et al., 2023). This brings supervision to learner-induced states rather than relying only on expert demonstrations (Ross et al., 2011). Teacher probabilities provide a dense learning signal throughout the response, making OPD a useful setting for studying how reasoning behavior is acquired during post-training.

The availability of dense feedback leaves a central design choice open: what should be the unit of credit assignment? A common sampled-token implementation assigns each token its own teacher–student log-probability difference (Fu et al., 2026). This treats feedback at neighboring positions separately, even when several tokens jointly express a short mathematical operation or a transition in a solution. We ask: can sharing teacher feedback across a short token block improve on-policy reasoning distillation?

Our intuition is that a local decision can have several textual components. Their individual probability differences may fluctuate with the realization of the expression, while their aggregate can capture a shared preference. A short block provides an intermediate unit between a single token and a complete response: it allows neighboring choices to learn together while retaining local supervision.

We introduce Block3 Mean, a block-level update that reuses the feedback already produced by an OPD teacher. After the student generates a response, we divide it into disjoint three-token blocks, average the distillation advantages within each block, and optimize a clipped objective with a joint block importance ratio and valid-length weighting (Figure 1). The procedure requires no additional reward model or process annotations. It changes how teacher scores guide the update, while retaining full-response sampling and the same teacher-scoring pipeline.

We evaluate on MATH500, AIME24, AIME25 and AMC23, with matched Token OPD controls and checkpoint-level reporting. In the historical Qwen3-1.7B comparison, Block3 improves Step200 Avg@8 by 8.90 percentage points on MATH500 and 5.42 on AMC23. The official Qwen3-8B-to-1.7B Instruct comparison also yields positive Step200 Avg@8 differences on all four benchmarks. To understand the update beyond endpoint scores, we connect its algebraic structure to training trajectories, including entropy, token overlap, gradient norms and local credit redistribution.

Our contributions are threefold. We formulate a local block-level OPD objective combining shared teacher feedback and a joint importance ratio. We derive its cross-token credit terms, gradient scale and clipping behavior, and characterize when local averaging improves a signal estimate. We then evaluate reasoning performance across teacher–student pairs and saved checkpoints, using trajectory diagnostics to connect the observed results to credit allocation and generation behavior.

2 Background

2.1 On-Policy Distillation for Autoregressive Models

A language model factorizes a response as $\pi(y \mid x) = \prod_t \pi(y_t \mid x, y_{<t})$. Unlike distillation on fixed responses, OPD asks the teacher to evaluate student-generated prefixes (Ross et al., 2011; Agarwal et al., 2023). This choice of prefix distribution is distinct from the choice of divergence: student prefixes can support forward KL, reverse KL or other matching losses.

Full-vocabulary distillation compares next-token distributions; sampled-token distillation uses probabilities only at the generated tokens, supplying scalar feedback for policy-gradient optimization (Gu et al., 2023; Fu et al., 2026). We study the allocation of these sampled signals across neighboring positions with a fixed external teacher.

2.2 Sampled Feedback and Clipped Policy Updates

For prompt x, let the rollout student π_{old} generate $y = (y_1, \ldots, y_T)$, and write $h_t = (x, y_{<t})$. The teacher feedback and the current-to-old student ratio are

$$a_t = \log \pi_{\text{T}}(y_t \mid h_t) - \log \pi_{\text{old}}(y_t \mid h_t), \quad r_t(\theta) = \frac{\pi_\theta(y_t \mid h_t)}{\pi_{\text{old}}(y_t \mid h_t)}. \tag{1}$$

The distillation advantage a_t compares teacher and rollout-student probabilities, not verified mathematical correctness. Both a_t and the old probabilities are detached during optimization.

Let $\mathcal{C}(r, a)$ be the PPO clipped surrogate (Schulman et al., 2017), with negative-advantage dual clipping specified in Appendix B. The baseline averages $-\mathcal{C}(r_t, \text{sg}(a_t))$ over valid tokens, assigning each position its own signal and ratio. We replace this unit with a short block. Here T counts valid tokens in one response; padding is masked and production reduction occurs within each actor micro-batch.

3 Methods

3.1 From Individual Feedback to Local Credit

Token-wise updates follow each teacher preference independently; block-wise updates aggregate the preferences and update the expression's components with one signal. In Figure 2, averaging $(2, -1, 2)$ gives positive feedback at all three positions. This couples adjacent updates. A fixed-width block is an operational unit, not an annotated reasoning step.

The student generates a complete response; the frozen teacher scores its prefixes; detached feedback is aggregated locally; and gradients update only the current student. No new teacher responses, reward model or process labels are needed. Teacher scoring can process the generated sequence in a batched forward pass, but grouping does not reduce the number of conditional probabilities scored.

3.2 Block-Level On-Policy Distillation

Shared feedback and joint ratio. Group consecutive response tensor positions into disjoint blocks B_j of width k, without packing across mask holes. Below, sums within a block select valid positions, n_j is their count, and $T = \sum_j n_j$. Empty blocks have zero weight. The ideal real-arithmetic reduction, omitting the implementation's 10^{-8} denominator stabilizer, is

$$\bar{a}_j = \frac{1}{\max(1, n_j)} \sum_{t \in B_j} a_t, \qquad R_j(\theta) = \exp\left(\sum_{t \in B_j} \log r_t(\theta) \right), \tag{2}$$

$$\mathcal{L}_{\text{block}} = -\frac{1}{T} \sum_{j : n_j > 0} n_j \mathcal{C}(R_j, \text{sg}(\bar{a}_j)), \qquad \mathcal{L}_{\text{token}} = -\frac{1}{T} \sum_{t=1}^{T} \mathcal{C}(r_t, \text{sg}(a_t)). \tag{3}$$

Block3 Mean uses $k = 3$, retained after the exploratory $k \in \{1, 3, 5, 10\}$ comparison described in Section 4.3.2. Its joint ratio $R_j = \prod_{t \in B_j} r_t$ measures the product of policy changes across the sampled block, rather than their average. Valid-length weighting gives a two-token tail weight 2/3 relative to a full block. The code implements the same reduction using fractional block masks n_j/k; Appendix B gives the exact stabilized denominator and micro-batch convention.

Optimization. Each collected batch supplies shared advantages and current-to-old block ratios for the clipped update. This uses sampled blocks, not enumeration of a $|V|^k$ distribution or optimizer steps after every three tokens. Relative to Token OPD, the intervention changes advantage sharing, ratio granularity and loss reduction together.

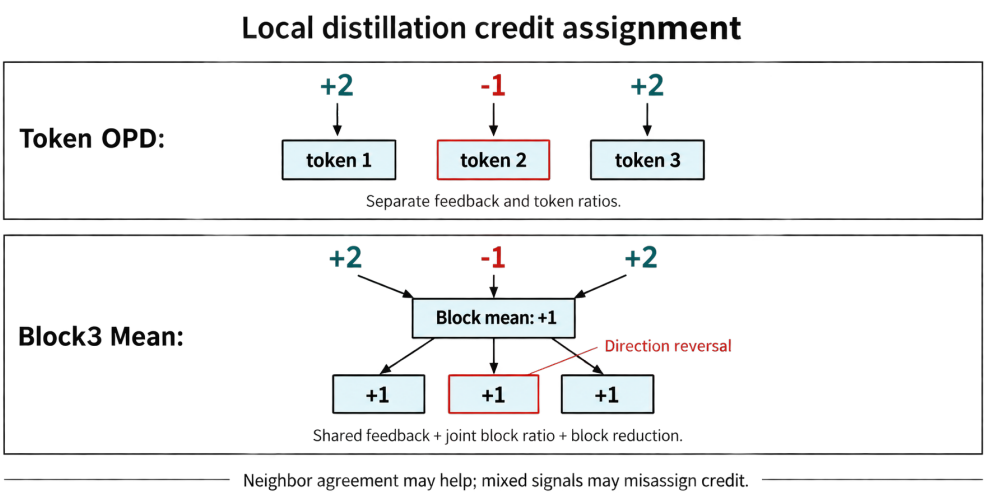

Figure 2: Illustrative feedback sharing, not an observed rollout. Averaging $(2, -1, 2)$ gives $(1, 1, 1)$, reversing the middle token's local signal. The implemented method additionally changes ratios and loss reduction. A local sign reversal does not by itself establish an error in task-level credit assignment.

3.3 Understanding the Block Update

Cross-token credit. Consider equal-sized blocks, $T = kJ$, at the old policy, where every ratio is one and clipping is inactive. Let $g_t = \nabla_\theta \log \pi_\theta(y_t \mid h_t)$. The policy-gradient components satisfy

$$\nabla \mathcal{L}_{\text{token}} = -\frac{1}{T} \sum_t a_t g_t,$$

$$\nabla \mathcal{L}_{\text{block}} = -\frac{1}{T} \sum_j \sum_{i \in B_j} \sum_{t \in B_j} a_i g_t = \nabla \mathcal{L}_{\text{token}} - \frac{1}{T} \sum_j \sum_{\substack{i,t \in B_j \\ i \neq t}} a_i g_t. \tag{4}$$

The additional terms send token i's signal to token t's score gradient. They can align or conflict with the baseline; their sign alone does not tell us whether downstream mathematical reasoning benefits.

Proposition 1 (Equal feedback does not imply equal updates). At $\theta = \theta_{\text{old}}$, with inactive clipping, full blocks of size k, and $a_t = a$ for all valid tokens, both policy losses equal $-a$, but $\nabla \mathcal{L}_{\text{block}} = k \nabla \mathcal{L}_{\text{token}}$.

Substituting $\bar{a}_j = a$ in Equation 3 proves the statement. The factor concerns the policy-loss derivative, not the subsequent Adam update after gradient clipping and other terms. It reveals a scale change even when all local feedback agrees.

Clipping geometry. Away from the old policy, the joint ratio also changes which updates are clipped. Three token ratios of 1.1 produce a block ratio of $1.1^3 = 1.331$: each token can remain inside a conventional 0.2 clipping band while the block leaves it. Learning-rate rescaling alone therefore cannot generally recover the token-level update.

When averaging could help. To separate the averaging hypothesis from the implementation above, consider a hypothetical token-ratio, token-normalized intervention using Wa instead of a, where W is deterministic and each row sums to one. Suppose, only as a local signal model, $a = \mu + \varepsilon$, $\mathbb{E}\varepsilon = 0$, and $\text{Cov}(\varepsilon) = \Sigma$. Then

$$\mathbb{E}\|Wa - \mu\|_2^2 = \|(W - I)\mu\|_2^2 + \text{tr}(W \Sigma W^\top). \tag{5}$$

Averaging helps this estimator when noise reduction exceeds the bias introduced by combining different latent signals. With a constant signal and independent equal-variance noise

within a block, the per-position variance becomes σ^2/k; correlated noise or heterogeneous signals can remove this benefit. This explains the motivation for short-range sharing, but is a statement about signal estimation, not a guarantee of higher task reward. The implemented objective additionally changes ratios and reduction. Appendix B provides the derivation, a counterexample to unconditional smoothing benefit, and a restricted interpretation in which block cross terms recover within-block future credit.

4 Experiments

We organize the experiments around three questions: how block-level updates affect mathematical reasoning relative to sampled-token OPD; how the comparison changes across teacher–student pairs and checkpoints; and what trajectory diagnostics reveal about credit sharing, block length and generation quality.

4.1 Experimental Setup

Models and training data. The Qwen3-4B-Base-GRPO teacher is the public checkpoint released with Rethinking On-Policy Distillation of Large Language Models: Phenomenology, Mechanism, and Recipe (Li et al., 2026a). It adapts Qwen3-4B-Base with GRPO on DAPO-Math-17k-Processed; it is not an official Qwen instruction model.[1] Historical pairs include this teacher with Qwen3-1.7B-Base and Qwen3-0.6B-Base students, DeepSeek/JustRL 1.5B, and Llama-3.2-3B-Instruct teaching Llama-3.2-1B-Instruct. Two further comparisons have fully retained trajectories: Qwen3-4B-Base with the same GRPO teacher, and official Qwen3-8B teaching Qwen3-1.7B instruction-tuned models (Yang et al., 2025; Grattafiori et al., 2024; Meta, 2024). The latter repositories omit "Instruct" from their names. Exact identities and protocol versions appear in Appendices C, E and H.

The two fully retained comparisons share a DAPO-derived pool (Yu et al., 2025): 17,917 question–answer rows repeated 100 times, with 17,237 exact unique questions. Training uses the questions without outcome reward. The 4B Base arms each encounter 800 prompt occurrences and 783 distinct questions, including one normalized-string MATH500 match once per arm. Appendix C gives the overlap audit and a separate 496-question sensitivity analysis.

Baseline and implementation. The baseline follows sampled-token OPD in Revisiting OPD (Fu et al., 2026). Paired arms independently initialize the same student and share teacher, inputs, optimizer and evaluation. The fully retained pairs use seed 21, 200 updates, four prompts with eight responses each, learning rate 2×10^{-6} and one PPO epoch per update: 6,400 trajectories per arm, with full states at Steps 50/100/150/200. Generation uses temperature 1.0, top-$p = 0.9$, prompt limit 2,048 tokens and response cap 16,384. Separately derived request seeds and input identities are checked for all trajectories. Appendix C gives micro-batch, precision and clipping settings.

Prompting. The 4B Base comparison uses the bare question, a boxed-answer instruction and a Solution: suffix, without ChatML. The official Instruct pair uses native Qwen chat with the same answer instruction, explicit non-thinking mode and no Solution: suffix. Each of these pairs uses matching training and evaluation inputs, checked at the token-ID level. Older experiments retain their original protocols, including the historical 1.7B train–evaluation prompt difference. Exact prompt strings and stop tokens are given in Appendices C, F and H; changing model families is not treated as a prompt-only ablation.

Evaluation. The two fully retained pairs use MATH500 (500 questions), AIME24 (30), AIME25 (30) and AMC23 (83), reported separately. Each model view generates 5,144 responses: eight per question, seeds 21–28, temperature 1.0, top-$p = 0.9$ and response cap 16,384. Both arms share the pinned historical grader; earlier studies retain their individually

[1] Model card: https://huggingface.co/Thinking-Space/Qwen3-4B-Base-GRPO. The archived repository ID is lllyx/Qwen3-4B-Base-GRPO; Appendix C records the retained revision.

documented protocols. For question i, let c_i count correct responses. We compute

$$\text{Avg}\,@8 = \frac{1}{N}\sum_{i=1}^{N}\frac{c_i}{8}, \qquad \text{Pass}\,@8 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[c_i > 0]. \tag{6}$$

These are mean sampled accuracy and eight-attempt coverage, not majority voting. We report Step200 and all saved checkpoints rather than separate best checkpoints; the endpoint choice is retrospective. The 95% paired-question intervals are conditional on trained checkpoints, without multiple-comparison correction or training-seed variability. Eight evaluation samples do not constitute eight training replications.

Training diagnostics. We record student and teacher entropy, entropy gaps, top-16 overlap and overlap mass, overlap-token advantage, policy loss, pre-clipping gradient norm, response length and cap incidence. Sign disagreement and advantage redistribution measure the effect of sharing on local feedback. Missing boxed-answer markers and engine length stops are recorded separately from correctness. Definitions and every-step position maps appear in Appendices D and I.

4.2 Main Results

Reasoning transfer across teacher–student pairs. Table 1 illustrates two historical settings with positive Step200 Avg@8 differences. With Qwen3-1.7B-Base and the public GRPO teacher from Rethinking OPD (Li et al., 2026a), Block3 improves Avg@8 on all four benchmarks: MATH500 rises from 45.85 to 54.75, AIME24 from 5.42 to 8.75, AIME25 from 2.50 to 5.42, and AMC23 from 21.84 to 27.26. These observations motivate local feedback sharing. DeepSeek/JustRL 1.5B also has positive endpoint differences, smaller on MATH500 and AMC23 than on AIME.

These are illustrative positive cases, not an exhaustive cross-model comparison. Other historical pairs favor Token; Table 6 in Appendix E retains all four endpoint comparisons and their protocol context, followed by every checkpoint result.

Table 1: Selected positive historical Step200 Avg@8 examples (%). Compare methods only within each student stratum. The complete comparison, including negative outcomes, appears in Table 6.

Student	Method	MATH500	AIME24	AIME25	AMC23
Qwen 1.7B	Token	45.85	5.42	2.50	21.84
Qwen 1.7B	Block3	54.75	8.75	5.42	27.26
DeepSeek 1.5B	Token	84.70	39.58	29.17	73.34
DeepSeek 1.5B	Block3	84.75	43.75	30.83	73.80

Official Qwen3-8B to 1.7B Instruct. Table 2 tests both methods under native non-thinking chat with matched training and evaluation inputs. Relative to Token, Block3's Step200 Avg@8 differences are +0.10, +2.50, +0.83 and +0.30 points on MATH500, AIME24, AIME25 and AMC23. Both methods exceed the original student's four Avg@8 scores. All pointwise 95% paired-question intervals for the method differences include zero, so the positive estimates do not resolve a statistical advantage at this endpoint.

Pass@8 separates average success from question coverage. Its Block3-minus-Token differences are +0.20, −6.67, 0.00 and −1.20 points, and both methods reduce the original student's MATH500 coverage from 91.20%. Thus higher mean accuracy can coexist with lower eight-attempt coverage. Appendix H reports the original student, all eight trained checkpoints and the associated behavior metrics.

Qwen3-4B Base with the public GRPO teacher. The matched completion-prompt comparison gives a different endpoint pattern (Table 3). At Step200, Block3's Avg@8 differences are −0.52, −2.92, +0.42 and −3.01 points. Token is higher on three tasks, while all pointwise

Table 2: Official Qwen3-8B to Qwen3-1.7B Instruct: Step200 scores (%) and Block3 minus Token differences (percentage points).

Benchmark	Original	Token	Block3	Δ [95% CI]
		Avg@8		
MATH500	71.45	72.12	72.22	+0.10 [-1.18, 1.38]
AIME24	10.00	14.17	16.67	+2.50 [-2.50, 8.33]
AIME25	10.42	12.08	12.92	+0.83 [-3.75, 5.83]
AMC23	40.51	43.07	43.37	+0.30 [-3.16, 3.92]
		Pass@8		
MATH500	91.20	89.60	89.80	+0.20 [-2.00, 2.40]
AIME24	26.67	43.33	36.67	-6.67 [-20.00, 6.67]
AIME25	23.33	30.00	30.00	+0.00 [-16.67, 16.67]
AMC23	69.88	75.90	74.70	-1.20 [-9.64, 7.23]

Original denotes the untrained-for-this-experiment Instruct student, not a Base model. Brackets are pointwise 95% paired-question bootstrap intervals (10,000 resamples); one training seed, no correction for checkpoint or task selection.

Table 3: Qwen3-4B paired Step200 evaluation (seed21; queue COMPLETE).

	Avg@8			Pass@8		
Benchmark	Block3	Token	Δ [95% CI]	Block3	Token	Δ [95% CI]
MATH500	75.40	75.92	-0.52 [-1.57, 0.55]	90.60	91.00	-0.40 [-2.40, 1.60]
AIME24	13.33	16.25	-2.92 [-7.08, 0.83]	33.33	30.00	+3.33 [-6.67, 13.33]
AIME25	13.75	13.33	+0.42 [-2.92, 3.75]	30.00	30.00	+0.00 [-13.33, 13.33]
AMC23	46.23	49.25	-3.01 [-6.17, 0.30]	74.70	77.11	-2.41 [-10.84, 6.02]

Scores are percentages; Δ is Block3 minus Token in percentage points. Brackets: pointwise 95% paired-question bootstrap within each benchmark, conditional on one training pair, not a multi-training-seed CI. NA: unavailable.

95% paired-question intervals cross zero; neither superiority nor equivalence is established. On AIME24, Block3 has lower Avg@8 but solves one more question at least once, giving Pass@8 of 33.33% versus 30.00%. Together, the paired results show that credit granularity affects the learning trajectory, with benefits depending on the teacher–student setting and the metric.

4.3 Analysis and Discussion

4.3.1 Checkpoint Dynamics

Figure 3 reveals checkpoint-dependent effects. At Step150, Block3 improves 4B Base MATH500 Avg@8 by 1.85 points (pointwise interval $[0.70, 3.00]$) and AIME24 by 2.92 ($[0.42, 5.83]$), but reduces AIME25 by 4.17. At Step100, the MATH500 difference is -1.40 ($[-2.57, -0.25]$); at Step50, only AMC23 favors Block3. The official Instruct pair also reverses: Step50 favors Token on three tasks, and Step150 favors Token on AIME24. The intervention changes the learning path, not a fixed performance offset. We retain all checkpoints rather than treating their maximum as a validated selection rule.

4.3.2 Why Three Tokens? Block-Length Selection

Table 4 gives the exploratory sweep that motivated retaining $k = 3$. Block3 exceeds Block5 on all four Avg@8 metrics, while Block10 degenerates. Relative to Token, Block3 improves three Avg@8 metrics but decreases AIME24; Pass@8 is mixed. Thus three tokens is an empirical short-window default, not a universally optimal width.

Larger blocks add cross-credit terms and change clipping (Equation 4). Subsequent Block10 runs also degenerate with finite gradient maxima of 867.50 and 865.21; their built-in AMC23 grading has a caveat (Appendix E). These observations do not isolate a cause. Ran-

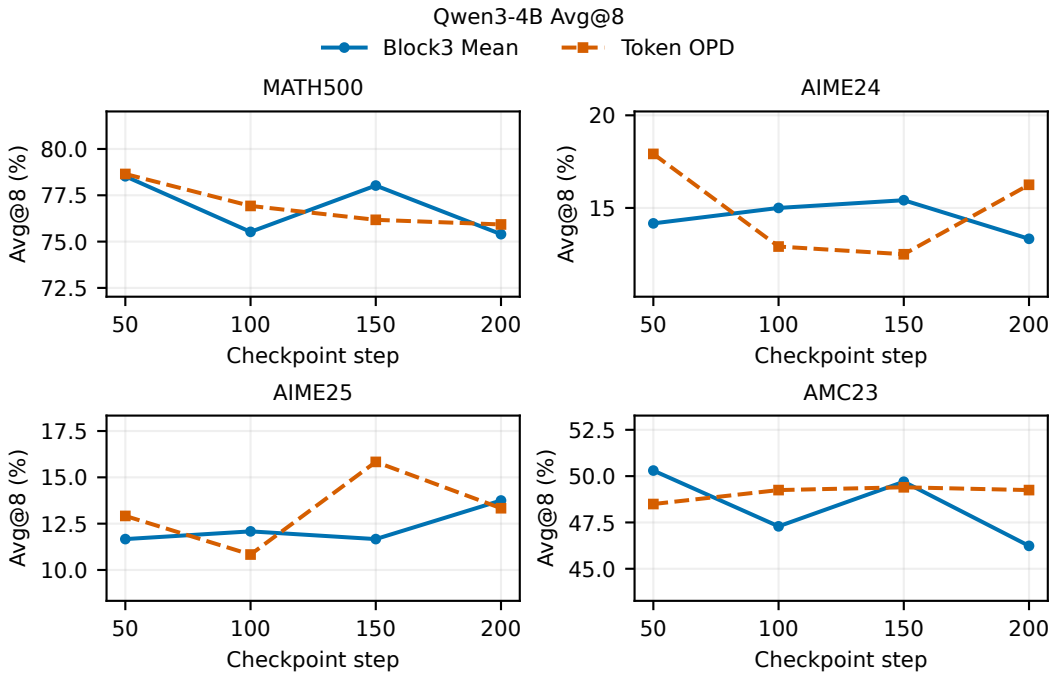

Figure 3: Current matched 4B comparison at all four saved checkpoints. Avg@8 is reported separately for each benchmark, with task-specific vertical ranges shared by both methods. Lines connect measured checkpoints, not additional evaluations. One training seed per method.

Table 4: Historical block-length sweep: Qwen3-1.7B-Base student, 4B-GRPO teacher, DAPO, 200 updates, seed 21. Cells show Avg@8 / Pass@8 (%); this is not the later same-host comparison.

Method (k)	MATH500	AIME24	AIME25	AMC23
Token (1)	69.12 / 88.60	13.33 / 26.67	6.67 / 16.67	36.75 / 62.65
Block3 (3)	69.27 / 87.80	9.58 / 30.00	8.75 / 26.67	39.91 / 61.45
Block5 (5)	67.40 / 87.60	9.17 / 20.00	7.92 / 23.33	34.19 / 61.45
Block10 (10)	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00

Token/Block3/Block5 used A800; Block10 used A100. Evaluation request seeds were not fixed explicitly. Block5 restarted with a lower rollout-memory budget; its scores survive as a curated summary. Protocol details: Appendix E.6.

dom3/Sliding3 test partitioning alternatives, not additional fixed widths or isolated components.

4.3.3 Prompting and Local Feedback Compatibility

The first 4B attempt repeats and hits the length cap before any update, excluding the block loss as the cause of those initial defects. The repair removes ChatML and separates request seeds together; it is not a chat-delimiter ablation. Appendix F documents model-specific input probes.

Repaired 4B training-cap incidence is 263/6,400 (4.11%) for Token versus 247/6,400 (3.86%) for Block3. Instruct engine length stops are 94/6,400 (1.47%) versus 77/6,400 (1.20%), without generated thinking tags. Evaluation differences in missing boxes and length stops remain benchmark-dependent (Appendix H); length control is distinct from correctness.

4.3.4 Credit Redistribution and Training Dynamics

Across 200 updates, 4B Block3's median sign disagreement is 15.57%, or 6.68% with absolute-advantage weighting. Median relative redistribution is 1.148, a magnitude ratio that can exceed one; size-one Token values are zero by construction. These quantify the intervention, not feedback correctness.

Figure 7 shows low late-training entropy and reduced gradients without Block10-like degeneration: redistribution alone does not diagnose collapse. Appendix I maps entropy and overlap with occupancy masks; changing on-policy prefixes preclude a fixed-input causal interpretation. Separating shared signals, clipping and loss scale requires component controls and replication (Appendix A).

5 Related Work

On-policy and self-distillation. GKD supports learner-generated sequences and multiple divergences (Agarwal et al., 2023), following learner-state supervision in DAgger (Ross et al., 2011). MiniLLM incorporates reverse KL and future-reward credit (Gu et al., 2023); Revisiting OPD analyzes sampled-token failures and supplies our baseline (Fu et al., 2026). OPSD uses privileged solutions to construct a self-teacher (Zhao et al., 2026). We retain an external teacher and study how its sampled feedback is shared across positions.

Selecting and weighting teacher feedback. Rethinking OPD motivates our compatibility diagnostics (Li et al., 2026a). FiRe-OPD filters trajectories and weights tokens, TIP selects informative tokens, and R2-OPD filters spans by reasoning progress (Li et al., 2026b; Xu et al., 2026; Yang et al., 2026). Block3 instead pools adjacent teacher–student log ratios, without progress labels or an extra verifier, potentially mixing conflicting preferences.

Blockwise and sequence-level policy updates. Blockwise policy-drift gating uses detached, mean-normalized old–current student drift weights with a token objective (Zheng & Jiang, 2026). Block3 aggregates teacher feedback and changes the ratio itself. Unlike GSPO's length-normalized sequence ratio (Zheng et al., 2025), it multiplies token ratios without geometric-mean normalization. Our contribution is this specified credit-sharing update and its analysis, not the generic use of blocks. Experiments compare against sampled-token OPD, not all these methods.

6 Conclusion

We introduced Block3 Mean, an on-policy distillation update that combines shared feedback over adjacent tokens with a joint block ratio and valid-length weighting. The method reuses teacher scores while changing the unit through which they influence the student. Experiments show substantial gains in the historical Qwen3-1.7B setting and positive final Avg@8 differences for the official 8B-to-1.7B Instruct pair, alongside checkpoint reversals and settings that favor token-level updates. The analysis explains how the objective introduces cross-token credit, scale changes and a different clipping geometry. Together with the retained trajectory diagnostics, these results identify local credit assignment as a concrete design choice for reasoning distillation and provide a basis for separating its optimization components.

Reproducibility Statement

Section 3 and Appendix B specify the objectives and derivations. Appendices C and H give model identities, training and evaluation settings, data checks and retention details; Appendix E documents the historical experiments. Complete states, raw trajectories and input checks are retained for the two fully archived comparisons. Tables and figures are generated from provenance-linked summaries. The accompanying manuscript source contains the rendered-result inputs; it is not a substitute for a complete anonymous release of the experimental code and data.

Ethics Statement

This study uses mathematical reasoning benchmarks and pretrained language models, without collecting human-subject data. Teacher imitation does not itself establish truthfulness or safety. We report both improvements and degraded outcomes, and distinguish mathematical accuracy from formatting and numerical diagnostics. Distribution of artifacts must respect upstream model and dataset licenses and exclude credentials, private model weights and identifying infrastructure information.

AI Use Statement

Generative AI assistants were used to refine hypotheses, help design experiments and mathematical arguments, implement and debug methods, monitor experiments, retrieve and summarize literature, process evidence, interpret results, and draft, revise and translate the manuscript. AI tools also assisted with conceptual figures; numerical result plots were generated from retained experimental data. Language models generate the on-policy training trajectories as part of the method. Automated tests and artifact checks verify specified contracts, but do not constitute independent replication. The authors remain responsible for checking all AI-assisted claims, proofs, code, citations and disclosures in the final submission.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes, 2023. URL https://arxiv.org/abs/2306.13649v3. ICLR 2024; arXiv v3, 17 January 2024.

Yuqian Fu, Haohuan Huang, Kaiwen Jiang, Jiacai Liu, Zhuo Jiang, Yuanheng Zhu, and Dongbin Zhao. Revisiting On-Policy Distillation: Empirical Failure Modes and Simple Fixes, 2026. URL https://arxiv.org/abs/2603.25562v2. Audited 2603.25562v2; revised 2026-04-27.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models, 2024. URL https://arxiv.org/abs/2407.21783. Metadata accessed 21 September 2026; abbreviated author list.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: On-Policy Distillation of Large Language Models, 2023. URL https://arxiv.org/abs/2306.08543v6. ICLR 2024; title and text from arXiv v6, 31 January 2026.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset, 2021. URL https://arxiv.org/abs/2103.03874v2. NeurIPS 2021; audited arXiv v2, 8 November 2021.

Yaxuan Li, Yuxin Zuo, Bingxiang He, Jinqian Zhang, Chaojun Xiao, Cheng Qian, Tianyu Yu, Huan-ang 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. Audited 2604.13016v2; revised 2026-04-15.

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, 2026b. URL https://arxiv.org/abs/2606.02684v2. Audited 2606.02684v2; revised 2026-06-04.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step, 2023. URL https://arxiv.org/abs/2305.20050v1. MATH-500 split and answer grading verified in the accompanying official PRM800K repository; paper abstract and metadata only.

Meta. Llama 3.2 1B Instruct: Model Card. Hugging Face model repository, 2024. URL https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct. Release date 25 September 2024; accessed 21 September 2026; repository revision 9213176726f574b556790deb65791e0c5aa438b6.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, volume 15 of Proceedings of Machine Learning Research, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html. Publisher author names; arXiv first posted in 2010.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, 2017. URL https://arxiv.org/abs/1707.06347v2. Audited 1707.06347v2; revised 2017-08-28.

Yuanda Xu, Hejian Sang, Zhengze Zhou, Ran He, Zhipeng Wang, and Alborz Geramifard. TIP: Token Importance in On-Policy Distillation, 2026. URL https://arxiv.org/abs/2604.14084v4. Audited 2604.14084v4; revised 2026-05-21.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388v1. Audited 2505.09388v1; revised 2025-05-14.

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/2608.19408v1. Audited 2608.19408v1; revised 2026-08-19.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: An Open-Source LLM Reinforcement Learning System at Scale, 2025. URL https://arxiv.org/abs/2503.14476v2. Audited 2503.14476v2; revised 2025-05-20.

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, 2026. URL https://arxiv.org/abs/2601.18734v3. arXiv v3, 20 March 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group Sequence Policy Optimization, 2025. URL https://arxiv.org/abs/2507.18071v2. Audited 2507.18071v2; revised 2025-07-28.

Liwen Zheng and Haiyun Jiang. Blockwise Policy-Drift Gating for On-Policy Distillation, 2026. URL https://arxiv.org/abs/2606.24084v1. Audited 2606.24084v1; revised 2026-06-23.

A Experimental Scope and Further Analysis

Update design and attribution. The experimental comparison evaluates Block3 as an integrated block-level update: shared advantages, joint importance ratios and valid-length-weighted block reduction. Section 3 characterizes the resulting cross-token credit, gradient scale and clipping geometry. Accordingly, the measured performance differences concern the complete update rather than advantage averaging alone. Component-specific attribution can be examined separately by holding ratios and normalization fixed when varying feedback sharing. The signal-estimation analysis describes averaging under explicit assumptions; task-level accuracy is measured by the benchmark evaluations.

Statistical unit and checkpoint reporting. Each reported arm uses one training seed. Paired-question intervals characterize evaluation uncertainty conditional on the trained checkpoints, rather than variability across training seeds. They are pointwise and use no multiple-comparison correction. Each AIME set contains 30 questions, so one additional solved question changes Pass@8 by 3.33 points. We retain the checkpoint trajectories alongside Step200, a common scheduled endpoint adopted retrospectively for reporting rather than a preregistered selection rule. Window and weighting variants arose through exploratory experiments; independent runs would evaluate the reproducibility of their observed differences.

Protocols and task coverage. Historical results are organized by their original prompt, seed-handling, teacher and implementation settings. Each stratum supplies a method comparison under its own protocol. The two fully retained comparisons share a DAPO-derived training pool; the overlap audit checks the retained data rather than semantic equivalence or upstream model exposure. The 4B completion study compares the two trained methods without a matched original-student evaluation, while the official Instruct pair additionally includes that control. Independent datasets, training seeds and teacher–student pairs provide natural axes for extending this evaluation.

Complementary diagnostics. Prompt probes examine behavior before optimization, while retained training trajectories characterize subsequent evolution. Entropy maps describe the prefixes visited by the current policy; missing boxed markers measure one specific formatting property. Gradient norms and length-stop rates track numerical and generation behavior, and sign disagreement measures how feedback is redistributed. Benchmark accuracy provides a separate task-level assessment. Interpreting these measurements jointly distinguishes input compatibility, optimization dynamics and mathematical correctness without treating one as a proxy for all three.

Baseline and computational scope. Sampled-token OPD is the matched experimental baseline; other OPD recipes are discussed as related work. Block3 changes how the available teacher feedback drives the update and still scores every valid response position. The present evaluation therefore concerns the behavior and accuracy of this update, not a reduction in teacher-scoring work. Per-task, per-checkpoint reporting captures the observed variation. Further comparisons can examine component choices and accuracy–cost trade-offs under a common budget.

B Additional Derivations

The production reduction. Let $m_t \in \{0,1\}$ be the final actor mask, $n_b = \sum_{t\in b} m_t$, $N = \sum_t m_t$, and $\eta = 10^{-8}$. Empty blocks have $\bar{a}_b = 0$. The code minimizes the negative clipped surrogate with a fractional block mask:

$$L_{\text{blk}} = -\frac{\sum_b (n_b/k)\mathcal{C}(R_b, \bar{a}_b)}{N/k + \eta} = -\frac{\sum_b n_b \mathcal{C}(R_b, \bar{a}_b)}{N + k\eta}. \tag{7}$$

The token denominator is $N+\eta$. Invalid positions are zeroed before block aggregation; holes are not removed and do not move neighboring tokens into different blocks. Both empty

blocks and padding have zero weight. With actor micro-batch size one, the update averages independently normalized response losses. It is not the batch-global token mean when response lengths differ. Thus the configuration label token-mean alone does not establish the global weighting used in DAPO (Yu et al., 2025).

Clipping convention. For lower and upper ratios $d = 0.8$, $u = 1.2$, define $s(r, a) = \min\{ra, \mathrm{clip}(r, d, u)a\}$. The inherited objective is

$$\mathcal{C}(r, a) = \begin{cases} \max\{s(r, a), 3a\}, & a < 0, \\ s(r, a), & a \geq 0. \end{cases} \tag{8}$$

This is ordinary PPO clipping with an additional negative-advantage dual-clip rule. The block ratio is formed before applying this operation. Gradient-norm clipping at one is a separate optimizer operation. Neither operation is a guarantee that all future policy ratios or teacher KL values remain inside a prescribed region.

First-order derivatives and synthetic checks. On the active branch, for valid $t \in b$, the derivatives with respect to current sampled log probabilities are

$$\frac{\partial L_{\mathrm{blk}}}{\partial \log p_\theta(y_t \mid h_t)} = -\frac{n_b \bar{a}_b R_b}{N + k\eta}, \qquad \frac{\partial L_{\mathrm{tok}}}{\partial \log p_\theta(y_t \mid h_t)} = -\frac{a_t r_t}{N + \eta}. \tag{9}$$

At ratio one, their policy loss values agree in the $\eta = 0$ idealization for arbitrary advantages, yet their derivatives need not. For five tokens with $a_t = 1$, the token derivatives are all $-1/5$, whereas Block3 gives $(-3/5, -3/5, -3/5, -2/5, -2/5)$. For $(a_1, a_2, a_3) = (1, -1, 1)$, the token derivatives are $(-1/3, 1/3, -1/3)$ and the block derivatives are $(-1/3, -1/3, -1/3)$. These are synthetic derivative examples, not estimates of real parameter-update magnitudes.

A restricted credit-horizon interpretation. Assume a fixed horizon T, a deterministic contiguous partition, no masks or truncation, locally parameter-independent student support on which the fixed teacher q is positive, and sampling from the full current student distribution p_θ. Suppose log probabilities are differentiable and differentiation and expectation can be interchanged. Define $c_t = \log p_\theta(y_t \mid h_t) - \log q(y_t \mid h_t)$ and $g_t = \nabla \log p_\theta(y_t \mid h_t)$. Then

$$\nabla \mathrm{KL}(P_\theta \| Q) = \mathbb{E}\left[\left(\sum_t c_t\right)\left(\sum_t g_t\right)\right] \tag{10}$$

$$= \mathbb{E}\left[\sum_t g_t \sum_{u=t}^{T} c_u\right]. \tag{11}$$

The derivative of the log-density term has expectation zero. Past terms vanish because c_u for $u < t$ is measurable at h_t and $\mathbb{E}[g_t \mid h_t] = 0$. Define $H_{\mathrm{blk}} = \sum_b (\sum_{u \in b} c_u)(\sum_{t \in b} g_t)$ and $H_{\mathrm{seq}} = (\sum_u c_u)(\sum_t g_t)$. Applying the same argument to an unclipped block estimator at the old policy gives

$$\mathbb{E}H_{\mathrm{blk}} = \mathbb{E}\left[\sum_t g_t \sum_{u=t}^{e(t)} c_u\right], \qquad \mathbb{E}(H_{\mathrm{seq}} - H_{\mathrm{blk}}) = \mathbb{E}\left[\sum_t g_t \sum_{u > e(t)} c_u\right], \tag{12}$$

where $e(t)$ is the end of token t's block. Relative to purely local token feedback, block aggregation can restore within-block future credit; it omits credit across a block boundary. Thus cross terms should not all be called harmful leakage. Their omitted vector can cancel, so increasing block size need not monotonically reduce bias. MiniLLM already accounts for future rewards; this interpretation is not a claim to first introduce temporal credit assignment in distillation (Gu et al., 2023).

These identities do not establish that the production loss is an exact sequence-KL optimizer. Top-p sampling changes the sampling distribution relative to the scored student; a product over masked positions is not generally a block marginal likelihood ratio; clipping and frozen old-policy targets alter the objective away from the old policy; and a random response length in the normalization introduces a dependence on future actions. The identities are therefore explanatory idealizations with explicit assumptions, not guarantees for the experiment.

Averaging error decomposition. With $a = \mu + \varepsilon$ and $\mathbb{E}\varepsilon = 0$, expand $Wa - \mu = (W - I)\mu + W\varepsilon$. The expected cross term vanishes, and $\mathbb{E}\varepsilon^\top W^\top W\varepsilon = \mathrm{tr}(W\Sigma W^\top)$, proving Equation 5. For independent, equal-variance noise and a constant signal within a k-token block, each averaged noise variance is $k\sigma^2/k^2 = \sigma^2/k$. For a deterministic signal $(2, -1, 2)$ with zero noise, the unaveraged error is zero; replacing every component by its mean one creates squared error six. This is a counterexample to universal signal-estimation benefit, not a claim that the teacher's per-token ratio is always the correct task-level signal.

What sliding phases can average out. Let X contain shared parameters, the collected batch, masks, frozen targets and all other forward randomness. Let $L_r(X)$ be the fully clipped and reduced loss for offset $r \in \{0, 1, 2\}$, and $G_r(X) = \nabla_\theta L_r(X)$. For a conditionally uniform random offset U, the implemented phase average satisfies

$$G_{\mathrm{slide}}(X) = \frac{1}{3}\sum_{r=0}^{2} G_r(X) = \mathbb{E}[G_U(X) \mid X]. \tag{13}$$

With finite second moments and the same distribution of X, the law of total covariance gives

$$\mathrm{Cov}(G_U) - \mathrm{Cov}(G_{\mathrm{slide}}) = \mathbb{E}_X[\mathrm{Cov}(G_U \mid X)] \succeq 0. \tag{14}$$

This removes variance from offset randomization conditional on a common state; it does not compare Sliding3 with a fixed Block3 partition. PPO clipping may be active because it occurs inside each L_r. Gradient-norm clipping and Adam are nonlinear later operations and are outside this identity. Distinct training histories induce distinct distributions of X, so the identity is not a performance guarantee for the retained runs or an assertion about an exactly realized average under one fixed RNG seed. Averaging losses is also not the same as averaging advantages before a single clipping operation.

For $N > 0$ and matching phase-zero partitions, $L_{\mathrm{hist}} = NL_0/(N + k\eta)$, including partial-tail weights. The gradients have the same factor because the collected mask is frozen. At $\eta = 0$, the old reduction and phase-zero reduction are identical. Leading padding or holes can invalidate the matching-partition premise; floating-point operation order can introduce small numerical differences even when the ideal expressions agree.

C Configuration and Data Provenance

Current model identities. The original student is Qwen/Qwen3-4B-Base, obtained from ModelScope at revision bbd6fc8d23e8788d987b7b970cbb7bd31c826e38. The public teacher's retained model card identifies lllyx/Qwen3-4B-Base-GRPO, released with Rethinking OPD (Li et al., 2026a). That repository now redirects to Thinking-Space/Qwen3-4B-Base-GRPO. It is a researcher's GRPO-adapted checkpoint, not an official Qwen instruction model and not a privately trained checkpoint from this project. Download-cache metadata records revision 1f3b2966edfb75f2f98a00617588c1f748088422; the two weight-file ETags match the retained asset hashes. The uploader's model card describes GRPO adaptation on DAPO-Math-17k-Processed. This upstream teacher training is distinct from our student's 200-update OPD runs.

Exact completion prompt and stopping. Training and evaluation in the 4B Base comparison use the same completion function, without ChatML:

> [question]
> Please solve the problem step by step and put the final answer in \boxed{}.
> Solution:

Blank lines separate the components, with a final newline after Solution:. Thinking is explicitly disabled, while ordinary step-by-step text is allowed. EOS is 151643 with no additional stop-token IDs. Actual input token IDs are checked after checkpoint merging. Evaluation runs on four independent tensor-parallel-size-one workers, using memory fraction 0.9 rather than the training rollout fraction 0.6. CPU checkpoint merging does not modify the original training weights. Raw responses, input tokens, request seeds and engine stopping reasons are retained.

Table 5: Shared configuration for the current 4B completion comparison. Differences between training rollout inference and full evaluation are intentional and identical across arms.

Setting	Value
Training seed / data seed	21 / 21
Updates / prompts per update / samples	200 / 4 / 8
PPO mini-batch / epochs	32 responses / 1
Actor / rollout-logprob / teacher micro-batch	1 / 4 / 1 per GPU
Optimizer / learning rate	AdamW / 2×10^{-6}
Adam betas / epsilon / weight decay	$(0.9, 0.999)$ / 10^{-8} / 0.01
Schedule / warmup	Constant / zero warmup ratio
Gradient-norm clipping	1.0
PPO lower / upper clip / dual clip	0.2 / 0.2 / 3.0
Explicit entropy / extra KL loss	0 / disabled
Outcome reward / OPD future-return weight	0 / 0
Teacher target log-ratio clipping	Disabled
Special-token OPD mask	Disabled; retain upstream response mask
Prompt cap / response cap	2,048 / 16,384
Temperature / top-p / top-k	1.0 / 0.9 / unrestricted
Training rollout / eval memory fraction	0.6 / 0.9
Training max batched rollout tokens	18,432
Hardware / tensor parallelism	4 A100-80GB / 1
Training parameter / optimizer offload	Enabled / enabled
Gradient checkpointing	Enabled
Save steps / checkpoint deletion	50, 100, 150, 200 / none
Diagnostic interval / top-k / position bin	Every update / 16 / 128 tokens
Block sign threshold	10^{-4}

Effective batching and seeds. Data batching follows the frozen seed-21 loader. Pair validation compares the actual source and prompt identities after collection; relying on equal seed labels alone would be insufficient if the arms consumed RNG differently. The current request-seed rule uses a stable SHA-256 derivation including step and sample identity and does not consume the training RNG merely to log a response. Eight requests for one prompt have separate seed identities. Historical fixed-request-seed runs are explicitly separated in Appendix E. A separate window-seed value is retained in the run card but has no effect on a fixed-block treatment.

Training pool. A direct read of the retained parquet verifies exact question–answer sequence repetition: a 17,917-row period occurs 100 times. There are 17,237 unique raw question strings and 17,249 unique raw question–answer pairs, without additional Unicode or mathematical normalization. These are exact-string counts, not semantic deduplication. The original expansion command and upstream dataset revision were not recovered. The retained file's SHA-256 starts with cf359f257a320aecb; the internal manifest preserves the complete digest. No claim of comprehensive train–test semantic decontamination follows from hashing. Both methods nevertheless use exactly the same retained pool, rather than separately downloading whichever dataset version currently bears the same name.

The retained converter names BytedTsinghua-SIA/DAPO-Math-17k, file data/dapo-math-17k.parquet, as its public default input. However, the historical manifest records a copied, already converted parquet, not its original Hub revision or the command that created the 100-fold expansion. The current converter's default deduplication would change this pool. Its current defaults must not be presented as the exact historical preprocessing recipe.

Retrospective string-overlap audit. Comparing the original question fields, two of MATH500's 500 questions occur exactly in the retained training pool. Collapsing Unicode whitespace to single spaces, stripping boundaries and applying Unicode casefold yields four matches in total, including those two. This normalization does not rewrite mathematics or LaTeX. The exact held-out indices are 80 and 276; normalized-only additions are 285 and 412 (zero-based indices in the retained MATH500 artifact). The corresponding counts

for AIME24, AIME25 and AMC23 are zero under both rules. We use the previously verified 17,917-row period to cover the pool's string set and check evaluation-file hashes and question-field identities. This is a retrospective data audit, not a filtering intervention: all reported full-benchmark scores retain the original questions. Pool membership alone does not demonstrate actual exposure during these 200 updates; string-level zero overlap does not establish semantic decontamination or audit the teacher's training/pretraining corpus.

A separate audit streams the actual retained input metadata from all 400 compressed training archives, verifies their recorded hashes, and compares both methods' complete 6,400-entry input streams. Each arm contains 800 prompt occurrences, 783 exact unique question strings and 782 whitespace/casefold-normalized strings. There are zero exact evaluation-question exposures and one normalized-only occurrence: MATH500 index 412 at Step71, eight trajectory records per arm. The other three pool-overlap candidates are absent. Both arms encounter the same input at the same step; Step50 precedes that exposure. This observation does not audit semantic equivalents or upstream model training. The separate 496-question sensitivity analysis excludes all four pool-matched items, independently of which happened to be drawn, while retaining full MATH500 as the primary report.

Evaluation artifacts and grading. MATH500 uses the HuggingFaceH4/MATH-500 test artifact, related to the MATH and PRM800K evaluation lineage (Hendrycks et al., 2021; Lightman et al., 2023). AIME24 was converted from BytedTsinghua-SIA/AIME-2024. AIME25 and AMC23 were converted from the corresponding retained test parquet files in the public OPD repository. Their local counts are respectively 30 and 83; in particular, the AMC23 count must not be silently equated with another paper's 40-question subset. The converted JSONL hash prefixes are cc164b47b60771ee (MATH500), 6aee5ed24b811ebc (AIME24), daedf1e406b9d73c (AIME25), and a07c52c0098a536a (AMC23).

The same historical external grader, SHA-256 prefix 04f7a0328be18409, scores both arms. Eight independently seeded generations are retained for each question. Every finalized checkpoint checks the expected question identities, response counts, raw output archives, and agreement between summary and graded records. This validates the evaluation contract, not the logical infallibility of the symbolic grader. Grader errors or missing parseable answers must not be reinterpreted as known mathematical correctness. Missing boxed-answer markers and engine length stops remain separate measurements.

The public repository for the two retained parquet sources is https://github.com/thunlp/OPD, checkout 1fd6cca846126af90d82ef122e8af261f59d2d37, at datasets/test_data/AIME25/test.parquet and datasets/test_data/AMC23/test.parquet. The checked-out files match the historical source hashes. Their more remote upstream Hub namespaces were not recovered. This data-source repository is distinct from the Revisiting OPD training codebase.

Resource accounting and retention. The two current 200-step training runs took approximately 9.4 and 9.7 hours on four A100 GPUs, excluding evaluation and bounded acceptance probes. This is roughly 76 GPU-hours of training for this pair, not an estimate of the entire project's compute. Total historical compute cannot be reconstructed from the retained summaries with comparable precision. All four full states per arm, including optimizer, scheduler and RNG state, remain retained. Each current arm has 200 diagnostic records and 6,400 raw training trajectories. Some early experiments did not save raw training trajectories; sampling an old checkpoint now would not recover those original rollouts.

Software and precision. The retained runtime uses Python 3.12.13, PyTorch 2.8.0, vLLM 0.11.0, Transformers 4.57.6, Ray 2.55.1, NumPy 1.26.4, Datasets 4.8.5 and OmegaConf 2.3.0. These versions were read from the evaluation environment; they are not inferred from a current package installation guide. The unchanged FSDP actor defaults load the training model in float32 and use bfloat16 forward parameters with float32 reduction and buffers; the reference model and rollout inference use bfloat16. Both arms use the same frozen runtime and model/tokenizer bytes. A generic tokenizer-regex warning was retained, not patched in only one arm during evaluation; the acceptance gate checks the actual prompt token IDs for

the paired protocol. This check establishes matched behavior, not that arbitrary prompts are immune to tokenizer defects.

D Diagnostic Definitions

Entropy and occupied prefixes. At a valid student-generated prefix h_t, let $p_t(v)$ and $q_t(v)$ be the student and teacher softmax distributions over the shared vocabulary. We record full-vocabulary Shannon entropies in nats,

$$H_{\mathrm{S},t} = -\sum_v p_t(v) \log p_t(v), \qquad H_{\mathrm{T},t} = -\sum_v q_t(v) \log q_t(v). \tag{15}$$

The configured temperature is one. These are not entropies of a renormalized top-16 support, nor negative log probabilities of the one sampled token. The signed gap is $H_{\mathrm{T},t} - H_{\mathrm{S},t}$; the absolute gap is the absolute value at each position, averaged afterward. Therefore the mean absolute gap is not generally the absolute value of the mean signed gap. Student and teacher are evaluated on the same student prefix. Their training-time snapshots are taken before the update for that batch; appended optimizer and post-update ratio diagnostics describe a different phase of the same update.

Top-token alignment. Let S_t and Q_t be the respective top-16 token-ID sets and $O_t = S_t \cap Q_t$. We compute

$$\mathrm{Overlap}_t = |O_t|/16, \quad M_{\mathrm{S},t} = \sum_{v \in O_t} p_t(v), \quad M_{\mathrm{T},t} = \sum_{v \in O_t} q_t(v). \tag{16}$$

Probability masses use the original full-vocabulary probabilities. They are not percentages of the top-16 mass. When $O_t \neq \varnothing$, the implemented overlap-token advantage first renormalizes within the intersection:

$$\bar{p}_t(v) = p_t(v)/M_{\mathrm{S},t}, \quad \bar{q}_t(v) = q_t(v)/M_{\mathrm{T},t}, \quad U_t = \frac{1}{|O_t|} \sum_{v \in O_t} \bar{p}_t(v) \log \frac{\bar{q}_t(v)}{\bar{p}_t(v)}. \tag{17}$$

Thus $U_t = -\mathrm{KL}(\bar{p}_t \| \bar{q}_t)/|O_t| \leq 0$ in exact arithmetic; it is not the average raw sampled-token advantage. Empty intersections are excluded from its mean, while their overlap ratio and masses are zero. We state the implemented definitions rather than assuming every diagnostic with the same name in another paper uses identical normalization.

Local sign disagreement and redistribution. Let b_t be the block mean copied to valid token t, m_t the recorded response mask, and $\epsilon = 10^{-4}$. Define

$$e_t = m_t \mathbf{1}[|a_t| > \epsilon] \mathbf{1}[|b_t| > \epsilon], \tag{18}$$

$$f_t = e_t \mathbf{1}[\mathrm{sign}(a_t) \neq \mathrm{sign}(b_t)]. \tag{19}$$

The recorded sign-flip rate and its magnitude-weighted variant are

$$F = \frac{\sum_t f_t}{\max(1, \sum_t e_t)}, \qquad F_w = \frac{\sum_t |a_t| f_t}{\max(\epsilon, \sum_t |a_t| e_t)}. \tag{20}$$

Tokens with near-zero local or shared feedback are not classified as direction reversals. A batch with no eligible entries yields zero together with zero coverage, not evidence that all directions agree. The redistribution statistics are

$$D = \frac{\sum_t m_t |b_t - a_t|}{\max(1, \sum_t m_t)}, \qquad D_{\mathrm{rel}} = \frac{\sum_t m_t |b_t - a_t|}{\max(\epsilon, \sum_t m_t |a_t|)}. \tag{21}$$

The source calls these quantities leakage_magnitude and normalized_leakage. We use "redistribution" because neither difference from token feedback nor a sign reversal proves harmful task-level credit. These statistics also do not include the joint-ratio derivative or its effective scale. For a size-one Token treatment, $b_t = a_t$ and both families vanish by construction; this is not independent evidence of superior training behavior.

Position aggregation and missingness. For every update, the retained snapshot stores sums, squared sums and valid counts at generated response positions, using the diagnostic response mask. In the current completion run, the historical EOS mask includes the first EOS (151643), excludes subsequent positions and the prompt, and includes every position of a cap-length response without EOS. The figures combine consecutive positions into bins of 128 tokens using ratios of sums to counts, not an unweighted average of precomputed means. Alignment is sampled at stride one. Sign-flip heatmaps use eligible counts rather than all response-token counts. A displayed bin must have at least one original position with eight valid observations; the threshold is applied to the maximum per-position count inside the bin, not to its summed token count. A gray masked cell is not zero entropy or zero disagreement. A bin's denominator counts token-position observations and need not equal the number of distinct trajectories. The changing density of long responses is shown separately from the measured values. Comparing two bins or training steps conditions on different on-policy prefixes and therefore cannot on its own establish which process caused collapse.

Optimizer, lengths and format. The recorded gradient norm is the worker's norm before configured norm clipping, reduced by the training logger; it is not a measurement that every token-specific or layer-specific derivative remains bounded. Policy-gradient loss is the actual clipped objective, which need not be directly comparable in scale between Token and Block3. Training response-cap incidence counts trajectories that reach the configured cap. For evaluation we additionally retain the engine's length-stop reason; a cap statistic based on a mask and an engine stopping reason should not be assumed identical without checking.

Our format indicator is the evaluation wrapper's literal test that the response does not contain the string \boxed. It does not validate brace balance, extractability or whether the box contains the final answer. It is narrower than "the answer has any formatting defect," and it is separate from incorrect arithmetic, symbolic grading failure, malformed thinking tags and length termination. A response can contain the marker and still be incorrect; a length-stopped response may already contain it. All rates use the relevant benchmark's retained response count. We make no cross-benchmark aggregate formatting claim.

E Historical Evidence and Protocol Boundaries

Table 6: Complete historical Step200 Avg@8 comparison (%). Original prompts and protocols are retained; compare methods only within each student stratum. These are not extra training seeds of the two fully retained comparisons.

Student	Method	MATH500	AIME24	AIME25	AMC23
Qwen 1.7B	Token	45.85	5.42	2.50	21.84
Qwen 1.7B	Block3	54.75	8.75	5.42	27.26
DeepSeek 1.5B	Token	84.70	39.58	29.17	73.34
DeepSeek 1.5B	Block3	84.75	43.75	30.83	73.80
Qwen 0.6B	Token	42.88	2.50	0.83	18.98
Qwen 0.6B	Block3	0.57	0.00	0.00	0.45
Llama 1B	Token	15.55	0.42	0.00	6.02
Llama 1B	Block3	6.30	0.00	0.00	1.51

Inventory and evidence strength. The historical inventory contains 17 experimental series and 171 evaluation records, including alternative grader views and explicit missing results. It is not a count of 171 independent training runs. Below, percentages are reported separately for each benchmark. A dash means that no completed result was recovered, not zero accuracy. In the n=8 tables, cells contain Avg@8 / Pass@8. The early n=1 and 100-question pilot tables retain their original metric definitions. Regrading the same generations does not add samples, and re-evaluating an old checkpoint does not add a training seed.

The strongest retained evidence consists of original evaluation summaries with their acceptance records. Other tables come from finalized paired artifacts or curated summaries. The earliest block-sum and block-advantage results are available only as rounded values in historical reports; we retain that limitation rather than reconstructing exact counts from rounded percentages. Source paths and identifying infrastructure information remain outside the anonymous manuscript. A complete internal provenance map records which fields support each value.

H01–H05: clean-room exploration. The earliest student and teacher were the official post-trained Qwen3-0.6B and Qwen3-4B, not their Base variants and not the later GRPO teacher. The 3,400-prompt training pool combined 2,000 GSM8K training examples with 200 training examples from each of MATH's seven subjects. Small weighting pilots, outcome weighting, reference anchoring, naive block sums, and sum/mean/mixed variants were explored. Full evaluation meant GSM8K's 1,319 test examples and 500 MATH examples, n=1, with a 256-token output cap, temperature 0.7 and top-$p = 0.95$. The 100-example pilots are not full evaluations and their sample-count parameter is not recoverable from those summaries.

These runs use a different loss family from the main study. The clean-room implementation applies detached weights and advantages to summed current log probabilities without a PPO ratio, clamps its raw signal, and discards incomplete blocks. Its mixed mode interpolates token-local and block-mean signals; this is not the later block-sum/block-mean interpolation option. Exact settings for some named router or anchor variants are incomplete. Consequently, the old pilot tables document the research trajectory and negative outcomes, not replications of the current loss or implementations of similarly named external papers. Outcome weighting did not prevent endpoint degradation: the two reweighted 200-step arms attain primary-grader MATH500 accuracies of 16.60% and 16.40%, below the recorded 20.20% initial reference; the standard arm attains 14.80%. Missing intermediate reweighted evaluations prevent locating the onset of that degradation. The standard arm's Step50 entry in an older composite chart came from an earlier pilot and is not connected here as if it were the same training run.

H06–H09: size sweep, collapse and Qwen 1.7B. The DAPO-era Qwen3-1.7B-Base student uses the retained 4B GRPO teacher. The original block-size sweep mixed hardware contexts and did not explicitly fix evaluation seeds in every older invocation. Token, Block3 and Block5 used A800 hardware; Block10 used A100 hardware. The retained Step200 Avg@8 / Pass@8 table for $k \in \{1, 3, 5, 10\}$ appears in Appendix E.6. Block3 exceeds Block5 on every Avg@8 metric and on three Pass@8 metrics, tying AMC23 coverage. Its comparison to Token is mixed, so the sweep motivates retaining a short block rather than claiming universal superiority. The Block5 scores survive as a curated summary rather than recovered original evaluation files. The first Block5 attempt stopped at Step42 with a vLLM OOM; the successful restart lowered the rollout-memory fraction to 0.6. These resource changes must not be hidden when interpreting the sweep.

An earlier clean-room pilot compared $k \in \{1, 2, 3\}$ at 50 updates (Appendix E.4). Its recorded GSM8K/MATH500 accuracies were 45.94/20.00 for Token, 45.41/19.40 for Block2 and 46.85/20.20 for Block3. It used official 0.6B/4B post-trained models and a non-PPO loss, so this is preliminary motivation, not a $k = 2$ result for the later 1.7B PPO implementation. No completed $k = 2$ arm exists in the retained 200-step sweep.

The later same-host Token/Block3 comparison provides the cleaner historical method comparison. At Step200, its MATH500 Avg@8 changes from 45.85% to 54.75%, AIME24 from 5.42% to 8.75%, AIME25 from 2.50% to 5.42%, and AMC23 from 21.84% to 27.26%. Intermediate checkpoints are not uniformly favorable. Six later checkpoint evaluations regenerate outputs for the same Token/Block3 weights; these are evaluation sensitivity records, not six new training runs.

Both 1.7B arms use ChatML and request <think>...</think> during training without explicitly disabling thinking in the template kwargs. Evaluation uses ChatML with thinking disabled and no instruction requiring those tags, retaining the already-closed empty thinking-control prefix. The old collector resets the per-request training seed to 21, so

nominal eight-response groups may contain duplicates. The original full training-token trajectories were not retained, preventing retrospective measurement of their exact duplicate rate. These shared historical choices do not become the current completion protocol merely because the dataset and global seed match.

The Block10 diagnostic repetitions both use training seed 21, one on A800 hardware and one on A100 hardware. Both record zero scores on all four benchmarks at Step200 while available optimizer diagnostics remain finite. Their grader is the legacy built-in VERL path, not the pinned external path used for the current comparison. In particular, built-in AMC23 zero scores have a known compatibility caveat and cannot independently establish absence of AMC capability; we retain them as historical outputs, not validated capability measurements. The severe generation degradation and failures on the other tasks remain relevant observations. These are two hardware-context repetitions, not a multiple-training-seed confidence estimate. Neither the scalar summaries nor these endpoints prove a unique cause such as gradient explosion, teacher unreliability, or reverse-KL mode collapse.

Table 7 makes the limitation of a set-overlap diagnostic concrete: the two runs retain top-16 overlap ratios near 0.8 at Step200, while the shared tokens carry only about 0.17–0.24 of the full probability mass. High rank-set overlap need not imply concentrated agreement. The largest finite pre-clipping gradient norms occur at different steps (191 on A800 and 1 on A100), so these summaries do not support one reproducible temporal ordering of collapse.

Table 7: Retained Block10 diagnostic summaries. Entropies (nats), overlap ratio and intersection probability masses are measured at Step200; the last column is the maximum pre-clipping gradient norm over each run. Values come from the archived diagnostic report rather than a fresh training replication.

Hardware	H_S	H_T	Overlap	M_S	M_T	Max. grad
A800	7.146	6.745	0.837	0.226	0.237	867.495
A100	7.576	6.962	0.793	0.166	0.186	865.207

H10–H11: another distilled pair and alternative windows. The DeepSeek-R1-Distill-Qwen-1.5B/JustRL pair is a separate released-model comparison and still has a Qwen-derived backbone; it is not a new non-Qwen architecture. Its Step200 Avg@8 differences favor Block3 by 0.05, 4.17, 1.67 and 0.45 percentage points on MATH500, AIME24, AIME25 and AMC23, respectively. However, Step200 Pass@8 decreases on MATH500 and AIME24; Step50 AIME24/AMC23 and Step100 AIME25 also contain regressions. The archived protocol classifies this as a failed directional replication: its original criterion required positive equally weighted task-mean differences for both Avg@8 and Pass@8. We preserve that historical decision without using an aggregate score in the present tables. Positive Avg@8 endpoints should neither erase the failed criterion nor be described as uniformly negative. One training seed does not establish general transfer; all task/checkpoint views remain available.

Random3 and Sliding3 were trained and evaluated as their own pair at Steps 50, 100 and 200. Random3 chooses one of three partition offsets per update using a saved PCG64 RNG (seed 910021); Sliding3 averages the three complete offset-specific losses, after each phase's PPO clipping. Both retain partial edge windows. Their length-weighted phase-zero loss is algebraically equivalent to historical Block3 when valid responses start at position zero with right padding and the 10^{-8} denominator stabilizer is ignored. It would be misleading to call this a fundamentally different normalization. The implementation restarts coordinates at mask-separated valid regions and changes floating-point summation, so arbitrary masks need not give identical partitions. Each still performs one optimizer update per batch. A late advantage of Sliding3 over Random3 does not establish superiority over a newly matched fixed Block3 or Token control. Both alternatives can deteriorate during training. The archived grader recovery changes evaluation plumbing, not the frozen training run. Corrected external views support method comparisons; older built-in views remain explicitly separate, including the AMC23 grading caveat above.

H12–H15: small Base and instruction-tuned students. The old Qwen3-0.6B-Base attempt trained Token OPD but its Step50 evaluation was stopped during truncation diagnosis; no corresponding Block3 training is claimed. A later paired experiment explicitly disables thinking and retains ChatML, with both arms using the same repaired non-thinking protocol. This repair also changes the EOS stopping set and masks according to actual generated lengths; it is not a prompt-only intervention relative to the older run. It produces a strong negative transfer result for Block3: its Step200 MATH500 Avg@8 is 0.57%, versus 42.88% for Token. The 1.7B result is therefore not a universal argument for using blocks in smaller students.

The Llama pair is Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct, obtained as ModelScope mirrors of the same Meta release (Meta, 2024). It is not a 3.1 teacher paired with a 3.2 student, and neither checkpoint is a Base model. In the first native-prompt attempt, initial-model evaluation completed but Token training was deliberately stopped at Step146. A new pair then reproduced the historical 1.7B training/evaluation instruction difference while retaining Llama-native role boundaries. Token and Block3 complete 200 steps in this setting, with Block3 substantially worse on MATH500 and AMC23. The thinking flag is not a native Llama mode switch; an instruction asking for think tags is separate from the tokenizer's role format.

Both the 0.6B and Llama training series retain the old per-request fixed seed 21. Archived groups show that the eight nominal training responses to one prompt can be duplicates, so 6,400 retained trajectories do not imply 6,400 independent generations. The current completion comparison repairs this request-level seed rule. This training-sampling limitation is distinct from the eight independently seeded evaluation responses. It applies to interpreting effective exploration, not to retroactively changing any recorded benchmark score.

H16–H17: pre-update repetition and the completion restart. The 4B Base-model ChatML attempt was stopped after four updates. Repetition and length-cap hits were already present in its initial pre-update outputs, so those initial failures cannot be caused by its later Block3 gradients. Bounded prompt diagnostics examined native input token IDs, thinking instructions, EOS handling and request-seed reuse. The subsequent completion protocol simultaneously removes ChatML and uses independently derived per-request seeds. It is an engineering repair with matched new Token/Block3 arms, not a single-factor causal ablation of role tokens.

A first formal restart failed before its first update with a temporary-directory cleanup error on network storage. The retry moved compiler/temporary caches to executable shared memory while retaining the frozen training code and model/data contract. The failed attempt and its first trajectories remain retained. CPU/GPU acceptance probes that save and resume Steps1–2 are recorded as technical acceptance, not as extra benchmark experiments. The current pair's final results appear in Appendix G, replacing the pre-evaluation snapshot in this historical inventory.

Archive inclusion. The tables report completed evaluations with retained result artifacts; diagnostic probes and interrupted attempts are identified separately. Later checkpoint sampling is recorded as a new evaluation, not as recovery of the original training trajectories. The interpretation of the integrated update and directions for further comparisons are discussed in Appendix A. Table 6 includes all four historical endpoint comparisons; Table 1 selects two positive Avg@8 examples for the main text. The following figures and tables retain the complete checkpoint trajectories.

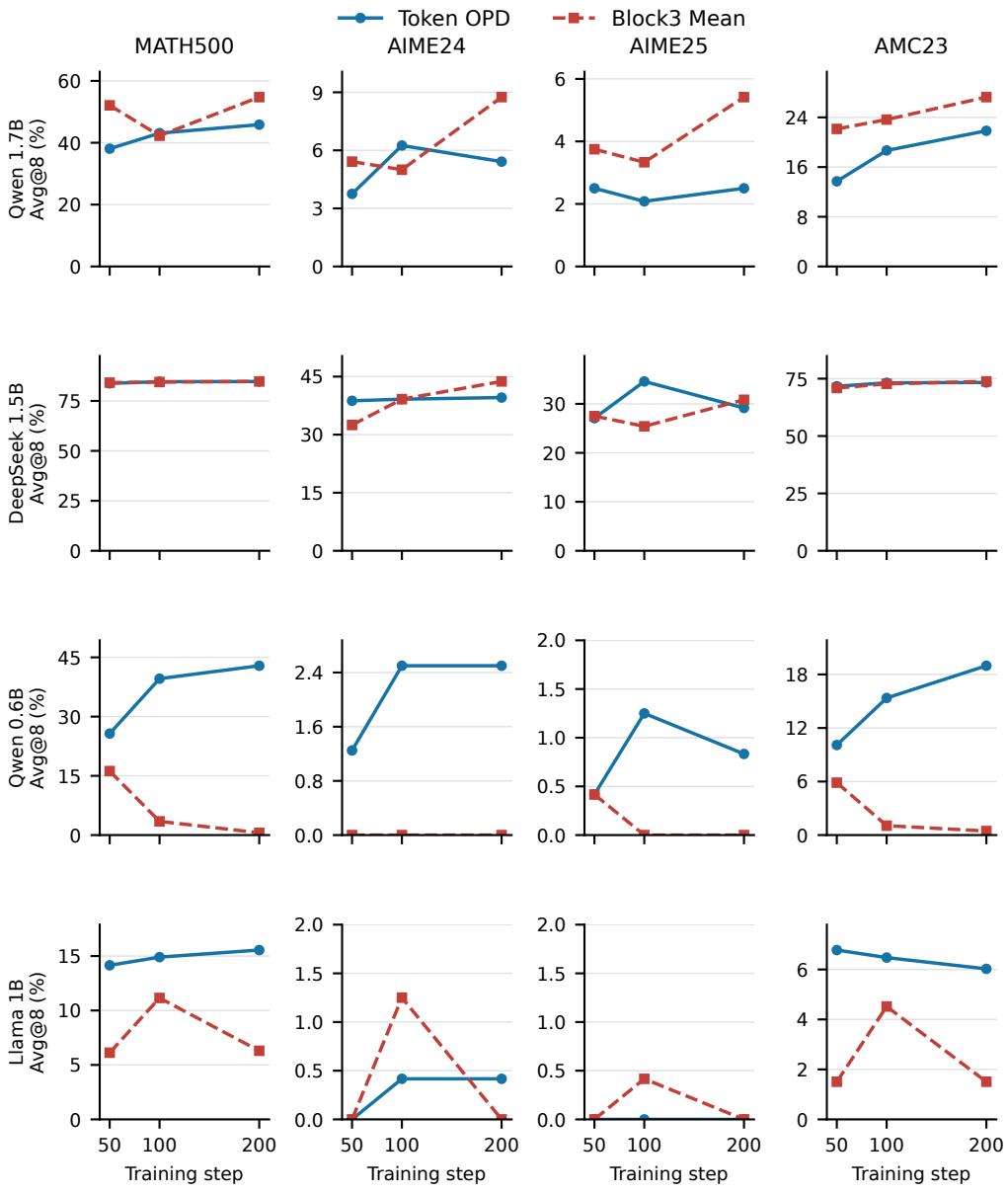

Figure 4: Historical Avg@8 checkpoint trajectories, using the audited external view for Qwen 1.7B and the retained historical external view for the other strata. Each row keeps its own protocol; all use one training seed, not independent multi-seed replications. Panel scales differ. Lines only connect the three measured checkpoints and do not represent measurements between them.

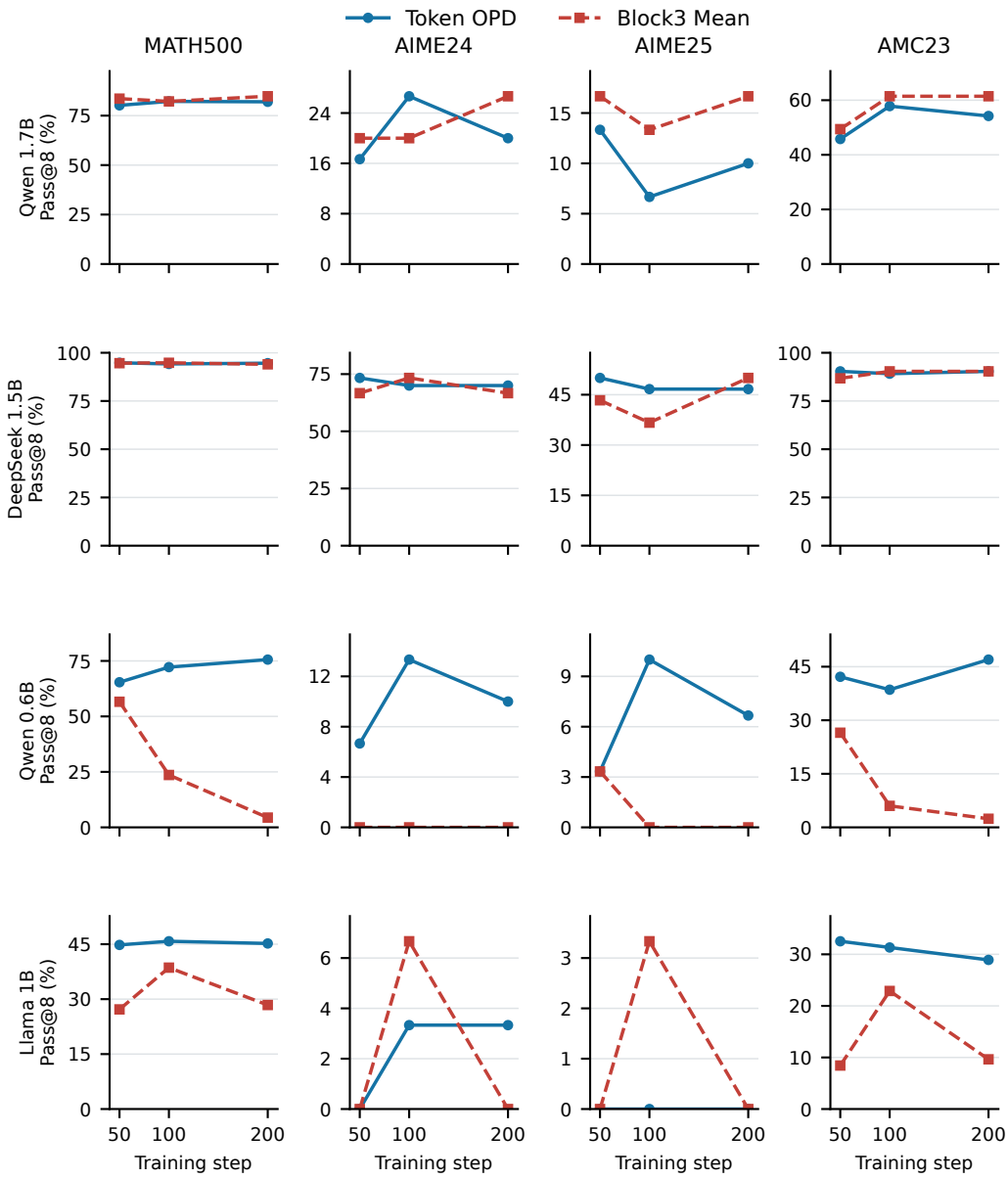

Figure 5: Historical Pass@8 for the same checkpoints and protocol strata as Figure 4. Each question contributes one if at least one of its eight evaluation responses is correct. This coverage statistic is distinct from averaging the eight correctness indicators. No interval over independent training seeds is implied.

E.1 H01: Early weighting pilots

Training seeds: 7, 13; recorded training steps: 50.

100-question pilot: mean score / Pass@k, k unrecorded

Method	Step	Seed	GSM8K	MATH500
adaptive_exopd	50	7	39.00 / 39.00	22.00 / 22.00
base	0	–	43.00 / 43.00	22.00 / 22.00
routed_adaptive	50	7	40.00 / 40.00	26.00 / 26.00
seed13_standard	50	13	43.00 / 43.00	22.00 / 22.00
seed13_topk_default	50	13	38.00 / 38.00	27.00 / 27.00
seed7_routed_lam110	50	7	44.00 / 44.00	25.00 / 25.00
seed7_topk_strict	50	7	41.00 / 41.00	20.00 / 20.00
standard	50	7	46.00 / 46.00	25.00 / 25.00
topk_router	50	7	44.00 / 44.00	25.00 / 25.00

Full n=1: original simple grader, accuracy

Method	Step	Seed	GSM8K	MATH500
base	0	–	41.70	16.00
routed_lam110	50	7	41.93	16.00
standard	50	7	41.77	16.80
topk_router	50	7	41.70	16.40

Full n=1: post-hoc primary grader, accuracy

Method	Step	Seed	GSM8K	MATH500
base	0	–	46.25	20.20
routed_lam110	50	7	47.23	20.20
standard	50	7	46.17	20.20
topk_router	50	7	46.17	19.40

E.2 H02: Outcome weighting and long-run pilots

Training seeds: 7; recorded training steps: 200.

Full n=1: original simple grader, accuracy

Method	Step	Seed	GSM8K	MATH500
outcome_reweight	200	7	33.36	12.60
outcome_topk_router	200	7	33.21	12.20
standard	100	7	35.25	13.80
standard	150	7	33.43	12.40
standard	200	7	32.15	11.80
topk_router	100	7	35.25	14.20
topk_router	150	7	34.87	12.00
topk_router	200	7	33.89	13.20

Full n=1: post-hoc primary grader, accuracy

Method	Step	Seed	GSM8K	MATH500
outcome_reweight	200	7	37.83	16.60
outcome_topk_router	200	7	38.51	16.40

Method	Step	Seed	GSM8K	MATH500
standard	100	7	39.95	16.20
standard	150	7	38.06	15.80
standard	200	7	36.16	14.80
topk_router	100	7	40.26	16.20
topk_router	150	7	39.27	14.40
topk_router	200	7	38.82	16.20

E.3 H03: Reference-anchor pilots

Training seeds: unrecorded; recorded training steps: 100.

Full n=1: original simple grader, accuracy

Method	Step	Seed	GSM8K	MATH500
anchored_standard	100	–	35.41	12.40
anchored_topk_router	100	–	37.15	13.40
standard	100	–	35.48	13.40
topk_router	100	–	34.57	14.00

Full n=1: post-hoc primary grader, accuracy

Method	Step	Seed	GSM8K	MATH500
anchored_standard	100	–	40.03	15.20
anchored_topk_router	100	–	41.93	17.80
standard	100	–	40.71	16.80
topk_router	100	–	39.42	17.60

E.4 H04: Naive block-sum pilot

Training seeds: 7; recorded training steps: 50.

Full n=1: documented rounded accuracy only

Method	Step	Seed	GSM8K	MATH500
block_k2	50	7	45.41	19.40
block_k3	50	7	46.85	20.20
token_opd	50	7	45.94	20.00

E.5 H05: Block sum, mean and mixed pilot

Training seeds: 7; recorded training steps: 50.

Full n=1: documented rounded accuracy only

Method	Step	Seed	GSM8K	MATH500
block3_mean	50	7	46.93	19.80
block3_mixed_lam05	50	7	47.08	18.80
block3_sum	50	7	46.55	19.20

E.6 H06: Qwen 1.7B block-size sweep

Training seeds: 21; recorded training steps: 200.

Legacy external grader: Avg@8 / Pass@8, evaluation seeds unspecified

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
block10_mean	200	21	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
block3_mean	200	21	69.27 / 87.80	9.58 / 30.00	8.75 / 26.67	39.91 / 61.45
block5_mean	200	21	67.40 / 87.60	9.17 / 20.00	7.92 / 23.33	34.19 / 61.45
token_opd	200	21	69.12 / 88.60	13.33 / 26.67	6.67 / 16.67	36.75 / 62.65

E.7 H07: Block10 two-host diagnostic repetitions

Training seeds: 21; recorded training steps: 200. Historical built-in AMC23 grading has a compatibility caveat; zeros are not independently validated capability measurements.

Built-in grader: Avg@8 / Pass@8 (separate view)

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
Block10: A100 host	50	21	37.75 / 78.60	4.17 / 20.00	2.08 / 13.33	0.00 / 0.00
Block10: A100 host	100	21	0.97 / 7.20	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
Block10: A100 host	200	21	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
Block10: A800 host	50	21	45.20 / 79.20	2.50 / 13.33	2.92 / 10.00	0.00 / 0.00
Block10: A800 host	100	21	2.70 / 16.60	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00
Block10: A800 host	200	21	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00	0.00 / 0.00

E.8 H08: Qwen 1.7B same-host paired study

Training seeds: 21; recorded training steps: 200. Historical built-in AMC23 grading has a compatibility caveat; zeros are not independently validated capability measurements.

Built-in grader: Avg@8 / Pass@8 (separate view)

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
block3_mean	50	21	50.25 / 81.00	5.42 / 20.00	3.75 / 16.67	0.00 / 0.00
block3_mean	100	21	40.80 / 80.40	5.00 / 20.00	3.33 / 13.33	0.00 / 0.00
block3_mean	200	21	52.83 / 82.20	8.75 / 26.67	5.42 / 16.67	0.00 / 0.00
token_opd	50	21	37.05 / 78.00	3.75 / 16.67	2.50 / 13.33	0.00 / 0.00
token_opd	100	21	41.80 / 80.20	6.25 / 26.67	2.08 / 6.67	0.00 / 0.00
token_opd	200	21	44.45 / 80.00	5.42 / 20.00	2.50 / 10.00	0.00 / 0.00

Audited historical external grader: Avg@8 / Pass@8

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
block3_mean	50	21	52.10 / 83.60	5.42 / 20.00	3.75 / 16.67	22.14 / 49.40
block3_mean	100	21	42.27 / 82.20	5.00 / 20.00	3.33 / 13.33	23.64 / 61.45
block3_mean	200	21	54.75 / 84.80	8.75 / 26.67	5.42 / 16.67	27.26 / 61.45
token_opd	50	21	38.07 / 80.20	3.75 / 16.67	2.50 / 13.33	13.70 / 45.78
token_opd	100	21	43.10 / 82.20	6.25 / 26.67	2.08 / 6.67	18.67 / 57.83
token_opd	200	21	45.85 / 82.00	5.42 / 20.00	2.50 / 10.00	21.84 / 54.22

E.9 H09: Qwen 1.7B checkpoint re-evaluation

Training seeds: 21; recorded training steps: –.

Historical external grader: Avg@8 / Pass@8

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
block3_mean	50	21	52.05 / 84.60	5.00 / 16.67	3.33 / 16.67	23.80 / 51.81
block3_mean	100	21	42.65 / 83.00	5.00 / 20.00	2.50 / 13.33	23.49 / 56.63
block3_mean	200	21	54.15 / 84.00	7.08 / 26.67	5.83 / 16.67	29.22 / 65.06
token_opd	50	21	38.30 / 81.80	2.50 / 16.67	3.75 / 16.67	14.61 / 51.81
token_opd	100	21	43.35 / 83.00	6.25 / 20.00	2.08 / 6.67	18.52 / 55.42
token_opd	200	21	45.92 / 83.00	6.67 / 16.67	3.75 / 10.00	23.34 / 60.24

E.10 H10: DeepSeek/JustRL paired study

Training seeds: 21; recorded training steps: 200.

Historical external grader: Avg@8 / Pass@8

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
block3_mean	50	21	84.15 / 94.60	32.50 / 66.67	27.50 / 43.33	70.93 / 86.75
block3_mean	100	21	84.45 / 94.80	39.17 / 73.33	25.42 / 36.67	72.74 / 90.36
block3_mean	200	21	84.75 / 94.00	43.75 / 66.67	30.83 / 50.00	73.80 / 90.36
token_opd	50	21	83.78 / 94.80	38.75 / 73.33	27.08 / 50.00	71.69 / 90.36
token_opd	100	21	84.62 / 94.20	39.17 / 70.00	34.58 / 46.67	73.19 / 89.16
token_opd	200	21	84.70 / 94.60	39.58 / 70.00	29.17 / 46.67	73.34 / 90.36

Historical built-in AMC23 grading has a compatibility caveat; zeros are not independently validated capability measurements.

Built-in grader: Avg@8 / Pass@8 (separate view)

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
block3_mean	50	21	81.80 / 91.60	32.50 / 66.67	27.50 / 43.33	0.00 / 0.00
block3_mean	100	21	82.23 / 91.80	39.17 / 73.33	25.42 / 36.67	0.00 / 0.00
block3_mean	200	21	82.25 / 90.80	43.75 / 66.67	30.83 / 50.00	0.00 / 0.00
token_opd	50	21	81.47 / 91.60	38.75 / 73.33	27.08 / 50.00	0.00 / 0.00
token_opd	100	21	82.12 / 91.00	39.17 / 70.00	34.58 / 46.67	0.00 / 0.00
token_opd	200	21	82.25 / 91.60	39.58 / 70.00	29.17 / 46.67	0.00 / 0.00

E.11 H11: Random3 and Sliding3 paired study

Training seeds: 21; recorded training steps: 200.

Historical external grader: Avg@8 / Pass@8

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
random3	50	21	45.23 / 81.60	5.00 / 20.00	2.92 / 10.00	18.22 / 48.19
random3	100	21	36.70 / 82.40	2.08 / 13.33	1.67 / 10.00	14.91 / 48.19
random3	200	21	21.18 / 73.60	3.33 / 16.67	0.00 / 0.00	12.20 / 46.99
sliding3	50	21	41.33 / 81.80	3.75 / 23.33	1.67 / 10.00	15.66 / 48.19
sliding3	100	21	52.08 / 85.20	4.58 / 20.00	2.92 / 10.00	24.85 / 57.83
sliding3	200	21	34.75 / 81.00	5.00 / 20.00	2.08 / 10.00	15.81 / 48.19

Historical built-in AMC23 grading has a compatibility caveat; zeros are not independently validated capability measurements.

Built-in grader: Avg@8 / Pass@8 (separate view)

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
random3	50	21	43.43 / 79.00	5.00 / 20.00	2.92 / 10.00	0.00 / 0.00
random3	100	21	35.27 / 79.60	2.08 / 13.33	1.67 / 10.00	0.00 / 0.00
random3	200	21	20.50 / 71.60	3.33 / 16.67	0.00 / 0.00	0.00 / 0.00
sliding3	50	21	40.02 / 79.20	3.75 / 23.33	1.67 / 10.00	0.00 / 0.00
sliding3	100	21	50.42 / 82.80	4.58 / 20.00	2.92 / 10.00	0.00 / 0.00
sliding3	200	21	33.70 / 78.00	5.00 / 20.00	2.08 / 10.00	0.00 / 0.00

E.12 H12: Qwen 0.6B old-prompt interrupted attempt

Training seeds: 21; recorded training steps: 200.

Incomplete planned evaluation: no full benchmark claim

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
token_opd	50	21	–	–	–	–
token_opd	100	21	–	–	–	–
token_opd	200	21	–	–	–	–

E.13 H13: Qwen 0.6B non-thinking paired study

Training seeds: 21; recorded training steps: 200.

Historical external grader: Avg@8 / Pass@8

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
block3_mean	50	21	16.20 / 56.60	0.00 / 0.00	0.42 / 3.33	5.87 / 26.51
block3_mean	100	21	3.48 / 23.60	0.00 / 0.00	0.00 / 0.00	1.05 / 6.02
block3_mean	200	21	0.57 / 4.40	0.00 / 0.00	0.00 / 0.00	0.45 / 2.41
student_base	0	–	10.45 / 45.20	0.42 / 3.33	0.00 / 0.00	3.46 / 24.10
teacher	0	–	83.43 / 93.40	23.33 / 43.33	18.75 / 26.67	57.68 / 77.11
token_opd	50	21	25.70 / 65.40	1.25 / 6.67	0.42 / 3.33	10.09 / 42.17
token_opd	100	21	39.60 / 72.20	2.50 / 13.33	1.25 / 10.00	15.36 / 38.55
token_opd	200	21	42.88 / 75.60	2.50 / 10.00	0.83 / 6.67	18.98 / 46.99

E.14 H14: Llama 3.2 original-prompt interrupted attempt

Training seeds: 21; recorded training steps: 146.

Historical external grader: Avg@8 / Pass@8

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
student_base	0	–	13.80 / 45.20	0.42 / 3.33	0.00 / 0.00	4.07 / 21.69
token_opd	50	21	–	–	–	–
token_opd	100	21	–	–	–	–
token_opd	200	21	–	–	–	–

E.15 H15: Llama 3.2 historical-prompt paired study

Training seeds: 21; recorded training steps: 200.

Historical external grader: Avg@8 / Pass@8

Method	Step	Seed	MATH500	AIME24	AIME25	AMC23
block3_mean	50	21	6.12 / 27.20	0.00 / 0.00	0.00 / 0.00	1.51 / 8.43
block3_mean	100	21	11.15 / 38.60	1.25 / 6.67	0.42 / 3.33	4.52 / 22.89
block3_mean	200	21	6.30 / 28.40	0.00 / 0.00	0.00 / 0.00	1.51 / 9.64
student_base	0	–	14.00 / 46.40	0.42 / 3.33	0.00 / 0.00	5.27 / 28.92
token_opd	50	21	14.15 / 44.80	0.00 / 0.00	0.00 / 0.00	6.78 / 32.53
token_opd	100	21	14.90 / 45.80	0.42 / 3.33	0.00 / 0.00	6.48 / 31.33
token_opd	200	21	15.55 / 45.20	0.42 / 3.33	0.00 / 0.00	6.02 / 28.92

E.16 H16: Qwen 4B ChatML interrupted attempt

Training seeds: 21; recorded training steps: 4. No completed benchmark result is recorded for this attempt.

E.17 H17: Qwen 4B completion study

Training seeds: 21; recorded training steps: 200. See Appendix G for the final current-pair results.

F Bounded Prompt Diagnostics

These are inference probes on selected training questions, not additional held-out benchmark scores. The Qwen3-4B and Llama prompt studies use 16 fixed questions with two independently seeded responses per condition (seeds 21 and 22). Qwen 0.6/1.7 primary and first-16 ablation cells retain their separate sample counts; overlapping questions are not summed as independent observations. The tables contain integer counts: length stops, periodic tails, missing literal boxed markers and correctly graded responses. A dash indicates an unrecorded statistic. A periodic tail requires at least 256 retained tail tokens and an offset of at most 128 with at least 99% token agreement within the last 2,048 tokens; it is a detector, not a general semantic repetition measure.

The historical training user message is:

> Math problem: [question]
> Please carefully reason through the math problem step by step and derive the correct answer. You must conduct reasoning inside <think> and </think> and give the final answer within \boxed{}.

The native model template wraps that content with an assistant generation prefix. “No think” deletes only the phrase conduct reasoning inside <think> and </think> and; it is not the same as bypassing a chat template. Historical evaluation instead uses the question and Please reason step by step, and put your final answer within \boxed{}., with thinking explicitly disabled. Thus asking for reasoning tags and disabling a native thinking mode are distinct controls.

For Qwen3-4B, “plain” is the bare question followed by two newlines and Solution:; “candidate” is the final matched completion protocol in Section 4.1. For Llama, “completion” retains the non-think historical instruction and a single native BOS, removes chat role boundaries, and appends Solution:; it is not exactly the Qwen plain condition. The “eval” cell uses Llama’s native non-thinking evaluation template. Removing chat raises Llama student length stops from 1/32 to 4/32 in this probe, while Qwen3-4B’s historical-to-candidate change reduces them from 18/32 to 1/32. These observations motivate model-specific validation, not a universal rule to remove chat from instruction-tuned models.

Llama prompt

Model	Condition	n	Length	Periodic	No box	Correct
student	historical	32	1	1	10	0
student	no_think	32	1	0	16	1
student	completion	32	4	1	3	1
student	eval	32	1	1	11	0
teacher	historical	32	2	1	6	3
teacher	no_think	32	4	2	9	2
teacher	completion	32	4	2	2	1
teacher	eval	32	0	0	6	2

Qwen3-4B prompt

Model	Condition	n	Length	Periodic	No box	Correct
student	historical	32	18	17	21	1
student	plain	32	0	0	9	2
student	candidate	32	1	1	0	4
teacher	historical	32	2	0	2	6
teacher	plain	32	0	0	0	5
teacher	candidate	32	3	2	2	8

Qwen 0.6/1.7 primary32

Model	Condition	n	Length	Periodic	No box	Correct
q06_base	–	32	12	10	–	0
q06_step50	–	32	10	4	–	1
q17_base	–	32	16	10	–	1
q17_step50	–	32	4	0	–	4

Qwen 0.6/1.7 prompt_ablation_first16

Model	Condition	n	Length	Periodic	No box	Correct
q06_base_no_think	–	16	5	4	–	1
q06_base_plain	–	16	1	1	–	1
q06_base_training	–	16	6	6	–	0
q17_base_no_think	–	16	1	1	–	1
q17_base_plain	–	16	0	0	–	0
q17_base_training	–	16	6	4	–	1

G Complete 4B Base Checkpoint Results

The current pair uses the matched completion protocol in Appendix C. Scores are percentages, not fractions. Differences subtract Token from Block3 at the same checkpoint and on the same benchmark. Missing-box-marker rates and engine length-stop rates use all eight responses per question; Avg@8 and Pass@8 have the definitions in Equation 6. Checkpoint comparisons describe a single training trajectory per method and should not be interpreted as independent repetitions.

Paired-question bootstrap intervals, where reported, resample the same question identities for both arms, separately within each benchmark. They condition on the recorded eight responses per question and on the trained weights. They do not resample training seeds, estimate teacher uncertainty or provide simultaneous coverage over multiple checkpoints and benchmarks. The percentile procedure uses 10,000 resamples with seed 20260921. A confidence interval crossing zero does not establish equivalence between the methods.

Table 34: Qwen3-4B completion, seed21, COMPLETE. All scores in percent; NA = unavailable.

Benchmark	Method	Step	Avg@8	Pass@8	No box	Trunc.
MATH500	Block3 Mean	200	75.40	90.60	0.90	0.90
MATH500	Block3 Mean	150	78.03	91.20	0.83	0.83
MATH500	Block3 Mean	100	75.52	90.80	1.25	1.18
MATH500	Block3 Mean	50	78.53	91.80	0.85	0.75
MATH500	Token OPD	200	75.92	91.00	0.85	0.80
MATH500	Token OPD	150	76.17	91.40	0.95	0.97
MATH500	Token OPD	100	76.92	90.40	1.07	0.95
MATH500	Token OPD	50	78.65	92.20	0.85	0.80
AIME24	Block3 Mean	200	13.33	33.33	12.08	11.25
AIME24	Block3 Mean	150	15.42	33.33	9.58	8.75
AIME24	Block3 Mean	100	15.00	30.00	13.75	12.08
AIME24	Block3 Mean	50	14.17	30.00	7.08	6.25
AIME24	Token OPD	200	16.25	30.00	7.92	6.67
AIME24	Token OPD	150	12.50	26.67	10.42	8.33
AIME24	Token OPD	100	12.92	26.67	12.08	9.17
AIME24	Token OPD	50	17.92	36.67	10.42	8.75
AIME25	Block3 Mean	200	13.75	30.00	4.17	3.33
AIME25	Block3 Mean	150	11.67	26.67	5.42	5.42
AIME25	Block3 Mean	100	12.08	33.33	7.50	6.25
AIME25	Block3 Mean	50	11.67	23.33	7.50	5.42
AIME25	Token OPD	200	13.33	30.00	3.75	3.33
AIME25	Token OPD	150	15.83	40.00	5.42	6.25
AIME25	Token OPD	100	10.83	23.33	5.83	5.42
AIME25	Token OPD	50	12.92	23.33	7.08	4.17
AMC23	Block3 Mean	200	46.23	74.70	4.52	4.07
AMC23	Block3 Mean	150	49.70	77.11	3.46	3.01
AMC23	Block3 Mean	100	47.29	74.70	3.92	3.01
AMC23	Block3 Mean	50	50.30	74.70	4.37	3.31
AMC23	Token OPD	200	49.25	77.11	3.61	3.01
AMC23	Token OPD	150	49.40	77.11	2.56	2.86
AMC23	Token OPD	100	49.25	78.31	2.56	2.56
AMC23	Token OPD	50	48.49	78.31	3.16	3.01

No box: no literal \boxed; truncation: engine length stop.

Table 35: Post-hoc overlap sensitivity: MATH500 subset (496 questions; COMPLETE).

	Avg@8			Pass@8		
Step	Block3	Token	Δ	Block3	Token	Δ
200	75.73	76.31	-0.58 [-1.66, 0.50]	90.73	91.33	-0.60 [-2.62, 1.21]
150	78.40	76.56	+1.84	91.53	91.53	+0.00
100	75.83	77.27	-1.44	90.93	90.52	+0.40
50	78.93	79.08	-0.15	91.94	92.34	-0.40

Excludes four audit whitespace+casefold matches, including two exact matches. This is not a new benchmark or a primary result; official full-500 scores are unchanged. Scores are percentages; Δ is Block3 minus Token in percentage points. NA: paired question scores unavailable. Brackets at Step200 only: pointwise 95% paired-question bootstrap on 496 questions; conditional on one training pair, not a multi-training-seed CI.

Post-hoc overlap sensitivity. Removing the four training-pool string matches described in Appendix C leaves 496 MATH500 questions. At Step200, the Avg@8 difference becomes -0.58 points ($[-1.66, 0.50]$), versus -0.52 on the full benchmark. The Step50/100/150 difference signs also remain unchanged. This sensitivity result does not remove semantic or upstream-model contamination, establish its causal effect, or create a new held-out benchmark. Full-500 scores remain primary.

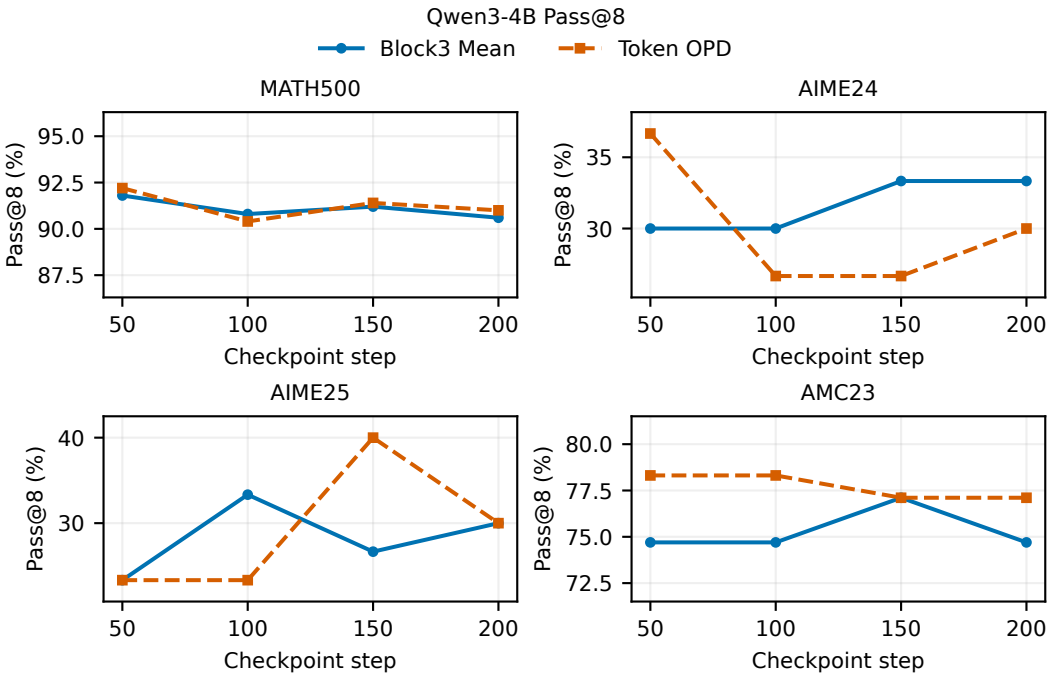

Figure 6: Current paired Pass@8 at all four checkpoints. Task-specific vertical ranges are shared by both methods. Each point uses all questions and eight responses per question, not a selected subset of outputs.

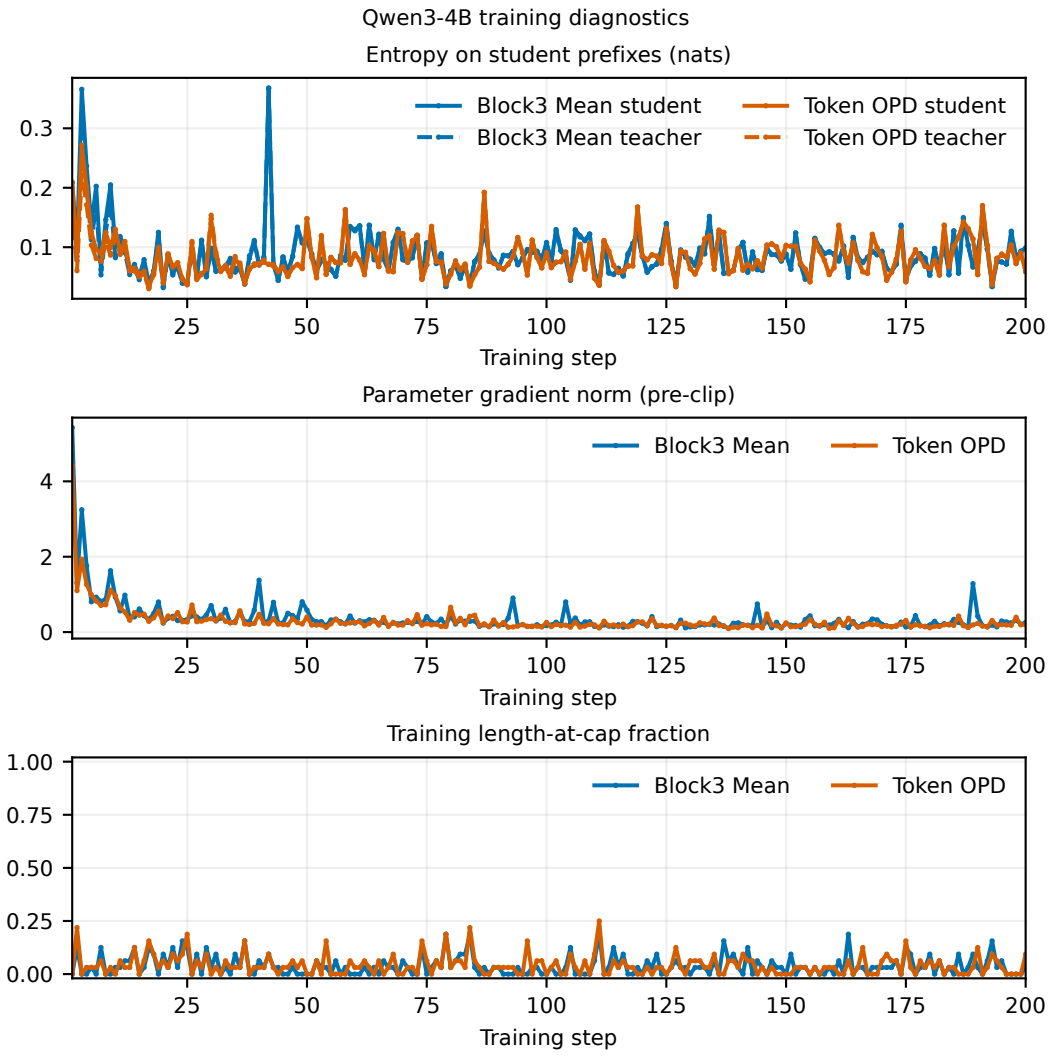

Figure 7: Every-step 4B training diagnostics. Entropies are full-vocabulary, pre-update quantities at student prefixes; gradient norms are pre-clipping. The bottom panel measures response length at the training cap, not evaluation engine stopping reasons. These curves do not establish the cause of performance differences.

H Official Qwen3-8B to 1.7B Instruct Comparison

A separate official Instruct-model stratum. The teacher is Qwen/Qwen3-8B, ModelScope revision 26028140be3ee69b82b1d1450179ab71bb1121b9; the student is Qwen/Qwen3-1.7B, revision 4855588ea1a12789f2e965e5f52a9e4a24c94b2a. Both are official post-trained models whose repository names omit the word Instruct. Neither is a Base checkpoint. This pair is distinct from the historical 1.7B student/public-GRPO-teacher stratum in Appendix E. Token and Block3 independently start from the same original student, using the same retained DAPO pool, seed21, 200 updates, four prompts and eight responses per prompt, learning rate 2×10^{-6}, and the four saved checkpoints. The optimizer, clipping and micro-batch settings follow the shared configuration in Appendix C. The historical Block3 objective is unchanged; this is not an advantage-only ablation. Complete input and request-seed audits match all 6,400 training trajectories per arm. Reusing the same data does not constitute an independent dataset replication or remove its overlap limitations.

Native, non-thinking prompt. Unlike the 4B Base completion stratum, this pair retains the original Qwen chat template. Its single user message consists of the stripped question, two newlines, and Please solve the problem step by step and put the final answer in \boxed{}.. There is no Solution: suffix. Both training and evaluation call the template with enable_thinking=False and add_generation_prompt=True. The resulting empty, already-closed <think>...</think> control prefix is part of the input, not generated reasoning. EOS is 151645 and the stop-token set is $\{151645, 151643\}$; recorded effective response lengths exclude padded suffixes. Actual collector, teacher and evaluation input token IDs are checked. Ordinary textual step-by-step solutions remain allowed.

Evaluation and recovery provenance. The original student and both arms at Steps50/100/150/200 each complete all four benchmarks with the same pinned historical grader, eight seeds21–28, temperature1.0, top-p = 0.9, and a 16,384-token output limit: nine model views and 46,296 evaluated responses. The archive label student_base means the original Instruct student, not a Base-model variant. An initial Token launch failed during Ray startup before any formal update. The successful retry preserved the frozen training implementation and independently initialized the original student; already accepted Block3 and original-student evaluations were reused, not regenerated or counted as additional runs. All raw training and evaluation trajectories are retained.

All checkpoints, not a selected best. Tables 36 and 37 retain every saved checkpoint. Step50 favors Token in three of four Avg@8 tasks, and Step150 favors Token on AIME24. The highest MATH500 Avg@8 in this stratum is Token Step100, not Block3 Step200. At the final endpoint, both methods exceed the original student on all four Avg@8 tasks, yet both have lower MATH500 Pass@8 (89.60% and 89.80% versus 91.20%). This does not support uniform improvement in eight-attempt coverage. The endpoint table in the main text uses 10,000 paired-question bootstrap resamples with bootstrap seed20260923. Intervals condition on one trained seed and the recorded responses, and are neither training-seed intervals nor corrected for multiple checkpoint/benchmark comparisons. All Step200 Avg@8 and Pass@8 intervals include zero; this is not proof of equivalence.

Table 36: Official 8B to 1.7B Instruct: complete Avg@8 (%), including the original student. Each benchmark is scored separately.

Model / Step	MATH500	AIME24	AIME25	AMC23
Original student	71.45	10.00	10.42	40.51
Token 50	72.47	13.33	8.75	45.48
Block3 50	71.50	12.50	9.58	43.67
Token 100	72.92	15.00	12.08	43.98
Block3 100	72.55	16.25	13.33	44.43
Token 150	72.47	15.42	10.00	43.83
Block3 150	72.62	13.33	12.08	46.84
Token 200	72.12	14.17	12.08	43.07
Block3 200	72.22	16.67	12.92	43.37

Table 37: The same model views evaluated with Pass@8 (%). Eight attempts per question are not eight independent training runs.

Model / Step	MATH500	AIME24	AIME25	AMC23
Original student	91.20	26.67	23.33	69.88
Token 50	90.60	33.33	26.67	73.49
Block3 50	89.00	26.67	16.67	73.49
Token 100	89.00	30.00	30.00	69.88
Block3 100	91.00	36.67	33.33	68.67
Token 150	90.60	30.00	26.67	71.08
Block3 150	91.60	30.00	36.67	74.70
Token 200	89.60	43.33	30.00	75.90
Block3 200	89.80	36.67	30.00	74.70

Behavior is not uniformly improved. During training, Token has 94/6,400 engine length stops (1.47%), versus 77/6,400 for Block3 (1.20%); neither generates thinking tags. The maximum recorded pre-clipping gradient norms are 5.13 and 6.90, respectively. These finite diagnostics do not prove better mathematical reasoning or reduced gradient variance. At evaluation, all nine views have zero generated thinking-tag counts, but Table 38 shows benchmark-dependent differences in missing boxed markers and length stops. In AIME24, Block3 has more of both than Token; in AIME25 and AMC23, it has fewer. A missing literal boxed marker is the historical format metric, not a full syntax check, answer-extraction audit or correctness label.

Table 38: Step200 behavior rates (%) for the official Instruct pair. Missing box and engine length stops are separate metrics with all responses as denominators.

Benchmark	Token missing box	Block3 missing box	Token length	Block3 length
MATH500	0.125	0.150	0.175	0.150
AIME24	1.667	2.083	1.667	2.500
AIME25	0.417	0.000	0.417	0.000
AMC23	0.602	0.000	0.602	0.000

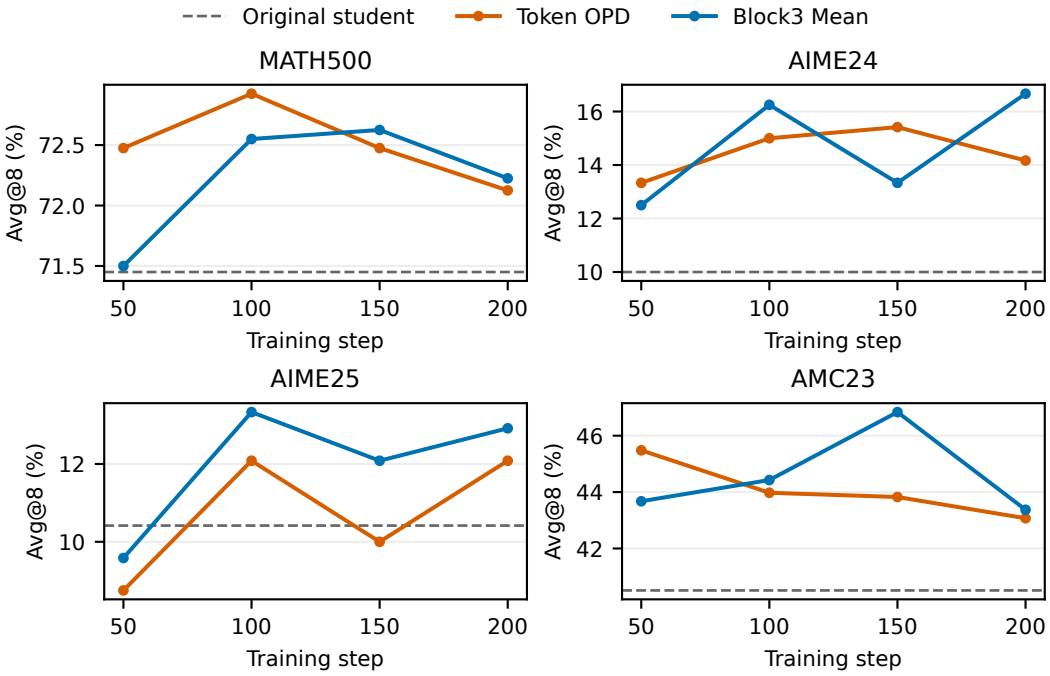

Figure 8: Official Instruct pair, Avg@8. Dashed lines denote the original student; curves connect only measured checkpoints. Each task has its own vertical range.

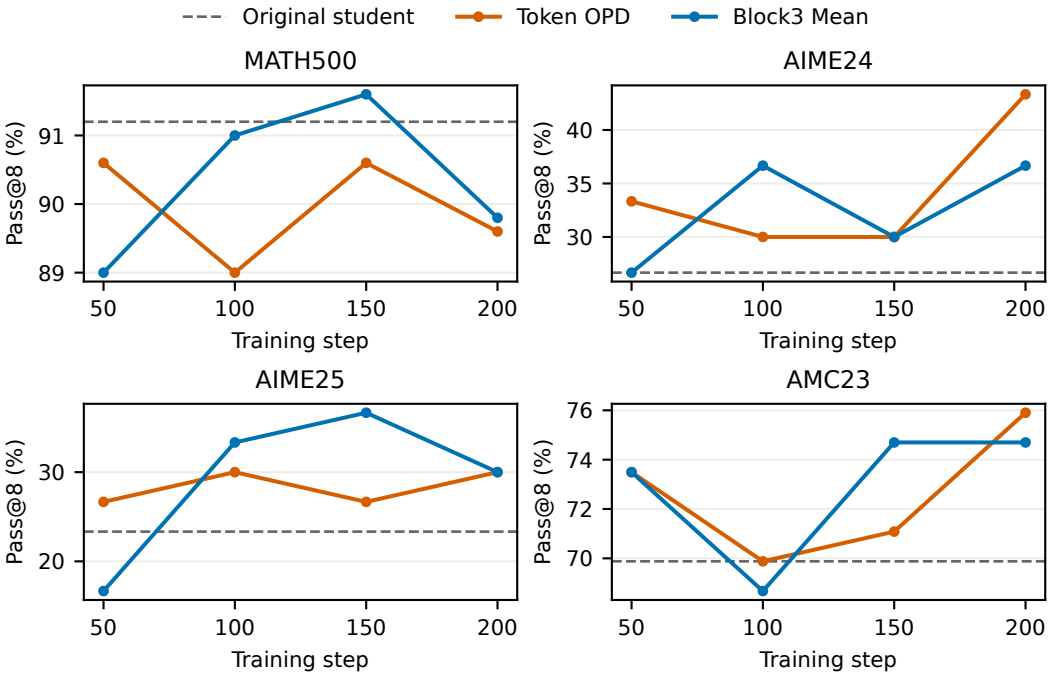

Figure 9: Official Instruct pair, Pass@8. Higher Avg@8 does not imply higher eight-attempt coverage. No across-task average is reported.

I Every-Step Position Diagnostics

These figures use all 200 retained diagnostic snapshots per arm, not interpolated checkpoint evaluations. Each panel pair uses one common color scale. Bins span 128 response-token positions and apply the coverage rule in Appendix D; gray is masked, not a zero measurement. The prefixes themselves change with the policy, so this is an occupancy-dependent training diagnostic rather than a fixed-input causal experiment. Scalar and positional reductions need not be identical because they weight observations differently.

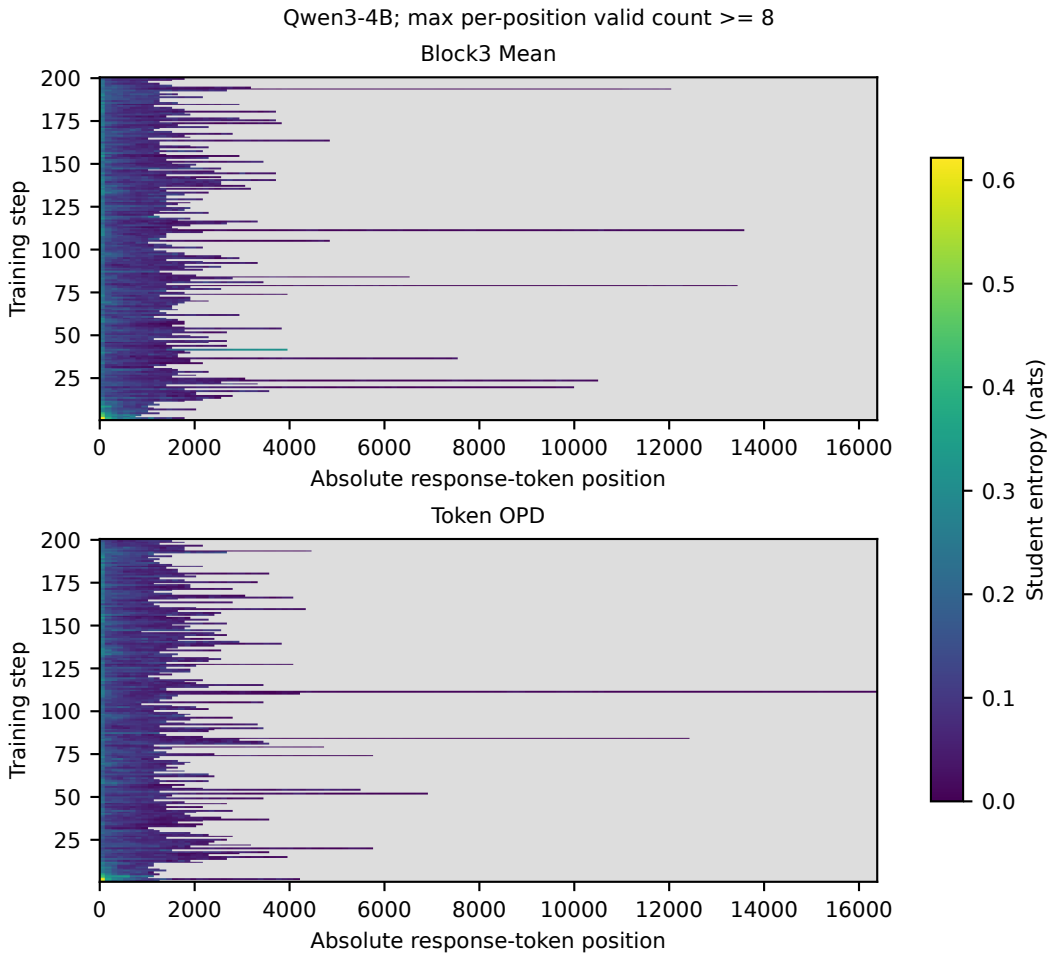

Figure 10: Student full-vocabulary entropy. Both arms remain far below the high-entropy historical Block10 endpoints. This observation does not establish that lower entropy is intrinsically better.

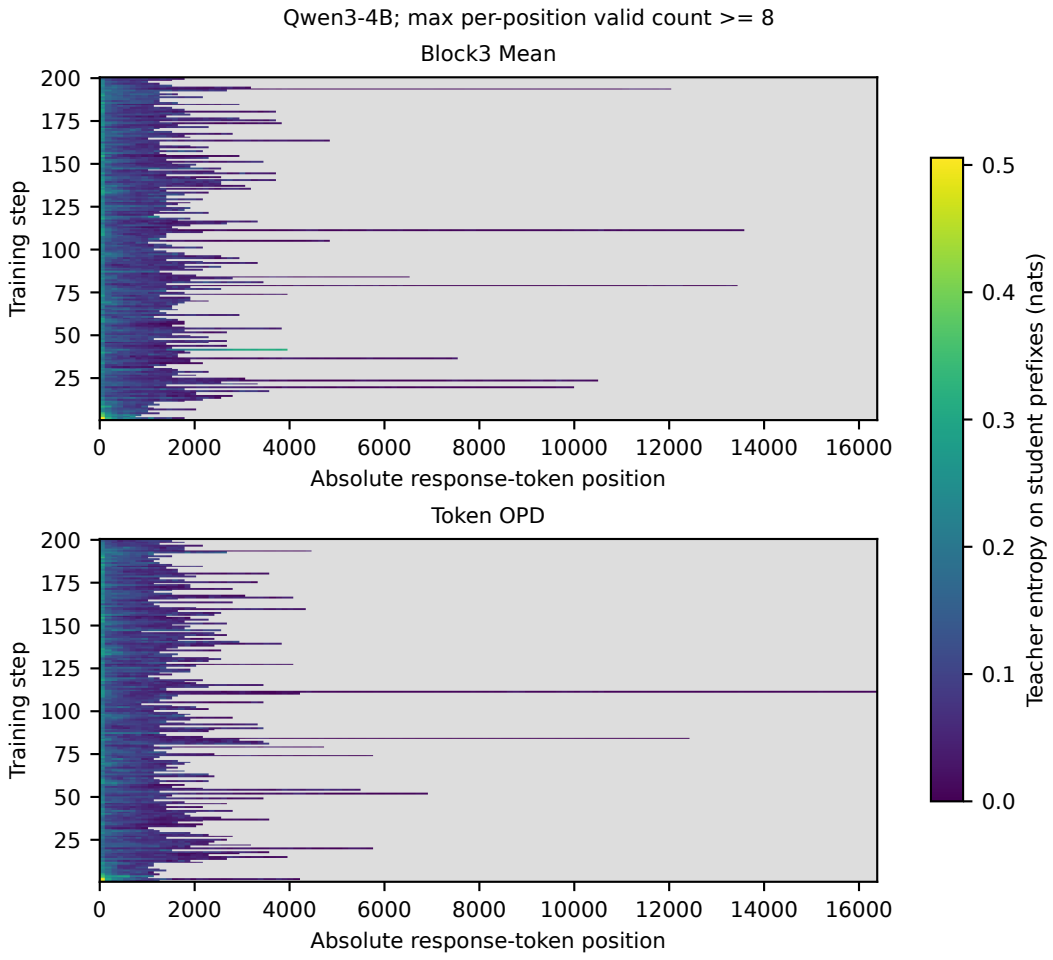

Figure 11: Teacher full-vocabulary entropy on student-generated prefixes. The frozen teacher is evaluated on each arm's own states; variation does not indicate that teacher weights changed.

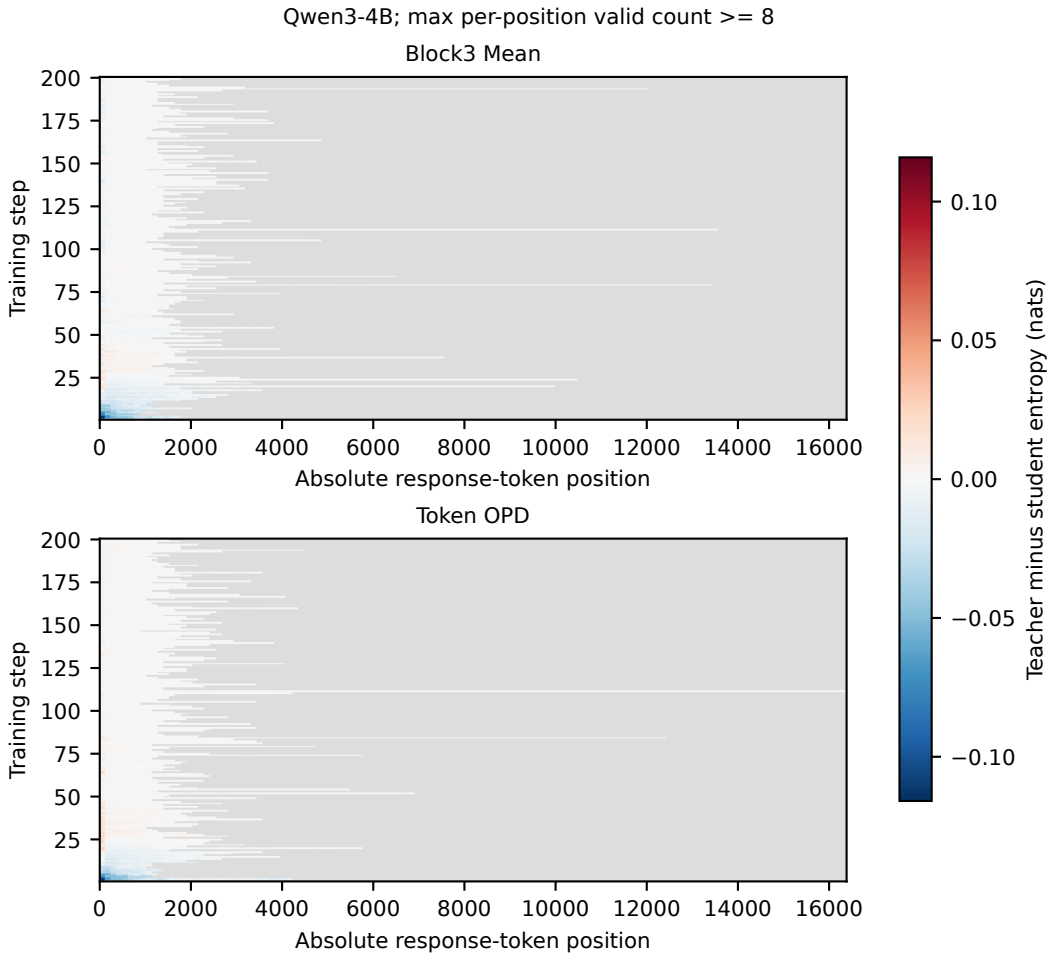

Figure 12: Signed entropy gap, teacher minus student. Positive and negative values describe different relative uncertainty, not correctness labels.

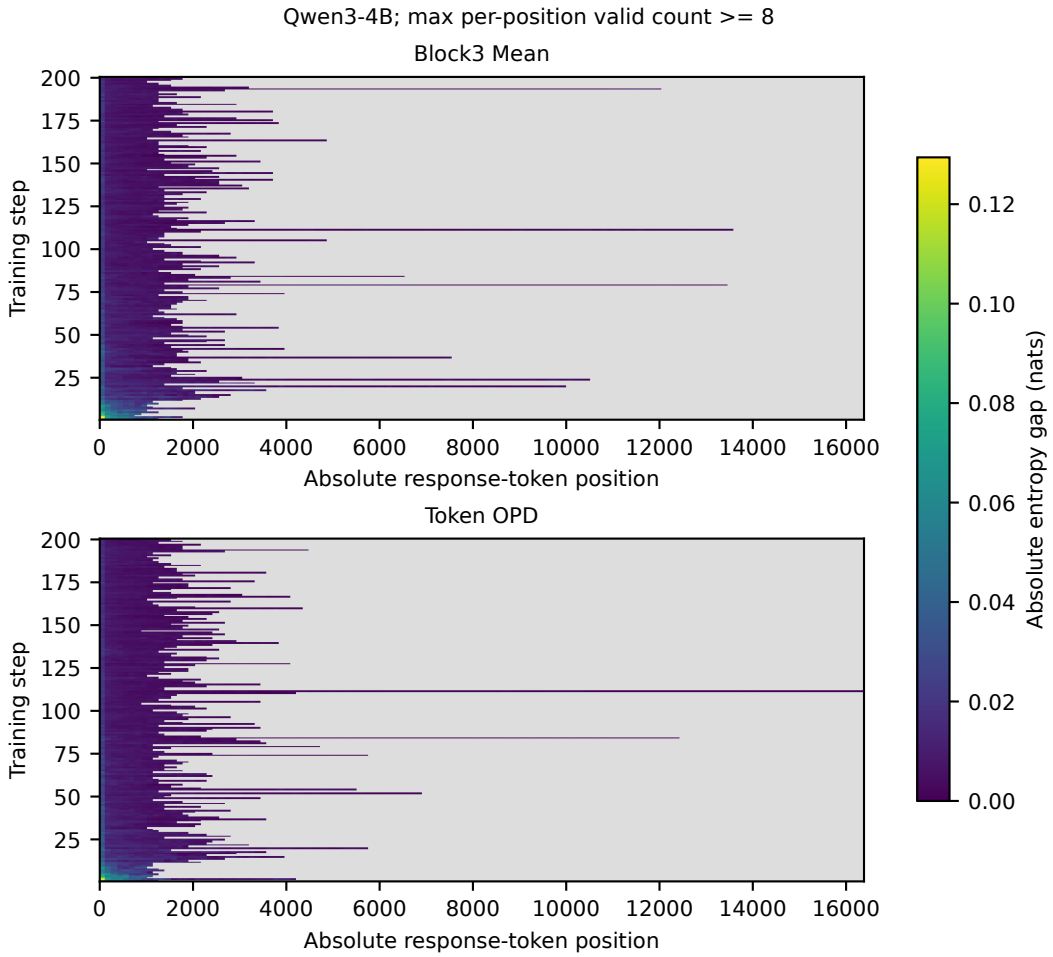

Figure 13: Absolute per-position entropy gap. Taking absolute values before averaging avoids cancellation; it is not the absolute value of the signed-bin mean.

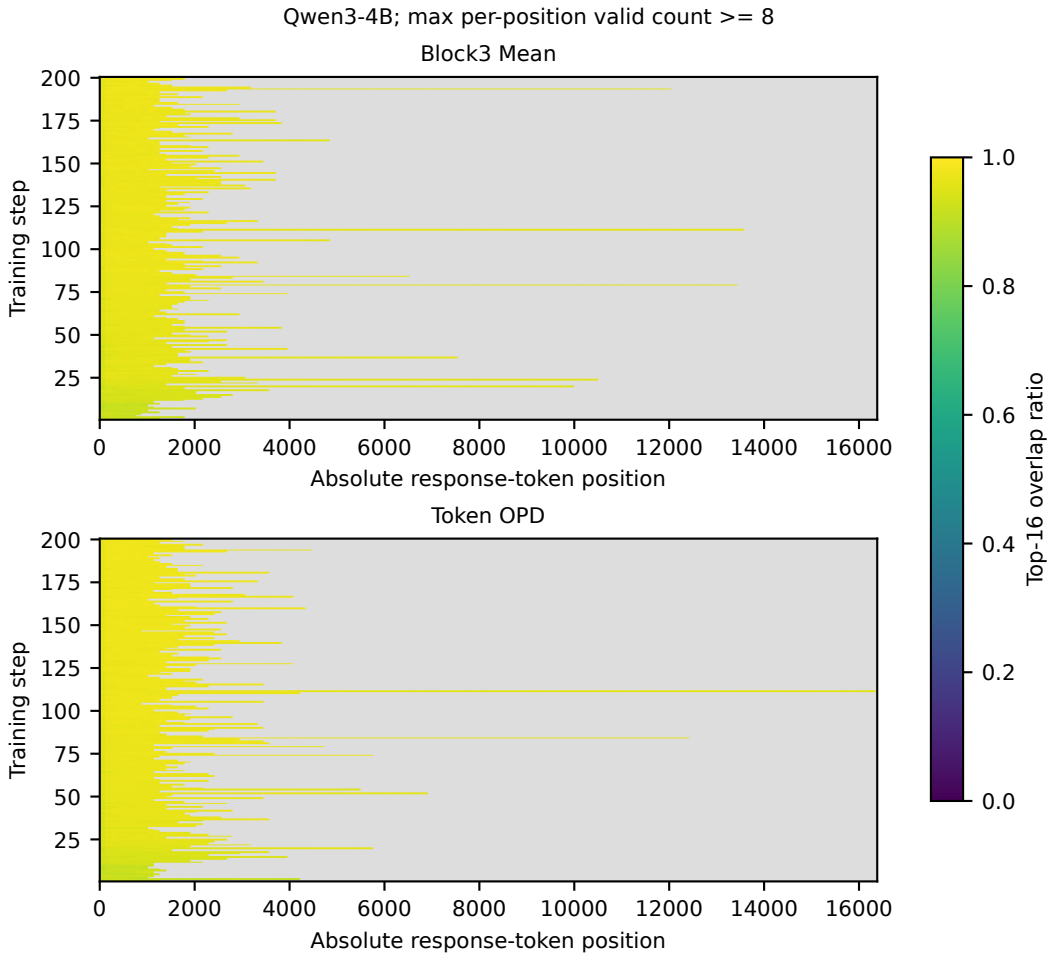

Figure 14: Top-16 set overlap ratio. Rank-set agreement does not determine probability mass or task accuracy.

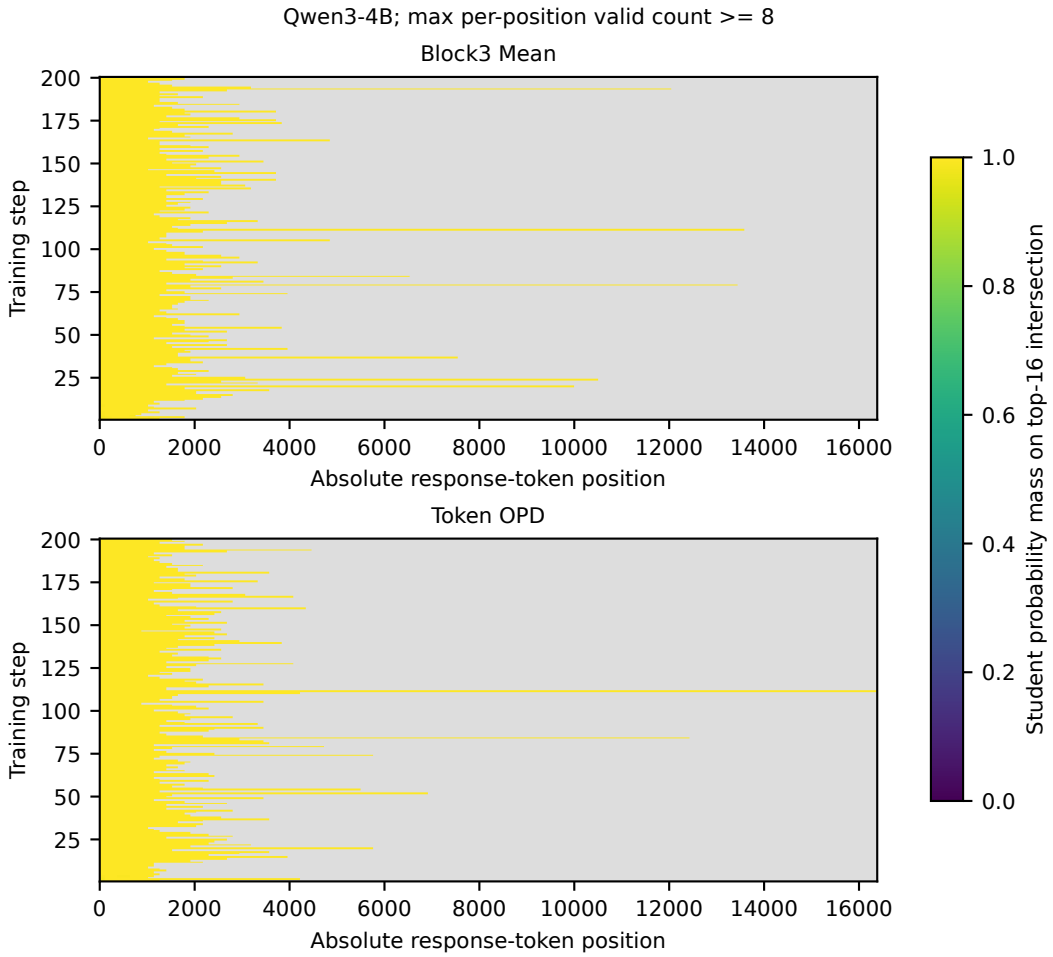

Figure 15: Student probability mass on the top-16 intersection. Mass is measured using the full student vocabulary distribution, not a top-16 renormalization.

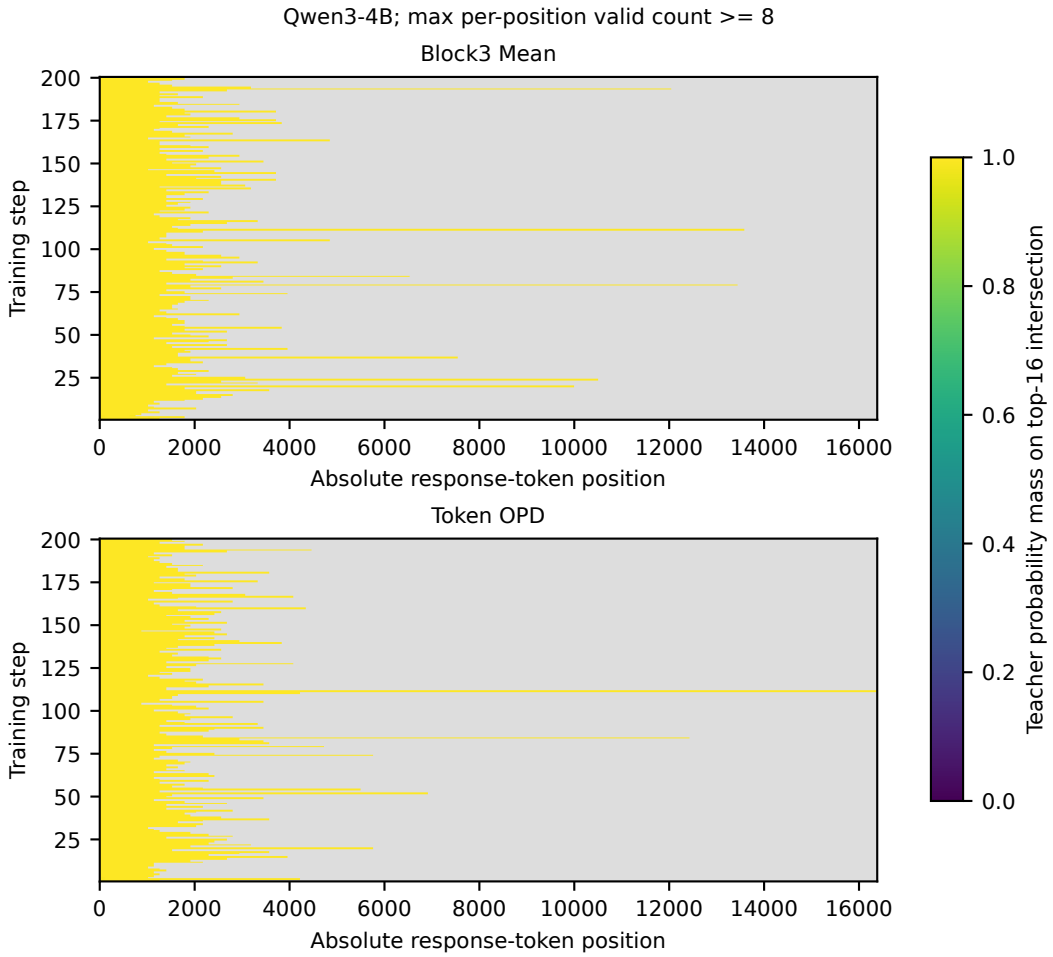

Figure 16: Teacher probability mass on the top-16 intersection. The intersection is the same set as in the student-mass map; its two probability measures can differ.

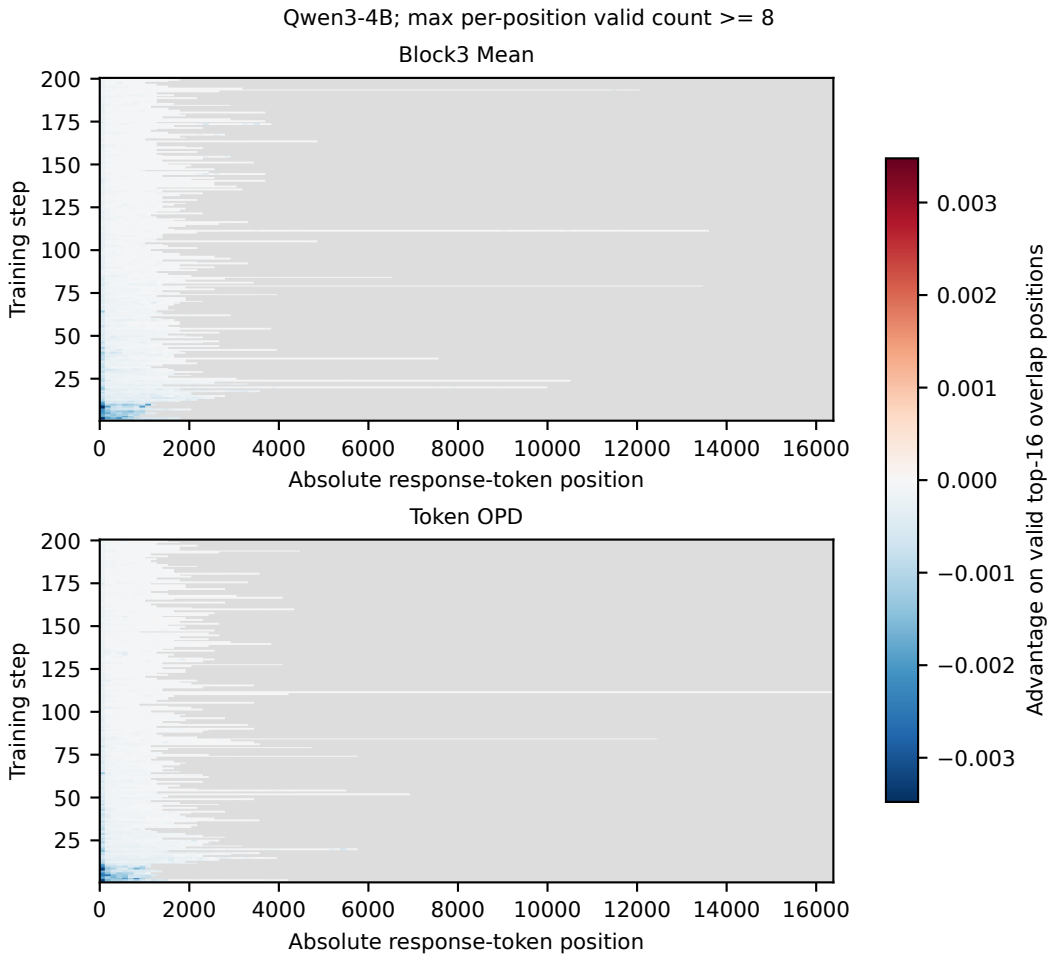

Figure 17: Implemented overlap-token advantage. The within-intersection renormalization yields a nonpositive KL-derived statistic as defined in Appendix D; empty intersections are excluded.

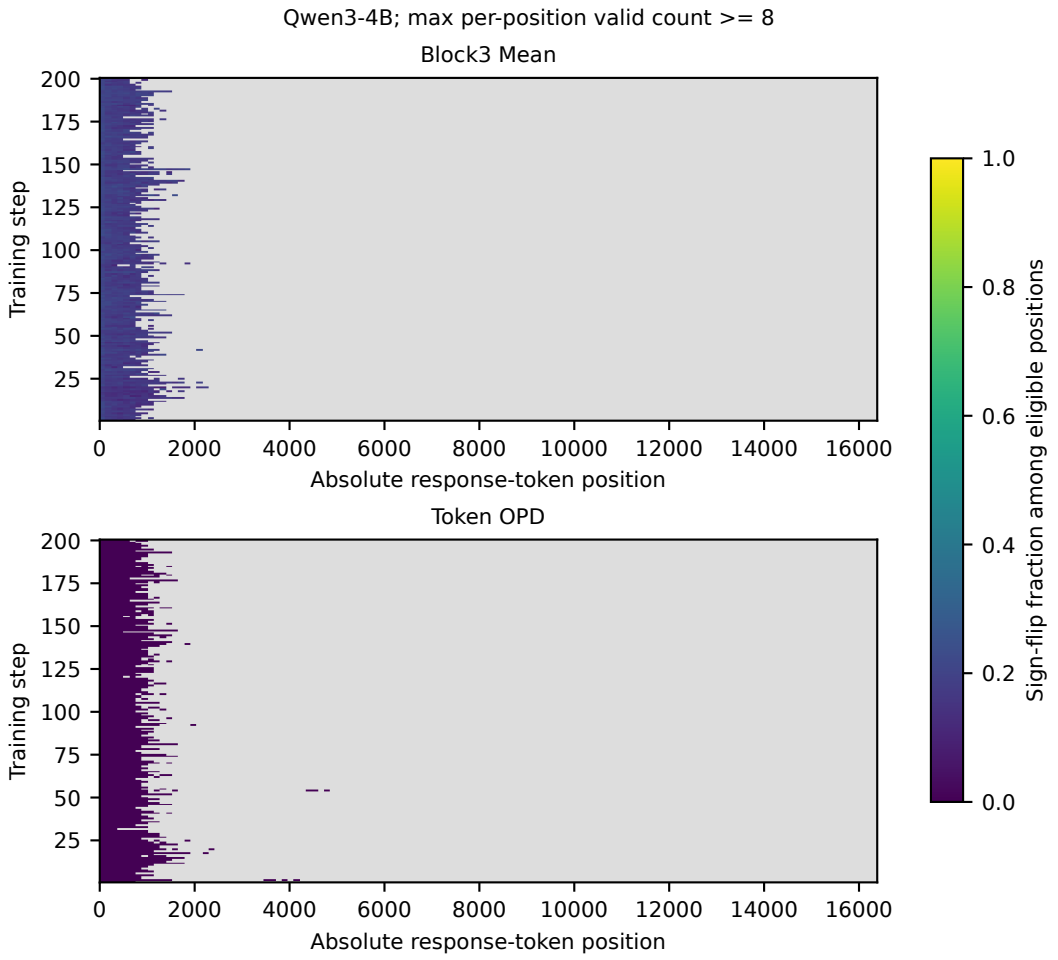

Figure 18: Local versus shared-advantage sign disagreement. Only positions passing both magnitude thresholds enter the denominator. The size-one Token arm is zero by construction.

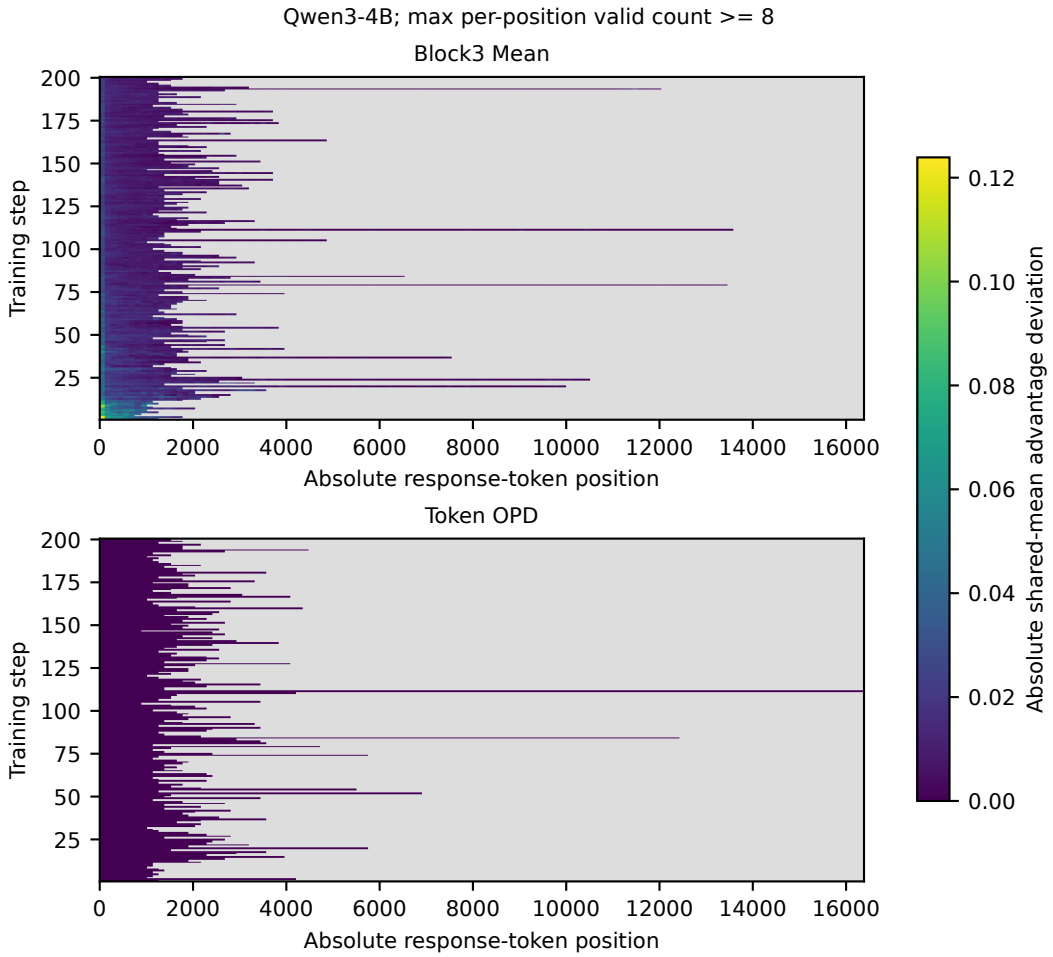

Figure 19: Absolute advantage redistribution magnitude. This is the mean absolute change in local feedback, not a probability or proof of harmful task-level credit.

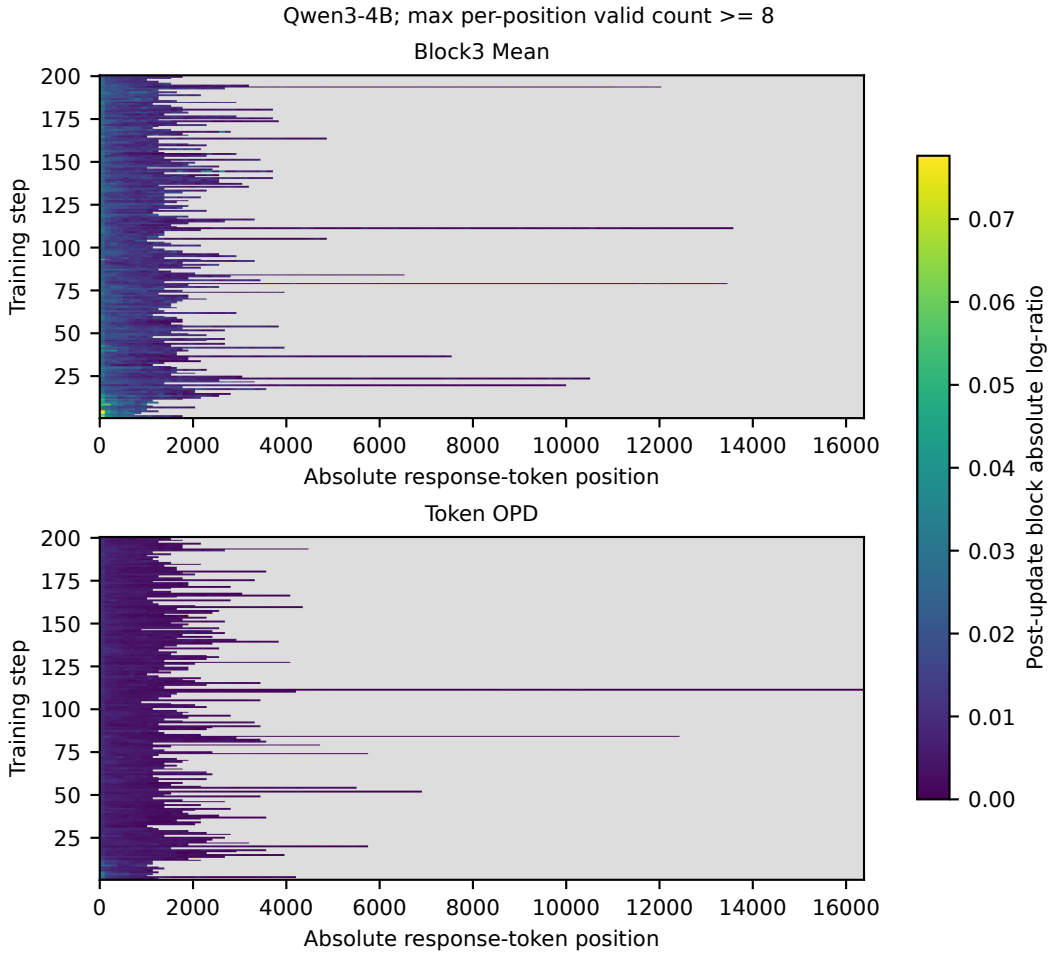

Figure 20: Absolute post-update block log-ratio. Block3 uses a product over three tokens while Token uses one; the two numerical scales are not the same likelihood-ratio unit. These measurements follow the update, unlike the entropy snapshots.

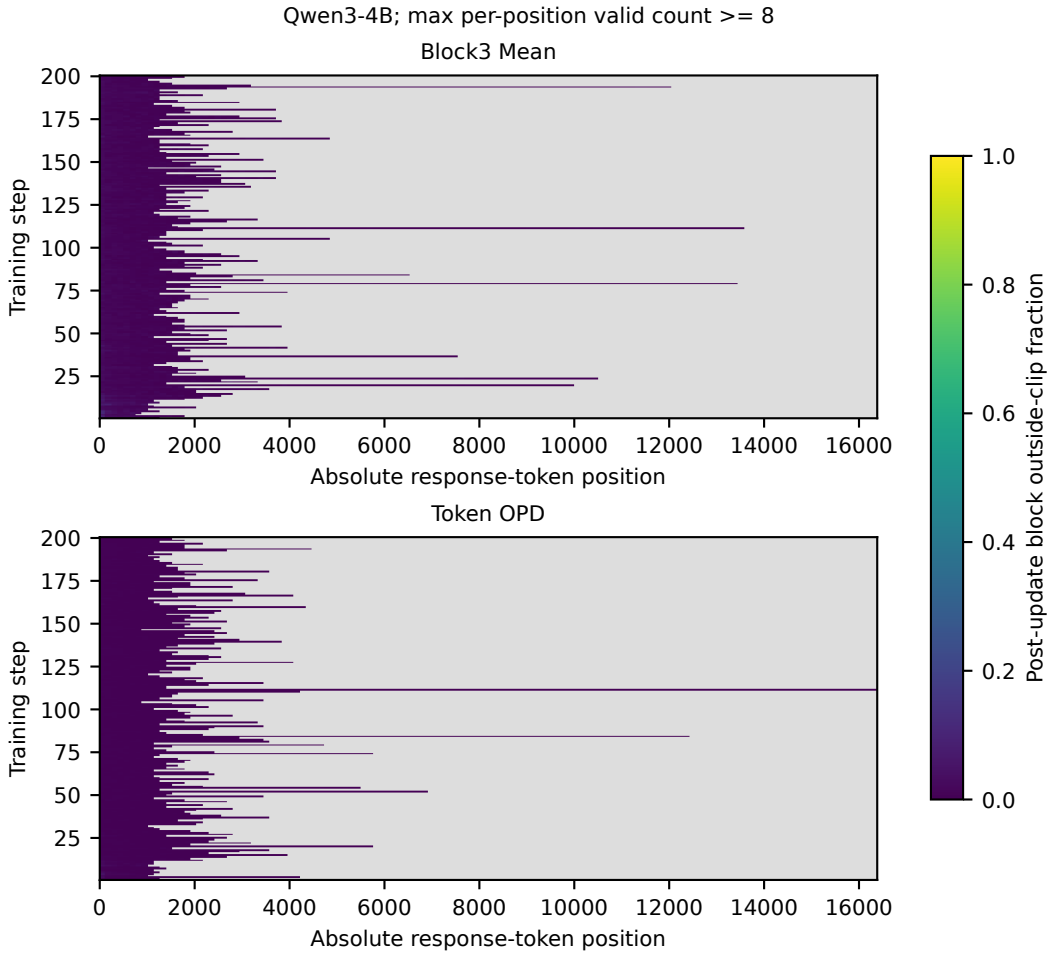

Figure 21: Post-update block ratios outside the configured PPO interval. This is not the fraction of optimization terms actually clipped: clipping also depends on advantage sign and the dual-clip rule.

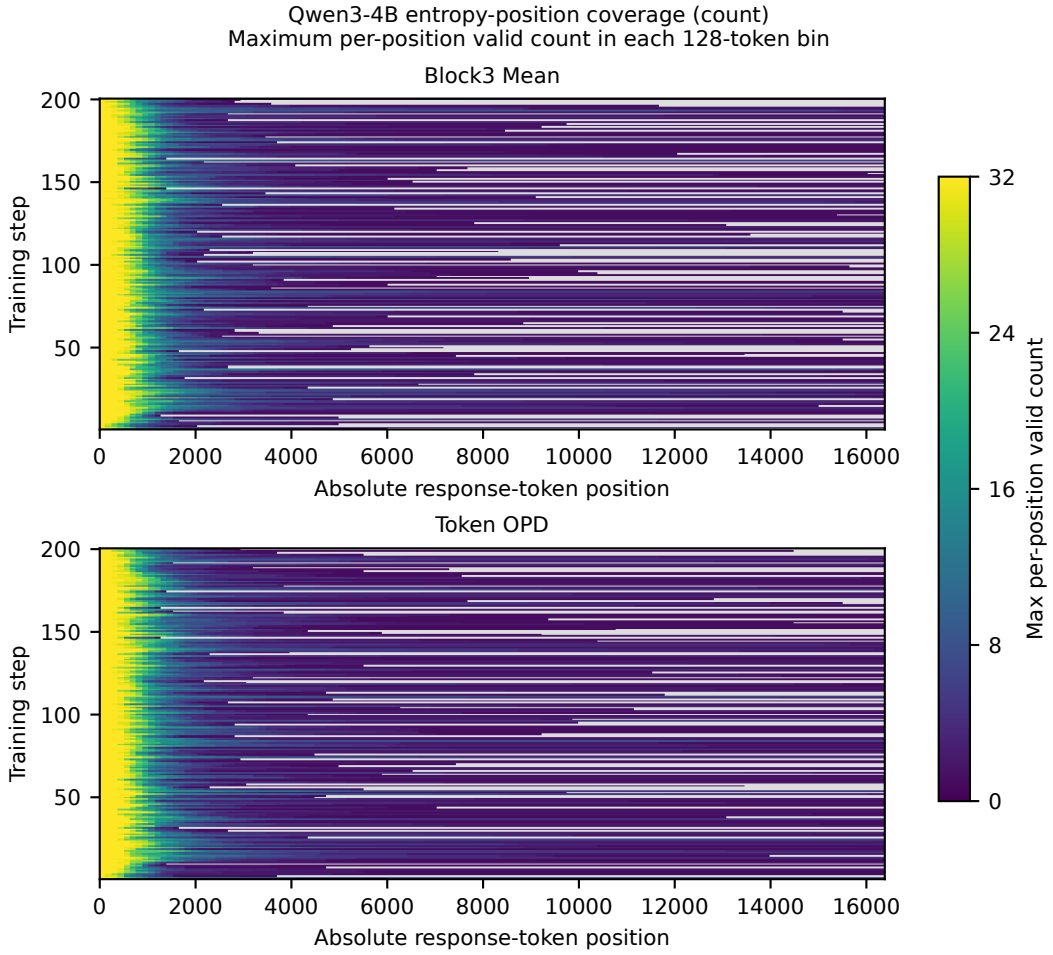

Figure 22: Coverage of the entropy maps: maximum valid response count at any original position within each 128-token bin, out of 32 responses per update. It is not the sum of observations over the bin. Low coverage explains many masked tail cells and must be distinguished from low entropy.