DECOUPLING OUTCOME-OBSERVATION AND ACTOR-UPDATE HORIZONS IN REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long rollout horizons are increasingly used in reinforcement learning with verifiable rewards (RLVR), yet their effect on policy optimization remains unclear. Existing methods mainly modify how boundary-truncated responses are filtered or penalized, while retaining the assumption that outcome observation and policy optimization should share the same horizon. We find that longer rollouts substantially delay the final answer without consistently improving task performance, while earlier response prefixes can already support correct answers. At the same time, short horizons may miss useful outcome feedback beyond the truncation boundary. Motivated by this mismatch, we propose **Natural-Continuation Boundary Return (NCBR)**, which naturally continues truncated responses for verification while updating only the original responses. Under a frozen continuation policy, the extended outcome estimates the continuation value of the original response, providing a basis for short-response updates. Across three mathematical reasoning benchmarks, NCBR consistently improves both DAPO and GRPO. With DAPO, it achieves a 57.4% relative improvement in AIME'24 Pass@1 over the 2K DAPO baseline, while outperforming Full 8K training with 54.5% fewer GPU-hours.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has emerged as a powerful post-training paradigm for strengthening reasoning capabilities in language models on tasks with automatically verifiable outcomes. Its effectiveness has been demonstrated across mathematical reasoning and competitive programming (Shao et al., 2024; Guo et al., 2025; Kimi Team, 2025), software engineering (Wei et al., 2025a), and interactive web agents (Wei et al., 2025b). During RLVR training, each response is generated under a maximum response-length budget, which we refer to as the *rollout horizon*. This horizon determines both how far the policy can generate to obtain an outcome and how much of the generated response can participate in policy optimization.

As reasoning trajectories become longer, responses increasingly reach the rollout boundary before terminating naturally. How such boundary-truncated responses should be handled remains unsettled. DAPO (Yu et al., 2025) argues that directly treating them as incorrect can introduce noisy supervision and therefore considers filtering and soft overlong punishment. In contrast, Skywork-OR1 (He et al., 2025) reports that retaining their negative advantages can improve token efficiency. These different treatments raise a more fundamental question: *what information does a boundary-truncated response actually provide for policy optimization?*

In this paper, we first study how different rollout horizons affect RLVR training and uncover two observations. **First, longer rollouts do not necessarily translate into better performance, but they substantially delay final-answer generation.** As shown in Figure 1 (a-b), under the same 8K evaluation budget, increasing the training horizon from 2K to 8K decreases AIME'24 Pass@1 from 10.52% to 9.90%, while shifting the mean final-answer position from 1036 to 1928 tokens. Thus, longer-horizon training changes *when* the model produces its final answer without correspondingly improving *what* it can solve.

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

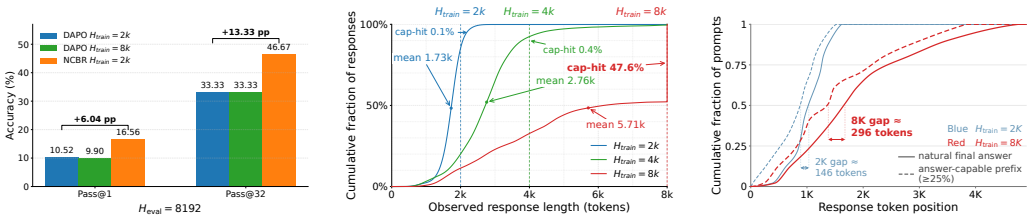


Figure 1: **Rollout horizons, task performance, and answer generation.** Results for Qwen3-4B trained with DAPO. (a) AIME'24 Pass@1 and Pass@32 for DAPO and DAPO with NCBR under a common evaluation horizon of $H_{\text{eval}} = 8192$. (b) Observed response-length distributions. (c) Distributions of natural final-answer positions (solid) and answer-capable prefix positions (dashed; probe accuracy $\geq 25\%$).

Second, a later final answer does not imply later answer capability. The comparable performance of the 2K- and 8K-trained policies raises a natural question: does the longer-horizon policy already have sufficient information to answer correctly before its final answer eventually appears? We test this using prefix probes, which truncate a response before its natural final answer and ask the same frozen policy to directly produce a short answer. As shown in Figure 1 (c), we find that earlier prefixes can already support verifier-correct answers, even when the original response continues substantially further before producing its final answer (Section 4.2). This reveals a separation between *answer capability* and *final-answer generation*.

However, this does not imply that the remaining generation is useless. A short rollout horizon may terminate outcome observation before the eventual task outcome becomes visible. When boundary-truncated responses at 2K are naturally continued to 8K under the same frozen policy, up to 33.56% of responses judged incorrect at the short boundary become verifier-correct after continuation (Section 4.1). Longer generation can therefore provide useful *outcome information*, even when the model may already possess the capability to answer earlier. These observations expose a mismatch in conventional RLVR: the number of tokens needed to *observe* an informative outcome need not equal the number of tokens that must directly participate in the *actor update*.

Our theoretical analysis connects this mismatch to feedback and policy optimization. Extended outcome feedback can weaken the relative preference for earlier answers when later correct outcomes become observable, providing a mechanism for the observed shift in final-answer position. Moreover, under frozen policy, natural continuation provides a Monte Carlo estimate of the continuation value of the original response. Holding this continuation value fixed allows extended outcome feedback to guide updates involving only the original-response token probabilities. This provides a principled basis for separating outcome observation from actor optimization (Section 5.4).

Motivated by this separation, we propose **Natural-Continuation Boundary Return (NCBR)**, which decouples the *outcome-observation horizon* L from the *actor-update horizon* $H < L$. NCBR naturally continues responses truncated at H under the same policy snapshot, up to a total response length of L, and uses the resulting verifier outcomes to reassess the original responses. Actor optimization uses only tokens from the original H-bounded responses. Empirically, NCBR improves DAPO and GRPO on Qwen3-4B and GRPO on Qwen3-8B, yielding consistent gains in Pass@1 and Pass@32 across AIME'24, AIME'25, and AMC'23, including relative AIME'24 Pass@1 gains of 57.4% with DAP on Qwen3-4B. Horizon ablations show that NCBR with 2K actor updates and 8K outcome observation outperforms Full 8K on both metrics across all benchmarks, using 54.5% fewer accounted GPU-hours and 62.4% fewer actor-update tokens.

2 Related Work

Policy optimization for RLVR. Critic-free language-model RL includes ReMax (Li et al., 2024) and RLOO (Ahmadian et al., 2024); GRPO (Shao et al., 2024) uses group-relative rewards. Subsequent variants revise normalization (Dr. GRPO (Liu et al., 2025)), clipping and sampling (DAPO (Yu et al., 2025)), sequence-level importance ratios (GSPO (Zheng et al., 2025)), or geometric-mean

aggregation (GMPO (Zhao et al., 2026)). DAPO also mitigates truncation-induced reward noise through overlong reward shaping.

Selective updates and credit assignment. Our work is related to methods that refine token selection and credit assignment in RLVR. Wang et al. (2025) restrict policy-gradient updates to high-entropy tokens, while D^{3}S (Wang et al., 2026) combines entropy and advantage magnitude for token selection; related work also studies divergence-weighted advantages (Meng et al., 2026). VinePPO (Kazemnejad et al., 2025) refines step-level credit through Monte Carlo estimation, whereas PPPO (Sun et al., 2026) estimates prefix-level learning signals from repeated continuations and progressively optimizes selected prefixes. Compared to PPPO (Sun et al., 2026), which estimates prefix-level learning signals from repeated continuations and progressively optimizes selected prefixes, NCBR targets verifier feedback missed under a fixed actor-update horizon, continuing each truncated response only once while retaining the original actor-update tokens.

Our study complements work on *where credit is assigned* by examining the distinct roles of the *outcome-observation horizon* and the *actor-update horizon*. We treat outcome observation and actor optimization as separate design decisions in RLVR, enabling delayed verifier feedback to inform learning without extending actor updates.

3 PRELIMINARIES

3.1 RLVR UNDER BOUNDED ROLLOUTS

Reinforcement learning with verifiable rewards (RLVR) optimizes a language model using automatically verifiable task outcomes. For a prompt $x \sim \mathcal{D}$, the rollout policy $\pi_{\theta_{\text{old}}}$ samples a group of G responses under a maximum response-length budget or rollout horizon, H:

$$y_i^H = (a_{i,1}, \ldots, a_{i,T_i}) \sim \text{Rollout}(\pi_{\theta_{\text{old}}}, x; H), \qquad i = 1, \ldots, G. \tag{1}$$

Here, $T_i = \min\{\tau_i, H\}$, where τ_i is the natural stopping time. Thus, generation ends either naturally or at the externally imposed rollout boundary.

A *strict boundary hit* reaches H before natural termination. For prompt x, we denote the set of such responses by

$$\mathcal{B}_x := \{i \in \{1, \ldots, G\} \mid T_i = H < \tau_i\}.$$

The task-specific verifier evaluates the observed response and returns the binary verifier reward $r_i^H = \mathcal{V}(x, y_i^H) \in \{0, 1\}$, indicating correctness observable under horizon H. The base recipe constructs the training score $s_i^H = f(r_i^H) + \ell_i^H$, where f is its fixed verifier-reward encoding and ℓ_i^H is auxiliary shaping on the original response, such as a length penalty. The binary-feedback analysis uses $f(r) = r$ and $\ell_i^H = 0$; the audited DAPO recipe uses $f(r) = 2r - 1$ (Appendix C.3).

For $i \in \mathcal{B}_x$, y_i^H is a boundary-truncated prefix. Accordingly, r_i^H describes only the task outcome observable from that prefix and need not match the outcome obtained by naturally continuing the same prefix under a larger observation budget $L > H$.

3.2 GROUP-RELATIVE POLICY UPDATES

Group-relative optimization constructs a learning signal by comparing training scores within each prompt group. Let $\bar{s}^H = \text{Mean}(s_1^H, \ldots, s_G^H)$ and $\sigma_s^H = \text{Std}(s_1^H, \ldots, s_G^H)$ denote the group mean and standard deviation, respectively. The sequence-level advantage is

$$A_i^H = \begin{cases} \dfrac{s_i^H - \bar{s}^H}{\sigma_s^H}, & \sigma_s^H > 0, \\ 0, & \sigma_s^H = 0. \end{cases} \tag{2}$$

These advantages provide relative learning signals for policy updates on the sampled responses. Because the normalization statistics are shared within a prompt group, changing one response's training score can also change the advantages assigned to other responses. The choice of policy objective is separate from score construction; algorithm-specific policy objectives are described in Appendix C.2, while DAPO-specific auxiliary shaping and optional dynamic sampling are detailed in Appendix C.3.

4 The Mismatch Between Outcome Observation and Actor Updates

4.1 The Horizon Dilemma: Balancing Update Efficiency and Feedback Completeness

The rollout horizon simultaneously controls how far the policy generates during training and how much of each response is included in the actor update. Increasing this horizon may expose later task outcomes, but it also increases rollout cost and places more generated tokens inside the optimization trajectory. Conversely, a short horizon is computationally attractive, but may terminate outcome observation before the response naturally reaches its eventual task outcome. We study this trade-off using Qwen3-4B trained with DAPO on DAPO-Math-17K under $H \in \{2048, 4096, 8192\}$; detailed training configurations are provided in Appendix A.1, the boundary-audit protocol in Appendix B.1, and complete training dynamics and cost in Appendix G.

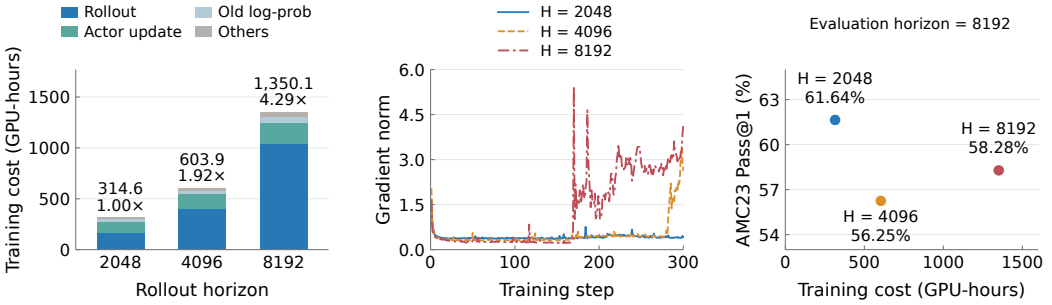


Figure 2: **Effect of rollout horizon on training efficiency, optimization stability, and downstream performance.** Left: training cost under $H \in \{2048, 4096, 8192\}$. Middle: gradient-norm dynamics during training, where longer horizons exhibit pronounced late-stage increases and spikes. Right: AMC23 Pass@1 of the Step-300 checkpoints under the evaluation budget $H_{\text{eval}} = 8192$.

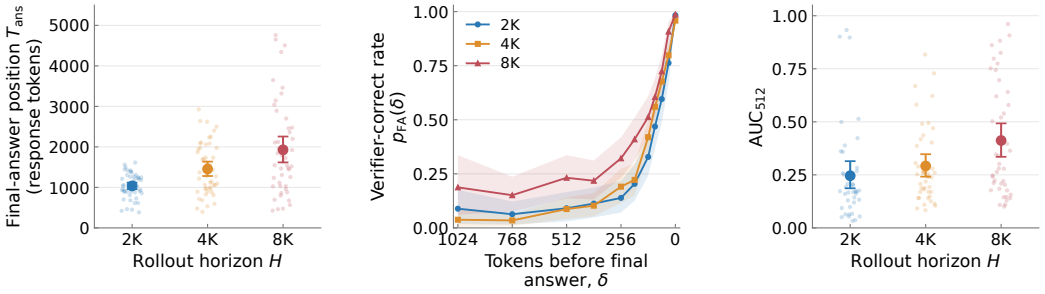
Long horizons increase cost and update burden without commensurate gains. As shown in Figure 2, increasing the rollout horizon from $H = 2048$ to $H = 8192$ raises training cost by $4.29\times$ and actor-update tokens from 0.788B to 2.058B. The larger update burden is accompanied by less stable optimization, with longer horizons exhibiting pronounced late-stage gradient increases and spikes. However, this additional computation and optimization does not translate into better downstream performance: under the common evaluation budget $H_{\text{eval}} = 8192$, AMC23 Pass@1 is 61.64%, 56.25%, and 58.28% for $H = 2048$, 4096, and 8192.

Finding 1. *Longer rollout horizons increase actor-update burden and optimization instability, yet do not yield commensurate downstream performance gains.*

Table 1: Boundary feedback revealed by natural continuation from $H = 2048$ to $L = 8192$. An entry $a \to b$ denotes the verifier outcome before and after continuation, with 0 for incorrect and 1 for correct. Recovery is measured among verifier-incorrect boundary responses; unlocked groups are all-zero candidate groups that become mixed-outcome groups under extended observation.

Checkpoint	$0 \to 0$	$0 \to 1$	$1 \to 1$	$1 \to 0$	Recovery	Unlocked groups
Step 0	9,058	183	1	0	1.98%	2.27%
Step 100	673	148	3	0	18.03%	7.87%
Step 200	8,194	3,041	49	7	27.07%	11.81%
Step 300	487	246	5	0	33.56%	12.84%

Short horizons reduce cost but can truncate informative feedback. Although short rollout horizons are computationally efficient, they may terminate generation before the eventual task outcome is observed. To examine this effect, we extend boundary-hit responses generated under $H = 2048$



(a) Final-answer position T_{ans}. (b) Prefix probe $p_{\text{FA}}(\delta)$. (c) Prefix-probe AUC_{512}.

Figure 3: **Final-answer position and prefix probes across training rollout horizons.** (a) Mean response-token position T_{ans} at which the verifier-selected final answer ends. (b) Verifier-correct rate $p_{\text{FA}}(\delta)$ of short completions sampled from response prefixes located δ tokens before T_{ans}. (c) AUC_{512} summarizes $p_{\text{FA}}(\delta)$ over the 512 response tokens preceding T_{ans}. Error bars and shaded regions denote prompt-level bootstrap 95% confidence intervals.

to a total observation budget of $L = 8192$ using the same frozen policy, and re-evaluate the concatenated responses with the same verifier. As shown in Table 1, a substantial fraction of responses judged incorrect at the short boundary can recover after continuation. By Step 300, 33.56% of such responses become correct, while 12.84% of all-zero candidate groups become mixed-outcome groups. The latter would otherwise be discarded by DAPO dynamic sampling. These results indicate that short-horizon truncation can underestimate the continuation potential of a response and alter both within-group relative advantages and the samples retained for policy optimization.

Finding 2. *Short horizons reduce training cost, but boundary truncation can alter boundary outcomes and group selection under dynamic sampling.*

4.2 Final-Answer Position and Earlier Answer Capability

The limited gains from longer rollout horizons raise a question about what the additional generation achieves: does a later final answer mean that the model could not have answered earlier? We distinguish *final-answer emission* from *earlier answer capability*: where the policy actually states its answer, and whether a preceding response prefix already supports a correct short answer when the same policy is prompted to answer directly.

Longer horizons delay answer emission. Under a common evaluation budget, longer-horizon policies place their final answers substantially later. On the shared prompt set with verifier-correct responses, the mean final-answer position T_{ans} increases from about 1.04K tokens for the 2K-trained policy to 1.93K for the 8K-trained policy (Figure 3a). This shift establishes when the answer is expressed, but leaves open what the preceding response already supports.

Earlier prefixes already support correct answers. To examine this distinction, we take prefixes δ tokens before each response's final answer and ask the same frozen policy to answer directly. Let $p_{\text{FA}}^{H}(\delta)$ denote the correctness rate of these short completions for the H-trained policy. Although the 8K-trained policy gives its final answer later, its preceding prefixes support correct short answers more readily over much of the relative-position range (Figure 3b). Its average probe accuracy over the 512-token pre-answer window, summarized by AUC_{512}, is 0.412, compared with 0.246 for the 2K-trained policy (Figure 3c). Thus, later answer emission can coexist with stronger answer capability before that answer, rather than requiring the policy to follow the entire original response before it can answer correctly. Appendix B.2 gives the probe protocol, matched-sample construction, and statistical results.

Finding 3. *Longer rollout horizons delay final-answer emission, yet earlier response prefixes can already support correct short answers. The final-answer position therefore does not by itself locate the response's earlier answer capability.*

5 METHOD

Natural-Continuation Boundary Return (NCBR) separates the outcome-observation horizon L from the actor-update horizon H. It naturally continues strict boundary hits to $L > H$, replaces their verifier feedback, and updates only the original H-bounded responses.

5.1 NATURAL CONTINUATION AT THE ROLLOUT BOUNDARY

For each prompt x, sample G original responses $\{y_i^H\}_{i=1}^G$ as in Section 3.1. For each strict boundary hit $i \in \mathcal{B}_x$, sample one continuation:

$$u_i \sim \mathrm{Rollout}\big(\pi_{\theta_{\mathrm{old}}}, (x, y_i^H); L - H\big),$$
$$y_i^L := y_i^H \oplus u_i, \qquad r_i^L := \mathcal{V}(x, y_i^L).$$

Continuation uses the exact original tokens, the same policy snapshot, and matched sampling settings. Generation stops naturally or at total response length L, and the same verifier evaluates the concatenated response. Responses outside $\mathcal{B}_x$ retain their original outcomes.

5.2 BOUNDARY-RETURN REPLACEMENT

NCBR replaces only the verifier-dependent part of the training score:

$$\widetilde{s}_i^L := \begin{cases} f(r_i^L) + \ell_i^H, & i \in \mathcal{B}_x, \\ s_i^H, & i \notin \mathcal{B}_x. \end{cases} \tag{3}$$

The reward encoding f and original-response shaping ℓ_i^H remain unchanged. Replacement uses the extended outcome whether correctness improves, regresses, or is preserved. The base algorithm recomputes advantages over the complete reassessed group, yielding $\widetilde{A}_i^L$ under Eq. (2). Shared group statistics can therefore change the advantage of a non-boundary response whose own score is unchanged.

5.3 PREFIX-ONLY OPTIMIZATION WITH THE BASE ALGORITHM

Let $\mathcal{L}_{\mathrm{base}}$ denote the native actor loss, including its importance ratios, clipping, aggregation, and configured regularization. NCBR supplies the reassessed advantages to this loss:

$$\mathcal{L}_{\mathrm{base+NCBR}}(\theta) := \mathcal{L}_{\mathrm{base}}\Big(\theta; \{(x, y_i^H, \widetilde{A}_i^L)\}_{x,i}\Big). \tag{4}$$

All policy-loss terms, likelihood ratios, and response-length normalization use only valid original-response tokens, with $T_i \le H$; returns and advantages are held constant during the update. The interface applies to GRPO and DAPO (Shao et al., 2024; Yu et al., 2025); Appendix C.2 gives their instantiations. When correctness-based group filtering is used, reassessment precedes filtering: reassessed outcomes determine eligibility, while reassessed scores determine advantages (Appendix C.3). Continuation adds generation and verification cost, but its outputs are rollout tokens, not actor-update tokens.

5.4 WHY SEPARATE THE TWO HORIZONS?

We first interpret the answer-position and relative prefix-probe observations in Section 4.2, then show how extended outcome feedback can guide updates to the original response.

Extended feedback changes pressure to answer early. Fix a response group observed to L, where $i \in \{1, \dots, G\}$ indexes the responses and each response has a defined final-answer position $T_{\mathrm{ans},i}$. For each $h \in \{H, L\}$, compute the group-relative advantages A_i^h from the corresponding response prefixes. For reweighting strength $\eta \ge 0$, let $w_i^h(\eta) \propto e^{\eta A_i^h}$ be normalized group weights and $\mathcal{T}_h(\eta) := \sum_i w_i^h(\eta) T_{\mathrm{ans},i}$. Then

$$\frac{d}{d\eta}(\mathcal{T}_L - \mathcal{T}_H)\bigg|_{\eta=0} = \mathrm{Cov}_G(T_{\mathrm{ans}}, A^L - A^H), \tag{5}$$

where Cov_G is the within-group covariance. A positive covariance means that extended feedback favors later answers relative to short feedback for sufficiently small positive η. For example, under unshaped binary feedback, if all responses finish correctly by L but only some reveal the correct answer by H, long feedback removes the preference for early answers induced by short feedback. Appendix D.1 gives the proof and finite-update example.

Answer emission and earlier capability can separate. The relative prefix probes in Section 4.2 examine whether response prefixes already support correct answering before final-answer emission. To interpret these probes, consider a two-level model: direct-answer success increases from q_0 to q_1, with $0 \le q_0 < q_1 \le 1$, at a transition preceding final-answer emission. Let $S_H \ge 0$ be the number of tokens from this transition to the final answer for the H-trained policy. For a fixed comparison population and common q_0, q_1 across responses and horizons, the average probe accuracy over a continuous relative-position window of width $D > 0$, denoted $\mathrm{AUC}^H_{D,\mathrm{cont}}$, is

$$\mathrm{AUC}^H_{D,\mathrm{cont}} = q_0 + \frac{q_1 - q_0}{D}\mathbb{E}[\min\{S_H, D\}]. \tag{6}$$

Under this model, a larger area corresponds to a longer average interval of higher answerability before final-answer emission, truncated at D. Appendix D.2 derives this relation and its discrete-grid counterpart. This distinction motivates NCBR's separation of the outcome-observation horizon from the actor-update horizon. The next analysis shows how outcomes observed through natural continuation can guide updates to the original response.

Outcome observation and actor updates have different value. Natural continuation gives a Monte Carlo evaluation of the original response. Fix x and the continuation policy $\pi_{\theta_{\mathrm{old}}}$. Let R_i be the reassessed binary outcome, equal to r_i^L for a strict boundary hit and r_i^H for a naturally terminated response, and define its continuation value as $v_x(y^H) := \mathbb{E}[R_i \mid x, y_i^H = y^H]$ under this frozen policy. Let $\pi_\theta(y^H \mid x)$ denote the probability of an original bounded response. Holding v_x fixed and evaluating all derivatives at $\theta = \theta_{\mathrm{old}}$ gives

$$\nabla_\theta \mathbb{E}_{y^H \sim \pi_\theta(\cdot|x)}[v_x(y^H)] = \mathbb{E}\left[R_i \nabla_\theta \log \pi_\theta(y_i^H \mid x)\right], \tag{7}$$

where the right-hand expectation uses the frozen rollout and continuation policy. The score term contains only original-response tokens (Schulman et al., 2015), so a later outcome can estimate this value gradient without including continuation-token score terms. For a policy-space comparison, consider the expected frozen continuation value penalized by the joint Kullback–Leibler (KL) divergence of original responses and continuations from the frozen rollout distribution. For any chosen original-response distribution, retaining the frozen continuation policy minimizes this divergence without changing the value term. Thus, this KL-regularized value objective admits a prefix-only policy-space optimum. Appendix D.3 proves both results.

NCBR further transforms reassessed outcomes into group-relative advantages. Under binary scores without auxiliary shaping, independent sampling, and no outcome-based group selection, one sampled boundary return favors higher-value original responses in expectation under group-relative response reweighting (Appendix D.4). Appendix E evaluates the reassessed advantages using independent continuation outcomes.

6 Experiments

Training and evaluation protocol. We compare DAPO and GRPO with and without NCBR under matched training settings. The main comparisons use actor-update horizon $H = 2048$ and checkpoints at Step 300. Vanilla baselines observe outcomes within H, whereas NCBR uses outcome-observation horizon $L = 8192$. We report Pass@1 and Pass@32 on AIME'24, AIME'25, and AMC'23 with an inference-time response budget $H_{\mathrm{eval}} = 8192$. Full experimental details are provided in Appendix A; training dynamics and cost appear in Appendices G, and H.

6.1 Main Results

Table 2 shows that NCBR improves Pass@1 and Pass@32 across all three benchmarks with DAPO and GRPO on Qwen3-4B, and with GRPO on Qwen3-8B. With DAPO on Qwen3-4B, AIME'24 Pass@1 increases from 10.52% to 16.56% (57.4% relative improvement), while Pass@32 rises from

Table 2: Performance (%) of pretrained reference models and RLVR checkpoints under $H_{\text{eval}} = 8192$. Qwen3 uses $H = 2048$ and NCBR $L = 8192$. Superscripts denote relative changes over the corresponding optimizer baseline; bold denotes the best result within each optimizer pair.

Model	Optimizer	Variant	AIME'25		AIME'24		AMC'23	
			Pass@1	Pass@32	Pass@1	Pass@32	Pass@1	Pass@32
Qwen3-4B	–	Base	0.21	6.67	2.08	26.67	9.14	80.00
	GRPO	Vanilla	6.67	26.67	12.92	33.33	46.95	82.50
		+ NCBR	**8.54**$^{+28.1\%}$	**33.33**$^{+25.0\%}$	**14.17**$^{+9.7\%}$	**36.67**$^{+10.0\%}$	**54.53**$^{+16.1\%}$	**87.50**$^{+6.1\%}$
	DAPO	Vanilla	11.46	36.67	10.52	33.33	61.64	85.00
		+ NCBR	**14.48**$^{+26.4\%}$	**40.00**$^{+9.1\%}$	**16.56**$^{+57.4\%}$	**46.67**$^{+40.0\%}$	**69.84**$^{+13.3\%}$	**95.00**$^{+11.8\%}$
Qwen3-8B	–	Base	2.29	23.33	2.71	40.00	11.95	90.00
	GRPO	Vanilla	14.69	40.00	18.12	50.00	64.38	90.00
		+ NCBR	**20.10**$^{+36.8\%}$	**46.67**$^{+16.7\%}$	**24.48**$^{+35.1\%}$	**60.00**$^{+20.0\%}$	**69.84**$^{+8.5\%}$	**97.50**$^{+8.3\%}$
	DAPO	Vanilla	18.54	36.67	**22.60**	**53.33**	**80.31**	**97.50**
		+ NCBR	**18.85**$^{+1.7\%}$	**43.33**$^{+18.2\%}$	18.96$^{-16.1\%}$	43.33$^{-18.8\%}$	71.72$^{-10.7\%}$	**97.50**

33.33% to 46.67%. GRPO on Qwen3-8B also achieves a 35.1% relative improvement in AIME'24 Pass@1, demonstrating benefits across model sizes with the actor-update horizon fixed at 2K.

Results with DAPO on Qwen3-8B are mixed: NCBR improves both metrics on AIME'25, lowers both on AIME'24, and decreases AMC'23 Pass@1 while preserving Pass@32. Figure 4 shows that the 8B model produces relatively short responses and rarely triggers continuation after early training. Although its boundary recovery rate is comparable to that of the 4B model, the low continuation rate limits the number of responses whose horizon-dependent feedback NCBR can correct.

Overall, NCBR is most effective when short-horizon truncation hides useful delayed outcomes, and provides less benefit when responses already terminate well before the rollout boundary.

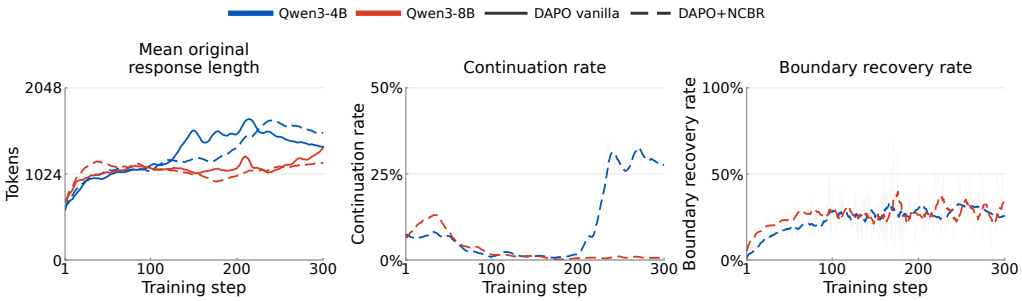


Figure 4: **Response length and boundary feedback under DAPO.** We compare Qwen3-4B and Qwen3-8B with and without NCBR at actor-update horizon $H = 2048$; NCBR uses outcome-observation horizon $L = 8192$. Left: mean original response length. Middle: fraction of candidate responses receiving natural continuation. Right: fraction of verifier-incorrect boundary responses that become correct after continuation.

6.2 Effect of Decoupling Update and Observation Horizons

We next examine the effect of separating the actor-update horizon H from the outcome-observation horizon L. Table 3 compares task performance, actor-update tokens, and accounted training cost across horizon configurations. Extending both horizons increases training cost and actor-update tokens without consistent performance gains. Variant 1 ($H = 2K, L = 8K$) achieves higher Pass@1 and Pass@32 than Full 8K on all benchmarks, while using 54.5% fewer accounted GPU-hours and 62.4% fewer actor-update tokens. At fixed H, extending L increases training cost while leaving cu-

Table 3: **Cost–performance trade-off across update and observation horizons.** All results use Step 300 and $H_{\text{eval}} = 8192$. Benchmark entries report Pass@1 / Pass@32 (%). H denotes the actor-update horizon and L the outcome-observation horizon. For Variant 4, -2048 denotes suffix updates on the final 2048 response tokens.

Variant	Training horizon		AIME'25	AIME'24	AMC'23	Actor-update tokens (B)	Accounted GPU-hours
	H	L	P@1 / P@32	P@1 / P@32	P@1 / P@32		
Full 2K	2048	2048	11.46 / 36.67	10.52 / 33.33	61.64 / 85.00	0.788	314.62
Full 4K	4096	4096	**16.25** / 30.00	15.83 / 40.00	56.25 / 82.50	1.311	603.89
Full 8K	8192	8192	11.35 / 26.67	9.90 / 33.33	58.28 / 87.50	2.058	1350.07
Variant 1	2048	8192	14.48 / 40.00	**16.56 / 46.67**	**69.84 / 95.00**	0.774	614.38
Variant 2	2048	4096	14.48 / **43.33**	15.42 / 36.67	68.75 / 92.50	0.767	420.03
Variant 3	1024	4096	10.94 / 30.00	14.90 / 33.33	68.75 / 87.50	0.362	199.54
Variant 4	-2048	4096	14.69 / 30.00	14.38 / 33.33	65.00 / 90.00	1.022	553.91

mulative actor-update tokens nearly unchanged. At fixed L, further reducing H lowers actor-update tokens but decreases performance on all three benchmarks. These results support separate control of the outcome-observation and actor-update horizons to balance performance and training cost.

Prefix versus suffix updates. With an outcome-observation horizon of $L = 4096$, Variant 4 updates the final 2K tokens, whereas Variant 2 updates the original 2K-bounded responses. Variant 2 achieves higher Pass@32 on all three benchmarks and higher Pass@1 on AIME'24 and AMC'23. Variant 2 also uses 25.0% fewer actor-update tokens and 24.2% fewer accounted GPU-hours. These results indicate that prefix updates make more effective use of training compute, achieving better overall performance at lower cost.

6.3 Training Cost and Actor-Update Burden

Lower cost than extending both horizons. Figure 5 reports training cost through Step 300. With actor-update horizon $H = 2048$ and outcome-observation horizon $L = 8192$, DAPO with NCBR uses 54.5% fewer accounted GPU-hours and 62.4% fewer actor-update tokens than DAPO with both horizons set to 8192, while achieving higher Pass@1 on all three benchmarks under the same evaluation budget $H_{\text{eval}} = 8192$.

Preserving the short actor update. Relative to DAPO with $H = 2048$, NCBR increases the accounted training cost from 314.6 to 614.4 GPU-hours ($1.95\times$), while cumulative actor-update tokens remain nearly unchanged (0.788B vs. 0.774B). The additional cost mainly comes from natural continuation for extended outcome observation. Appendix G details the accounting and execution configurations.

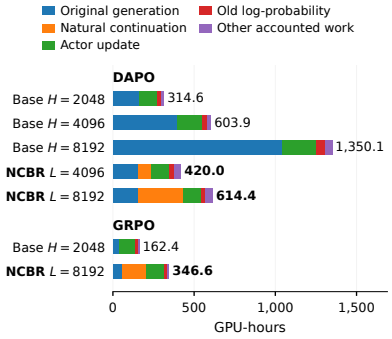


Figure 5: Training cost on Qwen3-4B through Step 300.

7 Conclusion

In this work, we explore the role of the rollout horizon in RLVR by examining what information a boundary-truncated response provides for policy optimization. We propose NCBR, which decouples outcome observation from actor optimization by extending verifier feedback while keeping the actor-update horizon short. With DAPO, NCBR improves AIME'24 by 57.4% while outperforming Full 8K training with 54.5% fewer accounted GPU-hours and 62.4% fewer actor-update tokens.

AI USE STATEMENT

AI tools were used to assist with data organization, table preparation, language polishing, and code development. The authors take full responsibility for the accuracy and integrity of the research and the final manuscript.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*, 2025. URL https://arxiv.org/abs/2505.22312.

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

Kimi Team. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.

Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29128–29163. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24cd.html.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=jLpC1zavzn.

Haoming Meng, Kexin Huang, Shaohang Wei, Chiyu Ma, Shuo Yang, Xue Wang, Guoyin Wang, Bolin Ding, and Jingren Zhou. Sparse but critical: A token-level analysis of distributional shifts in RLVR fine-tuning of LLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8vWIXno8LW.

John Schulman, Nicolas Heess, Theophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://arxiv.org/abs/1506.05254.

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

Yiliu Sun, Zicheng Zhao, Yang Wei, Yanfang Zhang, and Chen Gong. Well begun, half done: Reinforcement learning with prefix optimization for LLM reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):33144–33152, 2026. doi: 10.1609/aaai.v40i39.40598. URL https://ojs.aaai.org/index.php/AAAI/article/view/40598.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aBO5SvgSt1.

Chao Wang, Tao Yang, Hongtao Tian, Yunsheng Shi, Qiyao Ma, Xiaotao Liu, Ting Yao, and Wenbo Ding. Learning more with less: A dynamic dual-level down-sampling framework for efficient policy optimization. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xRgtzijvgc.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yuxiang Wei, Olivier Duchenne, Jade Copet, Quentin Carbonneaux, Lingming Zhang, Daniel Fried, Gabriel Synnaeve, Rishabh Singh, and Sida I. Wang. SWE-RL: Advancing LLM reasoning via reinforcement learning on open software evolution. *arXiv preprint arXiv:2502.18449*, 2025a.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. WebAgent-R1: Training web agents via end-to-end multi-turn reinforcement learning. *arXiv preprint arXiv:2505.16421*, 2025b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru 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, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/44a1f18afd6d5cc34d7e5c3d8a80f63b-Abstract-Conference.html.

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

A Training and Evaluation Protocols

A.1 Training Configuration

The main training experiments use Qwen3-4B, Qwen3-8B and DAPO-Math-17K. The main DAPO, GRPO comparisons use actor-update horizon $H = 2048$; vanilla baselines observe outcomes within H, whereas NCBR uses outcome-observation horizon $L = 8192$. The horizon ablation additionally considers $(H, L) = (2048, 4096)$ and $(1024, 4096)$, while the vanilla DAPO horizon study uses $H \in \{2048, 4096, 8192\}$ with matched outcome-observation and actor-update horizons.

DAPO samples 768 candidate prompts and retains 256 prompt groups through correctness-based dynamic sampling, whereas GRPO uses 256 prompt groups per iteration without this filtering. Within each base algorithm, the vanilla and NCBR variants use matched training settings; NCBR changes boundary-outcome reassessment before the native algorithm update. Tables 4 and 5 summarize the shared and algorithm-specific training hyperparameters, respectively. Hardware and software environments are reported in Appendix A.3.

Table 4: Shared optimization and rollout hyperparameters used in the training experiments. Algorithm-specific clipping and loss aggregation are reported in Table 5.

Setting	Value
Responses per prompt	$G = 8$
Train-batch size	256 (2,048 original responses)
Mini-batch size	16 prompt groups (128 responses)
Optimization epochs per update	1
Optimizer	AdamW
Learning rate	Constant 10^{-6}
Rollout sampling	Temperature 1.0, top-$p = 1.0$, no top-k truncation
Advantage normalization	Group-wise standard-deviation normalization enabled
Gradient clipping	Norm threshold 1.0

Table 5: Algorithm-specific training settings. The clipping interval is $[1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}]$. DAPO and GRPO use token-level importance ratios.

Optimizer	Correctness-based group filtering	ϵ_{low}	ϵ_{high}	Loss aggregation
DAPO	Enabled	0.2	0.28	Token mean
GRPO	Disabled	0.2	0.2	Sequence mean of token means

The algorithm-specific objectives, importance ratios, and NCBR score-replacement procedure are detailed in Appendix C.

A.2 Evaluation Protocol and Metrics

We use AIME'24, AIME'25, and AMC'23 for evaluation. For each problem, we independently sample 32 responses with an inference-time response budget of $H_{\text{eval}} = 8192$. Prompt formatting, answer extraction, and verifier are fixed across evaluations. For a benchmark with M problems, let $z_{j,k} \in \{0, 1\}$ denote verifier correctness for response k to problem j. We report

$$\text{Pass@1} = \frac{1}{32M} \sum_{j=1}^{M} \sum_{k=1}^{32} z_{j,k}, \qquad \text{Pass@32} = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\left\{ \sum_{k=1}^{32} z_{j,k} > 0 \right\}.$$

Pass@1 averages correctness over all sampled responses, while Pass@32 reports the fraction of problems solved by at least one of the 32 samples.

A.3 Hardware and Software

Training experiments are conducted on two nodes with eight NVIDIA A100-SXM4-40GB GPUs per node. The software environment uses VERL 0.8.0, vLLM 0.18.0, and PyTorch 2.10.0+cu129 with CUDA 12.9. Training uses PyTorch FSDP1. Computational-cost accounting is described in Appendix G.

B Supplementary Motivation Experiments

B.1 Natural-Continuation Boundary Audit

Protocol. We audit the $H = 2048$ checkpoints at Steps 0, 100, 200, and 300 on DAPO-Math-17K prompts. Each strict boundary-hit response is continued once from its exact original token prefix under the same frozen policy, to a total outcome-observation horizon of $L = 8192$. The verifier then re-evaluates the concatenated response. Naturally terminated responses are not continued, while all strict boundary hits are included regardless of their verifier outcome at H. Boundary detection and token alignment follow Appendix C.1.

We retain the original response groups so that outcome changes can be analyzed both at the response level and at the group level. This allows us to measure how often extended observation changes a boundary response's verifier outcome and how often it changes an all-zero group into a mixed-outcome group.

Metrics and denominators. Let n_{ab} denote the number of strict boundary-hit responses whose verifier outcome is a at H and b after continuation to L, for $a, b \in \{0, 1\}$. We define the recovery rate as

$$\frac{n_{01}}{n_{00} + n_{01}},$$

the fraction of verifier-incorrect boundary responses that become correct after continuation. For a candidate group x, let K_x^H and K_x^L denote the numbers of correct responses before and after reassessment. An all-zero group is unlocked when

$$K_x^H = 0 \qquad \text{and} \qquad 0 < K_x^L < G.$$

The unlock rate is the fraction of audited all-zero candidate groups that satisfy this condition. The number of boundary responses varies across checkpoints because strict boundary hits are determined separately under each frozen policy, and their frequency changes with the checkpoint's response-length and termination distribution.

Table 6: Boundary-outcome transitions under natural continuation from $H = 2048$ to $L = 8192$. An entry $a \to b$ denotes the verifier outcome before and after continuation. Recovery is computed over verifier-incorrect boundary responses, and the unlock rate is computed over all-zero candidate groups.

Checkpoint	$0 \to 0$	$0 \to 1$	$1 \to 1$	$1 \to 0$	Recovery	Unlock rate
Step 0	9,058	183	1	0	1.98%	2.27%
Step 100	673	148	3	0	18.03%	7.87%
Step 200	8,194	3,041	49	7	27.07%	11.81%
Step 300	487	246	5	0	33.56%	12.84%

Recovery increases from 1.98% at Step 0 to 33.56% at Step 300, where 246 of 733 verifier-incorrect boundary responses become correct after continuation. Over the same checkpoints, the unlock rate increases from 2.27% to 12.84%, showing that extended outcome observation can also change the within-group outcome structure used for correctness-based group selection. The $1 \to 0$ transitions at Step 200 further show that continuation can revise boundary outcomes in either direction, motivating symmetric outcome reassessment rather than a one-sided recovery bonus.

Because each continuation provides one stochastic outcome from the frozen policy, these statistics characterize sampled long-horizon boundary outcomes rather than a deterministic return associated with each prefix. Appendix E separately evaluates the reassessed feedback against an independent continuation outcome.

B.2 Final-Answer Position and Prefix Probes

Experimental Setup and Probe Protocol

We use the Qwen3-4B checkpoints trained with DAPO under $H \in \{2048, 4096, 8192\}$ at Step 100. Each checkpoint is evaluated with 32 independently sampled responses per problem on AIME'24, AIME'25, and AMC'23 using $H_{\text{eval}} = 8192$ following Appendix A.2. We retain prompts for which all three checkpoints produce at least two valid, verifier-correct responses, and analyze two responses per prompt and horizon. This yields 48 common prompts and 288 response instances across the three horizons.

As in Section 4.2, T_{ans} denotes the response-token position at which the verifier-selected final answer ends. The verifier selects the last `Answer:` match within the final 300 characters, and the end of that answer span is mapped to the original response tokens.

For each selected response, we probe prefixes at

$$\delta \in \{1024, 768, 512, 384, 256, 192, 128, 96, 64, 32, 0\},$$

where the prefix ends at

$$t = T_{\text{ans}} - \delta.$$

The exact token prefix is retained without re-encoding, and the following instruction is appended:

> Based on the reasoning above, give the final answer now. Do not restart or re-derive the solution. Put the final answer on its own line in the format `Answer: <final answer>`.

At each available prefix, we sample 16 short completions from the same frozen policy with at most 128 new tokens and normal EOS termination. Only the newly generated completion is scored by the verifier. The same selected responses are used across all probe offsets.

Metrics and Statistical Procedure

Let $z_{i,k}(\delta) \in \{0, 1\}$ denote the verifier outcome of completion k from response i at offset δ. We define the response-level prefix-probe accuracy as

$$p_{\text{FA},i}(\delta) = \frac{1}{16} \sum_{k=1}^{16} z_{i,k}(\delta).$$

For each horizon, the two selected responses are first averaged within each prompt, after which prompts receive equal weight. At a given offset, we use the common set of prompts for which both selected responses are available under all three horizons.

We summarize the 512 tokens preceding the final answer using AUC_{512}. For each response, let

$$\mathcal{D}_{512} = \{0, 32, 64, 96, 128, 192, 256, 384, 512\},$$

with offsets ordered as $0 = \delta_1 < \delta_2 < \cdots < \delta_9 = 512$. We compute

$$\text{AUC}_{512,i} = \frac{1}{512} \sum_{j=1}^{8} \frac{\delta_{j+1} - \delta_j}{2} \left[p_{\text{FA},i}(\delta_j) + p_{\text{FA},i}(\delta_{j+1}) \right].$$

Requiring all nine probe positions leaves 43 common prompts, comprising 7 from AIME'24, 8 from AIME'25, and 28 from AMC'23.

We report prompt-level 95% confidence intervals using 10,000 bootstrap resamples, with paired resampling for horizon comparisons.

Detailed Results and Endpoint Controls

Table 7 reports the final-answer positions, AUC_{512} values, and paired horizon differences underlying Figure 3. Table 8 reports results across all evaluated offsets. Prefix-probe accuracy increases as the probe approaches the original final-answer position, with the 8K-trained policy numerically highest across the reported offsets. The $\delta = 0$ and $t = 0$ rows serve as endpoint controls: probe accuracy is close to one when the prefix already contains the verifier-selected final answer, and no verifier-correct completion is observed when no response prefix is provided.

Table 7: Final-answer position and prefix-probe AUC_{512} at Step 100. Brackets show prompt-level bootstrap 95% confidence intervals. $\Delta(a \to b)$ denotes the paired difference from horizon a to b.

Horizon / contrast	T_{ans} $(n = 48)$	AUC_{512} $(n = 43)$
$H = 2\mathrm{K}$	$1036.8\ [956.1, 1117.9]$	$0.246\ [0.187, 0.315]$
$H = 4\mathrm{K}$	$1455.4\ [1281.2, 1637.0]$	$0.292\ [0.242, 0.347]$
$H = 8\mathrm{K}$	$1928.4\ [1617.7, 2256.4]$	$0.412\ [0.335, 0.493]$
$\Delta(2\mathrm{K} \to 4\mathrm{K})$	$+418.5\ [294.4, 555.5]$	$+0.0460\ [-0.0002, 0.0907]$
$\Delta(2\mathrm{K} \to 8\mathrm{K})$	$+891.6\ [639.0, 1165.3]$	$+0.1659\ [0.0984, 0.2367]$
$\Delta(4\mathrm{K} \to 8\mathrm{K})$	$+473.1\ [272.5, 691.9]$	$+0.1200\ [0.0557, 0.1878]$

Table 8: Complete prefix-probe results. Entries are $p_{\mathrm{FA}}(\delta)$ with prompt-level bootstrap 95% confidence intervals. At each offset, n denotes the common prompts available under all three horizons. The final row is the separate no-prefix control.

δ	n	$H = 2048$	$H = 4096$	$H = 8192$
1024	19	$0.089\ [0.018, 0.176]$	$0.038\ [0.005, 0.088]$	$0.188\ [0.061, 0.336]$
768	37	$0.063\ [0.016, 0.122]$	$0.035\ [0.009, 0.066]$	$0.151\ [0.075, 0.237]$
512	43	$0.091\ [0.030, 0.165]$	$0.087\ [0.041, 0.141]$	$0.232\ [0.137, 0.336]$
384	45	$0.113\ [0.048, 0.187]$	$0.103\ [0.054, 0.161]$	$0.218\ [0.132, 0.312]$
256	48	$0.139\ [0.071, 0.217]$	$0.191\ [0.116, 0.273]$	$0.322\ [0.230, 0.419]$
192	48	$0.204\ [0.126, 0.292]$	$0.222\ [0.151, 0.298]$	$0.410\ [0.311, 0.510]$
128	48	$0.328\ [0.242, 0.421]$	$0.419\ [0.341, 0.499]$	$0.512\ [0.415, 0.609]$
96	48	$0.469\ [0.367, 0.572]$	$0.560\ [0.488, 0.631]$	$0.605\ [0.517, 0.691]$
64	48	$0.596\ [0.486, 0.703]$	$0.676\ [0.606, 0.743]$	$0.724\ [0.643, 0.799]$
32	48	$0.762\ [0.666, 0.855]$	$0.798\ [0.736, 0.854]$	$0.908\ [0.854, 0.954]$
0	48	$0.983\ [0.961, 0.997]$	$0.958\ [0.940, 0.973]$	$0.987\ [0.967, 0.999]$
$t = 0$	48	$0.000\ [0.000, 0.000]$	$0.000\ [0.000, 0.000]$	$0.000\ [0.000, 0.000]$

C NCBR Implementation Details

C.1 Execution Semantics and Pseudocode

NCBR is a reward-stage module between verification of the original responses and the base optimizer's native training path. Its inputs are the original responses, verifier rewards, training scores, frozen rollout policy, actor-update horizon H, and outcome-observation horizon $L > H$. It returns the same original responses with reassessed training scores. When required by the base recipe, NCBR also returns reassessed correctness metadata for group selection.

Strict boundary detection distinguishes length termination from natural stopping. The implementation uses backend termination metadata first; if it is unavailable or ambiguous, a full response window without EOS is treated as a boundary hit. Each strict boundary hit receives one continuation, including those already correct at H.

The continuation uses the exact prompt and complete valid response-prefix tokens, the same policy snapshot, and matched sampling settings. The extended response is verified in a separate batch. In a token-reward implementation, the encoded-reward correction $f(r_i^L) - f(r_i^H)$ is added at the last valid token of the original response, preserving all other short-response reward contributions. Original response tokens, masks, and actor-batch dimensions are unchanged. Continuation tokens count as rollout tokens, not actor-update tokens, and do not enter likelihood ratios or response-length normalization.

The current implementation targets synchronous, single-turn text RLVR with group-relative outcome advantages. It uses the registered DAPO reward manager as the verifier/score interface, while the policy objective and group-selection procedure remain those of the configured base recipe. Reward-side KL penalties are disabled in the reported experiments; configured actor-loss regularization is retained.

Algorithm 1 RLVR training with Natural-Continuation Boundary Return (NCBR)

Require: Policy π_θ, base RLVR recipe $\mathcal{A}$, verifier $\mathcal{V}$, horizons $H < L$, group size G

for each rollout step **do**
  $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$; sample prompt batch $\mathcal{D}_b$
  for each $x \in \mathcal{D}_b$ **do**
    Sample G responses $\{y_i^H\}$ under $\pi_{\theta_{\text{old}}}$ and compute $\{r_i^H, s_i^H\}$
    Initialize $\widetilde{r}_i^L \leftarrow r_i^H$ and $\widetilde{s}_i^L \leftarrow s_i^H$
    for each $i \in \mathcal{B}_x$ **do**
      Sample $u_i \sim \text{Rollout}(\pi_{\theta_{\text{old}}}, (x, y_i^H); L - H)$
      $\widetilde{r}_i^L \leftarrow \mathcal{V}(x, y_i^H \oplus u_i)$ and reassess $\widetilde{s}_i^L$ by Eq. equation 3
    end for
    Apply native selection and advantage construction using the reassessed group
  end for
  Update π_θ with $\mathcal{A}$ using only tokens of $\{y_i^H\}$
end for

Ensure: Updated policy π_θ

Boundary continuations can have substantially different generation lengths because each request stops either at natural termination or at the outcome-observation horizon L. Executing them in fixed synchronous batches can therefore leave generation slots idle while waiting for longer requests to finish. We use a work-conserving continuation scheduler: once a continuation completes and releases its resources, the next pending boundary request is assigned to the freed slot. The scheduler preserves response identity, policy version, sampling settings, and L, and is used to reduce idle capacity during continuation generation. Its realized computation is included in the training-cost accounting described in Appendix G.

C.2 Instantiations with GRPO and DAPO

GRPO. For GRPO (Shao et al., 2024), the reassessed scores enter the native group-relative advantage computation. GRPO retains its token-level clipped surrogate, configured loss aggregation, and any supported actor-loss regularization.

DAPO. For DAPO (Yu et al., 2025), NCBR supplies reassessed scores to the same group-relative advantage computation and preserves the asymmetric token-level clipped objective and token-weighted aggregation. When DAPO's correctness-based dynamic sampling is enabled, reassessed correctness is also supplied to the existing filter.

Prefix-only importance ratios.

For a valid token of the original response, the token-level importance ratio is

$$\rho_{i,t}(\theta) := \frac{\pi_\theta(a_{i,t} \mid x, a_{i,<t})}{\pi_{\theta_{\text{old}}}(a_{i,t} \mid x, a_{i,<t})}, \qquad 1 \le t \le T_i. \tag{8}$$

C.3 DAPO-Specific Interactions

Preserving short-horizon auxiliary shaping.

When a base recipe uses auxiliary reward shaping, NCBR replaces only the verifier-dependent component of the training score while preserving shaping defined on the original response. We write

$$s_i^H = f(r_i^H) + \ell_i^H,$$

where f is the base recipe's verifier-reward encoding and ℓ_i^H denotes any shaping term computed from the original H-bounded response.

For the DAPO configuration used in our experiments, verifier correctness is encoded as

$$f(r) = 2r - 1,$$

and the length shaping is

$$\ell_i^H = \min\left\{-\frac{T_i - (H - C)}{C}, 0\right\},$$

where T_i is the original response length and C is the width of the length-penalty region. A strict boundary response is reassessed as

$$\widetilde{s}_i^L = f(r_i^L) + \ell_i^H.$$

Thus, continuation reassesses the verifier outcome without recomputing auxiliary shaping from the extended response. In particular, ℓ_i^H remains a function of the original response and is independent of the continuation length and the outcome-observation horizon L.

Optional post-replacement dynamic sampling.

When DAPO's correctness-based dynamic sampling is enabled, the correctness metadata is reassessed consistently with the boundary-return replacement.

$$\widetilde{r}_i^L := \begin{cases} r_i^L, & i \in \mathcal{B}_x, \\ r_i^H, & i \notin \mathcal{B}_x, \end{cases} \qquad K_x^L := \sum_{i=1}^{G} \widetilde{r}_i^L. \tag{9}$$

The DAPO recipe retains a group if and only if $0 < K_x^L < G$. Correctness determines group selection, whereas $\widetilde{s}_i^L$ determines advantage computation.

Continuation and score replacement precede correctness-based filtering, so the existing mixed-outcome criterion is applied to reassessed correctness. Additional candidate groups are sampled according to DAPO's existing batch-construction procedure until the required retained batch is obtained.

Additional effects on group selection.

Let $K_x^H := \sum_{i=1}^{G} r_i^H$ denote the short-horizon correctness count. An all-zero candidate group is *unlocked* when $K_x^H = 0$ but $0 < K_x^L < G$, so the existing DAPO filter retains a group that was homogeneous under short-horizon feedback. Conversely, reassessment can make a previously mixed group homogeneous or turn an all-correct group into a mixed group. Therefore, the retained groups under reassessed correctness need not form a superset of those retained under short-horizon correctness. Measured unlock rates are reported in Appendix B.1.

Group eligibility is determined by reassessed correctness, whereas advantages are computed from reassessed training scores over the retained group. Because the latter are group-normalized, score replacement can also change the advantages of non-boundary responses through the shared group statistics.

D Theoretical Analysis and Proofs

Appendices D.1–D.3 develop the three parts of Section 5.4: horizon-dependent answer feedback, relative prefix probes, and Monte Carlo boundary values. Appendix D.4 then establishes the selection effect of a sampled boundary return. Appendix D.5 derives its independent evaluation and comparison with short feedback, and Appendix D.6 examines auxiliary shaping. The corresponding empirical audit appears in Appendix E.

D.1 Horizon Feedback and Answer Position

A fixed-group comparison. Fix a prompt and a group of responses observed under the same frozen policy to horizon L. For this answer-position calculation, each response has a defined verifier-selected answer position $T_{\mathrm{ans},i}$. For $h \in \{H, L\}$, let y_i^h be its naturally terminated or h-truncated prefix, score it as $s_i^h = f(\mathcal{V}(x, y_i^h)) + \ell_i^h$, and compute A_i^h with Eq. (2). Both horizon comparisons use the same response group. For reweighting strength $\eta \geq 0$, define

$$w_i^h(\eta) = \frac{e^{\eta A_i^h}}{\sum_j e^{\eta A_j^h}}, \qquad \mathcal{T}_h(\eta) = \sum_i w_i^h(\eta) T_{\mathrm{ans},i}.$$

These weights solve the group-level KL-regularized problem given in Appendix D.4, with A_i^h as the score. They isolate the effect of changing feedback on a fixed set of trajectories. Here ℓ_i^h uses the corresponding horizon; NCBR instead preserves ℓ_i^H in Eq. (3).

Derivative and covariance. For any fixed statistic F_i,

$$\frac{dw_i^h}{d\eta} = w_i^h \left(A_i^h - \sum_j w_j^h A_j^h \right), \qquad \frac{d}{d\eta} \sum_i w_i^h F_i \bigg|_0 = \mathrm{Cov}_G(F, A^h),$$

where $\mathrm{Cov}_G(F, A) = G^{-1} \sum_i (F_i - \bar{F})(A_i - \bar{A})$. Subtracting the two horizon derivatives and taking $F_i = T_{\mathrm{ans},i}$ proves Eq. (5). For positive covariance, continuity gives $\mathcal{T}_L(\eta) > \mathcal{T}_H(\eta)$ for sufficiently small positive η. An equivalent pairwise expression is

$$\mathrm{Cov}_G(T_{\mathrm{ans}}, \Delta A) = \frac{1}{2G^2} \sum_{i,j} (T_{\mathrm{ans},i} - T_{\mathrm{ans},j})(\Delta A_i - \Delta A_j), \qquad \Delta A_i = A_i^L - A_i^H.$$

Thus a nondecreasing advantage change with answer position is a sufficient condition; strict increase for at least one pair makes the covariance positive. The sign depends on the advantage changes after within-group normalization.

A finite-update example. Consider unshaped binary feedback on a group whose responses each emit one correct final answer and terminate at that answer by L. Assume no correct answer is observable before that final answer. Let $n \in \{1, \ldots, G-1\}$ responses answer by H, and the remaining $G - n$ answer after H. Write their mean positions as $\bar{T}_{\mathrm{early}} \le H < \bar{T}_{\mathrm{late}}$. Long feedback is constant, so $w_i^L = 1/G$. Short feedback gives advantages $a_+ > 0$ to early responses and $a_- < 0$ to late responses. Their total short-feedback weight is

$$W_{\mathrm{late}}^H(\eta) = \frac{(G-n)e^{\eta a_-}}{ne^{\eta a_+} + (G-n)e^{\eta a_-}} < \frac{G-n}{G}, \qquad \eta > 0.$$

Consequently,

$$\mathcal{T}_L(\eta) - \mathcal{T}_H(\eta) = \left(\frac{G-n}{G} - W_{\mathrm{late}}^H(\eta) \right) (\bar{T}_{\mathrm{late}} - \bar{T}_{\mathrm{early}}) > 0. \tag{10}$$

Here extending the horizon removes a preference for early emission rather than giving an explicit reward for length. The next subsection separately relates answer emission to the relative prefix probe. Auxiliary shaping and correctness regressions are covered by the fixed-group covariance identity.

D.2 Earlier Answer Capability and Relative Prefix Probes

Probe model. Consider a fixed comparison population for which all offsets in $[0, D]$ are available, with the same prompt weights across horizons. For the reported $D = 512$ summary, this is the complete-window prompt set in Appendix B.2. Within a horizon, an original response is selected by drawing a prompt with these weights and then one of its two selected responses uniformly.

Let $t_{\mathrm{cap}} \in [0, T_{\mathrm{ans}}]$ be a modeled transition to higher answerability. Conditional on a selected response, the direct-answer probe succeeds with probability q_0 before the transition and with probability q_1 at or after it, where $0 \le q_0 < q_1 \le 1$. These probabilities are common across responses and horizons. Define

$$S_H := T_{\mathrm{ans}} - t_{\mathrm{cap}}, \tag{11}$$

the number of response tokens from the transition to final-answer emission for the H-trained policy.

Relative-position curve and continuous-window area. At offset δ, the prefix ends at $T_{\mathrm{ans}} - \delta$, and

$$T_{\mathrm{ans}} - \delta \ge t_{\mathrm{cap}} \quad \Longleftrightarrow \quad S_H \ge \delta.$$

Conditioning on this event gives

$$p_{\mathrm{FA}}^H(\delta) = q_0 + (q_1 - q_0) \Pr(S_H \ge \delta). \tag{12}$$

Define the continuous-window area as the average probe accuracy over $[0, D]$. Then

$$\begin{aligned} \mathrm{AUC}_{D,\mathrm{cont}}^H &:= \frac{1}{D} \int_0^D p_{\mathrm{FA}}^H(\delta)\, d\delta \\ &= q_0 + \frac{q_1 - q_0}{D} \mathbb{E}\left[\int_0^D \mathbf{1}\{S_H \ge \delta\}\, d\delta \right] \\ &= q_0 + \frac{q_1 - q_0}{D} \mathbb{E}[\min\{S_H, D\}]. \end{aligned} \tag{13}$$

The interchange follows from the bounded, nonnegative integrand. This proves Eq. (6). With shared q_0, q_1, a larger area corresponds to a larger expected interval of higher answerability before final-answer emission, truncated at D.

Discrete probe grid. For offsets $0 = \delta_1 < \cdots < \delta_J = D$, let b_j be the trapezoidal-integration weights:

$$b_1 = \frac{\delta_2 - \delta_1}{2}, \qquad b_J = \frac{\delta_J - \delta_{J-1}}{2}, \qquad b_j = \frac{\delta_{j+1} - \delta_{j-1}}{2} \quad (1 < j < J).$$

They sum to D. Defining the grid statistic by $D^{-1} \sum_j b_j p^H_{\mathrm{FA}}(\delta_j)$ gives

$$\mathrm{AUC}^H_{D,\mathrm{grid}} = q_0 + \frac{q_1 - q_0}{D} \mathbb{E}[K_D(S_H)], \tag{14}$$

where

$$K_D(s) := \sum_{j=1}^{J} b_j \mathbf{1}\{s \geq \delta_j\}. \tag{15}$$

The function K_D is the discrete-grid counterpart of $\min\{s, D\}$. The reported AUC_{512} first integrates each response's observed probe rates, then averages within prompts and across the fixed complete-window prompt set. Under the shared-q_0, q_1 probe model, Eqs. (13) and (14) interpret the relative prefix probe in terms of the interval between higher answerability and final-answer emission. This distinction motivates separating the outcome-observation horizon from the actor-update horizon. Appendix D.3 establishes how outcomes observed through natural continuation can guide updates to the original response.

D.3 Monte Carlo Boundary Values and Prefix Updates

1. Frozen continuation evaluation. Fix x, horizons $H < L$, and the rollout policy $\pi_{\theta_{\mathrm{old}}}$. Let P_x be the distribution of original H-bounded responses, and let $K_x(u \mid y^H)$ be the frozen natural-continuation law. For a naturally terminated response, u is empty and K_x is a point mass. For a strict boundary hit, K_x continues the exact original tokens until natural termination or total response length L. The joint rollout distribution is

$$P^L_x(y^H, u) = P_x(y^H) K_x(u \mid y^H).$$

Let $R_x(y^H, u) \in \{0, 1\}$ denote the reassessed verifier outcome. The continuation value defined in Section 5.4 is

$$v_x(y^H) = \sum_u K_x(u \mid y^H) R_x(y^H, u).$$

The finite vocabulary and bounded horizons give a finite response space. We assume differentiable policy probabilities on a common support and keep K_x fixed while varying the original-response policy.

2. Monte Carlo value gradient. Expanding the expected continuation value and applying the score-function identity (Schulman et al., 2015) yields

$$\begin{aligned} \nabla_\theta \mathbb{E}_{y^H \sim \pi_\theta(\cdot|x)}[v_x(y^H)] &= \sum_{y^H} \nabla_\theta \pi_\theta(y^H \mid x) \sum_u K_x(u \mid y^H) R_x(y^H, u) \\ &= \sum_{y^H, u} \pi_\theta(y^H \mid x) K_x(u \mid y^H) R_x(y^H, u) \nabla_\theta \log \pi_\theta(y^H \mid x). \end{aligned} \tag{16}$$

At $\theta = \theta_{\mathrm{old}}$, this is the expectation in Eq. (7) under the actual frozen rollout and continuation policy. For $y^H_i = (a_{i,1}, \ldots, a_{i,T_i})$,

$$\nabla_\theta \log \pi_\theta(y^H_i \mid x) = \sum_{t=1}^{T_i} \nabla_\theta \log \pi_\theta(a_{i,t} \mid x, a_{i,<t}).$$

The original response includes its natural termination token when present; external truncation introduces no sampled action. A continuation supplies the outcome sample for evaluating the original response, while the derivative acts on the original-response distribution.

3. Minimum KL change. Let $Q_x(y^H, u) = Q_x^H(y^H)Q_x(u \mid y^H)$ be any candidate joint distribution with finite KL divergence from P_x^L. Here Q_x^H is its original-response marginal. Define the distribution with that same marginal and frozen continuation law by

$$\overline{Q}_x(y^H, u) := Q_x^H(y^H)K_x(u \mid y^H).$$

For $\eta > 0$, consider the KL-regularized policy-space objective (Tomar et al., 2022)

$$\mathcal{F}_x(Q_x) := \mathbb{E}_{Q_x^H}[v_x(y^H)] - \eta^{-1} D_{\mathrm{KL}}(Q_x \| P_x^L). \tag{17}$$

Its value term evaluates original responses using the frozen K_x; the KL term measures the joint distribution change. Write $p_x := \mathbb{E}_{P_x}[v_x(y^H)]$ for the baseline value.

Proposition 1 (Prefix sufficiency under frozen continuation evaluation)**.** *For every candidate Q_x above,*

$$\mathcal{F}_x(\overline{Q}_x) - \mathcal{F}_x(Q_x) = \eta^{-1}\mathbb{E}_{Q_x^H}\left[D_{\mathrm{KL}}\big(Q_x(\cdot \mid y^H) \| K_x(\cdot \mid y^H)\big)\right] \geq 0. \tag{18}$$

Equality holds exactly when the two continuation laws agree for Q_x^H-almost every original response. An optimum retains K_x and has original-response marginal

$$Q_x^{H,*}(y^H) = \frac{P_x(y^H)e^{\eta v_x(y^H)}}{\mathbb{E}_{P_x}[e^{\eta v_x(y^H)}]}. \tag{19}$$

Its objective exceeds p_x whenever v_x is nonconstant under P_x.

Proof. Expand the joint likelihood ratio as

$$\log \frac{Q_x(y^H, u)}{P_x^L(y^H, u)} = \log \frac{Q_x^H(y^H)}{P_x(y^H)} + \log \frac{Q_x(u \mid y^H)}{K_x(u \mid y^H)}.$$

Taking expectations gives the KL chain rule:

$$\begin{aligned} D_{\mathrm{KL}}(Q_x \| P_x^L) = {} & D_{\mathrm{KL}}(Q_x^H \| P_x) \\ & + \mathbb{E}_{Q_x^H}\left[D_{\mathrm{KL}}\big(Q_x(\cdot \mid y^H) \| K_x(\cdot \mid y^H)\big)\right]. \end{aligned} \tag{20}$$

The conditional term is zero for $\overline{Q}_x$, and both candidates have the same value term. Subtraction proves Eq. (18), including the equality condition. To optimize the marginal, substitute Eq. (19) into the following identity:

$$\begin{aligned} \mathbb{E}_{Q_x^H}[v_x] - \eta^{-1} D_{\mathrm{KL}}(Q_x^H \| P_x) = {} & \eta^{-1} \log \mathbb{E}_{P_x}[e^{\eta v_x}] \\ & - \eta^{-1} D_{\mathrm{KL}}(Q_x^H \| Q_x^{H,*}). \end{aligned}$$

The maximum is therefore $\eta^{-1} \log \mathbb{E}_{P_x}[e^{\eta v_x}]$, attained at $Q_x^{H,*}$. Strict Jensen inequality makes this larger than p_x when v_x is nonconstant. If v_x is constant, the original distribution attains the maximum. □

The proposition also gives the comparison with a continuation-only change. If $Q_x^H = P_x$, the value term remains p_x and

$$\mathcal{F}_x(Q_x) = p_x - \eta^{-1}\mathbb{E}_{P_x} D_{\mathrm{KL}}\big(Q_x(\cdot \mid y^H) \| K_x(\cdot \mid y^H)\big) \leq p_x.$$

For any joint change, retaining its original-response marginal and restoring K_x preserves the value term at no larger KL cost. Thus the frozen-continuation objective admits a prefix-only optimum.

4. Sampled feedback and actor updates. The preceding statements use the continuation value v_x. NCBR observes sampled outcomes and passes their reassessed scores to the native group-relative loss. Appendix D.4 proves that one sampled binary return can favor original responses with higher continuation value, and Appendix D.5 derives its independent evaluation. The native objective retains its shaping, filtering, clipping, and length aggregation (Appendix C). A prefix-only loss restricts direct score terms to original-response tokens; shared parameters can still change future continuation probabilities. Proposition 1 concerns the specified frozen-continuation value objective.

D.4 Learning Signals from Boundary Returns

Appendix D.3 analyzes improvement using the continuation value v_x. We now show how one sampled return can favor higher-value original responses through group-relative response reweighting.

Sampling assumptions and reweighting. Fix x, H, L and the rollout/continuation policy. Original responses $Y_i = y_i^H$ are independent draws from P_x, with mean continuation value $p_x = \mathbb{E}_{P_x}[v_x(Y_i)]$. Draw one reassessed binary return $R_i^{(1)}$ per response, independently across responses conditional on the original group, with conditional mean $v_x(Y_i)$. Naturally terminated responses retain their deterministic verifier outcomes. Assume $G \geq 2$, no auxiliary shaping, and no outcome-based group selection. Construct $\widetilde{A}_i^{L,(1)}$ from the binary returns by Eq. (2), using zero advantages for constant-return groups. For $\eta \geq 0$, define

$$w_i(\eta) := \frac{e^{\eta \widetilde{A}_i^{L,(1)}}}{\sum_j e^{\eta \widetilde{A}_j^{L,(1)}}}, \qquad \mathcal{J}_x(\eta) := \mathbb{E}\left[\sum_i w_i(\eta) v_x(Y_i)\right]. \tag{21}$$

The weights act on response occurrences within each group, and $\mathcal{J}_x(0) = p_x$. For a realized group and $\eta > 0$, these weights solve

$$\max_{\boldsymbol{w} \in \Delta_G} \left\{ \sum_i w_i \widetilde{A}_i^{L,(1)} - \eta^{-1} D_{\mathrm{KL}}(\boldsymbol{w} \| \boldsymbol{u}) \right\}, \qquad u_i = 1/G,$$

where Δ_G is the probability simplex over the G responses. A Lagrange multiplier for $\sum_i w_i = 1$ gives the exponential weights. The parameter η controls this analysis, not the native actor learning rate.

Proposition 2 (Idealized policy improvement)**.** *Under the preceding assumptions, for* $0 < p_x < 1$ *and every* $\eta > 0$ *there is a coefficient* $\beta_G(p_x, \eta) > 0$ *such that*

$$\mathcal{J}_x(\eta) - p_x = \beta_G(p_x, \eta) \operatorname{Var}_{P_x}[v_x(Y_i)] \geq 0. \tag{22}$$

Expected weights and proof. Suppress x and write $p = \mathbb{E}_P[v]$. Integrating out the independent original responses gives independent first returns with distribution $\mathrm{Bernoulli}(p)$. For $a \in \{0, 1\}$, let $b_a = \mathbb{E}[w_i(\eta) \mid R_i^{(1)} = a]$. Conditional on $R_i^{(1)}$, the other responses and returns are independent of Y_i, so

$$\mathbb{E}[w_i(\eta) \mid Y_i = y] = b_0 + (b_1 - b_0) v(y).$$

Exchangeability and $\sum_i w_i = 1$ give $pb_1 + (1-p)b_0 = 1/G$. Substitution yields the key expected-weight identity:

$$\mathbb{E}[w_i(\eta) \mid Y_i = y] = \frac{1}{G} + (b_1 - b_0)(v(y) - p). \tag{23}$$

In a mixed group, each successful response receives weight above $1/G$ and each unsuccessful response receives weight below $1/G$; constant groups have uniform weights. Mixed groups occur with positive probability, giving $b_1 > 1/G > b_0$. Therefore,

$$\begin{aligned} \mathcal{J}_x(\eta) &= G\, \mathbb{E}_P[v(Y) \mathbb{E}[w_i(\eta) \mid Y_i = Y]] \\ &= p + G(b_1 - b_0)\{\mathbb{E}_P[v(Y)^2] - p^2\}. \end{aligned}$$

This proves Eq. (22) with $\beta_G(p, \eta) = G(b_1 - b_0) > 0$. If v is constant, the gain vanishes. This includes $p = 0$ and $p = 1$.

Explicit finite-group coefficients. Under population-standard-deviation normalization, let $K \sim \mathrm{Binomial}(G-1, p)$ and $n = K + a$. Set $W_a(K) = 1/G$ when $n \in \{0, G\}$. Otherwise define

$$a_+(n) = \sqrt{\frac{G-n}{n}}, \qquad a_-(n) = -\sqrt{\frac{n}{G-n}},$$

$$W_a(K) = \frac{\exp\{\eta a_a(n)\}}{n \exp\{\eta a_+(n)\} + (G-n)\exp\{\eta a_-(n)\}},$$

where $a_1 = a_+$ and $a_0 = a_-$. Then $b_a = \mathbb{E}[W_a(K)]$. Other positive group-standard-deviation denominators change these coefficients but preserve the weight ordering and gain identity.

D.5 Independent Outcomes and Comparison with Short Feedback

An independent evaluation of the first return. Use the assumptions of Appendix D.4 and abbreviate $A_i = \widetilde{A}_i^{L,(1)}$, $p = p_x$, and $v = v_x$. Let $\mathcal{F}_1$ contain the complete original response group and its first-realization returns. Draw a second reassessed binary return vector $\{R_i^{(2)}\}_{i=1}^G$ conditionally independently of the first, so $\mathbb{E}[R_i^{(2)} \mid \mathcal{F}_1] = v(Y_i)$. Centered advantages give $w_i'(0) = A_i/G$, and therefore

$$\mathcal{J}_x'(0) = \mathbb{E}\left[\frac{1}{G}\sum_i A_i v(Y_i)\right] = \mathbb{E}\left[\frac{1}{G}\sum_i A_i R_i^{(2)}\right].$$

For $0 < p < 1$, write $m_a = \mathbb{E}[A_i \mid R_i^{(1)} = a]$. Since $pm_1 + (1-p)m_0 = 0$,

$$\mathbb{E}[A_i \mid Y_i = y] = c_G(p)(v(y) - p), \qquad c_G(p) = m_1 - m_0.$$

Let $N \sim \mathrm{Binomial}(G, p)$, $q_n = n/G$, and d_n be the group advantage denominator for a group with n successes. Constant groups contribute zero. For a realized mixed group,

$$\frac{1}{G}\sum_i A_i R_i^{(1)} = \frac{q_N(1-q_N)}{d_N}.$$

Taking expectations, the left-hand side is $pm_1 = p(1-p)c_G(p)$, which yields

$$c_G(p) = \frac{\mathbb{E}_N[q_N(1-q_N)/d_N]}{p(1-p)} > 0.$$

Consequently,

$$\mathcal{J}_x'(0) = \mathbb{E}\left[\frac{1}{G}\sum_i A_i R_i^{(2)}\right] = c_G(p)\,\mathrm{Var}_P[v(Y)] \geq 0. \tag{24}$$

For population standard deviation, $d_n = \sqrt{q_n(1-q_n)}$, giving $c_G(p) = \mathbb{E}_N[\sqrt{N(G-N)}/G]/\{p(1-p)\}$. For the audit's sample standard deviation with stabilizer ϵ, use $d_n = \sqrt{Gq_n(1-q_n)/(G-1)} + \epsilon$. Without a stabilizer, sample-standard-deviation normalization scales the population coefficient by $\sqrt{(G-1)/G}$.

Contrast with short feedback. Define $\mathcal{J}_x^H$ by replacing A_i in the reweighting with A_i^H, while retaining $v(Y_i)$ as the evaluation quantity. More generally, for any centered first-realization advantages, conditional independence gives

$$\frac{d}{d\eta}(\mathcal{J}_x - \mathcal{J}_x^H)\Big|_0 = \mathbb{E}\left[\frac{1}{G}\sum_i (\widetilde{A}_i^{L,(1)} - A_i^H) R_i^{(2)}\right]. \tag{25}$$

This is the population quantity estimated by $\widehat{\Delta C}$ in Appendix E.

For unshaped binary short feedback, set $p_H = \mathbb{E}_P[r^H(Y)]$. When $0 < p_H < 1$, the same conditional-weight argument gives

$$\mathcal{J}_x^H(\eta) - p = \beta_G(p_H, \eta)\,\mathrm{Cov}_P(v(Y), r^H(Y)),$$
$$(\mathcal{J}_x^H)'(0) = c_G(p_H)\,\mathrm{Cov}_P(v(Y), r^H(Y)).$$

Thus the first-order increment is

$$c_G(p)\,\mathrm{Var}_P[v(Y)] - c_G(p_H)\,\mathrm{Cov}_P(v(Y), r^H(Y)).$$

A constant short reward gives zero short-feedback gain. In general the increment depends on both the continuation-value variation and the information already captured by short feedback, so the nonnegative absolute gain does not by itself order the two signals. Equation (25) also applies to shaped advantages; the audit measures its sign under the actual score construction.

D.6 AUXILIARY SHAPING

Auxiliary shaping changes the relation between correctness and relative advantage. The following independent-sampling example isolates this effect. Let $G = 2$ and draw each original response independently from two types with equal probability. Type a terminates incorrectly without a length penalty, so $v(a) = 0$ and its score is -1. Type b is a strict boundary hit with $v(b) = q \in (0, 1/2)$ and penalty -1. Under the audited encoding $2R - 1 + \ell^H$, its score is 0 on continuation success and -2 on failure.

With population-standard-deviation normalization, mixed-type groups have advantages $+1$ and -1. Let $u = 1/(1 + e^{-2\eta}) > 1/2$. Type b receives expected weight $qu + (1 - q)(1 - u)$. Relative to uniform weighting, the conditional value change in a mixed-type group is

$$q\left\{qu + (1-q)(1-u) - \frac{1}{2}\right\} = q(2q-1)\left(u - \frac{1}{2}\right) < 0.$$

Mixed-type groups occur with probability $1/2$, while homogeneous-type groups have constant continuation value and zero gain. The unconditional gain is therefore also negative. This preserves independent response sampling and isolates auxiliary shaping as the changed assumption. Appendix E evaluates the resulting signal under the recipe-faithful scores.

E INDEPENDENT AUDIT OF BOUNDARY-RETURN SIGNALS

We test whether boundary reassessment assigns higher advantages to responses that succeed under an independent continuation. The audit compares the original short-feedback advantage with the reassessed advantage using the same held-out outcome, following Appendix D.5.

E.1 DATA AND INDEPENDENT NATURAL CONTINUATIONS

We reuse the DAPO baseline responses and first continuations from Appendix B.1 at Steps 100 and 300. Each checkpoint supplies $N = 17{,}383$ pre-filter candidate prompt groups, with $G = 8$ responses per group and original horizon $H = 2048$. Each strict boundary hit receives two conditionally independent natural continuations under the same frozen policy to $L = 8192$. The first return $R_i^{(1)}$ constructs $\widetilde{A}_i^{L,(1)}$, and the second return $R_i^{(2)}$ provides the independent evaluation outcome. Naturally terminated responses retain their original outcomes, $R_i^{(1)} = R_i^{(2)} = r_i^H$. Table 9 summarizes coverage.

Table 9: Independent-audit coverage at Steps 100 and 300.

Quantity	Step 100	Step 300
Complete candidate prompt groups	17,383	17,383
Groups containing a strict boundary hit	567	579
Boundary responses: $r^H = 0$ / $r^H = 1$	821 / 3	733 / 5
Boundary responses with both continuation outcomes	824	738

E.2 SCORE CONSTRUCTION AND STATISTICAL ESTIMANDS

The binary construction uses r_i^H and $R_i^{(1)}$ as short-feedback and reassessed scores. The recipe-faithful construction uses the DAPO scores

$$s_i^H = f(r_i^H) + \ell_i^H, \qquad \widetilde{s}_i^{L,(1)} = f(R_i^{(1)}) + \ell_i^H,$$

with the reward encoding f and original-response shaping ℓ_i^H from Appendix C.3. Advantages use the full original group, the training recipe's sample-standard-deviation normalization, and zero advantages for constant-score groups. Both score constructions are evaluated against the same binary held-out outcomes.

Let $\mathcal{X}$ be the candidate prompt-group set. We compute

$$\begin{aligned}\widehat{C}_H &= \frac{1}{NG}\sum_{x\in\mathcal{X}}\sum_{i=1}^{G} A^H_{x,i} R^{(2)}_{x,i},\\ \widehat{C}_{\mathrm{NCBR}} &= \frac{1}{NG}\sum_{x\in\mathcal{X}}\sum_{i=1}^{G} \widetilde{A}^{L,(1)}_{x,i} R^{(2)}_{x,i},\\ \widehat{\Delta C} &= \widehat{C}_{\mathrm{NCBR}} - \widehat{C}_H.\end{aligned} \tag{26}$$

The two absolute signals average the product of each advantage and its independent outcome. A larger value indicates greater advantage assigned to independently successful responses. Their difference estimates the response-reweighting contrast in Eq. (25).

To separate the frequency of boundary reassessment from its effect within affected groups, let $\mathcal{X}_B \subseteq \mathcal{X}$ contain the groups with at least one strict boundary hit and let $N_B = |\mathcal{X}_B|$. Define

$$\widehat{\Delta C}_{\mathrm{boundary}} = \frac{1}{N_B G}\sum_{x\in\mathcal{X}_B}\sum_{i=1}^{G}(\widetilde{A}^{L,(1)}_{x,i} - A^H_{x,i}) R^{(2)}_{x,i}.$$

Groups without boundary hits have unchanged advantages, so

$$\widehat{\Delta C} = \frac{N_B}{N}\widehat{\Delta C}_{\mathrm{boundary}}.$$

We report 95% confidence intervals from 10,000 paired bootstrap resamples at the prompt-group level.

E.3 Results and Interpretation

Table 10: Independent selection signals over all 17,383 candidate groups per checkpoint. Binary uses verifier correctness; Recipe uses the DAPO scores $f(r) + \ell^H$. Increments are scaled by 10^3; brackets give paired prompt-group bootstrap 95% confidence intervals.

Step	Score	$\widehat{C}_H$	$\widehat{C}_{\mathrm{NCBR}}$	$10^3\widehat{\Delta C}$
100	Binary	0.19009	0.19058	0.499 [0.330, 0.681]
100	Recipe	0.18854	0.18948	0.945 [0.702, 1.212]
300	Binary	0.15923	0.16063	1.405 [1.140, 1.681]
300	Recipe	0.15670	0.15909	2.389 [1.964, 2.847]

Both checkpoints and score constructions yield positive increments with 95% confidence intervals above zero (Table 10). At Step 300, the recipe-faithful signal increases from 0.15670 to 0.15909, with increment 2.389×10^{-3}. Thus, the reassessed advantages are more strongly aligned with independent outcomes than the short-feedback advantages.

Boundary-containing groups account for 3.26% and 3.33% of the candidate groups at Steps 100 and 300. Table 11 reports increments on these groups; the other groups contribute zero to the aggregate increment.

Table 11: Boundary-conditional independent selection-signal increments. Values are scaled by 10^3; brackets give paired prompt-group bootstrap 95% confidence intervals.

Step	Boundary groups	Binary	Recipe
100	567 (3.26%)	15.286 [10.243, 20.673]	28.985 [21.854, 36.744]
300	579 (3.33%)	42.184 [34.856, 49.718]	71.719 [60.086, 84.044]

The boundary-conditional increments are positive under both score constructions at both checkpoints. At Step 300, the recipe-faithful increment is 0.071719 within boundary-containing groups. Together, the aggregate and conditional results show that boundary reassessment improves advantage alignment with held-out continuation outcomes, with the added signal concentrated in the groups whose outcomes are reassessed.

F Supplementary Evaluation Results

F.1 Vanilla DAPO Horizon Study

Table 12 reports Step-300 performance for vanilla DAPO when the outcome-observation and actor-update horizons are increased together. All checkpoints are evaluated under the common inference-time response budget $H_{\text{eval}} = 8192$.

Table 12: Vanilla DAPO horizon-study results on Qwen3-4B. Outcome-observation and actor-update horizons are both set to H. Evaluation uses 32 independently sampled responses per problem with $H_{\text{eval}} = 8192$. Entries report Pass@1 / Pass@32 (%).

Step	H	AIME'25	AIME'24	AMC'23
300	2048	11.46 / 36.67	10.52 / 33.33	61.64 / 85.00
300	4096	16.25 / 30.00	15.83 / 40.00	56.25 / 82.50
300	8192	11.35 / 26.67	9.90 / 33.33	58.28 / 87.50

Under the common evaluation budget, Step-300 performance is not monotonic in the training rollout horizon. Among the compared settings, $H = 4096$ achieves the highest Pass@1 on both AIME benchmarks, while $H = 2048$ achieves the highest Pass@1 on AMC'23.

F.2 Outcome-Observation and Actor-Update Horizons

Table 13 reproduces the reported Step-300 DAPO+NCBR horizon comparisons in Table 3, evaluated with $H_{\text{eval}} = 8192$. At fixed $H = 2048$, reducing L from 8192 to 4096 lowers accounted cost from 614.38 to 420.03 GPU-hours (31.6%), with less than 1% change in cumulative actor-update tokens. Pass@1 is unchanged on AIME'25 and lower by about 1.1 percentage points on AIME'24 and AMC'23. Pass@32 is not uniformly preserved: it increases on AIME'25 but decreases on AIME'24 and AMC'23. Nevertheless, both metrics remain above the short DAPO baseline on all three benchmarks.

At fixed outcome-observation horizon $L = 4096$, reducing the actor-update horizon H from 2048 to 1024 lowers cumulative actor-update tokens from 0.767B to 0.362B (52.8%) and accounted training cost from 420.03 to 199.54 GPU-hours (52.5%). Pass@32 decreases on all three benchmarks, while Pass@1 decreases on AIME'25 and AIME'24 and remains unchanged on AMC'23. Together with the fixed-H comparison, these Step-300 results illustrate the cost–performance trade-offs associated with independently varying the outcome-observation and actor-update horizons. These comparisons use a fixed number of training steps and do not establish superiority under equal compute budgets.

Table 13: DAPO+NCBR horizon results. Benchmark entries are Pass@1 / Pass@32 (%) under $H_{\text{eval}} = 8192$. Step-300 values reproduce Table 3.

Step	H	L	AIME'25	AIME'24	AMC'23
300	2048	8192	14.48 / 40.00	16.56 / 46.67	69.84 / 95.00
300	2048	4096	14.48 / 43.33	15.42 / 36.67	68.75 / 92.50
300	1024	4096	10.94 / 30.00	14.90 / 33.33	68.75 / 87.50

G Training Cost

Table 14 summarizes training cost and actor-update tokens. Accounted GPU-hours are computed from recorded training-step and periodic online-validation time, multiplied by the allocated GPU count. Checkpointing is included in the step timer. Actor-update tokens count valid original response tokens used in the actor objective; continuation tokens are rollout tokens and are not included in this count.

Table 14: Training costs for Qwen3-4B through Step 300. H and L denote the actor-update and outcome-observation horizons. Relative cost is normalized to each optimizer's vanilla baseline with $H = 2048$. B denotes billion tokens.

Method	H	L	Accounted GPU-hours	Relative cost	Actor-update tokens (B)
DAPO	2048	2048	314.62	1.000×	0.788
DAPO	4096	4096	603.89	1.919×	1.311
DAPO	8192	8192	1,350.07	4.291×	2.058
DAPO + NCBR	2048	8192	614.38	1.953×	0.774
DAPO + NCBR	2048	4096	420.03	1.335×	0.767
DAPO + NCBR	1024	4096	199.54	0.634×	0.362
GRPO	2048	2048	162.40	1.000×	0.628
GRPO + NCBR	2048	8192	346.56	2.134×	0.828

H Training Dynamics

Figures 6 and 7 show the training trajectories for the vanilla DAPO horizon study and the baseline/NCBR optimizer comparisons, respectively.

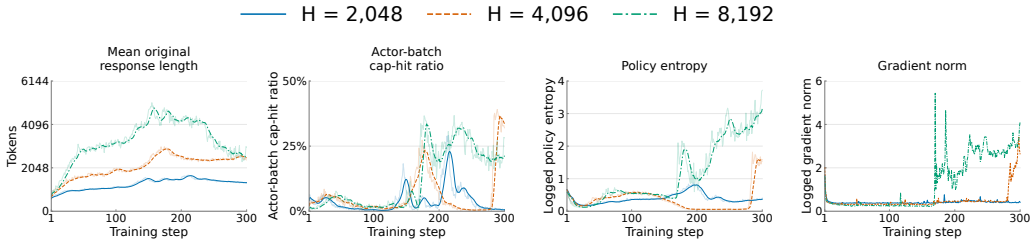


Figure 6: Vanilla DAPO training on Qwen3-4B with matched outcome-observation and actor-update horizons $H \in \{2048, 4096, 8192\}$.

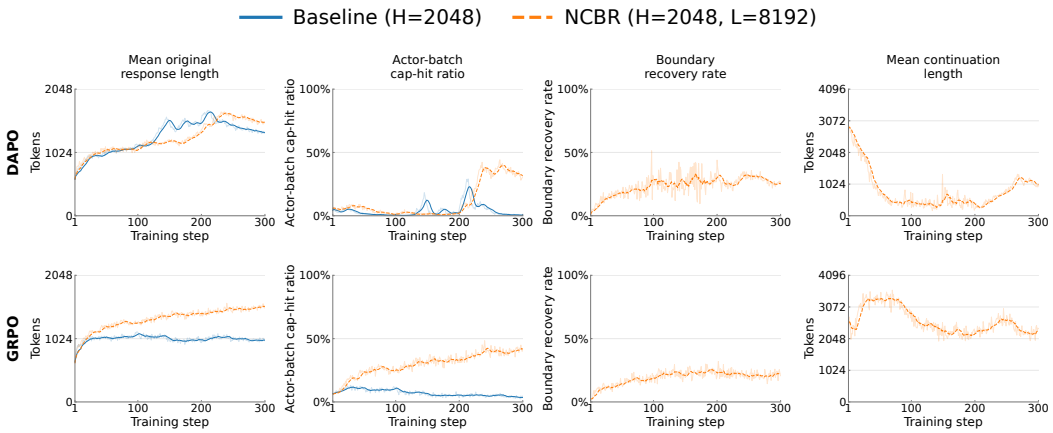


Figure 7: **Training dynamics with and without NCBR.** Qwen3-4B trained with DAPO and GRPO. Both baselines and NCBR use actor-update horizon $H = 2048$; NCBR uses outcome-observation horizon $L = 8192$. From left to right: mean original response length; fraction of actor-batch responses reaching the rollout boundary; fraction of verifier-incorrect boundary responses that become correct after natural continuation; and mean number of continuation tokens generated beyond the original response. Blue solid curves denote baselines, and orange dashed curves denote NCBR. The last two columns report NCBR only. Faint curves show per-step values, while darker curves show smoothed trends.