

PRIOR LOCKING IN RLVR: PRIOR-CORRECTED OPTIMIZATION FOR BROADER REASONING COVERAGE

Anonymous authors

Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has become a powerful paradigm for improving reasoning in large language models (LLMs). However, increasing aggregate success probability does not necessarily imply broad coverage over distinct correct reasoning modes. We show that, under the common binary-reward setting, even idealized KL-regularized policy improvement can exhibit *prior locking*, whereby the allocation among correct reasoning modes remains tied to initialization and limits coverage. To overcome this limitation, we propose *Prior-Corrected RLVR* (PC-RLVR), which dynamically rebalances inherited probability allocation; its oracle dynamics converge to the coverage-optimal uniform allocation over correct modes. To bridge the gap between the mode-level oracle and practical LLM training, we develop a historical-reference-based response-level prior correction with bounded weights, yielding a more stable practical realization under limited rollout budgets. We further derive a finite-time stationarity bound separating structural approximation residuals from controllable optimization and historical-reference estimation errors, with the controllable terms decaying as $O(T^{-2/5})$. Experiments on mathematical reasoning benchmarks show improved Avg@64 and Pass@64, together with stronger log-likelihood gains for initially lower-prior correct responses, providing complementary evidence consistent with broader coverage of successful responses.

1 INTRODUCTION

Recent advances in mathematical and code reasoning have made reinforcement learning with verifiable rewards (RLVR) a central paradigm for post-training large language models (LLMs), where solution quality can be automatically verified (Shao et al., 2024; Wen et al., 2026; Jin et al., 2026). Methods based on Group Relative Policy Optimization (GRPO) and its variants have delivered substantial gains in reasoning performance (Guo et al., 2025; Yu et al., 2025), yet improvements in Pass@1 do not always translate into comparable gains at larger sampling budgets (Karan & Du, 2026; Li et al., 2026a). This distinction matters because many reasoning problems admit multiple valid solution strategies: improving aggregate success probability does not necessarily imply broad coverage of successful reasoning. We abstract these strategies as latent reasoning modes and study how RLVR allocates probability among correct modes under binary feedback. This allocation determines whether probability mass is concentrated on a few correct modes or distributed more broadly across them.

Empirical studies have reported reduced output diversity or weaker large-budget sampling gains during RLVR, motivating methods based on entropy- or exploration-preserving mechanisms, modified policy anchors, and probability-aware reweighting (Yu et al., 2025; Liang et al., 2026; Cui et al., 2025; Lee et al., 2026; He et al., 2025; Li et al., 2026b; Plyusov et al., 2026). Recent theoretical analyses further highlight the distinction between aggregate success amplification and probability allocation among correct modes. Mroueh (2026) characterizes the amplification of aggregate success under binary-reward RLVR, while *Rate or Fate?* (Rad et al., 2026) shows that competition among correct modes can drive concentration. These findings suggest that binary-reward RLVR can substantially reshape the output distribution, but leave a more specific issue unresolved: how the resulting allocation among correct modes depends on the initial policy, whether this dependence itself limits coverage, and, if so, how it can be broken.

In this work, we show that this dependence can be exact. Under idealized KL-regularized policy improvement, the relative allocation among correct modes remains tied to initialization even as their aggregate probability increases, a phenomenon we call *prior locking*. Thus, an initially imbalanced allocation can persist throughout optimization and limit mode-level coverage. To break this dependence, we introduce *Prior-Corrected RLVR* (PC-RLVR), which dynamically modulates mode-wise update strengths according to the evolving allocation. In the oracle setting, we prove that the prior-corrected dynamics break prior locking and converge to the coverage-optimal uniform allocation over correct modes.

The oracle formulation relies on latent mode information that is unavailable in practical LLM training. We therefore develop a historical-reference-based response-level prior correction with bounded weights as the practical realization of PC-RLVR. Built from historical rollouts, this reference yields a more stable response-level correction under limited rollout budgets without increasing the current sampling cost. We further derive a finite-time stationarity bound separating structural approximation residuals from controllable optimization and historical-reference estimation errors, with the controllable terms decaying as $O(T^{-2/5})$.

We evaluate PC-RLVR on mathematical reasoning benchmarks against GRPO and closely related RLVR methods. PC-RLVR improves both mean Avg@64 and mean Pass@64, with stronger improvements in Pass@64. Moreover, initially lower-prior correct responses receive greater per-token log-likelihood gains under PC-RLVR relative to GRPO. Together, these results provide complementary evidence consistent with broader coverage of successful responses.

Our main contributions are:

- We characterize *prior locking* under idealized KL-regularized policy improvement, showing that the relative allocation among correct modes remains tied to initialization, while prior-corrected oracle dynamics converge to the coverage-optimal uniform allocation.
- We propose PC-RLVR, a historical-reference-based response-level realization of prior correction, and derive a finite-time stationarity bound separating structural approximation residuals from controllable optimization and historical-reference estimation errors.
- Experiments on mathematical benchmarks show stronger gains under broader sampling and prior-dependent log-likelihood gains, consistent with broader coverage of successful responses.

2 RELATED WORK

RLVR dynamics and probability allocation. Recent work has begun to characterize the optimization dynamics underlying RLVR. Mroueh (2026) derives KL-regularized binary-reward RLVR dynamics and characterizes the amplification of aggregate success probability. *Rate or Fate?* (Rad et al., 2026) studies GRPO-style softmax policy-gradient dynamics and identifies winner-take-all behavior arising from competition among correct modes. Our work complements these analyses by identifying a distinct form of initialization dependence: under KL-regularized binary-reward policy improvement, the relative allocation among correct modes can remain tied to the initial policy, which we term *prior locking*. The idealized update underlying our analysis is closely related to classical relative-entropy-regularized and mirror-descent policy improvement (Abdolmaleki et al., 2018; Geist et al., 2019; Tomar et al., 2022).

Coverage-oriented and probability-aware RLVR. Recent work has explored several ways to mitigate probability concentration and improve reasoning coverage in RLVR. One line reshapes the learning signal using response likelihood, model confidence, group success statistics, or reward sign (He et al., 2025; Li et al., 2026b; Plyusov et al., 2026; Zhu et al., 2025). Another line increases rollout budgets to expose the policy to a wider range of candidate behaviors (Hou et al., 2025; Hu et al., 2026). A complementary direction modifies the reference distribution used for KL regularization: either by periodically refreshing the reference policy or by reshaping the anchor toward underexplored reasoning regions (Liu et al., 2025; Lee et al., 2026). PC-RLVR instead introduces response-level prior correction to counteract probability imbalance inherited from the initial policy.

Optimization refinements in GRPO-style RLVR. Recent studies have revisited the optimization mechanics of GRPO through asymmetric clipping, response-level optimization, constrained objec-

tives, or importance-ratio correction (Yu et al., 2025; Zheng et al., 2025; Li et al., 2025; Wang et al., 2025a). These methods primarily refine the policy optimization procedure, whereas PC-RLVR focuses on how binary feedback shapes the allocation among correct modes.

3 PRIOR LOCKING AND PRIOR CORRECTION

3.1 PRIOR LOCKING UNDER BINARY FEEDBACK

We begin with an idealized mode-level view that isolates the allocation induced by binary feedback from policy parameterization and sampling noise. Fix a problem x and suppress dependence on x in the notation. Let $\mathcal{M} = \{1, \dots, M\}$ denote latent reasoning modes and $C : \mathcal{Y} \rightarrow \mathcal{M}$ map responses to modes. Each mode has a binary reward $r_m \in \{0, 1\}$, and we denote the set of correct modes by $\mathcal{C}^+ := \{m : r_m = 1\}$, with $K := |\mathcal{C}^+| \geq 1$. For policy π_θ , let $q_\theta(m) := \Pr_{Y \sim \pi_\theta(\cdot|x)}(C(Y) = m)$ and $q_t := q_{\theta_t}$. Under binary feedback, all correct modes are equally valued; accordingly, we take the uniform allocation $q^*(m) = K^{-1} \mathbf{1}\{m \in \mathcal{C}^+\}$ as the coverage-optimal target. At the mode level, we consider the following KL-regularized policy improvement step:

$$q_{t+1} = \arg \max_{q \in \Delta(\mathcal{M})} \mathbb{E}_{m \sim q}[r_m] - \frac{1}{\eta} \text{KL}(q \| q_t), \quad \eta > 0. \quad (1)$$

However, under binary feedback, this update can increase the probability of the correct set without changing the relative allocation among correct modes, a phenomenon we call *prior locking*.

Theorem 3.1 (Prior locking). *More generally, replace the reward signal r_m in equation 1 at iteration t by a_t for $m \in \mathcal{C}^+$ and b_t otherwise. Suppose $q_0(\mathcal{C}^+) > 0$ and $a_t > b_t$ for all t . Then*

$$q_T(m | \mathcal{C}^+) = q_0(m | \mathcal{C}^+), \quad m \in \mathcal{C}^+, \quad T \geq 0. \quad (2)$$

Additionally, if $\sum_{t=0}^{\infty} (a_t - b_t) = \infty$, then $q_T(\mathcal{M} \setminus \mathcal{C}^+) \rightarrow 0$.

Theorem 3.1 shows that the conditional distribution over correct modes remains fixed at initialization, even as probability shifts toward the correct set. For binary rewards, $a_t = 1$ and $b_t = 0$; hence, unless $q_0(\cdot | \mathcal{C}^+)$ is already uniform, the dynamics cannot reach q^* . This raises the question of how to break this initialization dependence and reshape the allocation among correct modes.

A natural approach is to flatten the policy distribution through entropy regularization. The following result shows how entropy changes the relative allocation among correct modes.

Proposition 3.2 (Entropy reshapes mode allocation). *Suppose $q_0(m) > 0$ for all $m \in \mathcal{M}$, and let $\kappa_H := (1 + \eta \lambda_H)^{-1} \in (0, 1)$. For the entropy-regularized counterpart of equation 1, if the learning signal takes a common value on all correct modes and another common value on all incorrect modes at every iteration, then any two modes m, m' with the same reward satisfy*

$$\log \frac{q_T^H(m)}{q_T^H(m')} = \kappa_H^T \log \frac{q_0(m)}{q_0(m')}. \quad (3)$$

Since $\kappa_H^T \rightarrow 0$, the probability ratio between any two modes with the same reward converges to one rather than remaining fixed at its initial value. Entropy therefore breaks prior locking. However, this rebalancing is not reward-selective: the same flattening effect also acts on incorrect modes and can retain nonzero asymptotic mass on them; see Proposition B.1 in Appendix. As a result, entropy alone does not guarantee convergence to q^* . This motivates a correction mechanism that dynamically modulates mode-wise update strengths according to both reward information and current allocation, with the goal of reaching the coverage-optimal allocation.

3.2 PRIOR-CORRECTED OPTIMIZATION

To overcome prior locking while avoiding the limitations of entropy regularization, we introduce a prior-corrected KL-regularized policy improvement step. Let $\phi : (0, 1] \rightarrow (0, \infty)$ be positive, continuous, and strictly decreasing. For time-varying reward signals a_t and b_t assigned to correct and incorrect modes, respectively, the oracle prior-corrected policy improvement step is

$$\bar{q}_{t+1} = \arg \max_{p \in \Delta(\mathcal{M})} a_t \sum_{m \in \mathcal{C}^+} p(m) \phi(\bar{q}_t(m)) + b_t \sum_{m \notin \mathcal{C}^+} p(m) \phi(\bar{q}_t(m)) - \frac{1}{\eta} \text{KL}(p \| \bar{q}_t), \quad \bar{q}_0 = q_0. \quad (4)$$

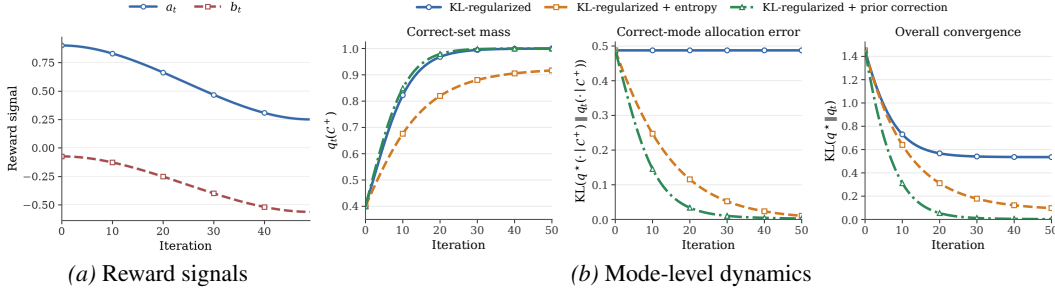


Figure 1: Controlled mode-level updates under time-varying reward signals. (a) Correct- and incorrect-mode signals a_t and b_t . (b) Evolution of correct-set probability mass, correct-mode allocation error, and distance to q^* under standard KL-regularized policy improvement, entropy regularization, and prior correction.

Equation 4 couples reward feedback with the current allocation multiplicatively. The reward signal determines the direction of the update, while $\phi(\bar{q}_t(m))$ modulates its strength according to the current allocation. For correct modes, since ϕ is decreasing, the positive reinforcement weakens as a mode accumulates probability, allowing the correct-mode allocation to progressively rebalance. For incorrect modes, $b_t\phi(\bar{q}_t(m))$ remains non-positive under the conditions below. Since ϕ is positive, it modulates only the magnitude of this non-positive signal without reversing its direction. Together, the update rebalances the allocation within the correct set while shifting probability mass away from incorrect modes. Under the conditions below, the resulting iterates converge to the coverage-optimal allocation q^* , as formalized in the following theorem.

Theorem 3.3 (Convergence of mode-level prior correction). *Suppose ϕ is continuously differentiable with $\phi'(u) < 0$ and $q_0(m) > 0$ for all $m \in \mathcal{C}^+$. Let $\rho_0 := \min_{m \in \mathcal{C}^+} q_0(m)$ and $L_\phi := \max_{u \in [\rho_0, 1]} \{-u\phi'(u)\}$. If $0 < a_t \leq a_{\max}$, $b_t \leq 0$, $\sum_{t=0}^{\infty} a_t = \infty$, and $0 < \eta < 1/(a_{\max}L_\phi)$, then $\bar{q}_t \rightarrow q^*$.*

The theorem includes binary feedback as the special case $a_t = 1$ and $b_t = 0$, showing that the prior-corrected dynamics break the initialization dependence identified in Theorem 3.1 and converge to the coverage-optimal allocation.

Remark 3.4 (Allocation under heterogeneous rewards). If modes with positive rewards carry heterogeneous scores r_m , an interior optimum $q^\dagger$ satisfies $r_m\phi(q^\dagger(m)) = \lambda$ for every active positive-reward mode. Hence, the correction does not enforce uniformity under heterogeneous scores: modes with larger rewards receive more probability mass, whereas modes with equal positive rewards receive equal allocation. Uniform coverage is recovered as the special case in which all positive-reward modes are equally valued.

To illustrate the distinction between success amplification and correct-mode allocation, we consider a deterministic problem with 20 modes, four of which are correct. We initialize the correct-mode conditional allocation to be highly nonuniform. We apply time-varying reward signals and compare the exact mode-level updates induced by standard KL-regularized policy improvement, entropy regularization, and prior correction.

As shown in Figure 1, standard KL-regularized policy improvement rapidly increases the total probability assigned to correct modes, but its internal allocation remains unchanged, illustrating prior locking. Entropy regularization progressively reduces this imbalance, but retains non-negligible mass on incorrect modes because its flattening effect is not reward-selective. Prior correction instead drives incorrect mass toward zero and rebalances the correct modes, moving the distribution toward q^* . Complete settings and correct-mode probability trajectories are provided in Appendix D.1.

4 PRACTICAL PRIOR-CORRECTED RLVR

4.1 RESPONSE-LEVEL PRIOR CORRECTION WITH HISTORICAL REFERENCE

The oracle construction in Section 3.2 relies on the current allocation $q_t(m)$ of latent reasoning modes, which is unavailable in practical RLVR. We therefore seek an observable response-level

realization of prior correction. For a response y belonging to mode m , write $\pi_t(y | x) = q_t(m)\rho_t(y | x, m)$. Its response surprisal decomposes as

$$S_t(y) := -\log \pi_t(y | x) = \underbrace{\log \frac{1}{q_t(m)}}_{\text{mode-allocation term}} + \underbrace{\sum_{j=1}^{H(y)} -\log \rho_t(y_j | x, y_{<j}, m)}_{\text{within-mode term}}, \quad (5)$$

where $H(y)$ is the response length. Thus, response surprisal provides an observable starting point for the practical correction, while inevitably mixing mode allocation with response-specific likelihood variation.

Directly mapping raw surprisal to a correction weight, however, would make the effective correction strength depend strongly on its scale. Its typical level grows with response length, while its fluctuation scale varies with both length and the evolving policy. A fixed reweighting rule would therefore implicitly change the effective modulation strength as these response-level statistics evolve, injecting unintended variability into the policy update. We instead standardize response surprisal relative to policy-level reference statistics.

For a fixed problem x , let $\ell_{t,j}(Y) := -\log \pi_t(Y_j | x, Y_{<j})$, and define the policy-level mean token surprisal $\mu_t := \mathbb{E}_t[H(Y)^{-1} \sum_j \ell_{t,j}(Y)]$, and dispersion $v_t := \mathbb{E}_t[H(Y)^{-1} \sum_j (\ell_{t,j}(Y) - \mu_t)^2]$, where $\mathbb{E}_t$ is taken over $Y \sim \pi_t(\cdot | x)$. We then define

$$Z_t(y) := \frac{S_t(y) - H(y)\mu_t}{\sqrt{H(y)v_t}}. \quad (6)$$

Centering removes the typical $O(H(y))$ accumulation of token surprisal, while scaling by $\sqrt{H(y)v_t}$ keeps the remaining fluctuation $O(1)$ under weak token dependence. This places the response-level statistic on a controlled scale for a fixed correction map.

This defines the ideal population-level response statistic. The remaining practical question is how to construct its reference statistics (μ_t, v_t) reliably from finite rollouts. Estimating (μ_t, v_t) from the current rollout group makes the resulting correction sensitive to group composition, especially under limited rollout budgets. Fortunately, KL-regularized policy improvement and GRPO-style clipped optimization both favor local policy updates, making recent completed rollouts informative about the evolving reference statistics. We therefore construct the reference from historical rollouts, trading statistical stability against staleness without increasing the current sampling budget.

At round t , the historical state $(\hat{\mu}_t, \hat{v}_t)$ is fixed before the current rollout group, and we compute

$$\hat{Z}_t(y) := \frac{S_t(y) - H(y)\hat{\mu}_t}{\sqrt{H(y)\hat{v}_t}}. \quad (7)$$

After completing the round and obtaining $(\tilde{\mu}_t, \tilde{v}_t)$, we update the historical reference via $\hat{\xi}_{t+1} = (1 - \beta)\hat{\xi}_t + \beta\tilde{\xi}_t$ for $\xi \in \{\mu, v\}$. The coefficient β controls the resulting stability–staleness tradeoff: smaller β increases historical smoothing, while larger β tracks the evolving policy more closely. A finite-window variant is deferred to Appendix B.4.

Finally, we map the historical-reference statistic to a bounded correction weight,

$$\hat{\omega}_t(y) := 1 + \frac{2\tau}{\pi} \arctan(\gamma \hat{Z}_t(y)), \quad 0 < \tau < 1. \quad (8)$$

The monotone, bounded map is centered at one, preserves the ordering induced by $\hat{Z}_t$, and limits the influence of extreme response-level statistics. Section 4.2 characterizes both the error introduced by the historical reference and the remaining gap between the response-level correction and the oracle mode-level update. We treat γ as fixed in theory and adapt it online in practice.

Conservative application in practical RLVR. In practical GRPO-style optimization, the correction weight multiplies the group-relative advantage $A_t(Y)$, yielding the update signal $A_t(Y)\hat{\omega}_t(Y)$. Since $\hat{\omega}_t(Y) > 1$ whenever $\hat{Z}_t(Y) > 0$, applying the full correction to a response with $A_t(Y) < 0$ would further increase the magnitude of its negative update. This behavior is appropriate when

the response-level statistic faithfully reflects the intended mode-level correction, but can be overly aggressive when it is partly driven by within-mode response variation. We therefore attenuate the correction only on this branch:

$$\tilde{\omega}_t(Y) = \begin{cases} 1 + \kappa(\hat{\omega}_t(Y) - 1), & A_t(Y) < 0, \hat{Z}_t(Y) > 0, \\ \hat{\omega}_t(Y), & \text{otherwise,} \end{cases} \quad \kappa \in [0, 1]. \quad (9)$$

Thus, $\kappa = 1$ recovers the symmetric correction, whereas $\kappa = 0$ disables prior correction on this potentially over-suppressive branch while leaving all other cases unchanged.

In implementation, the historical statistics $(\hat{\mu}_t, \hat{v}_t)$ are maintained from completed rollout batches pooled across prompts, providing a policy-level historical reference without increasing the current rollout-group size. The complete training procedure, including the historical-statistic updates and conservative weighting rule, is summarized in Appendix C.1. The analysis below concerns the underlying correction $\hat{\omega}_t$, while equation 9 is its GRPO-specific practical instantiation.

4.2 APPROXIMATION ANALYSIS OF RESPONSE-LEVEL PRIOR CORRECTION

We now analyze the gap introduced when the oracle mode-level correction in Section 3.2 is translated into stochastic response-level policy updates. The gap separates into structural approximation residuals and controllable errors arising from stochastic optimization and historical-reference estimation.

Response-level approximation of the oracle update. We condition on a fixed problem x . Because the verifier evaluates sampled responses rather than latent modes, we associate mode m with the oracle reward $r_m := \sup_{\rho \in \mathcal{P}_m} \mathbb{E}_{Y \sim \rho(\cdot|x,m)}[R(Y)]$, the best expected verification value achievable within that mode. Here $\mathcal{P}_m$ denotes an admissible family of within-mode response distributions. This recovers the binary mode reward used in Section 3.1 when verification is mode-consistent, while allowing responses within the same mode to receive different observed rewards. Let $\psi_t(y) := \nabla_{\theta} \log \pi_t(y | x)$. For any correction ϕ in Section 3.2, the oracle gradient admits the equivalent mode-level and response-level representations

$$g_t^* := \sum_{m \in \mathcal{M}} r_m q_t(m) \phi(q_t(m)) \nabla_{\theta} \log q_t(m) = \mathbb{E}_t[r_{m(Y)} \phi(q_t(m(Y))) \psi_t(Y)]. \quad (10)$$

This identity places the oracle update in the same response-space form as practical policy optimization. Relative to this oracle representation, the response-level update analyzed below makes two substitutions: $r_{m(Y)}$ is replaced by $R(Y)$, while $\phi(q_t(m(Y)))$ is replaced by $\hat{\omega}_t(Y)$. Accordingly, define the correction-weight mismatch

$$\mathcal{E}_t := \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\left(\frac{\hat{\omega}_t(Y)}{\phi(q_t(m))} - 1 \right)^2 \right], \quad (11)$$

and the reward mismatch $\delta_{t,\phi}^r := \mathbb{E}_t[\phi(q_t(m(Y))) (R(Y) - r_{m(Y)}) \psi_t(Y)]$. The following result shows how these two substitutions propagate to the oracle stationarity error.

Theorem 4.1 (Stationarity under response-level approximation). *Suppose Assumptions A.2, A.6, A.7 hold. For an optimization horizon of T update rounds, let $N \geq 2$ denote the rollout-group size. If $0 < \alpha \leq \min \left\{ 1/(L(1 - 1/N)), 3/(8L(\frac{1}{4}(1 - 1/N) + \frac{Q}{NM^2})) \right\}$, then*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 \leq \mathcal{O} \left(\frac{\Delta_{\phi,0}}{\alpha T} + \frac{G^2 + M^2 \sigma^2}{M} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \mathcal{E}_t + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\delta_{t,\phi}^r\|^2 + \frac{L \alpha \sigma^2 Q}{N} \right), \quad (12)$$

where $\Delta_{\phi,0}$ denotes the initial oracle-objective gap defined in the appendix, and L, G, σ , and Q are constants specified in the assumptions.

The first and last terms are controllable through the optimization horizon, stepsize, and rollout-group size: choosing $\alpha_T \rightarrow 0$ with $\alpha_T T \rightarrow \infty$ makes both vanish. The reward mismatch is structural and need not vanish because response-level verification need not coincide with the oracle mode reward. The remaining correction-dependent term is $\mathcal{E}_t$, which we next decompose into a structural response-to-mode residual and the additional error introduced by the historical reference.

Decomposing the correction mismatch. The correction mismatch has two sources: a structural response-to-mode mismatch and an additional error introduced by historical-reference estimation. The structural mismatch persists even when the current-policy statistics (μ_t, v_t) are known exactly, since $Z_t(Y)$ remains response dependent and need not coincide with a correction that depends only on the mode allocation $q_t(m)$.

To align the oracle correction with the arctangent form used by our response-level correction, we specialize the oracle correction in Theorem 4.1 to $\phi = \phi_{c,d}$, where $\phi_{c,d}(q) := 1 + \frac{2\tau}{\pi} \arctan(c \log q^{-1} - d)$, for fixed $c > 0$ and $d \in \mathbb{R}$. For a response y assigned to mode m , define the response-to-mode residual $\varepsilon_t(y, m) := \gamma Z_t(y) - c \log q_t(m)^{-1} + d$. This residual captures the structural discrepancy between the population response-level argument $\gamma Z_t(y)$ and the mode-level oracle argument $c \log q_t(m)^{-1} - d$, and remains even when (μ_t, v_t) are known exactly. The analytical choice of (c, d) and the detailed decomposition of this residual are deferred to Appendix B.3. With the structural response-to-mode residual isolated, the remaining mismatch comes from replacing (μ_t, v_t) with the historical estimates $(\hat{\mu}_t, \hat{v}_t)$. The following result quantifies how these two sources contribute to the correction mismatch $\mathcal{E}_t$.

Proposition 4.2 (Historical-reference correction error). *Suppose Assumptions A.2, A.6, A.7, A.11, and A.15 hold. Then there exists a constant $C > 0$ such that*

$$\mathbb{E}\mathcal{E}_t \leq C \left(\mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + (1 - \beta)^{2t} + \frac{\beta}{N} + \frac{\alpha^2}{\beta^2} \right). \quad (13)$$

The bound separates the structural response-to-mode residual from the additional error due to the historical reference. The transient decays geometrically, while β/N captures estimation noise and α^2/β^2 captures staleness as the policy evolves. Their opposing dependence on β captures the stability–staleness tradeoff: a smaller β provides stronger historical smoothing but incurs greater staleness, while a larger β tracks the evolving policy more closely at the cost of higher estimation noise. Combining Proposition 4.2 with Theorem 4.1 yields the final guarantee for the response-level prior correction.

Corollary 4.3 (Stationarity rate with historical-reference correction). *Under the conditions of Theorem 4.1 and Proposition 4.2, choosing $\alpha_T = \Theta(T^{-3/5})$ and $\beta_T = \Theta(T^{-2/5})$ yields, for some constant $C > 0$,*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 \leq \frac{C}{T} \sum_{t=0}^{T-1} \left(\mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \mathbb{E} \|\delta_{t,\phi}^r\|^2 \right) + O(T^{-2/5}). \quad (14)$$

The controllable optimization and historical-reference estimation errors therefore decay as $O(T^{-2/5})$. When both structural terms—the response-to-mode residual and the reward mismatch—vanish, the bound reduces to $O(T^{-2/5})$; otherwise, their time-averaged contributions determine the residual level of the oracle stationarity measure. All theoretical proofs and supporting lemmas are deferred to Appendix A.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models and training data. We evaluate PC-RLVR on Qwen3-1.7B and Qwen3-4B (Yang et al., 2025). All methods use the same 32K-problem subset of DeepMath (He et al., 2026) and the same training budget; under our schedule, each run observes fewer than 10K training examples.

Baselines. We compare PC-RLVR with the base model, GRPO (Guo et al., 2025), DAPO (Yu et al., 2025), GRPO with entropy regularization (GRPO+Entropy), and ProGRPO (Li et al., 2026b). Across methods, we keep the backbone, training data, rollout budget, and evaluation protocol fixed while retaining method-specific components and hyperparameters. Complete implementation details are provided in Appendix C.

Table 1: Main results on mathematical reasoning benchmarks. We report both Avg@64 and Pass@64. Mean denotes the unweighted average across the five benchmarks. Bold highlights PC-RLVR when it achieves or ties the best result; otherwise, the best competing result is bolded. Δ denotes the difference between PC-RLVR and the best non-base baseline in each column.

Method	AIME24		AIME25		AIME26		OlympiadBench		AMC23		Mean	
	Avg@64	Pass@64	Avg@64	Pass@64	Avg@64	Pass@64	Avg@64	Pass@64	Avg@64	Pass@64	Avg@64	Pass@64
<i>Qwen3-1.7B-Base</i>	13.1	52.1	9.0	42.5	8.8	26.7	44.5	80.0	46.5	95.0	24.4	59.3
GRPO	13.0	56.7	9.3	46.7	9.7	36.7	45.6	78.2	45.1	95.0	24.5	62.7
DAPO	22.4	60.0	16.4	50.0	13.3	43.3	55.0	79.4	60.0	95.0	33.4	65.5
GRPO+Entropy	21.1	60.0	16.7	50.0	12.8	43.3	54.7	80.6	58.4	97.5	32.7	66.3
ProGRPO	23.3	60.0	19.7	50.0	15.3	36.7	57.7	80.6	60.1	97.5	35.2	65.0
PC-RLVR (Ours)	26.8	63.3	18.9	60.0	15.9	46.7	57.3	82.6	63.4	97.5	36.5	70.0
Δ vs. best	+3.5	+3.3	-0.8	+10.0	+0.6	+3.4	-0.4	+2.0	+3.3	0.0	+1.3	+3.7
<i>Qwen3-4B-Base</i>	21.1	56.7	17.9	53.3	16.2	50.0	57.6	84.1	67.3	97.5	36.0	68.3
GRPO	30.0	63.3	20.7	63.3	22.3	66.7	62.1	82.9	75.4	97.5	42.1	74.7
DAPO	38.5	76.7	29.7	70.0	32.8	63.3	67.4	87.1	82.2	97.5	50.1	78.9
GRPO+Entropy	38.0	73.3	31.0	73.3	30.6	66.7	67.8	84.7	82.9	97.5	50.1	79.1
ProGRPO	31.8	73.3	23.6	76.7	25.2	60.0	64.9	84.7	78.6	97.5	44.8	78.4
PC-RLVR (Ours)	39.5	80.0	29.1	76.7	34.2	70.0	67.8	86.5	83.0	97.5	50.7	82.1
Δ vs. best	+1.0	+3.3	-1.9	0.0	+1.4	+3.3	0.0	-0.6	+0.1	0.0	+0.6	+3.0

Evaluation details. We evaluate mathematical reasoning on AIME 2024–2026 (Zhang & Team Math-AI, 2024; 2025; 2026), AMC23 (Li et al., 2024), and OlympiadBench (He et al., 2024). We report both Avg@64 and Pass@64 in the main results to characterize performance under broader sampling. Avg@64 denotes the mean correctness across the 64 sampled responses per problem, while Pass@64 measures whether at least one of the 64 sampled responses is correct.

5.2 MAIN RESULTS

Table 1 summarizes the main reasoning results. PC-RLVR achieves the best mean Avg@64 and Pass@64 at both model scales. For Qwen3-1.7B, it improves over the respective best baselines by 1.3 and 3.7 percentage points in mean Avg@64 and Pass@64, respectively; for Qwen3-4B, the corresponding gains are 0.6 and 3.0 percentage points. On Qwen3-1.7B, PC-RLVR achieves the best or tied-best Pass@64 on all five benchmarks, and on Qwen3-4B it does so on four of the five benchmarks. In contrast, the margins in Avg@64 are generally smaller and PC-RLVR is not uniformly best on every individual benchmark. This pattern is aligned with the intended rebalancing effect of prior correction: reducing concentration on already favored successful responses can make a broader set of successful responses accessible under repeated sampling. The larger gains in Pass@64 than in Avg@64 are therefore consistent with broader coverage of successful responses.

Representative Pass@ k curves over $k \in \{1, 2, 4, 8, 16, 32, 64\}$ are provided in Appendix D.2. On Qwen3-1.7B, the advantage persists across multiple sampling budgets and becomes especially pronounced at larger k on AIME25.

5.3 PRIOR-DEPENDENT LOG-LIKELIHOOD GAIN

Our mode-level analysis identifies the uniform allocation q^* over correct modes as coverage-optimal. Moving from an initially imbalanced allocation toward q^* requires larger relative gains for underrepresented successful behaviors and smaller gains for those already favored by the initial policy. Since latent mode probabilities are not directly observable, we test this implication at the response level. For each prompt, we rank correct responses by their average token surprisal under the frozen base policy and partition them into high-, medium-, and low-prior groups. We then measure the relative per-token log-likelihood gain of PC-RLVR over GRPO. Figure 2 shows a consistent prior-dependent trend across AIME24–26 and OlympiadBench: responses with lower initial prior receive larger relative log-likelihood gains, with the low-prior group attaining the largest mean gain in each case. Moreover, the low-prior group receives a larger gain than the high-prior group on 88.9%, 76.9%, 100%, and 95.3% of eligible prompts, respectively. This pattern is consistent with the rebalancing effect predicted by prior correction: relative to GRPO, PC-RLVR produces larger average per-token log-likelihood gains for successful responses that are less favored by the initial policy. Complete

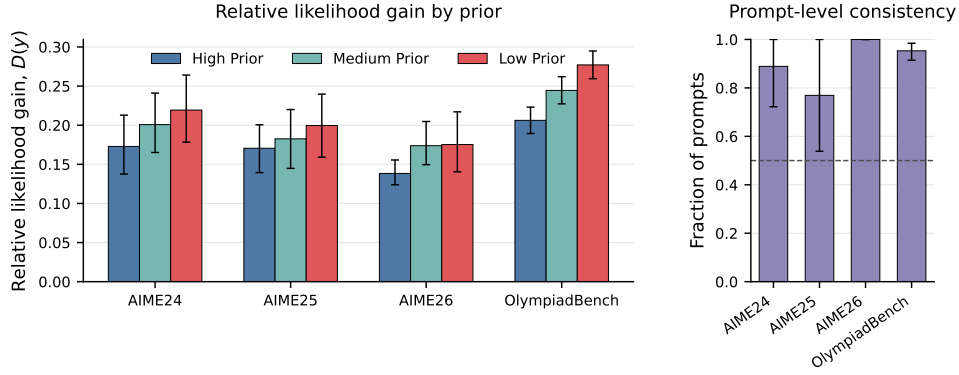


Figure 2: Prior-dependent log-likelihood gain of PC-RLVR relative to GRPO on Qwen3-1.7B. **Left:** prompt-averaged relative per-token log-likelihood gain across prior groups. **Right:** fraction of eligible prompts for which the low-prior group receives a larger gain than the high-prior group; the dashed line marks 50%. Error bars show 95% bootstrap confidence intervals over prompts.

definitions and diagnostics are provided in Appendix D.3. A complementary analysis of the final successful-response distribution under the initial-policy prior is given in Appendix D.4.

5.4 ABLATIONS AND SENSITIVITY ANALYSIS

We examine the main design choices and parameter sensitivity of PC-RLVR on Qwen3-4B.

Ablations. We evaluate controlled ablations on Qwen3-4B using $\tau = 0.2$ and report Avg@64 and Pass@64 across all five benchmarks, with full results and details in Appendix D.5. We first disable the correction weights while retaining the same asymmetric clipping range as PC-RLVR, thereby isolating prior correction from the clipping change. We then replace the bounded arctangent transformation with a slope-matched linear correction while keeping the remaining components unchanged. Removing prior correction reduces the mean Avg@64 and Pass@64 from 50.9/81.4 to 46.5/80.1, while the linear variant obtains 46.2/78.9. These results show that the gains are not explained by asymmetric clipping alone, and that prior correction also benefits from bounding the correction magnitude rather than applying an unrestricted linear transformation.

Sensitivity analysis. We study β , κ , and τ on Qwen3-4B by monitoring Pass@1 every 10 training steps on AIME24 and AIME25; full curves are provided in Appendix D.6. With $\tau = 0.2$ and $\kappa = 0$, $\beta = 0.7$ gives the strongest overall trajectory among $\beta \in \{0.5, 0.7, 0.9\}$, consistent with β controlling the balance between historical averaging and adaptation to policy changes. At $(\beta, \tau) = (0.7, 0.2)$, the $\kappa = 0$ trajectory also remains above the symmetric $\kappa = 1$ setting over most of training, supporting attenuation of the potentially over-suppressive branch in equation 9. We additionally compare $\tau \in \{0.2, 0.5\}$ at $(\beta, \kappa) = (0.7, 0)$; the $\tau = 0.5$ run yields stronger broader-sampling performance at its selected checkpoint and is used for the Qwen3-4B main configuration.

6 CONCLUSION AND DISCUSSION

We identify *prior locking* as a structural limitation of binary-reward RLVR: aggregate success probability can increase while the relative allocation among correct reasoning modes remains tied to initialization. We introduce prior correction and show that, under KL-regularized policy improvement, the oracle mode-level dynamics converge to the coverage-optimal uniform allocation over correct modes. For practical LLM training, PC-RLVR uses a historical-reference-based response-level correction with bounded weights. Our analysis separates structural approximation residuals from controllable optimization and historical-reference estimation errors, while experiments show improved Avg@64 and Pass@64 together with prior-dependent log-likelihood gains consistent with broader coverage of successful responses. Limitations and future directions are discussed in Appendix E.

AI USE STATEMENT

We used generative AI tools to assist with the development and refinement of theoretical arguments, formulation and checking of mathematical claims and proofs, feedback on research methodology and experimental design, and implementation and code development. We also used generative AI tools for language editing, improving the readability and organization of the manuscript, literature and reference checking, and formatting assistance.

All AI-assisted mathematical arguments were reviewed and verified by the authors. AI-assisted code was reviewed and tested by the authors before being used in our experiments. Experimental results reported in the paper were obtained from our own implementations and experimental pipeline, and all AI-assisted interpretations were checked against the underlying results. The authors reviewed all AI-assisted text, claims, code, and artifacts and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide complete proofs of the theoretical results in the appendix. The experimental setup, baseline implementations, training and evaluation protocols, and hyperparameters are also documented in the appendix. These details are intended to support reproduction of both the theoretical analysis and the empirical results. Source code for reproducing the experiments is provided as anonymous supplementary material.

REFERENCES

- Abbas Abdolmaleki, Jost Tobias Springenberg, Yuval Tassa, Remi Munos, Nicolas Heess, and Martin Riedmiller. Maximum a posteriori policy optimisation. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=S1ANxQW0b>.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.
- Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/geist19a.html>.
- 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(8081):633–638, 2025.
- Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25548–25560, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1298. URL <https://aclanthology.org/2025.emnlp-main.1298/>.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL <https://aclanthology.org/2024.acl-long.211/>.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and

- Dong Yu. DeepMath-103K: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=kHB5Te5IWm>.
- Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. T1: Advancing language model reasoning through reinforcement learning and inference scaling. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23976–24003. PMLR, 13–19 Jul 2025. URL <https://proceedings.mlr.press/v267/hou25e.html>.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Jian Hu, Mingjie Liu, Ximing Lu, Fang Wu, Zaid Harchaoui, Shizhe Diao, Yejin Choi, Pavlo Molchanov, Jun Yang, Jan Kautz, and Yi Dong. BroRL: Scaling reinforcement learning via broadened exploration. In *International Conference on Machine Learning (ICML)*, July 2026.
- Yiyang Jin, Kunzhao Xu, Hang Li, Xueting Han, Yanmin Zhou, Cheng Li, and Jing Bai. ReVeal: Self-evolving code agents via reliable self-verification. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=q56ZI1Co43>.
- Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 476–494, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/01498f17142547dadb60999c39bd0a09-Paper-Conference.pdf.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.
- Chanuk Lee, Minki Kang, and Sung Ju Hwang. SAGE: Shaping anchors for guided exploration in RLVR of LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=VzRUI5ZVkp>.
- Gang Li, Ming Lin, Tomer Galanti, Zhengzhong Tu, and Tianbao Yang. DisCO: Reinforcing large reasoning models with discriminative constrained optimization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 57304–57331. Curran Associates, Inc., 2025. doi: 10.52202/085713-1921. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/52d6c85793b3dd47636eb430710da915-Paper-Conference.pdf.
- Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. NuminaMath: The largest public dataset in AI4Maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 2024.
- Long Li, Zhijian Zhou, Jiaran Hao, Jason Klein Liu, Yanting Miao, Wei Pang, Xiaoyu Tan, Wei Chu, Zhe Wang, Shirui Pan, Chao Qu, and Yuan Qi. The choice of divergence: A neglected key to mitigating diversity collapse in reinforcement learning with verifiable reward. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL <https://openreview.net/forum?id=xPEsxc07F7>.
- Pengyi Li, Elizaveta Goncharova, Andrey Kuznetsov, and Ivan Oseledets. Back to basics: Revisiting exploration in reinforcement learning for LLM reasoning via generative probabilities. *arXiv preprint arXiv:2602.05281*, 2026b.

- Xiao Liang, Zhong-Zhi Li, Yeyun Gong, Yelong Shen, Yingnian Wu, Zhijiang Guo, and Weizhu Chen. Beyond Pass@1: Self-play with variational problem synthesis sustains RLVR. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 79443–79491, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/805cd624d38e39ca061e8fd431e20825-Paper-Conference.pdf.
- Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL <https://openreview.net/forum?id=YPsJha5HXQ>.
- Youssef Mroueh. Reinforcement learning with verifiable rewards: GRPO’s loss, dynamics, and success amplification. In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=z9Bo4eXSpW>.
- Daniil Plyusov, Alexey Gorbatovski, Boris Shaposhnikov, Viacheslav Sinii, Alexey Malakhov, and Daniil Gavrilov. F-GRPO: Don’t let your policy learn the obvious and forget the rare. In *The 1st Workshop on Scaling Post-training for LLMs*, 2026. URL <https://openreview.net/forum?id=DWS8VtJt6a>.
- Ali Rad, Khashayar Filom, Darioush Keivan, Peyman Mohajerin Esfahani, and Ehsan Kamalinejad. Rate or fate? $RLV^{\epsilon}R$: Reinforcement learning with verifiable noisy rewards. In *Forty-third International Conference on Machine Learning*, 2026. URL <https://openreview.net/forum?id=LwB2EacVT6>.
- Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3505–3506, 2020.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- 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=aBO5Svgt1>.
- Jiakang Wang, Runze Liu, Qingpeng Cai, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, Kun Gai, and Ling Pan. When importance sampling misallocates credit: Asymmetric ratios for outcome-supervised RL. *arXiv preprint arXiv:2510.06062*, 2025a.
- Weixun Wang, Shaopan Xiong, Gengru Chen, Wei Gao, Sheng Guo, Yancheng He, Ju Huang, Jiaheng Liu, Zhendong Li, Xiaoyang Li, et al. Reinforcement learning optimization for large-scale learning: An efficient and user-friendly scaling library. *arXiv preprint arXiv:2506.06122*, 2025b.
- Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhi-jian Xu, Xiao Liang, Junjie Li, Ziming Miao, Jiang Bian, and Mao Yang. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 49450–49483, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/517f9b9c227b9dd51dba4560f37165ed-Paper-Conference.pdf.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncui 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, Jiase 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 D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 113222–113244. Curran Associates, Inc., 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a4277440d50f1f15d2cb4c14f7e0c0d2-Paper-Conference.pdf.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (AIME) 2024, 2024.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (AIME) 2025, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (AIME) 2026, 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. *arXiv preprint arXiv:2507.18071*, 2025.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *2nd AI for Math Workshop @ ICML 2025*, 2025. URL <https://openreview.net/forum?id=cR8QXDD2dr>.

A PROOFS OF MAIN THEORETICAL RESULTS

A.1 PROOF OF THEOREM 3.1

Proof. At iteration t , the KL-regularized update in equation 1 has the closed-form solution

$$q_{t+1}(m) = \frac{q_t(m) \exp(\eta r_t(m))}{\sum_{m' \in \mathcal{M}} q_t(m') \exp(\eta r_t(m'))}, \quad r_t(m) = \begin{cases} a_t, & m \in \mathcal{C}^+, \\ b_t, & m \notin \mathcal{C}^+. \end{cases} \quad (15)$$

For any $m \in \mathcal{C}^+$, equation 15 gives

$$\begin{aligned} q_{t+1}(m | \mathcal{C}^+) &= \frac{q_t(m) \exp(\eta a_t)}{q_t(\mathcal{C}^+) \exp(\eta a_t)} \\ &= q_t(m | \mathcal{C}^+). \end{aligned} \quad (16)$$

Iterating over t therefore yields

$$q_T(m | \mathcal{C}^+) = q_0(m | \mathcal{C}^+), \quad m \in \mathcal{C}^+, \quad (17)$$

which proves the prior-locking property.

It remains to track the total probability mass assigned to the correct set. Let $p_t := q_t(\mathcal{C}^+)$. Summing equation 15 over $m \in \mathcal{C}^+$ gives

$$p_{t+1} = \frac{p_t \exp(\eta a_t)}{p_t \exp(\eta a_t) + (1 - p_t) \exp(\eta b_t)}. \quad (18)$$

If $p_0 = 1$, then $q_T(\mathcal{M} \setminus \mathcal{C}^+) = 0$ for every T , so the second claim is immediate. Otherwise, $p_0 \in (0, 1)$, and equation 18 implies

$$\frac{p_{t+1}}{1 - p_{t+1}} = \frac{p_t}{1 - p_t} \exp(\eta(a_t - b_t)). \quad (19)$$

Applying this recursion over $t = 0, \dots, T - 1$ gives

$$\frac{p_T}{1 - p_T} = \frac{p_0}{1 - p_0} \exp\left(\eta \sum_{t=0}^{T-1} (a_t - b_t)\right). \quad (20)$$

Hence, if $\sum_{t=0}^{\infty} (a_t - b_t) = \infty$, the right-hand side diverges and $p_T \rightarrow 1$. Equivalently,

$$q_T(\mathcal{M} \setminus \mathcal{C}^+) = 1 - p_T \rightarrow 0. \quad (21)$$

Thus, the update can drive all probability mass toward the correct set without changing the relative allocation among correct modes. Together, these establish both claims of Theorem 3.1. $\square$

A.2 PROOF OF PROPOSITION 3.2

We consider the entropy-regularized counterpart of equation 1,

$$q_{t+1}^H = \arg \max_{q \in \Delta(\mathcal{M})} \left\{ \mathbb{E}_{m \sim q} [r_t(m)] - \frac{1}{\eta} \text{KL}(q \parallel q_t^H) + \lambda_H \mathcal{H}(q) \right\}, \quad \lambda_H > 0, \quad (22)$$

where $\mathcal{H}(q) := -\sum_{m \in \mathcal{M}} q(m) \log q(m)$, $q_0^H = q_0$.

Proof. The optimality condition of equation 22 gives

$$q_{t+1}^H(m) = \frac{(q_t^H(m))^{\kappa_H} \exp(\eta \kappa_H r_t(m))}{\sum_{m' \in \mathcal{M}} (q_t^H(m'))^{\kappa_H} \exp(\eta \kappa_H r_t(m'))}, \quad \kappa_H = \frac{1}{1 + \eta \lambda_H} \in (0, 1). \quad (23)$$

Since $q_0(m) > 0$ for every $m \in \mathcal{M}$, equation 23 implies $q_t^H(m) > 0$ for all t .

Now consider any two modes m, m' receiving the same reward value. At every iteration they receive the same learning signal, i.e., $r_t(m) = r_t(m')$, whether this common value is a_t or b_t . Taking the log-ratio in equation 23 therefore gives

$$\begin{aligned} \log \frac{q_{t+1}^H(m)}{q_{t+1}^H(m')} &= \kappa_H \log \frac{q_t^H(m)}{q_t^H(m')} + \eta \kappa_H (r_t(m) - r_t(m')) \\ &= \kappa_H \log \frac{q_t^H(m)}{q_t^H(m')}. \end{aligned} \quad (24)$$

Iterating this recursion from $t = 0$ to $T - 1$ yields

$$\log \frac{q_T^H(m)}{q_T^H(m')} = \kappa_H^T \log \frac{q_0(m)}{q_0(m')}, \quad (25)$$

which proves Proposition 3.2. Since $\kappa_H \in (0, 1)$, the log-ratio converges to zero as $T \rightarrow \infty$, and hence $q_T^H(m)/q_T^H(m') \rightarrow 1$. Thus, entropy regularization progressively contracts the relative allocation imbalance among modes receiving the same reward. $\square$

A.3 PROOF OF THEOREM 3.3

We first record the closed-form oracle update and establish two properties used in the convergence proof: preservation of the correct-mode ordering and a uniform lower bound on correct-mode probabilities.

Closed-form oracle update. The KL-regularized update in equation 4 has the closed-form solution

$$\bar{q}_{t+1}(m) = \begin{cases} \frac{\bar{q}_t(m) \exp(\eta a_t \phi(\bar{q}_t(m)))}{\mathcal{Z}_t}, & m \in \mathcal{C}^+, \\ \frac{\bar{q}_t(m) \exp(\eta b_t \phi(\bar{q}_t(m)))}{\mathcal{Z}_t}, & m \notin \mathcal{C}^+, \bar{q}_t(m) > 0, \\ 0, & m \notin \mathcal{C}^+, \bar{q}_t(m) = 0, \end{cases} \quad (26)$$

where $\mathcal{Z}_t$ is the corresponding normalizing constant over the support of $\bar{q}_t$. The last case reflects the support-preserving property of the KL term and avoids requiring $\phi(0)$ to be defined.

Order preservation and correct-mode interiority. We next make explicit the quantities appearing in the step-size condition of Theorem 3.3. Recall that

$$\rho_0 := \min_{m \in \mathcal{C}^+} q_0(m) = \min_{m \in \mathcal{C}^+} \bar{q}_0(m) > 0, \quad (27)$$

where ρ_0 is the smallest initial probability assigned to a correct mode. Over the interval $[\rho_0, 1]$, define

$$L_\phi := \max_{u \in [\rho_0, 1]} \{-u\phi'(u)\}. \quad (28)$$

Since ϕ is continuously differentiable and $\rho_0 > 0$, $L_\phi < \infty$. The quantity L_ϕ bounds the scaled derivative $-u\phi'(u)$ that appears when differentiating the unnormalized update $u \exp(\eta a_t \phi(u))$. The step-size condition $\eta < 1/(a_{\max} L_\phi)$ therefore prevents the update from reversing the ordering of correct-mode probabilities.

Lemma A.1 (Order preservation and correct-mode interiority). *Under the conditions of Theorem 3.3, the ordering of the correct-mode probabilities is preserved. In particular, for any $m, m' \in \mathcal{C}^+$,*

$$\bar{q}_t(m) \leq \bar{q}_t(m') \implies \bar{q}_{t+1}(m) \leq \bar{q}_{t+1}(m'). \quad (29)$$

Moreover, the minimum probability among correct modes is nondecreasing:

$$\min_{m \in \mathcal{C}^+} \bar{q}_{t+1}(m) \geq \min_{m \in \mathcal{C}^+} \bar{q}_t(m). \quad (30)$$

Consequently, the initial lower bound ρ_0 is preserved along the entire oracle trajectory,

$$\bar{q}_t(m) \geq \rho_0, \quad m \in \mathcal{C}^+, \quad t \geq 0. \quad (31)$$

Proof. We proceed by induction. Suppose that, at iteration t , all correct-mode probabilities lie in $[\rho_0, 1]$, which holds at $t = 0$ by the definition of ρ_0 . For a correct-mode probability $u \in [\rho_0, 1]$, its unnormalized update in equation 26 is $u \exp(\eta a_t \phi(u))$, whose derivative satisfies

$$\begin{aligned} \frac{d}{du} [u \exp(\eta a_t \phi(u))] &= \exp(\eta a_t \phi(u)) [1 + \eta a_t u \phi'(u)] \\ &\geq \exp(\eta a_t \phi(u)) [1 - \eta a_{\max} L_\phi] > 0. \end{aligned} \quad (32)$$

Thus, the unnormalized update of a correct-mode probability is strictly increasing on $[\rho_0, 1]$. Since all correct-mode probabilities share the same normalizing constant $\mathcal{Z}_t$, this monotonicity preserves their ordering, which proves equation 29.

Now let $m_t \in \mathcal{C}^+$ be a correct mode with minimum probability at iteration t . Since ϕ is decreasing, m_t has the largest multiplicative factor among all correct modes. Moreover, every supported incorrect mode has multiplicative factor at most one because $b_t \leq 0$ and $\phi > 0$, whereas the factor of m_t is strictly larger than one because $a_t > 0$. Hence, the multiplicative factor of m_t is maximal over the support of $\bar{q}_t$. Since $\mathcal{Z}_t$ is the $\bar{q}_t$ -weighted average of these factors,

$$\mathcal{Z}_t \leq \exp(\eta a_t \phi(\bar{q}_t(m_t))). \quad (33)$$

Therefore,

$$\bar{q}_{t+1}(m_t) = \frac{\bar{q}_t(m_t) \exp(\eta a_t \phi(\bar{q}_t(m_t)))}{\mathcal{Z}_t} \geq \bar{q}_t(m_t). \quad (34)$$

By order preservation, m_t remains a minimum-probability correct mode after the update. Hence, equation 30 holds. Iterating this inequality gives $\bar{q}_t(m) \geq \rho_0$ for every $m \in \mathcal{C}^+$ and $t \geq 0$, which also closes the induction. $\square$

Completion of the proof. Having established order preservation and the uniform lower bound on correct-mode probabilities, we now prove convergence to q^* .

Proof of Theorem 3.3. By Lemma A.1, the ordering of the correct-mode probabilities is preserved and $\bar{q}_t(m) \geq \rho_0 > 0$ for every $m \in \mathcal{C}^+$ and $t \geq 0$. We prove convergence in two steps.

Vanishing probability on incorrect modes. Fix any correct mode $i \in \mathcal{C}^+$ and any incorrect mode $j \notin \mathcal{C}^+$. If $\bar{q}_0(j) = 0$, the support-preserving property of the KL-regularized update implies $\bar{q}_t(j) = 0$ for all t . It therefore suffices to consider $\bar{q}_0(j) > 0$.

From equation 26,

$$\log \frac{\bar{q}_{t+1}(i)}{\bar{q}_{t+1}(j)} = \log \frac{\bar{q}_t(i)}{\bar{q}_t(j)} + \eta [a_t \phi(\bar{q}_t(i)) - b_t \phi(\bar{q}_t(j))]. \quad (35)$$

Since $b_t \leq 0$, $\phi > 0$, and ϕ is decreasing,

$$a_t \phi(\bar{q}_t(i)) - b_t \phi(\bar{q}_t(j)) \geq a_t \phi(\bar{q}_t(i)) \geq a_t \phi(1).$$

Hence,

$$\log \frac{\bar{q}_{t+1}(i)}{\bar{q}_{t+1}(j)} \geq \log \frac{\bar{q}_t(i)}{\bar{q}_t(j)} + \eta a_t \phi(1). \quad (36)$$

Iterating from $t = 0$ to $T - 1$ gives

$$\log \frac{\bar{q}_T(i)}{\bar{q}_T(j)} \geq \log \frac{\bar{q}_0(i)}{\bar{q}_0(j)} + \eta \phi(1) \sum_{t=0}^{T-1} a_t. \quad (37)$$

Because $\sum_{t=0}^{\infty} a_t = \infty$, the right-hand side diverges, so

$$\frac{\bar{q}_T(j)}{\bar{q}_T(i)} \longrightarrow 0. \quad (38)$$

Since $\bar{q}_T(i) \leq 1$, this implies $\bar{q}_T(j) \rightarrow 0$. As the mode set is finite,

$$\bar{q}_T(\mathcal{M} \setminus \mathcal{C}^+) \longrightarrow 0, \quad (39)$$

and therefore

$$\bar{q}_T(\mathcal{C}^+) \longrightarrow 1. \quad (40)$$

Equalization among correct modes. If $K = 1$, equation 40 already implies $\bar{q}_T \rightarrow q^*$. Suppose therefore that $K \geq 2$.

Let $m_{\max}, m_{\min} \in \mathcal{C}^+$ be correct modes with the largest and smallest initial probabilities, respectively. By Lemma A.1, their ordering is preserved, so they remain maximum- and minimum-probability correct modes at every iteration. From equation 26,

$$\begin{aligned} \log \frac{\bar{q}_{t+1}(m_{\max})}{\bar{q}_{t+1}(m_{\min})} &= \log \frac{\bar{q}_t(m_{\max})}{\bar{q}_t(m_{\min})} \\ &\quad + \eta a_t [\phi(\bar{q}_t(m_{\max})) - \phi(\bar{q}_t(m_{\min}))]. \end{aligned} \quad (41)$$

Since $\bar{q}_t(m_{\max}) \geq \bar{q}_t(m_{\min})$ and ϕ is decreasing, the bracketed term is nonpositive. Therefore, the extreme correct-mode log-ratio is nonincreasing. It is also nonnegative, and hence converges to a finite nonnegative limit.

Suppose, for contradiction, that this limit is strictly positive. Then there exists $\delta > 0$ such that, for all sufficiently large t ,

$$\log \frac{\bar{q}_t(m_{\max})}{\bar{q}_t(m_{\min})} \geq \delta. \quad (42)$$

By Lemma A.1, both probabilities remain in $[\rho_0, 1]$. Since ϕ is continuous and strictly decreasing, compactness then implies the existence of a constant $\epsilon > 0$ such that, for all sufficiently large t ,

$$\phi(\bar{q}_t(m_{\min})) - \phi(\bar{q}_t(m_{\max})) \geq \epsilon. \quad (43)$$

Substituting this bound into equation 41 gives

$$\log \frac{\bar{q}_{t+1}(m_{\max})}{\bar{q}_{t+1}(m_{\min})} \leq \log \frac{\bar{q}_t(m_{\max})}{\bar{q}_t(m_{\min})} - \eta a_t \epsilon \quad (44)$$

for all sufficiently large t . Summing from such an index T_0 to $T - 1$ yields

$$\log \frac{\bar{q}_T(m_{\max})}{\bar{q}_T(m_{\min})} \leq \log \frac{\bar{q}_{T_0}(m_{\max})}{\bar{q}_{T_0}(m_{\min})} - \eta \epsilon \sum_{t=T_0}^{T-1} a_t. \quad (45)$$

Since $\sum_{t=0}^{\infty} a_t = \infty$, the right-hand side becomes negative for sufficiently large T , contradicting the nonnegativity of the log-ratio. Hence,

$$\frac{\bar{q}_t(m_{\max})}{\bar{q}_t(m_{\min})} \longrightarrow 1. \quad (46)$$

Finally, because $m_{\min}$ and $m_{\max}$ remain the extreme correct modes,

$$\bar{q}_t(m_{\min}) \leq \frac{\bar{q}_t(\mathcal{C}^+)}{K} \leq \bar{q}_t(m_{\max}). \quad (47)$$

By equation 40, the middle term converges to $1/K$. Moreover, equation 46 implies

$$\bar{q}_t(m_{\max}) - \bar{q}_t(m_{\min}) \longrightarrow 0. \quad (48)$$

Therefore,

$$\bar{q}_t(m_{\min}) \longrightarrow \frac{1}{K}, \quad \bar{q}_t(m_{\max}) \longrightarrow \frac{1}{K}. \quad (49)$$

Every correct-mode probability lies between these two extremes, so

$$\bar{q}_t(m) \longrightarrow \frac{1}{K}, \quad m \in \mathcal{C}^+. \quad (50)$$

□

A.4 MODE-LEVEL STATIONARITY UNDER APPROXIMATE CORRECTION

The preceding analysis studies the idealized mode-level dynamics in which the distribution q_t is updated directly. We now turn to the parameterized stochastic-optimization setting needed to analyze practical RLVR training, where the policy is represented by θ and updated from finite rollout samples. Our goal is to characterize how approximating the oracle correction affects stationarity of the corresponding mode-level objective.

As an intermediate step, we first consider the simpler setting in which the latent mode identity and mode reward are directly available, while the oracle correction $\phi(q_t(m))$ is replaced by an estimated mode-level weight $\omega_t(m)$. This removes the structural response-to-mode and reward mismatches and isolates the effect of correction-weight error. The response-level analysis below then extends this argument by incorporating the two additional sources of mismatch.

Define the primitive of the correction function by

$$\Phi(u) := \int_0^u \phi(v) dv. \quad (51)$$

and assume that Φ is finite on $[0, 1]$. The corresponding mode-level oracle objective is

$$F_\phi(\theta) := \sum_{m \in \mathcal{M}} r_m \Phi(q_\theta(m)). \quad (52)$$

Its gradient at θ_t is

$$g_t^* := \nabla_\theta F_\phi(\theta_t) = \sum_{m \in \mathcal{M}} r_m q_t(m) \phi(q_t(m)) \nabla_\theta \log q_t(m). \quad (53)$$

Let $\mathcal{F}_t$ denote the training history before the round- t samples are drawn, and write

$$\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot | \mathcal{F}_t].$$

Throughout this subsection and the response-level analysis below, $\mathbb{E}_t$ denotes conditional expectation over the fresh round- t sampling given $\mathcal{F}_t$.

Let $\omega_t(m) > 0$ be an $\mathcal{F}_t$ -measurable estimate of $\phi(q_t(m))$, and assume $q_t(m) > 0$ for every $m \in \mathcal{M}$ over the rounds under consideration. For conditionally i.i.d. mode samples $m_{t,1}, \dots, m_{t,N} \sim q_t$, define

$$\hat{g}_t^{\text{mode}} := \frac{1}{N} \sum_{i=1}^N r_{m_{t,i}} \omega_t(m_{t,i}) \nabla_{\theta} \log q_t(m_{t,i}), \quad \theta_{t+1} = \theta_t + \alpha \hat{g}_t^{\text{mode}}. \quad (54)$$

The corresponding mode-level correction mismatch is

$$\mathcal{E}_{t,\text{mode}}^{\omega} := \sum_{m \in \mathcal{M}} \left(\frac{\omega_t(m)}{\phi(q_t(m))} - 1 \right)^2. \quad (55)$$

Here r_m denotes the fixed oracle mode reward used in the response-level analysis, rather than the time-varying signals a_t and b_t considered in Section 3.

Assumption A.2 (Oracle-objective smoothness and boundedness). For every round under consideration, F_{ϕ} is L -smooth along the line segment between θ_t and θ_{t+1} , and

$$F_{\phi}^{\max} := \sup_{\theta} F_{\phi}(\theta) < \infty. \quad (56)$$

Moreover, there exists $G < \infty$ such that

$$\|g_t^{\star}\|^2 \leq G^2 \quad (57)$$

for all t .

For later use, define the contribution of mode m to the oracle gradient by

$$g_t^{\star}(m) := r_m q_t(m) \phi(q_t(m)) \nabla_{\theta} \log q_t(m), \quad g_t^{\star} = \sum_{m \in \mathcal{M}} g_t^{\star}(m). \quad (58)$$

Assumption A.3 (Bounded mode-gradient dispersion). There exists $\sigma_{\text{mode}}^2 < \infty$ such that

$$\left\| g_t^{\star}(m) - \frac{1}{M} g_t^{\star} \right\|^2 \leq \sigma_{\text{mode}}^2 \quad (59)$$

for every $m \in \mathcal{M}$ and every round under consideration.

Assumption A.4 (Bounded correction-weight second moment). There exists $Q_{\text{mode}} < \infty$ such that, almost surely,

$$\sum_{m \in \mathcal{M}} \frac{\omega_t(m)^2}{q_t(m) \phi(q_t(m))^2} \leq Q_{\text{mode}} \quad (60)$$

for every round under consideration.

Assumption A.2 provides the smoothness and bounded-improvement conditions needed for the one-step progress and telescoping argument. Assumption A.3 controls the variation of individual mode contributions around their average contribution, while Assumption A.4 bounds the variance amplification induced by mode sampling and estimated correction weights.

Lemma A.5 (Mode-level stationarity under approximate correction). *Suppose Assumptions A.2–A.4 hold and $N \geq 2$. If $0 < \alpha \leq \min \left\{ 1/(L(1 - 1/N)), 3/(8L(\frac{1}{4}(1 - 1/N) + \frac{Q_{\text{mode}}}{NM^2})) \right\}$, then*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^{\star}\|^2 \leq \frac{8\Delta_{\phi,0}}{3\alpha T} + \frac{8(G^2 + M^2\sigma_{\text{mode}}^2)}{3M} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \mathcal{E}_{t,\text{mode}}^{\omega} + \frac{8L\alpha\sigma_{\text{mode}}^2 Q_{\text{mode}}}{3N}, \quad (61)$$

where $\Delta_{\phi,0} := F_{\phi}^{\max} - F_{\phi}(\theta_0)$.

Proof. Define the conditional mean of the approximate mode-level estimator by

$$g_t^{\omega} := \mathbb{E}_t[\hat{g}_t^{\text{mode}}] = \sum_{m \in \mathcal{M}} q_t(m) r_m \omega_t(m) \nabla_{\theta} \log q_t(m). \quad (62)$$

We first control the conditional second moment and the bias of this estimator, and then apply the smoothness of F_{ϕ} .

Estimator second moment. Since $\hat{g}_t^{\text{mode}}$ is the average of N conditionally i.i.d. mode-sample contributions,

$$\mathbb{E}_t \|\hat{g}_t^{\text{mode}}\|^2 = \frac{1}{N} \sum_{m \in \mathcal{M}} q_t(m) \|r_m \omega_t(m) \nabla_{\theta} \log q_t(m)\|^2 + \left(1 - \frac{1}{N}\right) \|g_t^{\omega}\|^2. \quad (63)$$

By equation 58,

$$\sum_{m \in \mathcal{M}} q_t(m) \|r_m \omega_t(m) \nabla_{\theta} \log q_t(m)\|^2 = \sum_{m \in \mathcal{M}} \frac{\omega_t(m)^2}{q_t(m) \phi(q_t(m))^2} \|g_t^*(m)\|^2. \quad (64)$$

Assumption A.3 gives

$$\|g_t^*(m)\|^2 \leq 2\sigma_{\text{mode}}^2 + \frac{2}{M^2} \|g_t^*\|^2. \quad (65)$$

Combining this inequality with Assumption A.4 yields

$$\mathbb{E}_t \|\hat{g}_t^{\text{mode}}\|^2 \leq \frac{2Q_{\text{mode}}}{N} \left(\sigma_{\text{mode}}^2 + \frac{1}{M^2} \|g_t^*\|^2 \right) + \left(1 - \frac{1}{N}\right) \|g_t^{\omega}\|^2. \quad (66)$$

Bias from approximate correction. Using equation 58,

$$g_t^{\omega} - g_t^* = \sum_{m \in \mathcal{M}} \left(\frac{\omega_t(m)}{\phi(q_t(m))} - 1 \right) g_t^*(m). \quad (67)$$

Cauchy–Schwarz therefore gives

$$\|g_t^{\omega} - g_t^*\|^2 \leq \mathcal{E}_{t,\text{mode}}^{\omega} \sum_{m \in \mathcal{M}} \|g_t^*(m)\|^2. \quad (68)$$

Since $\sum_m g_t^*(m) = g_t^*$,

$$\begin{aligned} \sum_{m \in \mathcal{M}} \|g_t^*(m)\|^2 &= \sum_{m \in \mathcal{M}} \left\| g_t^*(m) - \frac{1}{M} g_t^* \right\|^2 + \frac{1}{M} \|g_t^*\|^2 \\ &\leq M\sigma_{\text{mode}}^2 + \frac{1}{M} \|g_t^*\|^2. \end{aligned} \quad (69)$$

Using $\|g_t^*\|^2 \leq G^2$, we obtain

$$\|g_t^{\omega} - g_t^*\|^2 \leq \frac{G^2 + M^2\sigma_{\text{mode}}^2}{M} \mathcal{E}_{t,\text{mode}}^{\omega}. \quad (70)$$

One-step ascent. By the L -smoothness of F_{ϕ} along the update segment and equation 66,

$$\begin{aligned} &\mathbb{E}_t[F_{\phi}(\theta_{t+1})] - F_{\phi}(\theta_t) \\ &\geq \alpha \langle g_t^*, g_t^{\omega} \rangle - \frac{L\alpha^2}{2} \left(1 - \frac{1}{N}\right) \|g_t^{\omega}\|^2 - \frac{L\alpha^2 Q_{\text{mode}}}{N} \left(\sigma_{\text{mode}}^2 + \frac{1}{M^2} \|g_t^*\|^2 \right). \end{aligned} \quad (71)$$

Using $2 \langle g_t^*, g_t^{\omega} \rangle = \|g_t^*\|^2 + \|g_t^{\omega}\|^2 - \|g_t^{\omega} - g_t^*\|^2$, we obtain

$$\begin{aligned} &\mathbb{E}_t[F_{\phi}(\theta_{t+1})] - F_{\phi}(\theta_t) \\ &\geq \frac{\alpha}{2} \|g_t^*\|^2 - \frac{\alpha}{2} \|g_t^{\omega} - g_t^*\|^2 + \frac{\alpha}{2} \left[1 - L\alpha \left(1 - \frac{1}{N}\right) \right] \|g_t^{\omega}\|^2 \\ &\quad - \frac{L\alpha^2 Q_{\text{mode}}}{N} \left(\sigma_{\text{mode}}^2 + \frac{1}{M^2} \|g_t^*\|^2 \right) \\ &\stackrel{(a)}{\geq} \alpha \left(\frac{3}{4} - L\alpha \left(\frac{1}{4} \left(1 - \frac{1}{N}\right) + \frac{Q_{\text{mode}}}{NM^2} \right) \right) \|g_t^*\|^2 \\ &\quad - \alpha \|g_t^{\omega} - g_t^*\|^2 - \frac{L\alpha^2 \sigma_{\text{mode}}^2 Q_{\text{mode}}}{N} \\ &\stackrel{(b)}{\geq} \frac{3\alpha}{8} \|g_t^*\|^2 - \alpha \|g_t^{\omega} - g_t^*\|^2 - \frac{L\alpha^2 \sigma_{\text{mode}}^2 Q_{\text{mode}}}{N} \\ &\stackrel{(c)}{\geq} \frac{3\alpha}{8} \|g_t^*\|^2 - \alpha \frac{G^2 + M^2\sigma_{\text{mode}}^2}{M} \mathcal{E}_{t,\text{mode}}^{\omega} - \frac{L\alpha^2 \sigma_{\text{mode}}^2 Q_{\text{mode}}}{N}. \end{aligned} \quad (72)$$

Here, (a) uses $\|g_t^\omega\|^2 \geq \frac{1}{2}\|g_t^*\|^2 - \|g_t^\omega - g_t^*\|^2$, together with $\alpha \leq 1/[L(1 - 1/N)]$, which ensures $1 - L\alpha(1 - 1/N) \in [0, 1]$. Inequality (b) follows from $\alpha \leq 3/\left(8L\left(\frac{1}{4}(1 - 1/N) + \frac{Q_{\text{mode}}}{NM^2}\right)\right)$, and (c) applies the bias bound in equation 70.

Telescoping. Taking total expectation and summing equation 72 from $t = 0$ to $T - 1$ gives

$$\begin{aligned} \frac{3\alpha}{8} \sum_{t=0}^{T-1} \mathbb{E}\|g_t^*\|^2 &\leq \mathbb{E}F_\phi(\theta_T) - F_\phi(\theta_0) + \alpha \frac{G^2 + M^2\sigma_{\text{mode}}^2}{M} \sum_{t=0}^{T-1} \mathbb{E}\mathcal{E}_{t,\text{mode}}^\omega \\ &\quad + \frac{L\alpha^2 T \sigma_{\text{mode}}^2 Q_{\text{mode}}}{N}. \end{aligned} \quad (73)$$

Using $\mathbb{E}F_\phi(\theta_T) \leq F_\phi^{\max}$ and $\Delta_{\phi,0} = F_\phi^{\max} - F_\phi(\theta_0)$, then dividing by $3\alpha T/8$, gives

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|g_t^*\|^2 \leq \frac{8\Delta_{\phi,0}}{3\alpha T} + \frac{8(G^2 + M^2\sigma_{\text{mode}}^2)}{3M} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\mathcal{E}_{t,\text{mode}}^\omega + \frac{8L\alpha\sigma_{\text{mode}}^2 Q_{\text{mode}}}{3N}, \quad (74)$$

which is equation 61. $\square$

A.5 PROOF OF THEOREM 4.1

We first introduce the response-level representation of the oracle gradient and the approximation terms used in the proof.

Response-level setup. Fix a problem instance x and suppose that the response space admits a fixed partition

$$\mathcal{Y} = \bigcup_{m \in \mathcal{M}} \mathcal{Y}_m, \quad \mathcal{Y}_m \cap \mathcal{Y}_{m'} = \emptyset \quad (m \neq m'). \quad (75)$$

For $y \in \mathcal{Y}_m$, write

$$\pi_t(y \mid x) = q_t(m)\rho_t(y \mid x, m). \quad (76)$$

where $\rho_t(\cdot \mid x, m)$ denotes the conditional response distribution within mode m . We assume $q_t(m) > 0$ for every $m \in \mathcal{M}$ over the rounds under consideration, so that the relative correction terms below are well defined.

Assume $|R(y)| \leq R_{\max}$. For each mode m , let $\mathcal{P}_m$ be a fixed admissible family containing $\rho_t(\cdot \mid x, m)$ for all rounds under consideration, and define

$$r_m := \sup_{\rho \in \mathcal{P}_m} \mathbb{E}_{\rho(\cdot \mid x, m)}[R(Y)]. \quad (77)$$

The current within-mode expected reward is

$$\bar{r}_t(m) := \mathbb{E}_{\rho_t(\cdot \mid x, m)}[R(Y)], \quad (78)$$

so $\bar{r}_t(m) \leq r_m$. The reward function, admissible families $\{\mathcal{P}_m\}$, and oracle mode rewards $\{r_m\}$ are independent of θ . Conditionally on $\mathcal{F}_t$, a fresh response is distributed as $Y \sim \pi_t(\cdot \mid x)$, equivalently by first sampling $m \sim q_t$ and then $Y \sim \rho_t(\cdot \mid x, m)$. Thus, $\mathbb{E}_t$ averages over the current-round response sampling, whereas unsubscripted $\mathbb{E}$ additionally averages over the randomness of the preceding training trajectory.

Response representation of the oracle gradient. For a fixed response y , define

$$\psi_t(y) := \nabla_\theta \log \pi_t(y \mid x).$$

For $y \in \mathcal{Y}_m$, the factorization $\pi_t(y \mid x) = q_t(m)\rho_t(y \mid x, m)$ gives

$$\psi_t(y) = \nabla_\theta \log q_t(m) + \nabla_\theta \log \rho_t(y \mid x, m). \quad (79)$$

Since the conditional score has zero mean,

$$\mathbb{E}_{\rho_t(\cdot \mid x, m)}[\nabla_\theta \log \rho_t(Y \mid x, m)] = 0,$$

we have

$$\begin{aligned}
& \mathbb{E}_t [r_{m(Y)} \phi(q_t(m(Y))) \psi_t(Y)] \\
&= \sum_{m \in \mathcal{M}} q_t(m) r_m \phi(q_t(m)) \mathbb{E}_{\rho_t(\cdot | x, m)} [\psi_t(Y)] \\
&= \sum_{m \in \mathcal{M}} r_m q_t(m) \phi(q_t(m)) \nabla_\theta \log q_t(m) = g_t^*.
\end{aligned} \tag{80}$$

This establishes the response-level representation in equation 10.

Response-level approximation. Conditionally on $\mathcal{F}_t$, let $Y_{t,1}, \dots, Y_{t,N} \stackrel{\text{i.i.d.}}{\sim} \pi_t(\cdot | x)$ and define

$$\hat{g}_t := \frac{1}{N} \sum_{i=1}^N R(Y_{t,i}) \omega_t(Y_{t,i}) \psi_t(Y_{t,i}). \tag{81}$$

Using the response representation of g_t^* , the difference between the conditional mean of $\hat{g}_t$ and the oracle gradient decomposes as

$$\mathbb{E}_t[\hat{g}_t] - g_t^* = \delta_{t,\phi}^\omega + \delta_{t,\phi}^r, \tag{82}$$

where

$$\begin{aligned}
\delta_{t,\phi}^\omega &:= \mathbb{E}_t [R(Y) (\omega_t(Y) - \phi(q_t(m(Y)))) \psi_t(Y)], \\
\delta_{t,\phi}^r &:= \mathbb{E}_t [\phi(q_t(m(Y))) (R(Y) - r_{m(Y)}) \psi_t(Y)].
\end{aligned} \tag{83}$$

The first term captures the mismatch between the response-level correction and the oracle mode-level correction, whereas the second collects the residual induced by replacing the oracle mode reward $r_{m(Y)}$ with the realized response reward $R(Y)$.

We quantify the correction mismatch by

$$\mathcal{E}_t := \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot | x, m)} \left[\left(\frac{\omega_t(Y)}{\phi(q_t(m))} - 1 \right)^2 \right]. \tag{84}$$

Assumption A.6 (Bounded response-gradient dispersion). There exists $\sigma^2 < \infty$ such that

$$\left\| q_t(m) \phi(q_t(m)) R(y) \psi_t(y) - \frac{1}{M} g_t^* \right\|^2 \leq \sigma^2 \tag{85}$$

for every t , every $m \in \mathcal{M}$, and $\rho_t(\cdot | x, m)$ -almost every y .

Unlike the mode-level case, an individual response-level gradient contribution need not average exactly to g_t^*/M , because it contains the realized response reward $R(y)$. The assumption therefore controls both within-mode response variability and its deviation from the oracle-gradient scale.

Assumption A.7 (Bounded response-level correction-weight second moment). There exists $Q < \infty$ such that, almost surely,

$$\sum_{m \in \mathcal{M}} \frac{\mathbb{E}_{\rho_t(\cdot | x, m)} [\omega_t(Y)^2]}{q_t(m) \phi(q_t(m))^2} \leq Q \tag{86}$$

for every round under consideration.

The first assumption controls the dispersion of individual response-level gradient contributions around the oracle gradient, while the second bounds the second-moment amplification caused by response sampling and correction reweighting. We next establish several auxiliary bounds before completing the proof of Theorem 4.1.

Lemma A.8 (Response-level correction bias and second moment). *Under Assumptions A.6 and A.7,*

$$\|\delta_{t,\phi}^\omega\|^2 \leq \frac{2}{M} (\|g_t^*\|^2 + M^2 \sigma^2) \mathcal{E}_t. \tag{87}$$

Moreover,

$$\mathbb{E}_t [\|R(Y) \omega_t(Y) \psi_t(Y)\|^2] \leq 2Q \left(\sigma^2 + \frac{1}{M^2} \|g_t^*\|^2 \right). \tag{88}$$

Proof. We first bound the bias induced by the response-level correction. By the definition of $\delta_{t,\phi}^\omega$,

$$\begin{aligned}
\delta_{t,\phi}^\omega &= \sum_{m \in \mathcal{M}} q_t(m) \mathbb{E}_{\rho_t(\cdot|x,m)} [R(Y) (\omega_t(Y) - \phi(q_t(m))) \psi_t(Y)] \\
&= \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\left(\frac{\omega_t(Y)}{\phi(q_t(m))} - 1 \right) q_t(m) \phi(q_t(m)) R(Y) \psi_t(Y) \right] \\
&= \frac{1}{M} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\frac{\omega_t(Y)}{\phi(q_t(m))} - 1 \right] \right) g_t^* \\
&\quad + \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\left(\frac{\omega_t(Y)}{\phi(q_t(m))} - 1 \right) \left(q_t(m) \phi(q_t(m)) R(Y) \psi_t(Y) - \frac{1}{M} g_t^* \right) \right]. \quad (89)
\end{aligned}$$

For the first term, Cauchy–Schwarz and Jensen’s inequality give

$$\begin{aligned}
&\left\| \frac{1}{M} \left[\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\frac{\omega_t(Y)}{\phi(q_t(m))} - 1 \right] \right] g_t^* \right\|^2 \\
&\leq \frac{1}{M} \|g_t^*\|^2 \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\left(\frac{\omega_t(Y)}{\phi(q_t(m))} - 1 \right)^2 \right] \\
&= \frac{1}{M} \|g_t^*\|^2 \mathcal{E}_t. \quad (90)
\end{aligned}$$

For the second term, Cauchy–Schwarz and Assumption A.6 yield

$$\begin{aligned}
&\left\| \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\left(\frac{\omega_t(Y)}{\phi(q_t(m))} - 1 \right) \left(q_t(m) \phi(q_t(m)) R(Y) \psi_t(Y) - \frac{1}{M} g_t^* \right) \right] \right\|^2 \\
&\leq M \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\left(\frac{\omega_t(Y)}{\phi(q_t(m))} - 1 \right)^2 \right] \\
&\quad \cdot \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\left\| q_t(m) \phi(q_t(m)) R(Y) \psi_t(Y) - \frac{1}{M} g_t^* \right\|^2 \right] \\
&\leq M \sigma^2 \mathcal{E}_t. \quad (91)
\end{aligned}$$

Applying $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$ to equation 89 and using equation 90 and equation 91 gives

$$\begin{aligned}
\|\delta_{t,\phi}^\omega\|^2 &\leq \frac{2}{M} \|g_t^*\|^2 \mathcal{E}_t + 2M \sigma^2 \mathcal{E}_t \\
&= \frac{2}{M} (\|g_t^*\|^2 + M^2 \sigma^2) \mathcal{E}_t, \quad (92)
\end{aligned}$$

which proves equation 87.

We next bound the second moment of a single response-level gradient sample. Conditioning on the latent mode,

$$\begin{aligned}
&\mathbb{E}_t \left[\|R(Y) \omega_t(Y) \psi_t(Y)\|^2 \right] \\
&= \sum_{m \in \mathcal{M}} q_t(m) \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\omega_t(Y)^2 R(Y)^2 \|\psi_t(Y)\|^2 \right] \\
&= \sum_{m \in \mathcal{M}} \frac{1}{q_t(m) \phi(q_t(m))^2} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\omega_t(Y)^2 \|q_t(m) \phi(q_t(m)) R(Y) \psi_t(Y)\|^2 \right]. \quad (93)
\end{aligned}$$

By Assumption A.6,

$$\begin{aligned} & \|q_t(m)\phi(q_t(m))R(Y)\psi_t(Y)\|^2 \\ & \leq 2 \left\| q_t(m)\phi(q_t(m))R(Y)\psi_t(Y) - \frac{1}{M}g_t^* \right\|^2 + \frac{2}{M^2}\|g_t^*\|^2 \\ & \leq 2\sigma^2 + \frac{2}{M^2}\|g_t^*\|^2. \end{aligned} \quad (94)$$

Substituting this bound into equation 93 and applying Assumption A.7 yields

$$\begin{aligned} & \mathbb{E}_t \left[\|R(Y)\omega_t(Y)\psi_t(Y)\|^2 \right] \\ & \leq 2 \left(\sigma^2 + \frac{1}{M^2}\|g_t^*\|^2 \right) \sum_{m \in \mathcal{M}} \frac{\mathbb{E}_{\rho_t(\cdot|x,m)} [\omega_t(Y)^2]}{q_t(m)\phi(q_t(m))^2} \\ & \leq 2Q \left(\sigma^2 + \frac{1}{M^2}\|g_t^*\|^2 \right), \end{aligned} \quad (95)$$

which proves equation 88. $\square$

Completion of the proof. Having established the correction-bias and second-moment bounds, we now combine them with the smoothness of F_ϕ .

Proof. Recall that

$$\theta_{t+1} = \theta_t + \alpha \hat{g}_t, \quad (96)$$

and define the conditional mean

$$\tilde{g}_t := \mathbb{E}_t[\hat{g}_t]. \quad (97)$$

By equation 82,

$$\tilde{g}_t - g_t^* = \delta_{t,\phi}^\omega + \delta_{t,\phi}^r. \quad (98)$$

Conditional independence of the N sampled responses gives

$$\begin{aligned} \mathbb{E}_t \|\hat{g}_t\|^2 &= \left(1 - \frac{1}{N} \right) \|\tilde{g}_t\|^2 \\ &+ \frac{1}{N} \mathbb{E}_t \left[\|R(Y)\omega_t(Y)\psi_t(Y)\|^2 \right]. \end{aligned} \quad (99)$$

Applying Lemma A.8 therefore yields

$$\mathbb{E}_t \|\hat{g}_t\|^2 \leq \left(1 - \frac{1}{N} \right) \|\tilde{g}_t\|^2 + \frac{2Q}{N} \left(\sigma^2 + \frac{1}{M^2}\|g_t^*\|^2 \right). \quad (100)$$

By Assumption A.2, the L -smoothness of F_ϕ along the update segment gives

$$\begin{aligned} & \mathbb{E}_t [F_\phi(\theta_{t+1})] - F_\phi(\theta_t) \\ & \geq \alpha \langle g_t^*, \tilde{g}_t \rangle - \frac{L\alpha^2}{2} \mathbb{E}_t \|\tilde{g}_t\|^2 \\ & \geq \alpha \langle g_t^*, \tilde{g}_t \rangle - \frac{L\alpha^2}{2} \left(1 - \frac{1}{N} \right) \|\tilde{g}_t\|^2 - \frac{L\alpha^2 Q}{N} \left(\sigma^2 + \frac{1}{M^2}\|g_t^*\|^2 \right) \\ & = \frac{\alpha}{2} \|g_t^*\|^2 - \frac{\alpha}{2} \|\tilde{g}_t - g_t^*\|^2 + \frac{\alpha}{2} \left(1 - L\alpha \left(1 - \frac{1}{N} \right) \right) \|\tilde{g}_t\|^2 - \frac{L\alpha^2 Q}{N} \left(\sigma^2 + \frac{1}{M^2}\|g_t^*\|^2 \right) \\ & \stackrel{(a)}{\geq} \alpha \left(\frac{3}{4} - L\alpha \left(\frac{1}{4} \left(1 - \frac{1}{N} \right) + \frac{Q}{NM^2} \right) \right) \|g_t^*\|^2 - \alpha \|\tilde{g}_t - g_t^*\|^2 - \frac{L\alpha^2 \sigma^2 Q}{N} \\ & \stackrel{(b)}{\geq} \frac{3\alpha}{8} \|g_t^*\|^2 - \alpha \|\tilde{g}_t - g_t^*\|^2 - \frac{L\alpha^2 \sigma^2 Q}{N}. \end{aligned} \quad (101)$$

Here, (a) uses $\|\tilde{g}_t\|^2 \geq \frac{1}{2}\|g_t^*\|^2 - \|\tilde{g}_t - g_t^*\|^2$ together with $\alpha \leq 1/(L(1 - 1/N))$, which ensures $1 - L\alpha(1 - 1/N) \in [0, 1]$. Inequality (b) follows from $\alpha \leq 3/(8L(\frac{1}{4}(1 - 1/N) + Q/(NM^2)))$.

By equation 98 and Young's inequality,

$$\begin{aligned}\|\tilde{g}_t - g_t^*\|^2 &\leq 2\|\delta_{t,\phi}^\omega\|^2 + 2\|\delta_{t,\phi}^r\|^2 \\ &\leq \frac{4}{M} (G^2 + M^2\sigma^2) \mathcal{E}_t + 2\|\delta_{t,\phi}^r\|^2,\end{aligned}\quad (102)$$

where the second inequality follows from Lemma A.8 and $\|g_t^*\|^2 \leq G^2$ from Assumption A.2.

Substituting equation 102 into equation 101 gives

$$\mathbb{E}_t[F_\phi(\theta_{t+1})] - F_\phi(\theta_t) \geq \frac{3\alpha}{8} \|g_t^*\|^2 - \frac{4\alpha}{M} (G^2 + M^2\sigma^2) \mathcal{E}_t - 2\alpha\|\delta_{t,\phi}^r\|^2 - \frac{L\alpha^2\sigma^2Q}{N}. \quad (103)$$

Taking total expectation and summing equation 103 over $t = 0, \dots, T-1$ yields

$$\begin{aligned}\frac{3\alpha}{8} \sum_{t=0}^{T-1} \mathbb{E}\|g_t^*\|^2 &\leq \mathbb{E}F_\phi(\theta_T) - F_\phi(\theta_0) + \frac{4\alpha}{M} (G^2 + M^2\sigma^2) \sum_{t=0}^{T-1} \mathbb{E}\mathcal{E}_t \\ &\quad + 2\alpha \sum_{t=0}^{T-1} \mathbb{E}\|\delta_{t,\phi}^r\|^2 + \frac{L\alpha^2T\sigma^2Q}{N}.\end{aligned}\quad (104)$$

Since $\mathbb{E}F_\phi(\theta_T) \leq F_\phi^{\max}$ and $\Delta_{\phi,0} = F_\phi^{\max} - F_\phi(\theta_0)$, dividing by $3\alpha T/8$ gives

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|g_t^*\|^2 \leq \frac{8\Delta_{\phi,0}}{3\alpha T} + \frac{32(G^2 + M^2\sigma^2)}{3M} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\mathcal{E}_t + \frac{16}{3} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|\delta_{t,\phi}^r\|^2 + \frac{8L\alpha\sigma^2Q}{3N}, \quad (105)$$

which proves Theorem 4.1. $\square$

A.6 PROPERTIES OF THE OBSERVABLE BOUNDED CORRECTION

Theorem 4.1 is stated for a generic response-level correction ω_t and depends on the correction mismatch $\mathcal{E}_t$ and a uniform correction-weight second-moment bound Q . We now specialize to $\omega_t = \hat{\omega}_t$ and $\phi = \phi_{c,d}$, the bounded arctangent correction used in the main text.

To separate historical estimation error, let $\bar{\omega}_t(y) := 1 + \frac{2\tau}{\pi} \arctan(\gamma Z_t(y))$ denote the same bounded correction constructed from the population statistic $Z_t(y)$. By construction,

$$1 - \tau < \bar{\omega}_t(y), \hat{\omega}_t(y), \phi_{c,d}(q) < 1 + \tau. \quad (106)$$

Moreover, the 1-Lipschitz property of $\arctan$ gives

$$|\hat{\omega}_t(y) - \bar{\omega}_t(y)| \leq \frac{2\tau\gamma}{\pi} \left| \hat{Z}_t(y) - Z_t(y) \right|. \quad (107)$$

Decomposition of the correction mismatch. Recall the response-to-mode residual $\varepsilon_t(y, m)$ introduced in the main text. It is the difference between the population response-level argument $\gamma Z_t(y)$ and the mode-level oracle argument $c \log q_t(m)^{-1} - d$ before the bounded arctangent transformation. Hence, the mismatch between the response-level correction $\hat{\omega}_t$ and the mode-level oracle correction $\phi_{c,d}$ has two sources: the response-to-mode residual ε_t and the additional estimation error $\hat{Z}_t - Z_t$ induced by the historical reference.

Lemma A.9 (Observable correction mismatch). *For every round t ,*

$$\mathcal{E}_t \leq \frac{8\tau^2}{\pi^2(1-\tau)^2} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} \left[\varepsilon_t(Y, m)^2 + \gamma^2 \left(\hat{Z}_t(Y) - Z_t(Y) \right)^2 \right]. \quad (108)$$

Proof. For $Y \in \mathcal{Y}_m$, the definition of $\varepsilon_t(Y, m)$ and the 1-Lipschitz property of $\arctan$ imply

$$\left| \frac{\bar{\omega}_t(Y)}{\phi_{c,d}(q_t(m))} - 1 \right| \leq \frac{2\tau}{\pi(1-\tau)} |\varepsilon_t(Y, m)|. \quad (109)$$

Moreover, by equation 106 and equation 107,

$$\frac{|\hat{\omega}_t(Y) - \bar{\omega}_t(Y)|}{\phi_{c,d}(q_t(m))} \leq \frac{2\tau\gamma}{\pi(1-\tau)} \left| \hat{Z}_t(Y) - Z_t(Y) \right|. \quad (110)$$

Hence, using $(u+v)^2 \leq 2u^2 + 2v^2$,

$$\left(\frac{\hat{\omega}_t(Y)}{\phi_{c,d}(q_t(m))} - 1 \right)^2 \leq \frac{8\tau^2}{\pi^2(1-\tau)^2} \left(\varepsilon_t(Y, m)^2 + \gamma^2 \left(\hat{Z}_t(Y) - Z_t(Y) \right)^2 \right). \quad (111)$$

Taking the within-mode expectation and summing over $m \in \mathcal{M}$ proves the result. $\square$

Uniform correction-weight second moment. The preceding lemma controls the correction mismatch $\mathcal{E}_t$. We next verify the uniform second-moment condition on the correction weights required by Theorem 4.1.

Lemma A.10 (Uniform second-moment bound for bounded corrections). *Suppose there exists $\underline{q} > 0$ such that*

$$q_t(m) \geq \underline{q}, \quad m \in \mathcal{M}, \quad (112)$$

over the rounds under consideration. Then

$$\sum_{m \in \mathcal{M}} \frac{\mathbb{E}_{\rho_t(\cdot|x, m)} [\hat{\omega}_t(Y)^2]}{q_t(m)\phi_{c,d}(q_t(m))^2} \leq \frac{M(1+\tau)^2}{\underline{q}(1-\tau)^2}. \quad (113)$$

Thus, $Q = M(1+\tau)^2/(\underline{q}(1-\tau)^2)$ is a valid deterministic choice in Theorem 4.1.

Proof. By equation 106, $\hat{\omega}_t(Y) \leq 1 + \tau$ and $\phi_{c,d}(q) \geq 1 - \tau$. Therefore,

$$\begin{aligned} \sum_{m \in \mathcal{M}} \frac{\mathbb{E}_{\rho_t(\cdot|x, m)} [\hat{\omega}_t(Y)^2]}{q_t(m)\phi_{c,d}(q_t(m))^2} &\leq \frac{(1+\tau)^2}{(1-\tau)^2} \sum_{m \in \mathcal{M}} \frac{1}{q_t(m)} \\ &\leq \frac{M(1+\tau)^2}{\underline{q}(1-\tau)^2}, \end{aligned} \quad (114)$$

which proves the result. $\square$

It remains to control the historical estimation error $\hat{Z}_t(Y) - Z_t(Y)$. The next subsection bounds this term for the historical-reference estimator used in the practical correction.

A.7 PROOF OF PROPOSITION 4.2

We first establish the auxiliary bounds needed to control the error introduced by the historical-reference statistics. Recall that μ_t and v_t are the policy-level mean and per-token dispersion of token surprisal used to standardize response surprisal. Since the policy evolves during training, these population statistics also drift over time. We first control this drift, then construct conditionally unbiased per-round estimates, and finally analyze the lagged EMA used in the practical method.

Assumption A.11 (Bounded surprisal and log-policy gradients). Along the parameter region under consideration, the response distribution has common support and differentiation may be interchanged with expectation. Moreover, there exist finite constants G_s and $\ell_{\max}$ such that, for every relevant θ , y , and $j \leq H(y)$,

$$\begin{aligned} \|\nabla_{\theta} \log \pi_{\theta}(y | x)\| &\leq G_s \\ \frac{1}{H(y)} \sum_{j=1}^{H(y)} \|\nabla_{\theta} \log \pi_{\theta}(y_j | x, y_{<j})\| &\leq G_s \\ 0 \leq -\log \pi_{\theta}(y_j | x, y_{<j}) &\leq \ell_{\max}. \end{aligned} \quad (115)$$

Lemma A.12 (Lipschitz continuity of population surprisal statistics). *Under Assumption A.11,*

$$\begin{aligned} |\mu(\theta') - \mu(\theta)| &\leq L_\mu \|\theta' - \theta\|, \\ |v(\theta') - v(\theta)| &\leq L_v \|\theta' - \theta\|, \end{aligned} \quad (116)$$

where

$$L_\mu = G_s(\ell_{\max} + 1), \quad L_v = G_s(\ell_{\max}^2 + 2\ell_{\max}).$$

Proof. By the definition of the population mean token surprisal,

$$\begin{aligned} \nabla_\theta \mu(\theta) &= \mathbb{E}_{\pi_\theta} \left[\left(\frac{1}{H(Y)} \sum_{j=1}^{H(Y)} -\log \pi_\theta(Y_j | x, Y_{<j}) \right) \nabla_\theta \log \pi_\theta(Y | x) \right] \\ &\quad - \mathbb{E}_{\pi_\theta} \left[\frac{1}{H(Y)} \sum_{j=1}^{H(Y)} \nabla_\theta \log \pi_\theta(Y_j | x, Y_{<j}) \right]. \end{aligned} \quad (117)$$

Assumption A.11 therefore gives

$$\|\nabla_\theta \mu(\theta)\| \leq G_s \ell_{\max} + G_s = L_\mu. \quad (118)$$

For the dispersion,

$$\begin{aligned} \nabla_\theta v(\theta) &= \mathbb{E}_{\pi_\theta} \left[\frac{1}{H(Y)} \sum_{j=1}^{H(Y)} (-\log \pi_\theta(Y_j | x, Y_{<j}) - \mu(\theta))^2 \nabla_\theta \log \pi_\theta(Y | x) \right] \\ &\quad - 2\mathbb{E}_{\pi_\theta} \left[\frac{1}{H(Y)} \sum_{j=1}^{H(Y)} (-\log \pi_\theta(Y_j | x, Y_{<j}) - \mu(\theta)) \nabla_\theta \log \pi_\theta(Y_j | x, Y_{<j}) \right] \\ &\quad - 2\mathbb{E}_{\pi_\theta} \left[\frac{1}{H(Y)} \sum_{j=1}^{H(Y)} (-\log \pi_\theta(Y_j | x, Y_{<j}) - \mu(\theta)) \right] \nabla_\theta \mu(\theta). \end{aligned} \quad (119)$$

The last term vanishes because the expectation inside it is zero by the definition of $\mu(\theta)$. Moreover, $0 \leq \mu(\theta) \leq \ell_{\max}$, and hence

$$|-\log \pi_\theta(Y_j | x, Y_{<j}) - \mu(\theta)| \leq \ell_{\max}.$$

Applying Assumption A.11 to the remaining terms yields

$$\|\nabla_\theta v(\theta)\| \leq G_s \ell_{\max}^2 + 2G_s \ell_{\max} = L_v. \quad (120)$$

The mean-value theorem proves equation 116. $\square$

This Lipschitz property allows us to control the round-to-round changes in μ_t and v_t through the parameter update $\theta_{t+1} - \theta_t$.

Population-statistic drift across policy updates. For $\xi \in \{\mu, v\}$, let L_ξ denote the corresponding Lipschitz constant in Lemma A.12. By Jensen's inequality, Lemma A.8, and $\|g_t^*\|^2 \leq G^2$ from Assumption A.2,

$$\mathbb{E}_t \|\hat{g}_t\|^2 \leq 2Q \left(\frac{G^2}{M^2} + \sigma^2 \right). \quad (121)$$

Since $\theta_{t+1} = \theta_t + \alpha \hat{g}_t$, Lemma A.12 implies

$$\|\xi_{t+1} - \xi_t\|_{L_2} \leq L_\xi \alpha \sqrt{2Q \left(\frac{G^2}{M^2} + \sigma^2 \right)}. \quad (122)$$

Here, $\|X\|_{L_2} := (\mathbb{E}|X|^2)^{1/2}$ denotes the L_2 norm of a scalar random variable X .

With the target drift controlled, we next construct unbiased round-wise observations of μ_t and v_t .

Per-round statistic estimates. Conditionally on $\mathcal{F}_t$, let $Y_{t,1}, \dots, Y_{t,N}$ be the N independently sampled responses at round t . Define

$$\tilde{\mu}_t := \frac{1}{N} \sum_{i=1}^N \frac{1}{H(Y_{t,i})} \sum_{j=1}^{H(Y_{t,i})} -\log \pi_t(Y_{t,i,j} \mid x, Y_{t,i,<j}). \quad (123)$$

For $N \geq 2$, define

$$\begin{aligned} \tilde{v}_t := & \frac{1}{N} \sum_{i=1}^N \frac{1}{H(Y_{t,i})} \sum_{j=1}^{H(Y_{t,i})} [-\log \pi_t(Y_{t,i,j} \mid x, Y_{t,i,<j})]^2 \\ & - \frac{1}{N(N-1)} \sum_{i \neq k} \left(\frac{1}{H(Y_{t,i})} \sum_{j=1}^{H(Y_{t,i})} -\log \pi_t(Y_{t,i,j} \mid x, Y_{t,i,<j}) \right) \\ & \cdot \left(\frac{1}{H(Y_{t,k})} \sum_{j=1}^{H(Y_{t,k})} -\log \pi_t(Y_{t,k,j} \mid x, Y_{t,k,<j}) \right). \end{aligned} \quad (124)$$

Lemma A.13 (Per-round estimation error). *Under Assumption A.11, the estimates in equation 123 and equation 124 are conditionally unbiased:*

$$\mathbb{E}[\tilde{\mu}_t \mid \mathcal{F}_t] = \mu_t, \quad \mathbb{E}[\tilde{v}_t \mid \mathcal{F}_t] = v_t. \quad (125)$$

Moreover, there exist finite constants σ_μ^2 and σ_v^2 such that

$$\mathbb{E}[(\tilde{\xi}_t - \xi_t)^2 \mid \mathcal{F}_t] \leq \frac{\sigma_\xi^2}{N}, \quad \xi \in \{\mu, v\}. \quad (126)$$

Consequently, $\tilde{\xi}_t - \xi_t$ is a martingale difference with respect to the training filtration.

Proof. The conditional unbiasedness of $\tilde{\mu}_t$ follows directly from the definition of μ_t and the conditional i.i.d. sampling.

For the dispersion,

$$v_t = \mathbb{E}_t \left[\frac{1}{H(Y)} \sum_{j=1}^{H(Y)} [-\log \pi_t(Y_j \mid x, Y_{<j})]^2 \right] - \mu_t^2. \quad (127)$$

For $i \neq k$, conditional independence gives

$$\begin{aligned} \mathbb{E} \left[\left(\frac{1}{H(Y_{t,i})} \sum_{j=1}^{H(Y_{t,i})} -\log \pi_t(Y_{t,i,j} \mid x, Y_{t,i,<j}) \right) \right. \\ \left. \cdot \left(\frac{1}{H(Y_{t,k})} \sum_{j=1}^{H(Y_{t,k})} -\log \pi_t(Y_{t,k,j} \mid x, Y_{t,k,<j}) \right) \middle| \mathcal{F}_t \right] = \mu_t^2. \end{aligned} \quad (128)$$

Substituting this identity into equation 124 proves $\mathbb{E}[\tilde{v}_t \mid \mathcal{F}_t] = v_t$.

Finally, the response-level quantities in equation 123 and equation 124 are uniformly bounded under Assumption A.11. The mean estimator is an average of N conditionally independent bounded response-level observations, while the second term of $\tilde{v}_t$ is a bounded second-order U-statistic. Standard conditional variance bounds therefore give equation 126. $\square$

These per-round estimates provide the observations for the lagged EMA. The next lemma bounds its tracking error due to sampling noise and population-statistic drift.

Lemma A.14 (Lagged-EMA historical-reference tracking error). *For $\xi \in \{\mu, v\}$, let $\tilde{\xi}_t$ denote the corresponding historical state before any projection:*

$$\tilde{\xi}_t := \begin{cases} \hat{\mu}_t, & \xi = \mu, \\ \bar{v}_t, & \xi = v. \end{cases} \quad (129)$$

After the round- t gradient has been formed, update

$$\check{\xi}_{t+1} = (1 - \beta)\check{\xi}_t + \beta\tilde{\xi}_t, \quad 0 < \beta \leq 1. \quad (130)$$

Then

$$\mathbb{E}[(\check{\xi}_t - \xi_t)^2] \leq 3(1 - \beta)^{2t} \mathbb{E}[(\check{\xi}_0 - \xi_0)^2] + \frac{3\beta\sigma_\xi^2}{N(2 - \beta)} + \frac{6L_\xi^2 Q \alpha^2}{\beta^2} \left(\frac{G^2}{M^2} + \sigma^2 \right). \quad (131)$$

Moreover,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[(\check{\xi}_t - \xi_t)^2] \leq \frac{3\mathbb{E}[(\check{\xi}_0 - \xi_0)^2][1 - (1 - \beta)^{2T}]}{T\beta(2 - \beta)} + \frac{3\beta\sigma_\xi^2}{N(2 - \beta)} + \frac{6L_\xi^2 Q \alpha^2}{\beta^2} \left(\frac{G^2}{M^2} + \sigma^2 \right). \quad (132)$$

Proof. Let

$$e_t := \check{\xi}_t - \xi_t.$$

Using equation 130,

$$e_{t+1} = (1 - \beta)e_t + \beta(\tilde{\xi}_t - \xi_t) + (\xi_t - \xi_{t+1}). \quad (133)$$

Unrolling the recursion gives

$$e_t = (1 - \beta)^t e_0 + \beta \sum_{s=0}^{t-1} (1 - \beta)^{t-1-s} (\tilde{\xi}_s - \xi_s) + \sum_{s=0}^{t-1} (1 - \beta)^{t-1-s} (\xi_s - \xi_{s+1}). \quad (134)$$

By Lemma A.13, the estimation errors form a martingale-difference sequence. Therefore,

$$\left\| \beta \sum_{s=0}^{t-1} (1 - \beta)^{t-1-s} (\tilde{\xi}_s - \xi_s) \right\|_{L_2}^2 \leq \frac{\beta^2 \sigma_\xi^2}{N} \sum_{j=0}^{t-1} (1 - \beta)^{2j} \leq \frac{\beta \sigma_\xi^2}{N(2 - \beta)}. \quad (135)$$

For the drift component, equation 122 and Minkowski's inequality give

$$\begin{aligned} & \left\| \sum_{s=0}^{t-1} (1 - \beta)^{t-1-s} (\xi_s - \xi_{s+1}) \right\|_{L_2} \\ & \leq L_\xi \alpha \sqrt{2Q \left(\frac{G^2}{M^2} + \sigma^2 \right)} \sum_{j=0}^{t-1} (1 - \beta)^j \leq \frac{L_\xi \alpha}{\beta} \sqrt{2Q \left(\frac{G^2}{M^2} + \sigma^2 \right)}. \end{aligned} \quad (136)$$

Combining the three terms in equation 134 and using $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ proves equation 131.

Finally,

$$\sum_{t=0}^{T-1} (1 - \beta)^{2t} = \frac{1 - (1 - \beta)^{2T}}{\beta(2 - \beta)}, \quad (137)$$

which gives equation 132. $\square$

The preceding result controls the estimation errors of the historical mean and dispersion statistics. To translate these bounds into an error bound for the standardized response score, we additionally control the quantities entering the standardization.

Assumption A.15 (Bounds for response standardization). There exist finite constants $H_{\max}$, R_S , and $0 < v_{\min} < v_{\max} < \infty$ such that, for all relevant responses and rounds,

$$\begin{aligned} 1 & \leq H(Y) \leq H_{\max}, \\ v_{\min} & \leq v_t \leq v_{\max}, \\ |S_t(Y) - H(Y)\mu_t| & \leq R_S. \end{aligned} \quad (138)$$

The per-round estimator $\hat{v}_t$ and its unprojected historical state $\bar{v}_t$ need not remain positive at finite sample size. Therefore, when constructing the standardized response score, we use

$$\hat{v}_t = \Pi_{[v_{\min}, v_{\max}]}(\bar{v}_t). \quad (139)$$

Since $v_t \in [v_{\min}, v_{\max}]$ under Assumption A.15, nonexpansiveness of Euclidean projection gives

$$\mathbb{E} [(\hat{v}_t - v_t)^2] \leq \mathbb{E} [(\bar{v}_t - v_t)^2]. \quad (140)$$

Thus, Lemma A.14 also controls the projected variance used in the standardized score.

Error propagation through response standardization. We next propagate the estimation errors in the historical mean and dispersion through the response-standardization map to bound $\hat{Z}_t - Z_t$.

Lemma A.16 (Historical-reference error in the response-level statistic). *Under Assumption A.15, for every $m \in \mathcal{M}$,*

$$\mathbb{E} \mathbb{E}_{\rho_t(\cdot|x, m)} [|\hat{Z}_t(Y) - Z_t(Y)|^2] \leq \frac{2H_{\max}}{v_{\min}} \mathbb{E} [(\hat{\mu}_t - \mu_t)^2] + \frac{R_S^2}{2v_{\min}^3} \mathbb{E} [(\hat{v}_t - v_t)^2]. \quad (141)$$

Proof. Direct subtraction gives

$$\hat{Z}_t(Y) - Z_t(Y) = -\frac{H(Y)(\hat{\mu}_t - \mu_t)}{\sqrt{H(Y)\hat{v}_t}} + (S_t(Y) - H(Y)\mu_t) \left(\frac{1}{\sqrt{H(Y)\hat{v}_t}} - \frac{1}{\sqrt{H(Y)v_t}} \right). \quad (142)$$

The coefficient of $\hat{\mu}_t - \mu_t$ is bounded by $\sqrt{H_{\max}/v_{\min}}$. Moreover,

$$\left| \frac{1}{\sqrt{H\hat{v}_t}} - \frac{1}{\sqrt{Hv_t}} \right| = \frac{|\hat{v}_t - v_t|}{\sqrt{H}\sqrt{\hat{v}_t}\sqrt{v_t}(\sqrt{\hat{v}_t} + \sqrt{v_t})} \leq \frac{|\hat{v}_t - v_t|}{2v_{\min}^{3/2}}. \quad (143)$$

Applying $(a+b)^2 \leq 2a^2 + 2b^2$ and equation 138, followed by expectation, proves equation 141. $\square$

Completion of the proof.

Proof. By Lemma A.9, Lemma A.16, Lemma A.14, and equation 140,

$$\begin{aligned} \mathbb{E} \mathcal{E}_t &\stackrel{(a)}{\leq} \frac{8\tau^2}{\pi^2(1-\tau)^2} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x, m)} (\varepsilon_t(Y, m)^2) \\ &\quad + \frac{8\tau^2\gamma^2}{\pi^2(1-\tau)^2} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x, m)} [|\hat{Z}_t(Y) - Z_t(Y)|^2] \\ &\stackrel{(b)}{\leq} \frac{8\tau^2}{\pi^2(1-\tau)^2} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x, m)} [\varepsilon_t(Y, m)^2] \\ &\quad + \frac{8\tau^2\gamma^2 M}{\pi^2(1-\tau)^2} \left(\frac{2H_{\max}}{v_{\min}} \mathbb{E}(\hat{\mu}_t - \mu_t)^2 + \frac{R_S^2}{2v_{\min}^3} \mathbb{E}(\hat{v}_t - v_t)^2 \right) \\ &\stackrel{(c)}{\leq} \frac{8\tau^2}{\pi^2(1-\tau)^2} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x, m)} [\varepsilon_t(Y, m)^2] \\ &\quad + \frac{24\tau^2\gamma^2 M}{\pi^2(1-\tau)^2} (1-\beta)^{2t} \left(\frac{2H_{\max}}{v_{\min}} \mathbb{E}(\hat{\mu}_0 - \mu_0)^2 + \frac{R_S^2}{2v_{\min}^3} \mathbb{E}(\bar{v}_0 - v_0)^2 \right) \\ &\quad + \frac{24\tau^2\gamma^2 M}{\pi^2(1-\tau)^2} \frac{\beta}{N(2-\beta)} \left(\frac{2H_{\max}}{v_{\min}} \sigma_\mu^2 + \frac{R_S^2}{2v_{\min}^3} \sigma_v^2 \right) \\ &\quad + \frac{48\tau^2\gamma^2 MQ\alpha^2}{\pi^2(1-\tau)^2\beta^2} \left(\frac{G^2}{M^2} + \sigma^2 \right) \left(\frac{2H_{\max}}{v_{\min}} L_\mu^2 + \frac{R_S^2}{2v_{\min}^3} L_v^2 \right) \\ &\stackrel{(d)}{\leq} C \left(\mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x, m)} [\varepsilon_t(Y, m)^2] + (1-\beta)^{2t} + \frac{\beta}{N} + \frac{\alpha^2}{\beta^2} \right). \end{aligned} \quad (144)$$

Here, (a) follows from Lemma A.9; (b) applies Lemma A.16 to each mode; (c) uses Lemma A.14 together with equation 140; and (d) uses $(2 - \beta)^{-1} \leq 1$ for $0 < \beta \leq 1$ and absorbs all fixed problem-dependent quantities into C .

One valid choice is

$$C := \max \left\{ \frac{8\tau^2}{\pi^2(1-\tau)^2}, \frac{24\tau^2\gamma^2 M}{\pi^2(1-\tau)^2} \left(\frac{2H_{\max}}{v_{\min}} \mathbb{E}(\hat{\mu}_0 - \mu_0)^2 + \frac{R_S^2}{2v_{\min}^3} \mathbb{E}(\bar{v}_0 - v_0)^2 \right), \right. \\ \left. \frac{24\tau^2\gamma^2 M}{\pi^2(1-\tau)^2} \left(\frac{2H_{\max}}{v_{\min}} \sigma_\mu^2 + \frac{R_S^2}{2v_{\min}^3} \sigma_v^2 \right), \right. \\ \left. \frac{48\tau^2\gamma^2 MQ}{\pi^2(1-\tau)^2} \left(\frac{G^2}{M^2} + \sigma^2 \right) \left(\frac{2H_{\max}}{v_{\min}} L_\mu^2 + \frac{R_S^2}{2v_{\min}^3} L_v^2 \right) \right\}, \quad (145)$$

which is independent of t , T , α , β , and N . This proves Proposition 4.2. $\square$

A.8 PROOF OF COROLLARY 4.3

Proof. By Theorem 4.1, there exists a constant $C_0 > 0$ such that

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 \leq C_0 \left(\frac{1}{\alpha T} + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \mathcal{E}_t + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\delta_{t,\phi}^r\|^2 + \frac{\alpha}{N} \right). \quad (146)$$

Applying Proposition 4.2 to $\mathcal{E}_t$ gives

$$\begin{aligned} & \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 \\ & \leq C_0 \left(\frac{1}{\alpha T} + \frac{C_1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + (1 - \beta)^{2t} + \frac{\beta}{N} + \frac{\alpha^2}{\beta^2} \right) \right. \\ & \quad \left. + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\delta_{t,\phi}^r\|^2 + \frac{\alpha}{N} \right) \\ & \leq C \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \|\delta_{t,\phi}^r\|^2 \right) + \frac{1}{\alpha T} + \frac{1}{T} \sum_{t=0}^{T-1} (1 - \beta)^{2t} \right. \\ & \quad \left. + \frac{\beta}{N} + \frac{\alpha^2}{\beta^2} + \frac{\alpha}{N} \right) \\ & \leq C \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \|\delta_{t,\phi}^r\|^2 \right) + \frac{1}{\alpha T} + \frac{1}{T\beta} + \frac{\beta}{N} + \frac{\alpha^2}{\beta^2} + \frac{\alpha}{N} \right). \end{aligned} \quad (147)$$

Here, the first inequality follows from Theorem 4.1 and the uniform correction-weight second-moment bound established in Lemma A.10; the second inequality follows from Proposition 4.2, after combining the resulting fixed constants into C ; and the last inequality uses $\frac{1}{T} \sum_{t=0}^{T-1} (1 - \beta)^{2t} = \frac{1 - (1 - \beta)^{2T}}{T\beta(2 - \beta)} \leq \frac{1}{T\beta}$.

Now choose $\alpha_T = \Theta(T^{-3/5})$ and $\beta_T = \Theta(T^{-2/5})$. For sufficiently large T , α_T is sufficiently small to satisfy the stepsize condition of Theorem 4.1, and $0 < \beta_T \leq 1$. Under these choices, the optimization and historical-tracking terms in equation 147 are dominated by $O(T^{-2/5})$. Hence,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 \leq \frac{C}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \|\delta_{t,\phi}^r\|^2 \right) + O(T^{-2/5}), \quad (148)$$

which proves the claimed convergence rate. $\square$

B ADDITIONAL THEORETICAL PROPERTIES AND EXTENSIONS

B.1 RESIDUAL INCORRECT-MODE MASS UNDER ENTROPY REGULARIZATION

Entropy regularization can mitigate prior concentration by encouraging exploration. However, since it does not explicitly distinguish correct and incorrect modes, the exploration benefit may also preserve non-negligible probability mass on incorrect modes. Under the same entropy-regularized update analyzed in Proposition 3.2, we characterize the limiting residual mass on incorrect modes under a bounded reward gap.

Proposition B.1 (Residual incorrect-mode mass under entropy regularization). *Under the setting of Proposition 3.2, suppose that at least one incorrect mode exists and that the reward gap is uniformly bounded:*

$$0 < a_t - b_t \leq \Delta_{\max} < \infty \quad \text{for all } t, \quad (149)$$

Then

$$\liminf_{T \rightarrow \infty} q_T^H(\mathcal{M} \setminus \mathcal{C}^+) \geq \frac{M - K}{M - K + K \exp(\Delta_{\max}/\lambda_H)} > 0. \quad (150)$$

If, in addition, $a_t - b_t \rightarrow \Delta_\infty$, then

$$q_T^H(\mathcal{M} \setminus \mathcal{C}^+) \rightarrow \frac{M - K}{M - K + K \exp(\Delta_\infty/\lambda_H)}. \quad (151)$$

Proof. Iterating the entropy-regularized update in equation 23 yields, for every $m \in \mathcal{M}$,

$$q_T^H(m) \propto q_0(m)^{\kappa_H^T} \exp\left(\eta \kappa_H \sum_{s=0}^{T-1} \kappa_H^{T-1-s} r_s(m)\right). \quad (152)$$

Because all correct modes receive a_s and all incorrect modes receive b_s at round s , summing equation 152 over the two groups gives

$$q_T^H(\mathcal{M} \setminus \mathcal{C}^+) = \frac{\sum_{m \notin \mathcal{C}^+} q_0(m)^{\kappa_H^T}}{\sum_{m \notin \mathcal{C}^+} q_0(m)^{\kappa_H^T} + \exp\left(\eta \kappa_H \sum_{s=0}^{T-1} \kappa_H^{T-1-s} (a_s - b_s)\right) \sum_{m \in \mathcal{C}^+} q_0(m)^{\kappa_H^T}}. \quad (153)$$

Under equation 149,

$$\eta \kappa_H \sum_{s=0}^{T-1} \kappa_H^{T-1-s} (a_s - b_s) \leq \eta \kappa_H \Delta_{\max} \sum_{j=0}^{T-1} \kappa_H^j \leq \frac{\eta \kappa_H}{1 - \kappa_H} \Delta_{\max} = \frac{\Delta_{\max}}{\lambda_H}, \quad (154)$$

where the last equality follows from $\kappa_H = (1 + \eta \lambda_H)^{-1}$.

Moreover, since $q_0(m) > 0$ and $\kappa_H^T \rightarrow 0$,

$$\begin{aligned} \sum_{m \in \mathcal{C}^+} q_0(m)^{\kappa_H^T} &\rightarrow K, \\ \sum_{m \notin \mathcal{C}^+} q_0(m)^{\kappa_H^T} &\rightarrow M - K. \end{aligned} \quad (155)$$

Combining equation 153–equation 155 proves equation 150.

If $a_t - b_t \rightarrow \Delta_\infty$, the geometrically discounted reward gap satisfies

$$\eta \kappa_H \sum_{s=0}^{T-1} \kappa_H^{T-1-s} (a_s - b_s) \rightarrow \frac{\Delta_\infty}{\lambda_H}. \quad (156)$$

Substituting this limit into equation 153 together with equation 155 gives equation 151. $\square$

This result shows that entropy regularization alone cannot guarantee full concentration on correct modes: although it reduces the dependence on the initial prior, a bounded reward gap limits the relative advantage of correct modes over incorrect modes, leaving a nonzero residual mass on incorrect modes.

B.2 OPTIMIZATION INTERPRETATION AND COVERAGE PROPERTIES OF PRIOR CORRECTION

The previous analysis establishes the convergence of the oracle prior-corrected dynamics to the uniform coverage-optimal allocation over correct modes. We next provide additional structural interpretations of this result. Specifically, we characterize the optimization structure underlying the oracle update, show that the induced objectives share the same coverage-optimal solution, and explain how the oracle update locally counteracts prior imbalance among correct modes.

To characterize the optimization structure underlying the oracle update, we identify an objective whose coordinate-wise gradient coincides with the mode-wise update coefficient in equation 4. Recall the primitive Φ defined in equation 51, which satisfies $\Phi'(u) = \phi(u)$. Since ϕ is positive, continuous, and strictly decreasing, Φ is strictly increasing and strictly concave.

For the time-varying reward signals a_t and b_t appearing in equation 4, define the following objective

$$f_{\phi,t}(q) := a_t \sum_{m \in \mathcal{C}^+} \Phi(q(m)) + b_t \sum_{m \notin \mathcal{C}^+} \Phi(q(m)). \quad (157)$$

For every coordinate with positive probability,

$$\frac{\partial f_{\phi,t}(q)}{\partial q(m)} = \begin{cases} a_t \phi(q(m)), & m \in \mathcal{C}^+, \\ b_t \phi(q(m)), & m \notin \mathcal{C}^+. \end{cases} \quad (158)$$

Hence, the oracle update in equation 4 is a KL-regularized first-order update obtained by linearizing $f_{\phi,t}$ at the current allocation $\bar{q}_t$: the reward signal determines whether a mode is promoted or suppressed, while the current mode allocation modulates the magnitude of the mode-wise coefficient through $\phi(q(m))$.

Although the reward signals determine the instantaneous update direction and may vary across iterations, this temporal variation does not alter the coverage target. We next show that all induced objectives share the same unique coverage-optimal solution.

Proposition B.2 (Shared coverage-optimal solution). *Suppose that $a_t > 0 \geq b_t$. Then, for every iteration t , the coverage-optimal allocation q^* is the unique maximizer of $f_{\phi,t}$ over $\Delta(\mathcal{M})$.*

Proof. We first show that a maximizer cannot assign positive probability to an incorrect mode. Suppose that $q \in \Delta(\mathcal{M})$ satisfies $q(j) > 0$ for some $j \notin \mathcal{C}^+$. Choose any $m \in \mathcal{C}^+$ and sufficiently small $\epsilon \in (0, q(j)]$, and let $q^\epsilon = q + \epsilon e_m - \epsilon e_j$. Then $q^\epsilon \in \Delta(\mathcal{M})$ and

$$\begin{aligned} f_{\phi,t}(q^\epsilon) - f_{\phi,t}(q) &= a_t [\Phi(q(m) + \epsilon) - \Phi(q(m))] \\ &\quad + b_t [\Phi(q(j) - \epsilon) - \Phi(q(j))]. \end{aligned} \quad (159)$$

Since Φ is strictly increasing,

$$\Phi(q(m) + \epsilon) - \Phi(q(m)) > 0, \quad (160)$$

whereas

$$\Phi(q(j) - \epsilon) - \Phi(q(j)) < 0. \quad (161)$$

Because $a_t > 0 \geq b_t$, the first term on the right-hand side of equation 159 is strictly positive and the second term on the right-hand side of equation 159 is not less than zero. Hence,

$$f_{\phi,t}(q^\epsilon) > f_{\phi,t}(q). \quad (162)$$

Therefore, every maximizer assigns zero probability to all incorrect modes.

It remains to optimize over the correct-mode face of the simplex. Since $\Phi(0) = 0$ and $a_t > 0$, this is equivalent to maximizing $\sum_{m \in \mathcal{C}^+} \Phi(q(m))$ subject to $\sum_{m \in \mathcal{C}^+} q(m) = 1$. By Jensen's inequality,

$$\frac{1}{K} \sum_{m \in \mathcal{C}^+} \Phi(q(m)) \leq \Phi\left(\frac{1}{K} \sum_{m \in \mathcal{C}^+} q(m)\right) = \Phi\left(\frac{1}{K}\right). \quad (163)$$

Because Φ is strictly concave, equality in equation 163 holds if and only if

$$q(m) = \frac{1}{K}, \quad m \in \mathcal{C}^+. \quad (164)$$

Together with zero probability on all incorrect modes, this uniquely identifies the maximizer as q^* . Since the argument holds for every t satisfying $a_t > 0 \geq b_t$, all induced objectives $f_{\phi,t}$ share the same unique maximizer q^* . $\square$

The above proposition identifies the global target of the oracle prior-corrected dynamics. We next characterize the local update behavior that moves probability mass among correct modes toward this target.

Proposition B.3 (One-step prior-correction effect). *Suppose that $a_t > 0$, and let $m, m' \in \mathcal{C}^+$ satisfy $0 < \bar{q}_t(m) < \bar{q}_t(m')$. Under the oracle prior-corrected update,*

$$\log \frac{\bar{q}_{t+1}(m)}{\bar{q}_{t+1}(m')} > \log \frac{\bar{q}_t(m)}{\bar{q}_t(m')}. \quad (165)$$

Proof. For $m, m' \in \mathcal{C}^+$, the closed-form update gives

$$\log \frac{\bar{q}_{t+1}(m)}{\bar{q}_{t+1}(m')} = \log \frac{\bar{q}_t(m)}{\bar{q}_t(m')} + \eta a_t [\phi(\bar{q}_t(m)) - \phi(\bar{q}_t(m'))]. \quad (166)$$

Since $\bar{q}_t(m) < \bar{q}_t(m')$ and ϕ is strictly decreasing, the bracketed term is strictly positive. Together with $a_t > 0$, this proves equation 165. $\square$

This proposition makes the local reallocation effect explicit: among correct modes receiving the same reward signal, a lower-allocation mode receives a larger relative amplification. This result characterizes only the local direction of reallocation among correct modes. The resulting global convergence behavior is established by the preceding dynamics analysis under the corresponding step-size condition.

Together, these properties provide a structural explanation of the coverage behavior induced by the oracle prior-corrected update: the shared maximizer of the induced objectives identifies the coverage-optimal allocation, while the one-step correction property characterizes the local reallocation mechanism that counteracts prior imbalance among correct modes.

B.3 SELECTION OF ANALYTICAL REFERENCE PARAMETERS

Recall the analytical reference form

$$\phi_{c,d}(q) = 1 + \frac{2\tau}{\pi} \arctan(c \log q^{-1} - d),$$

introduced in Section 4.2 to provide a fixed mode-level reference for analyzing the response-level correction. The parameters (c, d) specify the coefficient and offset in this mode-level reference form, and are used only for theoretical characterization rather than by the practical algorithm.

To understand the relation between the response-level statistic and this reference form, consider a response $y \in \mathcal{Y}_m$. Since

$$S_t(y) = -\log \pi_t(y | x) = \log \frac{1}{q_t(m)} + \log \frac{1}{\rho_t(y | x, m)},$$

the correction score admits the following decomposition

$$\gamma Z_t(y) = c_t(y) \log \frac{1}{q_t(m)} - d_t(y), \quad (167)$$

where

$$c_t(y) := \frac{\gamma}{\sqrt{H(y)v_t}}, \quad (168)$$

and

$$d_t(y) := \frac{\gamma}{\sqrt{H(y)v_t}} (\log \rho_t(y | x, m) + H(y)\mu_t). \quad (169)$$

Therefore, the response-level correction score contains a term proportional to $\log q_t(m)^{-1}$, but its effective coefficient and offset remain dependent on response-level quantities.

Recall the correction mismatch term $\varepsilon_t(Y, m)$ defined in the convergence analysis. Under the above decomposition, it can be written as

$$\begin{aligned} \varepsilon_t(Y, m) &= \gamma Z_t(Y) - \left(c \log \frac{1}{q_t(m)} - d \right) \\ &= (c_t(Y) - c) \log \frac{1}{q_t(m)} - (d_t(Y) - d). \end{aligned} \quad (170)$$

This decomposition separates two components of the mismatch: the deviation of the response-dependent coefficient of $\log q_t(m)^{-1}$ from the fixed coefficient c , and the deviation of the response-dependent offset from the fixed offset d .

For theoretical analysis, one may define the best fixed reference by minimizing the accumulated mismatch along the optimization trajectory:

$$(c^*, d^*) \in \arg \min_{\substack{c > 0 \\ d \in \mathbb{R}}} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x, m)} \left[\left((c_t(Y) - c) \log \frac{1}{q_t(m)} - (d_t(Y) - d) \right)^2 \right]. \quad (171)$$

This optimization identifies the fixed reference that minimizes the accumulated residual within this family, but does not eliminate the structural mismatch in general. The mismatch vanishes in the idealized case where the response-level quantities induce the same fixed coefficient and offset across modes and iterations. For example, if each latent mode contains a single response with a common response length and the population statistics remain fixed, then $c_t(Y) \equiv c$ and $d_t(Y) \equiv d$ for a suitable fixed pair (c, d) , yielding $\varepsilon_t(Y, m) = 0$.

Such a selection, however, relies on population statistics and latent mode information that are unavailable to the practical algorithm. Therefore, (c, d) should be viewed as analytical reference parameters used to characterize the response-level correction and its residual mismatch, rather than as parameters required by the practical algorithm.

B.4 FINITE-WINDOW HISTORICAL REFERENCE ANALYSIS

The preceding analysis establishes the convergence guarantee for the lagged EMA adopted in the main algorithm. We next study an alternative strictly historical estimator based on a finite window of completed rounds. Although not used in the proposed algorithm, this analysis provides a complementary characterization of historical reference statistics and reveals the same tradeoff between statistical stability and tracking the evolving policy through the window length K_h .

Lemma B.4 (Finite-window historical-reference tracking error). *For $K_h \geq 1$ and $t \geq K_h$, define the finite-window historical reference statistic*

$$\hat{\xi}_t^{(K_h)} := \frac{1}{K_h} \sum_{j=1}^{K_h} \tilde{\xi}_{t-j}, \quad \xi \in \{\mu, v\}. \quad (172)$$

Then

$$\mathbb{E} \left[\left(\hat{\xi}_t^{(K_h)} - \xi_t \right)^2 \right] \leq 2 \frac{\sigma_\xi^2}{K_h N} + 2 \left(\frac{K_h + 1}{2} L_\xi \alpha \sqrt{2Q \left(\frac{G^2}{M^2} + \sigma^2 \right)} \right)^2. \quad (173)$$

Proof. Decomposing the historical estimation error gives

$$\begin{aligned} & \mathbb{E} \left[\left(\hat{\xi}_t^{(K_h)} - \xi_t \right)^2 \right] \\ & \leq 2 \mathbb{E} \left[\left(\frac{1}{K_h} \sum_{j=1}^{K_h} (\tilde{\xi}_{t-j} - \xi_{t-j}) \right)^2 \right] + 2 \mathbb{E} \left[\left(\frac{1}{K_h} \sum_{j=1}^{K_h} (\xi_{t-j} - \xi_t) \right)^2 \right] \\ & \leq \frac{2\sigma_\xi^2}{K_h N} + 2 \left\| \frac{1}{K_h} \sum_{j=1}^{K_h} (\xi_{t-j} - \xi_t) \right\|_{L_2}^2 \\ & \leq \frac{2\sigma_\xi^2}{K_h N} + 2 \left[\frac{1}{K_h} \sum_{j=1}^{K_h} \sum_{r=1}^j \|\xi_{t-r+1} - \xi_{t-r}\|_{L_2} \right]^2 \\ & \leq \frac{2\sigma_\xi^2}{K_h N} + 2 \left[\frac{K_h + 1}{2} L_\xi \alpha \sqrt{2Q \left(\frac{G^2}{M^2} + \sigma^2 \right)} \right]^2. \end{aligned} \quad (174)$$

The first inequality follows from $(a + b)^2 \leq 2a^2 + 2b^2$. The second inequality uses the martingale-difference property of $\hat{\xi}_t - \xi_t$ established in Lemma A.13. The third inequality follows from the telescoping decomposition of $\xi_{t-j} - \xi_t$ and Minkowski's inequality. The final inequality applies the population-statistic drift bound in equation 122 and $\frac{1}{K_h} \sum_{j=1}^{K_h} j = (K_h + 1)/2$. $\square$

The lemma decomposes the finite-window reference error into a statistical estimation component and a tracking component caused by the evolution of the policy across historical rounds. We next characterize how this reference-statistic error affects the correction-weight mismatch.

Proposition B.5 (Finite-window historical-reference correction error). *Suppose Assumptions A.2, A.6, A.7, A.11, and A.15 hold. Then there exists a constant $C > 0$ such that*

$$\mathbb{E}\mathcal{E}_t \leq C \left(\mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \frac{1}{K_h N} + K_h^2 \alpha^2 \right). \quad (175)$$

Proof. For the finite-window estimator, $\hat{Z}_t(Y)$ denotes the standardized response score obtained by replacing the EMA statistics with the historical-window statistics $(\hat{\mu}_t^{(K_h)}, \hat{v}_t^{(K_h)})$. By Lemma A.9, Lemma A.16, and Lemma B.4, we have

$$\begin{aligned} \mathbb{E}\mathcal{E}_t &\stackrel{(a)}{\leq} \frac{8\tau^2}{\pi^2(1-\tau)^2} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] \\ &\quad + \frac{8\tau^2\gamma^2}{\pi^2(1-\tau)^2} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [|\hat{Z}_t(Y) - Z_t(Y)|^2] \\ &\stackrel{(b)}{\leq} \frac{8\tau^2}{\pi^2(1-\tau)^2} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] \\ &\quad + \frac{8\tau^2\gamma^2 M}{\pi^2(1-\tau)^2} \left(\frac{2H_{\max}}{v_{\min}} \mathbb{E}[(\hat{\mu}_t^{(K_h)} - \mu_t)^2] + \frac{R_S^2}{2v_{\min}^3} \mathbb{E}[(\hat{v}_t^{(K_h)} - v_t)^2] \right) \\ &\stackrel{(c)}{\leq} \frac{8\tau^2}{\pi^2(1-\tau)^2} \mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \frac{16\tau^2\gamma^2 M}{\pi^2(1-\tau)^2} \frac{1}{K_h N} \left(\frac{H_{\max}}{v_{\min}} \sigma_\mu^2 + \frac{R_S^2}{4v_{\min}^3} \sigma_v^2 \right) \\ &\quad + \frac{32\tau^2\gamma^2 M Q}{\pi^2(1-\tau)^2} \left(\frac{G^2}{M^2} + \sigma^2 \right) \left(\frac{H_{\max}}{v_{\min}} L_\mu^2 + \frac{R_S^2}{4v_{\min}^3} L_v^2 \right) K_h^2 \alpha^2 \\ &\stackrel{(d)}{\leq} C \left(\mathbb{E} \sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \frac{1}{K_h N} + K_h^2 \alpha^2 \right). \end{aligned} \quad (176)$$

Here, (a) follows by taking expectations in Lemma A.9. Step (b) applies Lemma A.16. Step (c) substitutes the historical-window tracking bound from Lemma B.4 for both statistics and uses $(a+b)^2 \leq 2a^2 + 2b^2$ together with $(K_h + 1)^2/4 \leq K_h^2$. Step (d) absorbs all fixed problem-dependent quantities into the constant

$$C = \max \left\{ \frac{8\tau^2}{\pi^2(1-\tau)^2}, \frac{16\tau^2\gamma^2 M}{\pi^2(1-\tau)^2} \left(\frac{H_{\max}}{v_{\min}} \sigma_\mu^2 + \frac{R_S^2}{4v_{\min}^3} \sigma_v^2 \right), \right. \quad (177)$$

$$\left. \frac{32\tau^2\gamma^2 M Q}{\pi^2(1-\tau)^2} \left(\frac{G^2}{M^2} + \sigma^2 \right) \left(\frac{H_{\max}}{v_{\min}} L_\mu^2 + \frac{R_S^2}{4v_{\min}^3} L_v^2 \right) \right\}. \quad (178)$$

The constant C is independent of t, T, α, K_h , and N . $\square$

Thus, increasing K_h improves statistical stability by averaging more historical observations, while making the reference statistics less responsive to the evolving policy. Combining this bound with the stationarity result yields the following convergence guarantee.

Corollary B.6 (Stationarity rate with finite-window historical-reference correction). *Suppose the conditions of Theorem 4.1, Proposition B.5, and Lemma A.10 hold. Choosing*

$$\alpha_T = \Theta(T^{-3/5}), \quad K_{h,T} = \Theta(T^{2/5}), \quad (179)$$

yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 \leq \frac{C}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \|\delta_{t,\phi}^r\|^2 \right) + O(T^{-2/5}). \quad (180)$$

Proof. By the explicit bound established in the proof of Theorem 4.1 and Lemma A.10, there exists a constant $C_0 > 0$ such that

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 \leq C_0 \left(\frac{1}{\alpha T} + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \mathcal{E}_t + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\delta_{t,\phi}^r\|^2 + \frac{\alpha}{N} \right). \quad (181)$$

Since the finite-window estimator is defined for $t \geq K_h$, we treat the first K_h rounds as a burn-in period. By Assumption A.2,

$$\frac{1}{T} \sum_{t=0}^{K_h-1} \mathbb{E} \|g_t^*\|^2 \leq \frac{G^2 K_h}{T}. \quad (182)$$

Applying the same stationarity argument as in Theorem 4.1 over the remaining rounds, together with Proposition B.5, where $C_1 > 0$ denotes the constant in the historical-reference correction bound, gives

$$\begin{aligned} \frac{1}{T} \sum_{t=K_h}^{T-1} \mathbb{E} \|g_t^*\|^2 &\leq C_0 \left(\frac{1}{\alpha T} + \frac{C_1}{T} \sum_{t=K_h}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \frac{1}{K_h N} + K_h^2 \alpha^2 \right) \right. \\ &\quad \left. + \frac{1}{T} \sum_{t=K_h}^{T-1} \mathbb{E} \|\delta_{t,\phi}^r\|^2 + \frac{\alpha}{N} \right) \\ &\leq C \left(\frac{1}{T} \sum_{t=K_h}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \|\delta_{t,\phi}^r\|^2 \right) \right. \\ &\quad \left. + \frac{1}{\alpha T} + \frac{1}{K_h N} + K_h^2 \alpha^2 + \frac{\alpha}{N} \right). \end{aligned} \quad (183)$$

Combining equation 182 and equation 183, and using the nonnegativity of the structural terms, yields

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 &\leq C \left(\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \|\delta_{t,\phi}^r\|^2 \right) \right. \\ &\quad \left. + \frac{1}{\alpha T} + \frac{1}{K_h N} + K_h^2 \alpha^2 + \frac{\alpha}{N} + \frac{K_h}{T} \right). \end{aligned} \quad (184)$$

For sufficiently large T , the step-size condition in Theorem 4.1 is satisfied. Under the choices in equation 179, the optimization, historical-reference tracking, and burn-in terms are dominated by $O(T^{-2/5})$. Therefore,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|g_t^*\|^2 \leq \frac{C}{T} \sum_{t=0}^{T-1} \mathbb{E} \left(\sum_{m \in \mathcal{M}} \mathbb{E}_{\rho_t(\cdot|x,m)} [\varepsilon_t(Y, m)^2] + \|\delta_{t,\phi}^r\|^2 \right) + O(T^{-2/5}), \quad (185)$$

which proves the claim. $\square$

The window schedule $K_{h,T} = \Theta(T^{2/5})$ matches the effective memory scale $\beta_T^{-1} = \Theta(T^{2/5})$ of the lagged EMA estimator. Therefore, both historical-reference mechanisms achieve the same $O(T^{-2/5})$ rate while balancing statistical stability and tracking the evolving policy.

C ALGORITHM AND EXPERIMENTAL DETAILS

C.1 COMPLETE PRIOR-CORRECTED RLVR ALGORITHM

This section provides the complete training procedure for the practical prior-corrected RLVR method described in Section 4. Algorithm 1 combines the historical-reference statistics, response-level prior correction, conservative correction rule, and policy optimization. While the theoretical analysis considers N responses generated for a fixed prompt x , the practical implementation pools all BN responses in each completed rollout batch when estimating the historical reference statistics. Here $\text{sg}(\cdot)$ denotes the stop-gradient operator.

C.2 EXPERIMENTAL SETUP

All training experiments are conducted on $2 \times \text{NVIDIA A100-80GB GPUs}$. Our implementation is built on the ROLL (Wang et al., 2025b) training framework, with vLLM (Kwon et al., 2023) used for rollout generation and DeepSpeed (Rasley et al., 2020) used for policy optimization. Given our computational budget, we optimize only LoRA (Hu et al., 2022) parameters for all methods. We use a linearly decayed learning rate throughout training.

For controlled comparison, all compared methods are reimplemented within the same ROLL pipeline and share the same GRPO training backbone whenever applicable. We keep the model initialization, training data, rollout configuration, optimization budget, and common hyperparameters identical across methods, while retaining the method-specific components required by each method. No explicit KL regularization is used in the compared methods. For each method and model scale, we select the checkpoint with the highest Pass@1 on the combined AIME24–AIME25 monitoring set during training for final evaluation.

C.3 IMPLEMENTATION DETAILS OF COMPARED METHODS

DAPO uses its token-level loss aggregation, whereas GRPO, GRPO with entropy regularization (GRPO+Entropy), ProGRPO, and PC-RLVR use the standard GRPO loss aggregation. Unless otherwise specified, the rollout and optimization configurations are shared across methods.

For PPO clipping, vanilla GRPO uses symmetric clipping with $\epsilon_{\text{low}} = \epsilon_{\text{high}} = 0.20$. DAPO uses asymmetric clipping with $\epsilon_{\text{low}} = 0.20$ and $\epsilon_{\text{high}} = 0.28$. To control for the effect of asymmetric clipping when comparing the additional mechanisms introduced by different methods, we use the same 0.20/0.28 clipping thresholds for GRPO+Entropy, ProGRPO, and PC-RLVR.

For each baseline, we start from the hyperparameters recommended in the original paper whenever applicable. Method-specific hyperparameters are modified only when necessary for stable optimization under our common training setup, with the settings relevant to our implementation reported below.

GRPO. We use standard GRPO with group-relative advantage normalization and symmetric PPO clipping, with $\epsilon_{\text{low}} = \epsilon_{\text{high}} = 0.20$.

DAPO. We use the DAPO training variant with token-level loss aggregation and asymmetric PPO clipping, setting $\epsilon_{\text{low}} = 0.20$ and $\epsilon_{\text{high}} = 0.28$.

GRPO+Entropy. We augment the GRPO objective with token-level entropy regularization using coefficient λ_H . We tune λ_H over $\{0.02, 0.005, 0.0005\}$ and use $\lambda_H = 0.0005$ for the reported experiments.

ProGRPO. We implement the probability-aware reweighting proposed by ProGRPO on top of the common GRPO backbone, with $\epsilon_{\text{low}} = 0.20$ and $\epsilon_{\text{high}} = 0.28$.

PC-RLVR. Our method uses the historical-reference prior correction described in Section 4, together with the conservative correction rule. We use $\epsilon_{\text{low}} = 0.20$ and $\epsilon_{\text{high}} = 0.28$. For numerical stability, the historical variance estimate is projected onto $[v_{\min}, v_{\max}]$, with $v_{\min} = 0.01$

Table 2: Common training hyperparameters used across methods.

Hyperparameter	Value
Rollout batch size (prompts)	32
Rollout group size	8
Minibatch size (prompts)	8
Optimizer updates per rollout batch	4
Maximum prompt length	1,024
Maximum response length	8,192
Rollout temperature	1.0
Rollout top- p	1.0
Initial learning rate	2×10^{-5}
LoRA rank	32
PPO clip lower threshold	0.20
Explicit KL regularization	None

and $v_{\max} = 100$. For the main experiments, we use $(\beta, \tau, \kappa) = (0.7, 0.5, 0)$ for Qwen3-4B and $(\beta, \tau, \kappa) = (0.5, 0.2, 0)$ for Qwen3-1.7B. In practical training, we also adapt γ online to limit excessive saturation of the arctangent correction. Specifically, we maintain an EMA of the 95th percentile of $|\hat{Z}_t(Y)|$ over completed rollout batches and set γ such that this reference magnitude is mapped to 1.5.

C.4 HYPERPARAMETERS

Table 2 summarizes the hyperparameters shared across methods. Unless otherwise stated, these values are fixed for all models and baselines.

Evaluation hyperparameters. All models are evaluated under the same decoding configuration, with temperature 0.6, top- $p = 0.95$, and a maximum response length of 8,192 tokens. For each problem, we sample n responses once and use the same sample set to estimate pass@ k for different values of k . For a problem with c correct responses among n sampled responses, we use the standard unbiased estimator

$$\widehat{\text{pass@}k} = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}, \quad n - c \geq k, \quad (186)$$

with the value set to 1 when $n - c < k$. We report results using the same sampled responses for all $k \leq n$.

D MORE RESULTS

D.1 CONTROLLED MODE-LEVEL EXPERIMENTS

This subsection provides additional analyses of the controlled mode-level setting in Section 3.2. The experiments below use the same initialization and the same method-specific parameter settings, while differing in the reward-signal schedule.

We consider $M = 20$ modes with $\mathcal{C}^+ = \{1, 2, 3, 4\}$. The initial probabilities of the correct modes are

$$q_0(1 : 4) = (0.28, 0.07, 0.03, 0.02),$$

while the remaining probability mass 0.60 is distributed uniformly over the 16 incorrect modes. Hence, $q_0(\mathcal{C}^+) = 0.40$, with initial conditional allocation

$$q_0(\cdot | \mathcal{C}^+) = (0.70, 0.175, 0.075, 0.05).$$

The coverage-optimal allocation q^* assigns probability $1/K = 0.25$ to each correct mode and zero probability to incorrect modes. All methods use the same initialization and KL parameter $\eta = 0.2$. For entropy regularization, we set $\lambda_H = 0.2$. For prior correction, we use the bounded correction $\phi_{c,d}$ defined in Section 4.2, with $c = 1$, $d = \log 4$, and $\tau = 0.8$.

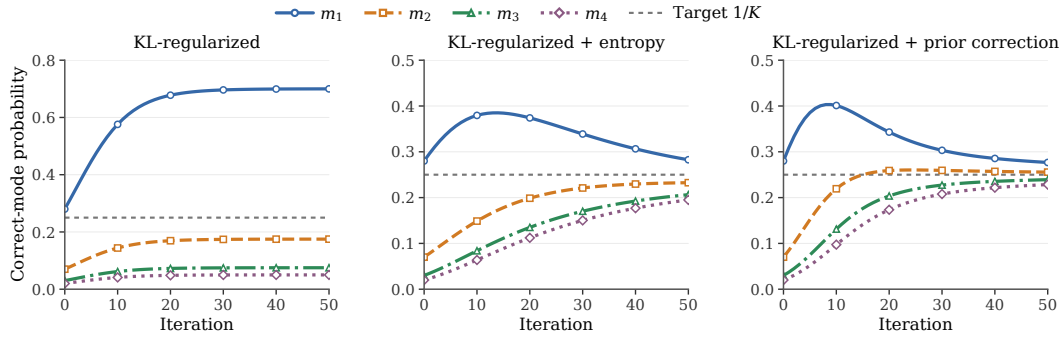


Figure 3: Mode-wise probability trajectories of the four correct modes under standard KL-regularized policy improvement, entropy regularization, and prior correction. The dashed line marks the correct-mode probability $1/K$ under the coverage-optimal allocation q^* .

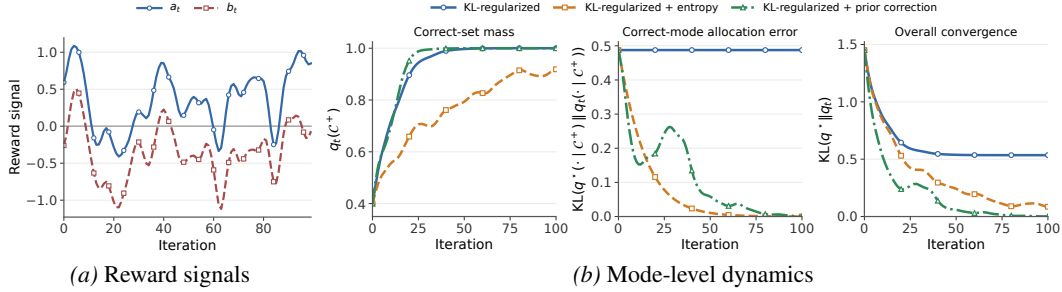


Figure 4: Controlled mode-level dynamics under time-varying reward signals satisfying $a_t > b_t$. Panel (a) shows the signals assigned to correct and incorrect modes. Panel (b) compares standard KL-regularized policy improvement, entropy regularization, and prior correction in terms of correct-set probability mass, correct-mode allocation error, and distance to q^* .

In all experiments, the mode distribution is evolved directly according to the exact update rule of each method; no stochastic sampling or gradient approximation is involved.

Correct-mode probability trajectories. Figure 3 complements Figure 1 by resolving the aggregate correct-mode allocation dynamics into the trajectories of the four individual correct modes. Under standard KL-regularized policy improvement, all four correct modes gain probability while preserving their relative allocation. Consequently, their probabilities approach

$$(0.70, 0.175, 0.075, 0.05),$$

which coincides with the initial conditional allocation over correct modes.

Both entropy regularization and prior correction progressively reduce the imbalance among correct modes. Prior correction drives the correct-mode probabilities toward $1/K = 0.25$, whereas entropy regularization equalizes the correct modes more gradually but need not reach q^* because it can retain nonzero probability on incorrect modes.

Prior locking under time-varying reward signals. The condition $a_t > b_t$ for prior locking is substantially weaker than the assumptions required for convergence of the prior-corrected policy updates in Theorem 3.3. To illustrate this distinction, we consider a second time-varying schedule satisfying $a_t > b_t$ at every iteration, while a_t and b_t may individually change sign. This experiment illustrates the prior-locking behavior under this broader signal condition, rather than asserting convergence of prior correction under these more general signals.

Figure 4 illustrates the resulting dynamics. Standard KL-regularized policy improvement increases the probability assigned to the correct set, but leaves the allocation among correct modes unchanged.

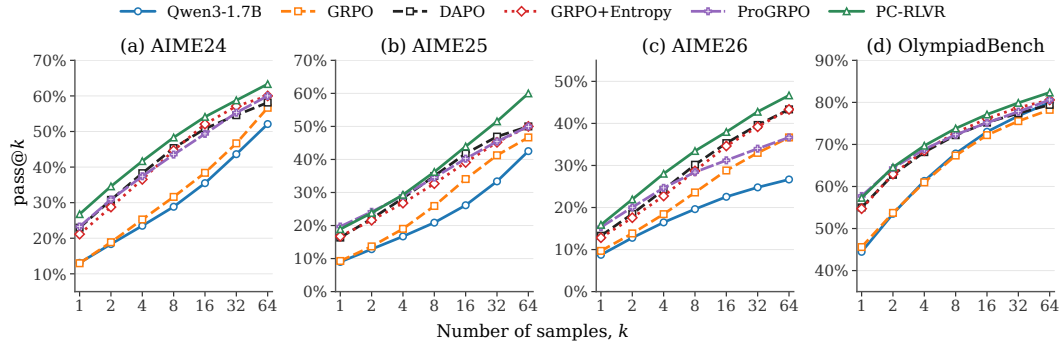


Figure 5: Pass@ k curves of Qwen3-1.7B under increasing sampling budgets, with $k \in \{1, 2, 4, 8, 16, 32, 64\}$. PC-RLVR maintains strong performance across the full sampling range on AIME24 and AIME26, exhibits an increasingly visible advantage at larger k on AIME25, and remains competitive with the strongest methods on OlympiadBench.

Accordingly, the correct-mode allocation error remains constant and the distance to q^* stays bounded away from zero. This is exactly the prior-locking behavior characterized by Theorem 3.1, which depends on $a_t > b_t$ rather than on the individual signs of the two signals.

For this particular schedule, entropy regularization and prior correction both reduce the correct-mode allocation error. Under prior correction, this error evolves non-monotonically as the reward signals change sign, but decreases to near zero together with the distance to q^* over the shown horizon. This observation is specific to the chosen schedule and should not be interpreted as an extension of Theorem 3.3. In particular, the signals considered here may violate the theorem’s conditions $0 < a_t \leq a_{\max}$ and $b_t \leq 0$, even though the chosen $\eta = 0.2$ satisfies the theorem’s bound on η .

D.2 PASS@ k UNDER INCREASING SAMPLING BUDGETS

To complement the high-budget results in Table 1, we examine how performance evolves as the sampling budget increases. Figure 5 reports Pass@ k for $k \in \{1, 2, 4, 8, 16, 32, 64\}$ on Qwen3-1.7B across four mathematical reasoning benchmarks. Across the three AIME benchmarks, PC-RLVR exhibits consistently strong performance over a broad range of sampling budgets, but the detailed patterns differ across datasets. On AIME24, PC-RLVR remains above the compared methods throughout the entire range of k , indicating that its advantage is not restricted to the largest sampling budget. A similar pattern appears on AIME26, where PC-RLVR maintains a visible margin over the strongest competing methods from small to large k .

The results on AIME25 show a more pronounced high-budget effect. At small sampling budgets, PC-RLVR is close to the strongest baselines, whereas the separation becomes increasingly visible as k grows. In particular, the gap is substantially larger at $k = 32$ and $k = 64$ than at the smallest sampling budgets. This behavior is consistent with the intended effect of prior correction: improvements become more apparent when a larger sampling budget provides greater opportunity to access successful responses that receive relatively little probability under the current policy.

On OlympiadBench, the strongest RLVR methods are much more closely clustered across the entire sampling range. PC-RLVR nevertheless remains among the strongest methods and attains the highest Pass@ k over most of the larger sampling budgets. The smaller separation is consistent with the relatively high success rates already achieved by several methods on this benchmark, leaving less room for additional gains as k increases.

Overall, these curves show that the high-budget improvements reported in Table 1 do not arise from an isolated Pass@64 operating point. PC-RLVR either maintains an advantage across sampling budgets or becomes increasingly competitive as the budget grows, with the latter behavior being especially clear on AIME25.

D.3 PRIOR-DEPENDENT LOG-LIKELIHOOD GAIN

This section provides the complete construction and additional diagnostics for the prior-dependent log-likelihood-gain analysis in Section 5.3. At the mode level, prior correction drives the allocation over correct modes toward the coverage-optimal target q^* . Starting from an imbalanced allocation, moving toward q^* requires underrepresented correct modes to gain relatively more probability than already dominant ones.

Latent mode probabilities are not directly observable in the evaluation data. We therefore examine the corresponding response-level signature: whether PC-RLVR, relative to GRPO, assigns larger log-likelihood gains to correct responses that receive lower log-likelihood under the initial policy.

Initial-prior grouping. For each prompt x , we rescore the deduplicated correct responses under the frozen base policy π_0 . For a response y , we define its average response-token surprisal as

$$\bar{S}_0(y | x) = -\frac{1}{H(y)} \log \pi_0(y | x). \quad (187)$$

Larger $\bar{S}_0(y | x)$ corresponds to lower average token log-likelihood under the initial policy. Within each prompt, correct responses are ranked by this score and partitioned into approximately equal-sized high-, medium-, and low-prior groups. Here, the prior-group terminology refers to this response-level ordering under the base policy rather than directly observed latent-mode probabilities. Prompts without sufficient distinct correct responses to form all three groups are excluded.

Relative log-likelihood gain. For a fixed correct response y , we measure the relative log-likelihood gain of PC-RLVR over GRPO using the average per-token log-likelihood difference

$$D(y) = \frac{1}{H(y)} [\log \pi_{\text{PC}}(y | x) - \log \pi_{\text{GRPO}}(y | x)]. \quad (188)$$

Thus, $D(y) > 0$ indicates that PC-RLVR assigns higher average token log-likelihood to y than GRPO, while a larger $D(y)$ indicates a stronger relative log-likelihood gain. The per-token normalization removes the direct linear accumulation of response log-likelihood with response length.

Prompt-balanced aggregation. For prior group $g \in \{\text{High}, \text{Medium}, \text{Low}\}$, let $\mathcal{Y}_g^+(x)$ denote the correct responses assigned to group g for prompt x . We first compute

$$D_g(x) = \frac{1}{|\mathcal{Y}_g^+(x)|} \sum_{y \in \mathcal{Y}_g^+(x)} D(y), \quad (189)$$

and then average $D_g(x)$ equally across eligible prompts. This prevents prompts with many correct responses from dominating the benchmark-level results.

We additionally report the fraction of eligible prompts satisfying $D_{\text{Low}}(x) > D_{\text{High}}(x)$ as a measure of prompt-level consistency and obtain 95% percentile bootstrap confidence intervals by resampling eligible prompts with replacement.

Prior-dependent gain. As shown in Figure 2, the benchmark-level results exhibit a consistent prior-dependent trend. The low-prior group attains the largest mean relative gain on each of the four analyzed benchmarks, while the medium-prior group generally lies between the high- and low-prior groups. The separation between the medium- and low-prior groups is small on AIME26, but the contrast between low- and high-prior responses remains consistent.

At the prompt level, the low-prior group receives a larger gain than the high-prior group for 88.9%, 76.9%, 100%, and 95.3% of eligible prompts on AIME24, AIME25, AIME26, and OlympiadBench, respectively. The agreement between the benchmark-level averages and the prompt-level comparison indicates that the observed effect is not driven by a small number of prompts with unusually large log-likelihood differences. Instead, correct responses with lower-initial-log-likelihood tend to receive larger relative gains under PC-RLVR than under GRPO.

This pattern is consistent with the rebalancing effect predicted by prior correction: relative to GRPO, PC-RLVR produces larger average per-token log-likelihood gains for successful responses that are less favored by the initial policy.

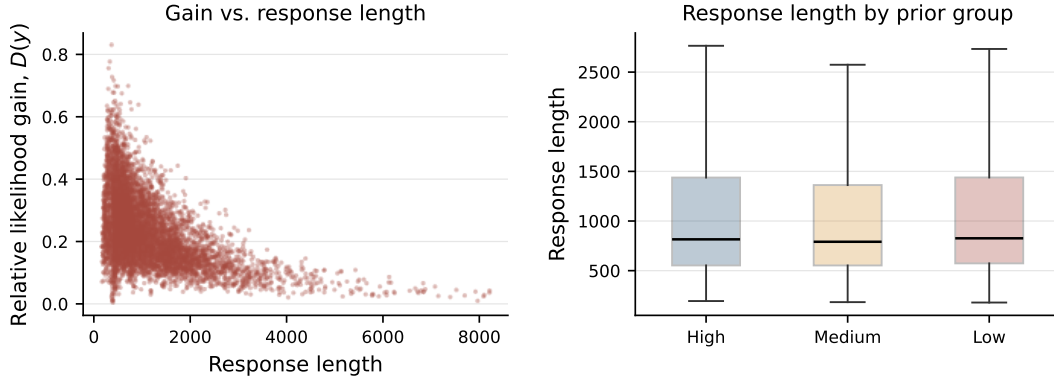


Figure 6: Length-related diagnostics for the prior-dependent log-likelihood gain analysis. **Left:** relationship between the relative per-token log-likelihood gain $D(y)$ and response length. **Right:** response-length distributions of correct responses in the high-, medium-, and low-prior groups. Although residual length dependence remains in $D(y)$, the three prior groups exhibit strongly overlapping length distributions.

Length-related diagnostics. Both quantities used in the analysis explicitly reduce the direct effect of response length. The initial-prior grouping uses average response-token surprisal, while $D(y)$ is an average per-token log-likelihood difference. These normalizations remove the direct linear accumulation with response length, but they do not imply statistical independence from response length.

We therefore examine the residual relationship between $D(y)$ and response length and compare the response-length distributions across the three prior groups. As shown in Figure 6, $D(y)$ retains a negative association with response length. However, the response-length distributions of the high-, medium-, and low-prior groups overlap substantially and have similar central tendencies. In particular, their length distributions do not reproduce the systematic increase in relative log-likelihood gain from high- to low-prior responses. This suggests that response length alone does not explain the observed prior-dependent gain.

Finally, this analysis operates at the response level and does not directly observe latent reasoning modes. We therefore interpret the prior-dependent log-likelihood gain as empirical evidence consistent with the rebalancing effect of prior correction, rather than as a direct measurement of mode-level allocation.

D.4 BASE-REFERENCE SURPRISAL ANALYSIS

The prior-dependent log-likelihood-gain analysis in Section 5.3 examines whether lower-prior correct responses receive larger relative gains under PC-RLVR. Here, we study the corresponding distributional outcome: whether the correct responses generated by the final policy extend further into regions that are less favored by the initial policy.

We use the saved Pass@64 generations from the corresponding evaluation runs and rescore all responses under the same frozen base policy π_0 . Following equation 187, we use the base-model average response-token surprisal $\bar{S}_0(Y \mid x)$ as a common response-level measure of initial log-likelihood. Larger values correspond to lower average token log-likelihood under the initial policy.

Distribution over correct responses. We restrict the distributional comparison to responses verified as correct. For an evaluated method h , let $\mathcal{Y}_h^+$ denote its collection of correct responses. We report the empirical cumulative distribution

$$\hat{F}_h(s) = \frac{1}{|\mathcal{Y}_h^+|} \sum_{Y \in \mathcal{Y}_h^+} \mathbf{1}\{\bar{S}_0(Y) \leq s\}. \quad (190)$$

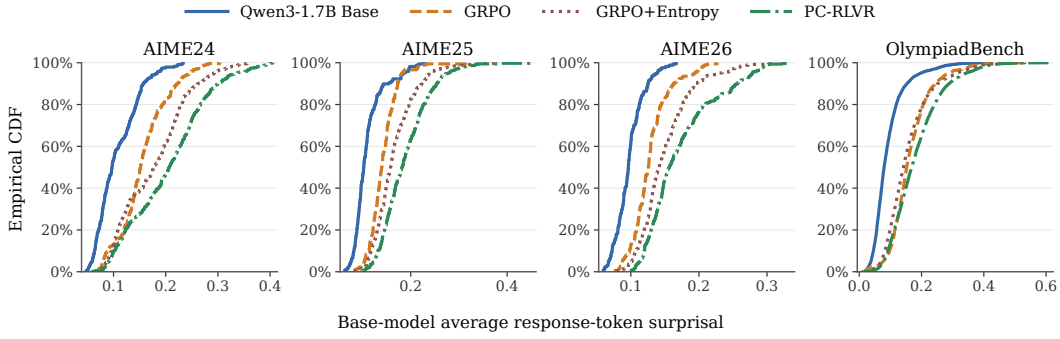


Figure 7: Empirical CDFs of base-model average response-token surprisal among correct responses from Qwen3-1.7B. All responses are rescored under the same frozen base policy π_0 . A rightward shift indicates that successful responses receive lower average token log-likelihood under the initial policy.

A rightward shift of the ECDF therefore indicates that a larger fraction of successful responses lies at higher base-model surprisal, or equivalently lower average token log-likelihood under the initial policy.

Figure 7 compares Qwen3-1.7B Base, GRPO, GRPO+Entropy, and PC-RLVR. We include GRPO+Entropy as the exploration-oriented alternative analyzed in Section 3, which can broaden the response distribution without explicitly correcting prior allocation. This comparison helps distinguish prior correction from a generic increase in exploration.

Relative to the base model, RLVR training generally shifts successful responses toward higher base-model surprisal. GRPO+Entropy produces a further rightward shift relative to vanilla GRPO, consistent with its broader exploration effect. PC-RLVR extends the distribution further toward lower-prior successful responses across the analyzed benchmarks.

Higher base-model surprisal is not itself a measure of better reasoning, since lower-log-likelihood regions may also contain noisy or redundant responses. The distributional result is therefore most informative when considered together with task performance: over the full evaluation suite, PC-RLVR improves Avg@64 and Pass@64 over GRPO+Entropy by 3.8 and 3.7 percentage points, respectively. Thus, the broader access to lower-prior successful responses is accompanied by stronger reasoning performance rather than merely a generic increase in exploration.

Controlling for asymmetric clipping. PC-RLVR uses a larger upper clipping threshold than vanilla GRPO in the main experiments. To determine whether the observed rightward shift can be explained by this optimization change alone, we perform a matched control on Qwen3-4B. The *w/o prior correction* variant uses the same clipping thresholds as PC-RLVR but disables the prior-correction weights. For this controlled analysis, the PC-RLVR model uses the $\tau = 0.2$ configuration, matching the controlled ablation setting in Appendix D.5. Figure 8 shows that enlarging the upper clipping threshold already shifts successful responses toward higher base-model surprisal relative to vanilla GRPO. PC-RLVR nevertheless exhibits a further rightward shift across all four analyzed benchmarks. This indicates that the observed distributional change cannot be explained solely by the enlarged upper clipping threshold.

The same comparison is also favorable in downstream performance: PC-RLVR improves the mean Avg@64 and Pass@64 over the matched *w/o prior correction* variant by 4.4 and 1.3 percentage points, respectively, over the full evaluation suite.

Quantitative summary. Table 3 quantifies the distributional shifts observed in Figures 7 and 8. We report changes in the mean and selected quantiles of $\bar{S}_0$ over correct responses. Here, Q_p denotes the p -th percentile of the corresponding surprisal distribution, and each Δ is computed as PC-RLVR minus the comparison method. Positive values indicate a shift toward responses with lower average token log-likelihood under the initial policy. For readability, all reported shifts are scaled by 10^2 .

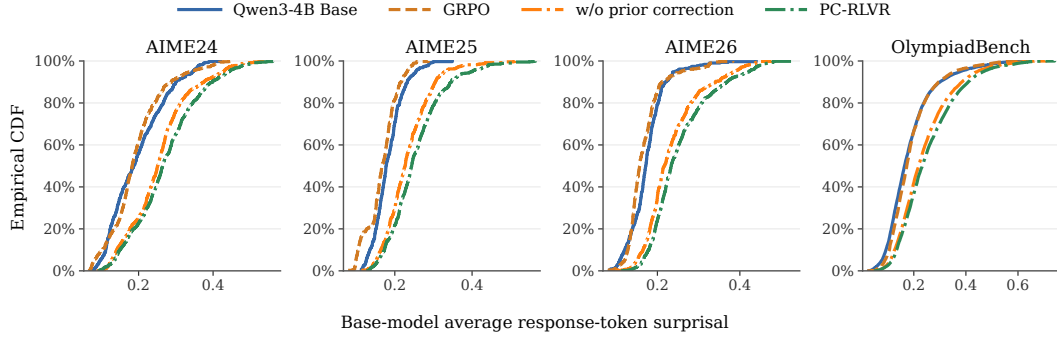


Figure 8: Matched clipping control on Qwen3-4B. The *w/o prior correction* variant uses the same clipping thresholds as PC-RLVR while disabling prior correction. PC-RLVR remains further shifted toward higher base-model average response-token surprisal across all four benchmarks.

Table 3: Quantitative summary of the shift in base-model average response-token surprisal over correct responses for the two controlled comparisons. All entries are scaled by 10^2 for readability.

Model	Benchmark	Δ Mean	ΔQ_{50}	ΔQ_{75}	ΔQ_{90}
<i>PC-RLVR vs. GRPO+Entropy</i>					
Qwen3-1.7B	AIME24	2.8	3.1	3.2	3.9
	AIME25	2.4	2.4	3.0	3.1
	AIME26	2.2	1.5	2.8	6.3
	OlympiadBench	2.5	2.6	3.6	3.6
<i>PC-RLVR vs. w/o prior correction</i>					
Qwen3-4B	AIME24	1.7	1.5	3.0	2.4
	AIME25	2.2	1.7	2.3	4.1
	AIME26	2.0	1.6	2.3	2.7
	OlympiadBench	1.5	1.5	2.7	2.4

These results provide response-level rather than direct mode-level evidence. Distinct responses need not correspond to distinct reasoning modes, and higher surprisal is not inherently preferable because low-log-likelihood regions may also contain low-quality or redundant responses. We therefore interpret the observed rightward shifts as response-level evidence consistent with broader coverage of successful responses beyond the high-prior region of the initial policy. Together with the prior-dependent log-likelihood-gain analysis in Section 5.3, this complementary distributional evidence is consistent with the rebalancing effect of prior correction and its goal of improving coverage of successful responses.

D.5 ABLATION ANALYSIS

We further analyze the two design choices ablated in the main text: prior correction and the bounded arctangent transformation. All variants use Qwen3-4B and otherwise follow the same training configuration as PC-RLVR. For controlled comparison, the ablations in this subsection use $\tau = 0.2$. We additionally report the $\tau = 0.5$ PC-RLVR configuration used for the Qwen3-4B main result.

Removing prior correction. The *w/o prior correction* variant disables the correction weights, equivalently setting the correction factor to one for every response, while retaining the same GRPO backbone and asymmetric PPO clipping as PC-RLVR, with $\epsilon_{\text{low}} = 0.20$ and $\epsilon_{\text{high}} = 0.28$. This comparison isolates the contribution of prior correction from the enlarged upper clipping range.

Linear correction. To evaluate the role of bounded saturation, we replace the arctangent transformation with

$$\hat{\omega}_t^{\text{lin}}(Y) = 1 + \frac{2\tau\gamma}{\pi} \hat{Z}_t(Y), \quad (191)$$

Table 4: Ablation of prior correction and the bounded arctangent transformation on Qwen3-4B. We report Avg@64 and Pass@64 on each benchmark. Mean denotes the unweighted average across the five benchmarks.

Method	AIME24		AIME25		AIME26		OlympiadBench		AMC23		Mean	
	Avg@64	Pass@64	Avg@64	Pass@64	Avg@64	Pass@64	Avg@64	Pass@64	Avg@64	Pass@64	Avg@64	Pass@64
PC-RLVR ($\tau = 0.5$, main)	39.5	80.0	29.1	76.7	34.2	70.0	67.8	86.5	83.0	97.5	50.7	82.1
PC-RLVR ($\tau = 0.2$)	39.5	80.0	31.6	73.3	33.2	66.7	67.2	89.4	82.9	97.5	50.9	81.4
w/o prior correction	34.7	76.7	26.2	70.0	27.5	70.0	65.5	86.5	78.8	97.5	46.5	80.1
Linear correction ($\tau = 0.2$)	34.9	73.3	25.6	70.0	25.7	66.7	66.1	87.1	78.7	97.5	46.2	78.9

while retaining the same historical-reference statistic and all other components. The linear form matches the slope of the arctangent correction at the origin, so the two transformations respond similarly to small correction signals, while the linear correction removes the saturation for larger $|\hat{Z}_t(Y)|$. This comparison therefore isolates the role of bounding large correction magnitudes.

Results. Table 4 shows that both design choices are important for the final performance. Removing prior correction lowers the mean Avg@64 from 50.9 to 46.5 and the mean Pass@64 from 81.4 to 80.1. The reduction in Avg@64 appears on all five benchmarks, showing that the improvement of PC-RLVR cannot be attributed to the larger upper clipping threshold alone.

The linear-correction variant retains prior correction but removes the bounded saturation of the arctangent transformation. It obtains mean Avg@64 and Pass@64 of 46.2 and 78.9, respectively, substantially below the full method. The linear form is chosen to match the slope of the arctangent correction at $Z = 0$, so the two transformations behave similarly for small correction signals. Their main difference appears for larger $|Z|$: the arctangent transformation saturates, whereas the linear correction continues to increase in magnitude without bound. The resulting performance gap therefore indicates that prior correction benefits from explicitly limiting the magnitude of large corrections, rather than extrapolating them linearly.

D.6 SENSITIVITY ANALYSIS

We study sensitivity to the historical update rate β , the conservative-correction parameter κ , and the correction bound τ on Qwen3-4B. During training, we monitor Pass@1 every 10 optimization steps on all problems from AIME24 and AIME25 using temperature 0.6. For visualization, the reported Pass@1 trajectories are smoothed over training steps.

Under the EMA convention used throughout the paper, larger β places more weight on the current statistic and therefore updates the historical reference more rapidly. We compare $\beta \in \{0.5, 0.7, 0.9\}$ at $(\kappa, \tau) = (0, 0.2)$, compare $\kappa \in \{0, 1\}$ at $(\beta, \tau) = (0.7, 0.2)$, and compare $\tau \in \{0.2, 0.5\}$ at $(\beta, \kappa) = (0.7, 0)$.

Effect of the historical update rate. Figure 9 shows that the choice of β has a clear effect on the training trajectory. On Qwen3-4B, the intermediate setting $\beta = 0.7$ performs best overall, while $\beta = 0.5$ and $\beta = 0.9$ are less favorable over substantial portions of training.

This behavior is consistent with the role of β in the historical-reference construction. Smaller β places greater emphasis on older statistics, whereas larger β responds more strongly to recent statistics. As suggested by our analysis, β therefore controls the balance between historical averaging and adaptation to the evolving policy. The appropriate balance can depend on the model and training regime, making β an important tunable parameter rather than a universally fixed choice.

Effect of conservative correction. The parameter κ only affects responses satisfying $A_t(Y) < 0$ and $\hat{Z}_t(Y) > 0$, for which the unattenuated correction increases the magnitude of an already negative update. Figure 9 shows that, with β and τ fixed, $\kappa = 0$ yields a stronger Pass@1 trajectory than the symmetric setting $\kappa = 1$ over most of training.

This result is consistent with the motivation for the conservative branch in equation 9. Because the response-level statistic may contain within-mode variation in addition to information about mode

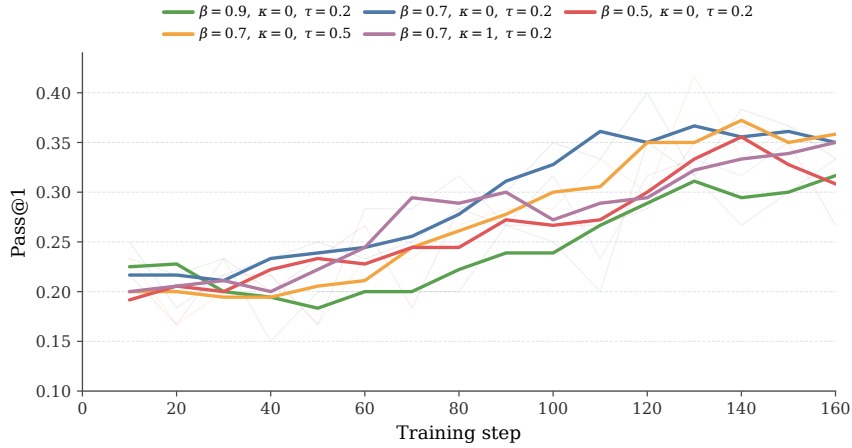


Figure 9: Sensitivity of PC-RLVR to β , κ , and τ on Qwen3-4B. We vary one parameter at a time around $(\beta, \kappa, \tau) = (0.7, 0, 0.2)$. Smoothed Pass@1 is shown throughout training.

allocation, fully amplifying negative updates on this branch can be overly aggressive. Attenuating this correction with $\kappa = 0$ performs better in the present setting.

Effect of the correction bound. At $(\beta, \kappa) = (0.7, 0)$, increasing τ from 0.2 to 0.5 produces a comparable Pass@1 trajectory, with the $\tau = 0.5$ run becoming competitive or stronger in the later stage of training. Its selected checkpoint further improves broader-sampling performance, and we therefore use $\tau = 0.5$ for the Qwen3-4B main configuration.

E LIMITATIONS AND FUTURE DIRECTIONS

Our current analysis still leaves several gaps between the idealized mode-level formulation and practical RLVR training. First, reasoning modes and their allocation probabilities are latent. PC-RLVR therefore implements prior correction using a response-level surprisal statistic rather than direct mode-level allocation information. Within-mode response variation can consequently introduce a discrepancy between the response-level correction and the idealized mode-level correction. Explicitly modeling reasoning modes could provide more direct allocation statistics and reduce this discrepancy.

Second, the correction acts multiplicatively on the group-relative advantage and therefore cannot create a learning signal when that advantage vanishes. In particular, if all sampled responses receive the same reward, prior correction alone cannot redistribute probability among them. The policy may therefore concentrate on a subset of correct modes before their relative allocation is sufficiently rebalanced, after which zero group-relative advantage prevents further redistribution even when the response-level statistic remains informative.

Finally, our mode-level theory conditions on a fixed prompt, whereas the practical historical reference is estimated across prompts within the training process. This introduces cross-prompt heterogeneity that is not explicitly represented in the current analysis. Explicit mode modeling, additional coverage-aware learning signals when group-relative advantage vanishes, and prompt-conditioned historical references are promising directions for extending the framework.

Algorithm 1 Prior-Corrected RLVR with Historical Reference

Require: Initial policy π_{θ_0} , prompt dataset $\mathcal{D}$, verifier R , rollout batch size B , group size N , training rounds T , correction parameters τ, κ , initial correction scale γ_0 , EMA coefficient β , variance bounds $v_{\min}, v_{\max}$, clipping thresholds $\epsilon_{\text{low}}, \epsilon_{\text{high}}$, and learning rate α .

- 1: Initialize historical reference statistics $\hat{\mu}_0$ and $\bar{v}_0$, set $\hat{v}_0 = \Pi_{[v_{\min}, v_{\max}]}(\bar{v}_0)$, and set $\gamma \leftarrow \gamma_0$,
- 2: **for** $t = 0, \dots, T - 1$ **do**
- 3: Sample a rollout prompt batch $\mathcal{B}_t = \{x_i\}_{i=1}^B \sim \mathcal{D}$.
- 4: For each x_i , sample $Y_{i,g} \sim \pi_{\theta_t}(\cdot | x_i)$, $g = 1, \dots, N$, and obtain rewards $r_{i,g} = R(x_i, Y_{i,g})$.
- 5: For each response token, compute $\ell_{i,g,j}^{(t)} := -\log \pi_{\theta_t}(Y_{i,g,j} | x_i, Y_{i,g,<j})$, and the response surprisal $S_t(Y_{i,g}) := \sum_{j=1}^{H(Y_{i,g})} \ell_{i,g,j}^{(t)}$.
- 6: Compute the group-relative advantages

$$A_{i,g} = \frac{r_{i,g} - \text{mean}_g(r_{i,g})}{\text{std}_g(r_{i,g})}.$$

- 7: Compute $\hat{Z}_t(Y_{i,g})$ using the historical reference statistics $(\hat{\mu}_t, \hat{v}_t)$ according to equation 7.
- 8: Compute the historical-reference correction $\hat{\omega}_t(Y_{i,g})$ according to equation 8, and apply the conservative correction rule in equation 9 to obtain $\tilde{\omega}_t(Y_{i,g})$.
- 9: Set

$$A_{i,g}^{\text{PC}} = \text{sg}(\tilde{\omega}_t(Y_{i,g})) A_{i,g}.$$

- 10: For each token, define the importance ratio

$$\varrho_{i,g,j}(\theta) := \frac{\pi_{\theta}(Y_{i,g,j} | x_i, Y_{i,g,<j})}{\pi_{\theta_t}(Y_{i,g,j} | x_i, Y_{i,g,<j})}.$$

- 11: For each token, define the clipped importance ratio

$$\bar{\varrho}_{i,g,j}(\theta) := \text{clip}(\varrho_{i,g,j}(\theta), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}).$$

- 12: Define the token-level clipped surrogate $L_{i,g,j}^{\text{PC}}(\theta) := \min\{\varrho_{i,g,j}(\theta) A_{i,g}^{\text{PC}}, \bar{\varrho}_{i,g,j}(\theta) A_{i,g}^{\text{PC}}\}$.
- 13: Form the policy objective

$$\mathcal{J}_t^{\text{PC}}(\theta) = \frac{1}{BN} \sum_{i=1}^B \sum_{g=1}^N \frac{1}{H(Y_{i,g})} \sum_{j=1}^{H(Y_{i,g})} L_{i,g,j}^{\text{PC}}(\theta).$$

- 14: Optimize $\mathcal{J}_t^{\text{PC}}$ over the rollout batch using the configured optimizer minibatches and learning rate α to obtain θ_{t+1} .
- 15: Using all BN responses in the completed rollout batch, define

$$a_{i,g}^{(t)} := \frac{S_t(Y_{i,g})}{H(Y_{i,g})}, \quad b_{i,g}^{(t)} := \frac{1}{H(Y_{i,g})} \sum_{j=1}^{H(Y_{i,g})} \left(\ell_{i,g,j}^{(t)}\right)^2.$$

- 16: Compute the rollout-batch statistics $\tilde{\mu}_t = \frac{1}{BN} \sum_{i=1}^B \sum_{g=1}^N a_{i,g}^{(t)}$, and

$$\tilde{v}_t = \frac{1}{BN} \sum_{i=1}^B \sum_{g=1}^N b_{i,g}^{(t)} - \frac{\left(\sum_{i=1}^B \sum_{g=1}^N a_{i,g}^{(t)}\right)^2 - \sum_{i=1}^B \sum_{g=1}^N \left(a_{i,g}^{(t)}\right)^2}{BN(BN - 1)}.$$

- 17: Update the historical reference statistics:

$$\hat{\mu}_{t+1} = (1 - \beta)\hat{\mu}_t + \beta\tilde{\mu}_t, \quad \bar{v}_{t+1} = (1 - \beta)\bar{v}_t + \beta\tilde{v}_t.$$

- 18: Project the variance used for response standardization: $\hat{v}_{t+1} = \Pi_{[v_{\min}, v_{\max}]}(\bar{v}_{t+1})$.
- 19: Update γ from the completed rollout statistics using the EMA-based adaptive rule described under PC-RLVR in Section C.3.
- 20: **end for**
- 21: **return** π_{θ_T} .