A Few Bad Apples Spoil the Bunch: Preventing Global Entropy Collapse Driven by a Small Set of Tokens in LLM Reasoning

Jaeeun Jang* **Hansle Lee** **Sangmin Kim**
AI Research Team, Hanwha Systems, Seongnam-si, Gyeonggi-do, Republic of Korea
{wkdwodms0779,hssarah13,kimsangmin603}@gmail.com

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) and Reinforcement Learning from Internal Feedback (RLIF) often fail to benefit from test-time compute due to *entropy collapse* and the resulting loss of reasoning diversity. We show that this collapse is driven not by uniform entropy decay, but by premature overconfidence at a small number of structurally critical decision points. Based on a token-level analysis of GRPO-style policy optimization, we propose SCOPE (Structural Collapse-aware Optimization via Partial Entropy control), which assigns each generated token a redistribution score and applies selective KL regularization to only the top $\sim 5\%$ of tokens under this score. Across model scales and architectures on math reasoning benchmarks, SCOPE consistently improves performance under both RLVR and RLIF settings, demonstrating that targeted entropy control at a vanishingly small subset of tokens is sufficient to sustain reasoning diversity and effective test-time scaling.

1 Introduction

Reinforcement learning (RL) has emerged as the dominant post-training paradigm for eliciting complex reasoning in large language models (LLMs) (Jaech et al., 2024; Guo et al., 2025). Unlike supervised fine-tuning (SFT), which imposes the surface form of human demonstrations, RL supplies only outcome-level feedback and lets the model discover its own reasoning trajectories (Guo et al., 2025), while exhibiting stronger robustness to distribution shifts (Chu et al., 2025). Crucially, RL does not implant fundamentally new capabilities; rather, it redistributes probability mass across reasoning trajectories that already lie within the base model's support (Yue et al., 2025; Tsilivis et al., 2025)—amplifying latent reasoning-promoting tokens that next-token pretraining leaves underweighted (Zhao et al., 2025a; Chen et al., 2025b; Ouyang et al.).

This redistribution, however, comes at a structural cost. RL systematically sharpens the policy, lifting single-sample accuracy (Pass@1) but eroding Pass@k by collapsing the diversity of reasoning paths the model is willing to explore (Zhu et al., 2025; Chen et al., 2025a). Diversity, far from a cosmetic concern, is the substrate on which test-time scaling—multi-sample decoding, self-consistency, and tree search—fundamentally relies. Whereas training-time scaling (Hoffmann et al., 2022) remains a reliable lever where scaling model size and compute yields predictable gains, test-time scaling is still far less understood: its success hinges on sustaining distributional breadth so that repeated attempts explore distinct solutions rather than redundantly resampling the same trajectory. Without this breadth, additional inference compute is wasted on redundant samples, and the model cannot extend into capability regions it has not yet covered. The central tension of RL post-training is therefore an exploration–exploitation dilemma at the token level: how do we raise the probability of high-quality reasoning trajectories while preserving the breadth of the underlying distribution?

A natural response to this tension is to directly regularize the policy's entropy during RL training. Existing methods in this family, however, apply entropy regularization uniformly across the decoding sequence (Yao et al., 2025)—wasting budget where determinism is desirable, while under-protecting the few positions that actually shape reasoning. Token-level analyses make this mismatch explicit: most tokens are ordinary continuations, and only a small subset of *branch-defining tokens* can alter the reasoning trajectory (Wang et al., 2025; Li et al., 2025). This implies that the true locus of reasoning diversity is far more concentrated than prevailing methods assume. Yet attempts to isolate these tokens rely on heuristic proxies—e.g., raw entropy thresholds—that are disconnected from the optimization dynamics driving collapse.

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 13134–13154

In this work, we provide a mathematically rigorous account of where and why entropy collapses under GRPO-style policy optimization. Through a fixed-context, token-level analysis of GRPO updates in logit coordinates, we derive an *exact, non-asymptotic decomposition* of the expected redistribution magnitude into a smooth leading geometry term and a token–outcome coupling residual. Building on this, we propose **SCOPE** (**S**tructural **C**ollapse-aware **O**ptimization via **P**artial **E**ntropy control): we turn the leading geometry into a batch-computable instance-level score and apply selective KL regularization to the top $\sim$5% of tokens ranked by it. Unlike heuristic selection rules or uniform entropy penalties, SCOPE intervenes precisely where the update rule itself predicts collapse-inducing redistribution. Across model scales and architectures on standard math reasoning benchmarks, SCOPE consistently improves both Pass@1 and Pass@k under RLVR and RLIF, outperforming prior entropy-control baselines. These results highlight that *which updates are regularized* is far more important than *how much entropy is preserved*.

2 Related Work

2.1 Reinforcement Learning for Reasoning and the Emergence of RLIF

Reinforcement learning with verifiable rewards (RLVR) has become a central approach for improving reasoning performance in LLMs, particularly in conjunction with test-time scaling (Jaech et al., 2024; Guo et al., 2025). While RLVR improves single-sample accuracy and robustness (Chu et al., 2025; Zhu et al., 2025), it often leads to a significant reduction in reasoning-path diversity, in some cases more severely than supervised fine-tuning (SFT) (Chen et al., 2025a). This loss of diversity limits the effectiveness of test-time scaling strategies such as self-consistency and majority voting.

To address the reliance on external reward design, Reinforcement Learning from Internal Feedback (RLIF) has been proposed, where models are optimized using intrinsic signals derived from their own outputs (Zhao et al., 2025b; Zhang et al., 2025; Prabhudesai et al., 2025). Prior to RLIF, several studies demonstrated that internal confidence and uncertainty can serve as effective proxies for reasoning quality at test time, including probability-disparity–based chain-of-thought selection (Wang and Zhou, 2024), confidence-weighted aggregation and voting (Taubenfeld et al., 2025; Fu et al., 2025; Razghandi et al., 2025), and entropy-based inference strategies (Zhou et al., 2023; Agarwal et al., 2025). While these approaches operate purely at inference time, they establish the feasibility of using internal signals without external supervision. Building on this insight, INTUITOR introduces a representative RLIF framework that uses self-certainty as a reward signal, achieving performance comparable to RLVR while improving out-of-distribution generalization (Zhao et al., 2025b). Despite this promise, RLIF inherits a key limitation of RLVR: intrinsic signals are inherently noisy and imperfectly calibrated, and confidence-based optimization often induces entropy collapse and overconfidence, leading to premature convergence and degraded reasoning diversity (Zhang et al., 2025).

2.2 Entropy Control for RL-based Reasoning

To mitigate the loss of reasoning diversity under RL-based post-training, several recent methods have introduced entropy-aware or diversity-aware optimization strategies. Wang et al. (Wang et al., 2025) analyze token-level entropy in reasoning models and show that entropy is highly unevenly distributed, with a small subset of high-entropy tokens contributing disproportionately to exploration. Notably, they demonstrate that restricting updates to these informative tokens can preserve entropy more effectively than uniformly updating all tokens, suggesting that entropy collapse depends not only on the magnitude of optimization but also on where it is applied. Another line of work focuses on explicit entropy regularization during RL optimization. Diversity-Aware Policy Optimization (DA-PO) introduces entropy-preserving regularization on reward-positive trajectories, encouraging broader exploration and mitigating premature convergence to a single reasoning path (Yao et al., 2025). More recent work provides a mechanistic perspective on entropy collapse. Cui et al. (Cui et al., 2025) show that entropy reduction in RL-based reasoning can be understood through the covariance between reward-weighted update signals and entropy-sensitive directions in the policy. Based on this analysis, KL-Cov selectively constrains updates on high-covariance actions (or token instances) using KL regularization, improving exploration and test-time scaling performance. This line of work is particularly important because it connects entropy collapse directly to the underlying optimization dynamics rather than treating it as a purely phenomenological issue.

2.3 Limitations of Existing Approaches

Despite these advances, existing methods exhibit several limitations. First, many approaches rely on heuristic or coarse structural assumptions, rather than directly identifying the token instances responsible for entropy collapse, and thus do not constitute a fundamental solution that generalizes consistently across different models. Second, entropy regularization is often applied uniformly or stochastically, without distinguishing between collapse-inducing and irrelevant tokens. Third, while recent work provides mechanistic insights, it does not fully translate these insights into a principled and well-justified token-level selection criterion. In particular, the connection between theoretical formulations and actionable token-level interventions remains underdeveloped, leaving room for selection rules that are more tightly grounded in the underlying optimization dynamics. These limitations highlight the need for a more precise, token-level characterization of entropy collapse, enabling selective regularization that directly addresses the underlying cause of premature convergence.

3 Proposed Method: SCOPE

We develop SCOPE from an *exact*, fixed-context analysis of GRPO-style policy optimization. Rather than appealing to heuristics or asymptotic approximations, we derive a non-asymptotic identity that cleanly decomposes the expected redistribution magnitude at every token position into a smooth geometric term and a token–outcome coupling residual. This identity tells us which token positions are structurally collapse-critical, motivating a principled instance-level surrogate score that SCOPE uses to select a small subset of tokens for targeted KL regularization. The complete step-by-step derivation is provided in Appendix A.

3.1 Preliminaries

Let x denote the input prompt and $\tau = (y_1, \ldots, y_T)$ a completion sampled from the policy $\pi_\theta(\cdot \mid x)$. For a fixed position t we write the context as $c_t = (x, y_{<t})$, the pre-softmax logits as $z_t = (z_{t,u})_{u \in \mathcal{V}}$, and the conditional probability as

$$p_u \;:=\; \pi_\theta(u \mid c_t) \;=\; \frac{e^{z_{t,u}}}{\sum_{k \in \mathcal{V}} e^{z_{t,k}}}, \quad 0 < p_u < 1. \tag{1}$$

Under GRPO, the same trajectory-level advantage $\hat{A}(\tau) = (r(\tau) - \mu_r)/\sigma_r$ is broadcast to every token of τ (Shao et al., 2024); crucially, $\hat{A}(\tau) \not\perp y_t$ in general, so our analysis will *not* assume independence between token identity and the advantage.

Local on-policy score. The unclipped, on-policy token-time score that drives a GRPO update is

$$\nabla_\theta \mathcal{J}_t(\theta) \;=\; \mathbb{E}_{\tau \sim \pi_\theta(\cdot|x)}\big[\, \hat{A}(\tau)\, \nabla_\theta \log \pi_\theta(y_t \mid c_t) \,\big]. \tag{2}$$

All subsequent statements concern ascent-direction signals induced by (2) in logit coordinates; they are independent of optimizer-specific transforms (Adam preconditioning, clipping, KL penalties, minibatch averaging).

Logit-coordinate contribution. A direct computation of the log-softmax derivative with respect to the logits (Appendix A.7) yields the identity

$$\frac{\partial}{\partial z_{t,v}} \log \pi_\theta(y_t \mid c_t) \;=\; \mathbf{1}[y_t = v] - p_v. \tag{3}$$

We define the *realized local ascent contribution* in logit coordinates at position t for token v as

$$\Delta^{\mathrm{loc}}_{t,v} \;:=\; \eta\, \hat{A}(\tau)\, \big(\mathbf{1}[y_t = v] - p_v\big), \tag{4}$$

where $\eta > 0$ is a step-size-like scalar. The quantity of interest is the expected absolute magnitude of this contribution, conditional on the context:

$$\begin{aligned} S_v \;&:=\; \mathbb{E}_t\big[\, \big|\Delta^{\mathrm{loc}}_{t,v}\big| \,\big] \\ &=\; \eta\, \mathbb{E}_t\big[\, |\hat{A}(\tau)|\, \big|\mathbf{1}[y_t = v] - p_v\big| \,\big], \end{aligned} \tag{5}$$

where $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid c_t]$ and, under the conditioning on c_t, the probability p_v is a *deterministic constant*. S_v is *not* a direct entropy decrement; it measures the exact conditional magnitude of local logit-coordinate redistribution induced on token v.

3.2 Exact Non-Asymptotic Decomposition

Define the token-conditioned advantage means

$$\begin{aligned} \alpha_v \;&:=\; \mathbb{E}\big[\, |\hat{A}(\tau)| \;\big|\; c_t, y_t = v \,\big], \\ \beta_v \;&:=\; \mathbb{E}\big[\, |\hat{A}(\tau)| \;\big|\; c_t, y_t \neq v \,\big], \end{aligned} \tag{6}$$

together with the global conditional scale $A_{\mathrm{scale}} := \mathbb{E}_t[|\hat{A}(\tau)|] = \sum_u p_u \alpha_u$. Our main structural result is an *exact* decomposition of S_v into a smooth leading term and a token–outcome coupling residual.

Proposition 1 (Exact local-redistribution decomposition)**.** *Under mild integrability* ($\mathbb{E}[|\hat{A}(\tau)| \mid c_t, y_t = u] < \infty$ *for all* u*), the following identity holds* exactly *(no asymptotic or small-step approximation) for every context* c_t *and token* $v \in \mathcal{V}$*:*

$$\boxed{\; S_v \;=\; L_v \;+\; R_v \;} \tag{7}$$

where the leading geometry term and the token–outcome coupling residual are defined by

$$L_v \;:=\; 2\eta\, A_{\text{scale}}\, p_v(1-p_v), \tag{8}$$

$$R_v \;:=\; \eta\,(1-2p_v)\,\mathrm{Cov}_t\big(|\hat{A}(\tau)|, \mathbf{1}[y_t = v]\big) \tag{9}$$

i.e., $\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v]) = p_v(1-p_v)(\alpha_v - \beta_v)$*, so* $R_v = \eta(1-2p_v)\,p_v(1-p_v)(\alpha_v - \beta_v)$.

The complete proof proceeds by a law-of-total-expectation decomposition of the absolute-value expectation into the events $\{y_t = v\}$ and $\{y_t \neq v\}$, followed by a weighted zero-mean simplification of deviation variables $\delta_u := \alpha_u - A_{\text{scale}}$. The full derivation, including intermediate lemmas and identities, is deferred to Appendix A, with the main proof chain summarized in Appendix A.19.

What the two terms mean. The leading term $L_v \propto p_v(1-p_v)$ is a smooth, symmetric logit-coordinate geometry factor that depends only on the current policy probability and the global advantage scale A_{scale}. It is maximized in the intermediate probability regime and vanishes at both $p_v \to 0$ and $p_v \to 1$. The residual R_v is modulated by the *token–outcome coupling* $\alpha_v - \beta_v$: it is essentially zero when sampling v versus not sampling it does not change the expected trajectory quality, and grows in magnitude precisely at *branch-defining* positions where $y_t = v$ meaningfully alters the downstream outcome. This provides a principled, analytic distinction between *ordinary continuation tokens* ($R_v \approx 0$, $S_v \approx L_v$) and *branch-defining tokens* (non-negligible R_v); see Appendix A.20.

3.3 Drivers of Entropy Collapse

Proposition 1 refines prior accounts of entropy collapse in two ways. First, it shows that local redistribution magnitude is driven *in expectation* by the product $p_v(1-p_v)$, not by low-probability tokens alone: although a low-probability token receives a large per-update signal when sampled, the sampling event itself occurs with probability p_v, and the two effects cancel to yield a net $p_v(1-p_v)$ scaling. Consequently, the dominant contributors lie in the *intermediate* probability regime—tokens sampled often enough to accumulate updates while still admitting non-trivial redistribution.

Second, the *sign* of the advantage determines the direction of the induced local dynamics (Appendix A.10). For $\hat{A}(\tau) > 0$, the sampled token is reinforced while all alternatives are suppressed (local one-step *sharpening*); for $\hat{A}(\tau) < 0$, probability mass is redistributed away from the sampled token (local *flattening*). Monotonic entropy decay is therefore driven specifically by the positive-advantage regime, in which repeated reinforcement of high-$p_v(1-p_v)$ tokens propagates certainty across the sequence and collapses exploration.

Taken together, entropy collapse is driven by tokens that (i) receive positive-advantage signals, (ii) have sufficiently high policy probability to be sampled repeatedly, and (iii) exhibit non-negligible expected update magnitude under the $p_v(1-p_v)\,|\hat{A}|$ scaling. These tokens form only a small subset of the action space; nevertheless, their repeated reinforcement induces global entropy decay.

3.4 Instance-Level Redistribution Score

Although (7) is exact, neither α_v, β_v, nor $\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v])$ is directly accessible during training. SCOPE therefore adopts the following *modeling choice* (not an approximation theorem):

$$S_v \;\approx\; L_v \;=\; 2\eta\, A_{\text{scale}}\, p_v(1-p_v), \tag{10}$$

i.e., we use the leading geometry as an instance-level surrogate. Because collapse-aware control targets the *sharpening* direction (Section 3.3), we further restrict to positive-advantage events. For a rollout k and step t with sampled token $y_t^{(k)}$, current policy probability $p_{k,t} := \pi_\theta(y_t^{(k)} \mid c_t^{(k)})$ where $c_t^{(k)} := (x, y_{<t}^{(k)})$, and trajectory-level advantage $\hat{A}_{k,t}$, we define the *Redistribution* score as

$$\boxed{s_{k,t} \;:=\; p_{k,t}\,(1-p_{k,t})\,\big[\hat{A}_{k,t}\big]_+} \tag{11}$$

where $[\hat{A}]_+ := \max(\hat{A}, 0)$. The score (11) simultaneously captures three properties: (i) the leading redistribution geometry $p_{k,t}(1-p_{k,t})$; (ii) restriction to the positive-advantage sharpening side; and (iii) computability at the level of individual sampled-token instances. The rationale for using $[\hat{A}]_+$, and the precise relationship between $s_{k,t}$ and the exact target S_v, is detailed in Appendix A.21.

3.5 Structure-Aware KL Regularization

SCOPE applies KL regularization selectively to the token instances with the largest redistribution score. Let $\mathcal{B}$ denote the set of valid generated token instances in a minibatch. We compute $s_{k,t}$ online for every $(k,t) \in \mathcal{B}$ and form the top-$q\%$ threshold

$$\tau_q \;=\; \mathrm{Quantile}_{1-q}\big(\{\, s_{k,t} : (k,t) \in \mathcal{B} \,\}\big), \tag{12}$$

with the corresponding binary mask

$$m_{k,t} \;:=\; \mathbf{1}\big[\, s_{k,t} \geq \tau_q \,\big]. \tag{13}$$

Denoting the token-wise KL divergence against the reference policy conditioned on the context $c_t^{(k)}$, as

$$D_{k,t} = D_{\mathrm{KL}}\big(\pi_\theta(\cdot \mid c_t^{(k)}) \,\big\|\, \pi_{\mathrm{ref}}(\cdot \mid c_t^{(k)})\big), \quad (14)$$

we incorporate the selective constraint into the GRPO objective by applying KL regularization:

$$\max_\theta \; \mathcal{J}_{\mathrm{GRPO}}(\theta) \; - \; \lambda \frac{\sum_{(k,t)\in\mathcal{B}} m_{k,t}\, D_{k,t}}{\sum_{(k,t)\in\mathcal{B}} m_{k,t} + \varepsilon}, \quad (15)$$

where $\lambda > 0$ is the penalty weight and $\varepsilon > 0$ stabilizes the denominator. The mask selects the token instances at which Proposition 1 predicts the largest positive-advantage redistribution magnitude.

4 Experimental Setup

Training setup. Experiments are conducted using the Open-R1 framework on the MATH training set (7,500 problems (Hendrycks et al., 2021)) with Qwen2.5 Base models (1.5B, 3B, 7B). Each update samples 8 candidate solutions for 128 problems. To ensure a fair comparison, RLVR and RLIF share identical hyperparameters and differ only in reward design (see Appendix B for compute details). RLVR employs verifiable rewards based on answer correctness, while RLIF uses *self-certainty* (Zhao et al., 2025b) as an intrinsic feedback signal.

Evaluation. We evaluate on five held-out benchmarks: **GSM8K** (Cobbe et al., 2021), **MATH500** (Lightman et al., 2023), **AMC**, **AIME2024**, and **AIME2025** (Li et al., 2024). For each problem, we generate 16 stochastic samples ($T = 1.0$) and report exact-match accuracy (Acc@16).

4.1 Redistribution Ablation

We conduct an ablation study to isolate the effect of redistribution–based token selection. Specifically, we apply KL regularization to the top-$q\%$ tokens ranked by the redistribution score (Eq. 11). We vary q from 0% to 100% in increments of 5%.

As shown in Fig. 1, performance consistently exhibits a non-monotonic trend across all benchmarks and model scales, peaking around $q \approx 5\%$. Applying KL regularization to a small subset of high redistribution tokens yields the best performance, while increasing q degrades accuracy due to over-regularization that suppresses beneficial updates beyond collapse-critical positions. This result demonstrates that collapse-inducing tokens are highly sparse, and that selective regularization based solely on redistribution score is sufficient to capture the dominant drivers of entropy collapse.

4.2 Does Token Selection Matter under a Fixed Regularization Budget?

The ablation in Section 4.1 reveals that concentrating KL regularization on the top-5% of tokens maximizes reasoning performance. We formalize this configuration as **SCOPE** and ask a question: under a fixed regularization budget, is downstream performance governed by *how many* tokens are regularized or by *which* tokens are selected?

Token-selection strategies. All variants regularize exactly 5% of tokens to ensure a strictly fair comparison; only the *selection criterion* differs. **Uniform** selects these tokens uniformly at random from the entire sequence. **Positive-Adv.** restricts random selection to positive-advantage tokens only. **Forking** samples randomly from the forking-token set (i.e., the top-20% highest-entropy tokens). **Fork + Adv.** further restricts random sampling to the intersection of forking and positive-advantage tokens. **Non-Forking** samples exclusively from non-forking positions. **KL-Cov** follows the covariance-aware entropy-control approach proposed in (Cui et al., 2025). Finally, **SCOPE (Ours)** selects the top-5% of tokens ranked by redistribution score (see Table 1).

Entropy Collapse Metric. We report d_{collapse}, a proxy for entropy collapse during RL training. Concretely, d_{collapse} quantifies the relative reduction in the fraction of forking tokens (top-20% by entropy) induced by RL optimization; lower values indicate greater preservation of reasoning diversity.

Results. Although every variant regularizes the same proportion of tokens, performance differs substantially—demonstrating that the entropy–performance trade-off is governed by *where* entropy is preserved, not *how much*. **(i)** SCOPE achieves the highest accuracy and the lowest d_{collapse} across all model scales and benchmarks under both RLVR and RLIF. **(ii)** Placement quality dominates quantity: Non-Forking, which avoids structurally critical positions, often underperforms even Uniform, while Fork + Adv. and KL-Cov attain comparable d_{collapse} to SCOPE yet fall short in accuracy—indicating that indiscriminate entropy preservation retains uninformative rather than reasoning-relevant diversity. **(iii)** RLIF variants exhibit higher d_{collapse} than RLVR, reflecting the stronger entropy-reducing pressure of confidence-based intrinsic rewards and underscoring the importance of targeted regularization under RLIF.

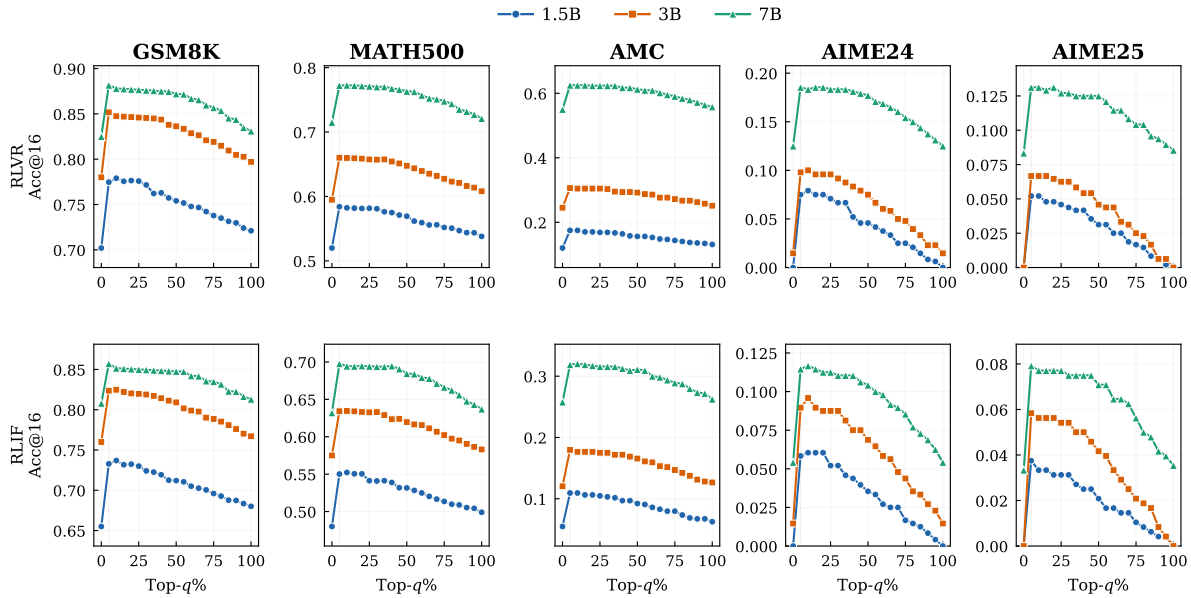

Figure 1: Ablation of redistribution–based token selection. KL regularization is applied to the top-$q\%$ tokens ranked by redistribution score, with q varied from 0% to 100%. Results are shown for RLVR (top) and RLIF (bottom).

4.3 SCOPE as a Plug-in Enhancement

Table 2 evaluates SCOPE as a drop-in addition to existing RLVR and RLIF methods on Qwen2.5 Base models. On the RLVR side, **Forking RL** (Wang et al., 2025) restricts policy-gradient updates to high-entropy forking-token positions while masking gradients elsewhere; **Div** (Yao et al., 2025) applies diversity-aware regularization across the full sequence; and **KL-Cov** (Cui et al., 2025) follows the covariance-based entropy-control objective. On the RLIF side, **Self-Certainty** (Zhao et al., 2025b) uses the KL divergence from a uniform distribution as a trajectory-level confidence signal; **Tok-Entropy** (Prabhudesai et al., 2025) applies a dense per-token entropy penalty; and **Traj-Entropy** (Agarwal et al., 2025) aggregates token entropies into a sequence-level objective. All methods are implemented identically to the original formulations in their respective prior works. Each **+ SCOPE** variant retains the original objective unchanged and only adds KL regularization on the top-5% of tokens ranked by redistribution score.

Two findings stand out. **(i)** SCOPE yields consistent gains across every base method, model scale, and paradigm, with the largest improvements appearing on harder benchmarks (AMC, AIME24/25) where preserving entropy at collapse-critical tokens has the greatest impact. **(ii)** Under RLIF, the gains from SCOPE are proportionally larger than under RLVR—e.g., Self-Certainty + SCOPE improves AIME24 from 0.054 to 0.115 at the 7B scale—reflecting that confidence-based intrinsic objectives impose stronger entropy-reducing pressure, making targeted regularization more critical.

4.4 Robustness under Stochastic Training

To evaluate whether redistribution-based token selection introduces instability under stochastic optimization, we train each configuration with 10 independent seeds and report per-seed Acc@16 in Figure 2. Three patterns emerge. **(i)** On benchmarks where both methods achieve non-negligible accuracy (GSM8K, MATH500, AMC, and AIME24/25 at the 7B scale), SCOPE seed distributions are tighter than the corresponding baselines. For example, at the 7B scale under RLIF, INTUITOR exhibits a standard deviation of 0.013 on AIME24 versus 0.009 for SCOPE, while the mean improves from 0.052 to 0.113. **(ii)** The stabilization effect is more pronounced under RLIF than RLVR, consistent with confidence-based intrinsic rewards amplifying trajectory-level stochasticity, making targeted regularization more impactful. **(iii)** On harder benchmarks at smaller scales (e.g., 1.5B AIME24/25), baseline seeds cluster at zero due to floor effects, making variance comparisons uninformative; SCOPE's wider spread is a natural consequence of achieving higher mean accuracy. Across all configurations, SCOPE improves every seed rather than relying on outliers, indicating a systematic effect of targeted entropy preservation.

Table 1: Comparison of token-selection strategies for KL regularization. All baseline methods apply KL regularization to the same proportion of tokens (≈5%). Best results are in **bold**.

Scale	Method	RLVR						RLIF					
		GSM8K	MATH500	AMC	AIME24	AIME25	$d_{\text{collapse}}\downarrow$	GSM8K	MATH500	AMC	AIME24	AIME25	$d_{\text{collapse}}\downarrow$
Qwen2.5-1.5B	Uniform	0.708	0.521	0.112	0.000	0.000	0.28	0.690	0.510	0.076	0.000	0.000	0.45
	Positive-Adv.	0.724	0.533	0.126	0.015	0.000	0.26	0.709	0.524	0.094	0.015	0.000	0.42
	Forking	0.722	0.535	0.126	0.015	0.000	0.23	0.712	0.522	0.094	0.015	0.000	0.42
	Fork + Adv.	0.737	0.560	0.139	0.033	0.015	0.21	0.714	0.532	0.102	0.026	0.015	0.36
	Non-Forking	0.688	0.500	0.096	0.000	0.000	0.33	0.656	0.481	0.055	0.000	0.000	0.48
	KL-Cov (Cui et al., 2025)	0.738	0.560	0.138	0.054	0.033	0.24	0.713	0.531	0.101	0.033	0.015	0.41
	SCOPE (Ours)	**0.773**	**0.585**	**0.173**	**0.075**	**0.050**	**0.19**	**0.732**	**0.549**	**0.108**	**0.058**	**0.038**	**0.38**
Qwen2.5-3B	Uniform	0.786	0.597	0.232	0.054	0.031	0.24	0.765	0.580	0.140	0.033	0.015	0.39
	Positive-Adv.	0.798	0.605	0.256	0.072	0.044	0.21	0.783	0.598	0.165	0.054	0.033	0.36
	Forking	0.800	0.610	0.261	0.071	0.044	0.17	0.788	0.596	0.165	0.062	0.038	0.35
	Fork + Adv.	0.813	0.624	0.270	0.086	0.054	0.15	0.802	0.618	0.174	0.072	0.044	0.32
	Non-Forking	0.767	0.576	0.215	0.012	0.000	0.28	0.761	0.566	0.121	0.009	0.000	0.40
	KL-Cov (Cui et al., 2025)	0.815	0.626	0.274	0.086	0.054	0.19	0.806	0.617	0.173	0.082	0.050	0.38
	SCOPE (Ours)	**0.851**	**0.661**	**0.305**	**0.096**	**0.067**	**0.14**	**0.824**	**0.634**	**0.180**	**0.090**	**0.058**	**0.32**
Qwen2.5-7B	Uniform	0.818	0.709	0.553	0.103	0.069	0.22	0.802	0.650	0.280	0.054	0.033	0.38
	Positive-Adv.	0.832	0.725	0.569	0.123	0.083	0.18	0.820	0.676	0.301	0.086	0.056	0.35
	Forking	0.829	0.723	0.575	0.123	0.081	0.16	0.818	0.676	0.301	0.086	0.054	0.33
	Fork + Adv.	0.846	0.747	0.591	0.160	0.110	0.13	0.839	0.690	0.312	0.092	0.060	0.30
	Non-Forking	0.800	0.690	0.527	0.095	0.063	0.25	0.789	0.633	0.259	0.033	0.015	0.39
	KL-Cov (Cui et al., 2025)	0.841	0.750	0.584	0.141	0.096	0.17	0.838	0.689	0.311	0.092	0.060	0.36
	SCOPE (Ours)	**0.882**	**0.773**	**0.628**	**0.185**	**0.131**	**0.12**	**0.856**	**0.697**	**0.319**	**0.115**	**0.081**	**0.31**

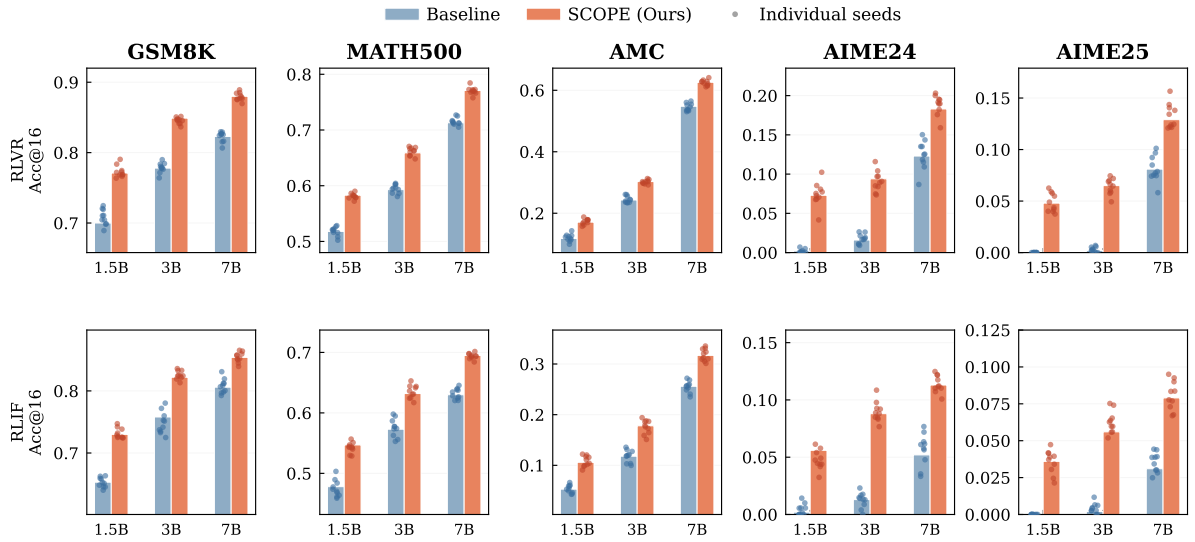

Figure 2: **Multi-seed evaluation ($N = 10$) under RLVR (top) and RLIF (bottom).** Bars denote mean Acc@16; dots denote individual seed results. RLVR baselines use GRPO; RLIF baselines use INTUITOR (Zhao et al., 2025b).

4.5 Test-Time Scaling Analysis

Figure 3 reports Pass@k curves on the 7B model to characterize how SCOPE shapes the reasoning distribution under varying test-time compute budgets; per-scale improvement curves (ΔPass@k = SCOPE − Baseline) across all model scales are provided in Appendix C. Three findings stand out. **(i)** SCOPE achieves the largest absolute gains at small k, indicating improved sampling efficiency: by preserving entropy at collapse-critical tokens, SCOPE concentrates probability mass on solution-relevant trajectories, yielding higher first-attempt success rates. **Notably, this Pass@1 improvement does not come at the expense of Pass@k at larger** k—a departure from the exploration–exploitation trade-off exhibited by standard RL post-training, where single-sample accuracy and sampling coverage are typically sacrificed against one another (Appendix C.1). **(ii)** On harder benchmarks (AMC, AIME24/25), SCOPE's advantage *grows* with k (Figure 4), demonstrating that redistribution-guided regularization sustains the reasoning diversity needed for effective multi-sample inference. On saturating benchmarks (MATH500), the gain naturally shrinks as both methods approach ceiling coverage. **(iii)** The dynamics differ markedly between paradigms. Under RLIF, the base model surpasses both SCOPE and INTUITOR at large k on AIME24/25 (Figure 3, bottom right), confirming that confidence-based intrinsic objectives reduce diversity beyond the point of utility. SCOPE

Table 2: **Integrating SCOPE with existing RLVR and RLIF methods on Qwen2.5 Base models.** Left: RLVR methods with verifiable rewards. Right: RLIF methods with intrinsic feedback only. Each base method is evaluated standalone and with + SCOPE. All + SCOPE variants apply KL regularization to the top-5% of tokens ranked by redistribution score. Best results per model scale and paradigm are in **bold**.

		RLVR						RLIF				
Scale	**Method**	GSM8K	MATH	AMC	AIME24	AIME25	**Method**	GSM8K	MATH	AMC	AIME24	AIME25
Qwen2.5-1.5B	Baseline†	0.702	0.520	0.120	0.000	0.000	Baseline‡	0.582	0.450	0.040	0.000	0.000
	Forking RL (Wang et al., 2025)	0.745	0.556	0.147	0.035	0.015	Self-Cert. (Zhao et al., 2025b)	0.655	0.480	0.055	0.000	0.000
	+ SCOPE	0.764	0.576	0.166	0.064	0.044	+ SCOPE	0.732	0.549	0.108	0.058	**0.038**
	Div (Yao et al., 2025)	0.716	0.533	0.129	0.000	0.000	Tok-Ent. (Prabhudesai et al., 2025)	0.672	0.493	0.068	0.015	0.000
	+ SCOPE	**0.768**	0.579	**0.170**	0.069	0.044	+ SCOPE	**0.738**	**0.552**	**0.109**	**0.060**	**0.038**
	KL-Cov (Cui et al., 2025)	0.738	0.560	0.138	0.054	0.033	Traj-Ent. (Agarwal et al., 2025)	0.648	0.462	0.049	0.014	0.000
	+ SCOPE	0.766	**0.581**	**0.170**	**0.075**	**0.046**	+ SCOPE	0.725	0.538	0.102	0.052	0.031
Qwen2.5-3B	Baseline†	0.780	0.595	0.245	0.015	0.000	Baseline‡	0.673	0.544	0.093	0.000	0.000
	Forking RL (Wang et al., 2025)	0.812	0.620	0.269	0.066	0.042	Self-Cert. (Zhao et al., 2025b)	0.760	0.575	0.120	0.015	0.000
	+ SCOPE	0.842	0.652	0.297	0.090	0.058	+ SCOPE	0.824	0.634	**0.180**	**0.090**	**0.058**
	Div (Yao et al., 2025)	0.793	0.603	0.249	0.015	0.000	Tok-Ent. (Prabhudesai et al., 2025)	0.775	0.588	0.131	0.036	0.015
	+ SCOPE	0.846	0.656	**0.300**	0.088	0.056	+ SCOPE	**0.826**	**0.637**	**0.180**	0.089	0.056
	KL-Cov (Cui et al., 2025)	0.815	0.626	0.274	0.086	0.054	Traj-Ent. (Agarwal et al., 2025)	0.746	0.568	0.112	0.024	0.000
	+ SCOPE	**0.847**	**0.660**	0.298	**0.094**	**0.060**	+ SCOPE	0.808	0.630	0.169	0.079	0.050
Qwen2.5-7B	Baseline†	0.825	0.715	0.550	0.125	0.083	Baseline‡	0.742	0.598	0.236	0.056	0.033
	Forking RL (Wang et al., 2025)	0.834	0.731	0.582	0.133	0.090	Self-Cert. (Zhao et al., 2025b)	0.808	0.632	0.258	0.054	0.033
	+ SCOPE	0.873	0.762	0.620	0.160	0.110	+ SCOPE	0.856	**0.697**	0.319	**0.115**	**0.081**
	Div (Yao et al., 2025)	0.827	0.716	0.553	0.125	0.083	Tok-Ent. (Prabhudesai et al., 2025)	0.821	0.640	0.270	0.048	0.027
	+ SCOPE	0.877	0.766	**0.622**	0.169	0.117	+ SCOPE	**0.858**	**0.697**	**0.320**	**0.115**	0.079
	KL-Cov (Cui et al., 2025)	0.841	0.750	0.584	0.141	0.096	Traj-Ent. (Agarwal et al., 2025)	0.801	0.614	0.246	0.044	0.023
	+ SCOPE	**0.884**	**0.770**	**0.622**	**0.171**	**0.123**	+ SCOPE	0.852	0.680	0.305	0.108	0.073

†GRPO with verifiable rewards. ‡Base model without RL training.

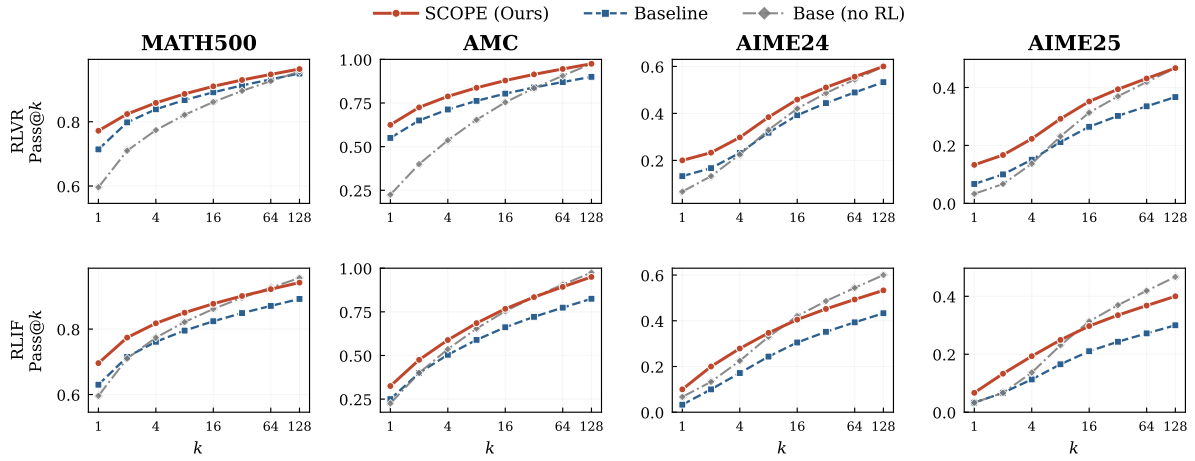

Figure 3: **Pass@k scaling on Qwen2.5-7B under RLVR (top) and RLIF (bottom).** SCOPE (solid) is compared against the trained baseline—GRPO for RLVR, INTUITOR for RLIF—(dashed) and the base model without RL training (dash-dotted). Results for 1.5B and 3B are in Appendix C.

counteracts this more effectively than INTUITOR—maintaining competitive Pass@128 while substantially improving Pass@1—but the base model's inherent diversity remains dominant at the 7B scale when the sampling budget is large enough. At smaller scales (1.5B, 3B), this crossover disappears: the base model lacks sufficient knowledge to benefit from diversity alone, and SCOPE outperforms all methods at every k (Appendix C). This scale-dependent disappearance of the crossover admits a broader implication: the k-range over which RL post-training confers benefit is itself scale-dependent, suggesting that the value of RL post-training must be evaluated jointly with model scale and sampling budget (Appendix C.2).

4.6 Generalization Across Architectures

To evaluate whether SCOPE generalizes beyond Qwen2.5-7B-Base, we conduct additional experiments on two architectures: Qwen2.5-Math-7B, a math-specialized variant, and LLaMA3.1-8B-Instruct, a general-purpose instruction-tuned model

Table 3: **Cross-architecture evaluation under RLVR and RLIF.** SCOPE is applied to Qwen2.5-Math-7B and LLaMA3.1-8B-Instruct in addition to the primary Qwen2.5-7B-Base model.

		RLVR					RLIF				
Model	**Method**	GSM8K	MATH500	AMC	AIME24	AIME25	GSM8K	MATH500	AMC	AIME24	AIME25
Qwen2.5-7B-Base	Baseline	0.825	0.715	0.550	0.125	0.083	0.808	0.632	0.258	0.054	0.033
	SCOPE	**0.882**	**0.773**	**0.628**	**0.185**	**0.131**	**0.856**	**0.697**	**0.319**	**0.115**	**0.081**
Qwen2.5-Math-7B	Baseline	0.844	0.753	0.624	0.263	0.147	0.838	0.692	0.378	0.144	0.077
	SCOPE	**0.863**	**0.771**	**0.641**	**0.284**	**0.156**	**0.846**	**0.712**	**0.394**	0.141	**0.081**
LLaMA3.1-8B-Inst.	Baseline	0.756	0.533	0.221	0.028	0.000	0.742	0.529	0.169	0.000	0.000
	SCOPE	**0.772**	**0.548**	**0.242**	0.028	0.000	**0.747**	**0.533**	0.167	0.000	0.000

from a different model family. For LLaMA, we use the instruction-tuned variant rather than the base model, as preliminary experiments revealed that the base model exhibited weak instruction-following, frequent repetitive generation, and near-zero rewards in early RL rollouts, preventing stable training. Table 3 reports Acc@16 on five benchmarks under both RLVR and RLIF. We include the primary Qwen2.5-7B-Base results for reference. Three observations emerge. **(i)** SCOPE yields consistent improvements on Qwen2.5-Math-7B across both paradigms, with gains on GSM8K, MATH500, AMC, and AIME25. The gains are smaller than on Qwen2.5-7B-Base, in part because the hyperparameters of SCOPE were tuned for the latter, but more fundamentally because Qwen2.5-Math-7B has already undergone extensive math-focused continued pre-training and thus enters RL with its math-relevant distribution already concentrated—leaving correspondingly less headroom for SCOPE to preserve (Appendix D.1). **(ii)** On LLaMA3.1-8B-Instruct, SCOPE provides modest improvements on easier benchmarks (GSM8K, MATH500, AMC) under RLVR, but shows no benefit on AIME24/25 where the baseline already achieves near-zero performance. Under RLIF, the pattern is similar: marginal gains on easier tasks, with no meaningful signal on harder benchmarks. This is consistent with recent findings that LLaMA-family models exhibit weaker search-like reasoning behaviors than Qwen-family models (Gandhi et al., 2025), and that RL tends to amplify existing reasoning structures rather than create new ones. More specifically, this benchmark-conditional pattern—SCOPE gains where the baseline is nonzero, null where it is zero—constitutes a direct empirical confirmation of the support-redistribution view of RL (Yue et al., 2025): SCOPE can only rescue trajectories already within the base policy's support, and the LLaMA-3.1 base policy evidently lacks AIME-level trajectories to rescue (Appendix D.2). **(iii)** Taken together, these results suggest that SCOPE is broadly applicable across architectures but its effectiveness depends on the structural properties of the initial policy. When a model already possesses search-like reasoning behavior, SCOPE effectively mitigates entropy collapse and yields consistent improvements. When such structure is insufficiently developed, the benefit of entropy-aware regularization is limited by the reasoning capacity of the initial policy itself.

5 Conclusion

In this work, we identified entropy collapse as a shared structural failure mode that limits the effectiveness of test-time scaling in both RLVR and RLIF. Through token-level analysis, we showed that entropy collapse is not a uniform phenomenon, but is driven by premature overconfidence at a small number of structurally critical decision points. Based on this insight, we proposed SCOPE (Structural Collapse-aware Optimization via Partial Entropy control), which assigns each generated token a redistribution score and applies selective KL regularization only to the top $\sim 5\%$ of tokens, mitigating global entropy collapse while allowing confident convergence elsewhere. Across model scales and architectures on math reasoning benchmarks, SCOPE outperforms prior entropy-control and diversity-based methods under an identical regularization budget in both RLVR and RLIF. These results indicate that the entropy–performance trade-off in LLM reasoning is governed not by *how much* uncertainty is preserved, but by *where* it is retained—offering a targeted mechanism for sustaining reasoning diversity under post-training.

Acknowledgment

This work was supported by the Korea Research Institute for Defense Technology Planning and Advancement (KRIT) under Grant KRIT-CT-23-021.

Limitations

This work focuses on mitigating entropy collapse in post-training alignment for mathematical reasoning tasks, and several limitations remain.

First, our evaluation is restricted to math reasoning benchmarks (GSM8K, MATH500, AMC, AIME24/25), where reasoning structure and branch-defining token behavior are relatively well-defined. While we expect SCOPE to generalize to other reasoning domains such as code generation, logical question answering, or multi-hop commonsense reasoning, additional validation is required to confirm its effectiveness under different distributional and structural characteristics. Domains with less explicit branching structure (e.g., open-ended dialogue) may exhibit qualitatively different token-level entropy dynamics that fall outside the scope of our analysis.

Second, there is a gap between the exact decomposition in Proposition 1 and the practical surrogate that SCOPE deploys. While $S_v = L_v + R_v$ is exact, the instance-level score $s_{k,t} = p_{k,t}(1-p_{k,t})[\hat{A}_{k,t}]_+$ captures only the leading geometry L_v and discards the token–outcome coupling residual R_v—the very term that distinguishes branch-defining tokens from ordinary continuations in our analysis. We explicitly adopt this as a modeling choice rather than a justified approximation (Appendix A.21). Whether recovering R_v via within-group estimation of $\alpha_v - \beta_v$, or empirically verifying that $|R_v| \ll |L_v|$ on realized rollouts, yields further improvements or justifies the L_v-only surrogate remains an open question for future work.

Ethics Statement

This work studies post-training optimization methods for large language models with the goal of improving reasoning performance and stability. The proposed method, SCOPE, operates solely at the level of training dynamics and does not introduce new model capabilities, data sources, or interaction modalities beyond those of existing LLMs.

All experiments are conducted on publicly available mathematical reasoning benchmarks, and no personal, sensitive, or proprietary data are used. The method does not involve human subjects, human feedback collection, or deployment in real-world decision-making systems.

While improved reasoning performance could potentially amplify both beneficial and harmful downstream uses of LLMs, SCOPE itself is a generic optimization technique that does not target any specific application domain. We therefore do not foresee direct ethical risks unique to this work beyond those commonly associated with large language models, and we encourage responsible use and evaluation when applying the method to safety-critical or real-world settings.

References

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.

Feng Chen, Allan Raventos, Nan Cheng, Surya Ganguli, and Shaul Druckmann. 2025a. Rethinking fine-tuning when scaling test-time compute: Limiting confidence improves mathematical reasoning. *arXiv preprint arXiv:2502.07154*.

Xingwu Chen, Tianle Li, and Difan Zou. 2025b. Reshaping reasoning in llms: A theoretical analysis of rl training dynamics through pattern selection. *arXiv preprint arXiv:2506.04695*.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *arXiv preprint arXiv:2508.15260*.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, and 1 others. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9.

Qingbin Li, Rongkun Xue, Jie Wang, Ming Zhou, Zhi Li, Xiaofeng Ji, Yongqi Wang, Miao Liu, Zheming Yang, Minghui Qiu, and 1 others. 2025. Cure: Critical-token-guided re-concatenation for entropy-collapse prevention. *arXiv preprint arXiv:2508.11016*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Siru Ouyang, Xinyu Zhu, Zilin Xiao, Minhao Jiang, Yu Meng, and Jiawei Han. Rast: Reasoning activation in llms via small-model transfer. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.

Ali Razghandi, Seyed Mohammad Hadi Hosseini, and Mahdieh Soleymani Baghshah. 2025. Cer: Confidence enhanced reasoning in llms. *arXiv preprint arXiv:2502.14634*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. *CoRR*.

Nikolaos Tsilivis, Eran Malach, Karen Ullrich, and Julia Kempe. 2025. How reinforcement learning after next-token prediction facilitates learning. *arXiv preprint arXiv:2510.11495*.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, and 1 others. 2025. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *arXiv preprint arXiv:2506.01939*.

Xuezhi Wang and Denny Zhou. 2024. Chain-of-thought reasoning without prompting. *Advances in Neural Information Processing Systems*, 37:66383–66409.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Jian Yao, Ran Cheng, Xingyu Wu, Jibin Wu, and Kay Chen Tan. 2025. Diversity-aware policy optimization for large language model reasoning. *arXiv preprint arXiv:2505.23433*.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. 2025. No free lunch: Rethinking internal feedback for llm reasoning. *arXiv preprint arXiv:2506.17219*.

Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025a. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. 2025b. Learning to reason without external rewards. *arXiv preprint arXiv:2505.19590*.

Chuyue Zhou, Wangjie You, Juntao Li, Jing Ye, Kehai Chen, and Min Zhang. 2023. Inform: Information entropy based multi-step reasoning for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3565–3576.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. The surprising effectiveness of negative reinforcement in llm reasoning. *arXiv preprint arXiv:2506.01347*.

A Full Derivation: Token-Logit Local Ascent Contribution and Exact Conditional Redistribution Decomposition

This appendix provides a complete, self-contained derivation of the results underlying Section 3. We make explicit every probability space, every conditioning step, and every algebraic manipulation, so that Proposition 1 and the practical surrogate in (11) are justified without reliance on heuristic arguments or asymptotic approximations.

A.1 Motivating Ambiguities

Prior discussions of the entropy-collapse mechanism often cite an informal identity of the form

$$\mathbb{E}\big[|\Delta z_v|\big] \;\propto\; p_v(1-p_v)\,|\hat{A}(\tau)|.$$

However, this expression leaves several questions unanswered: (i) over which probability space is the expectation taken? (ii) is p_v a constant or a random variable? (iii) is $\hat{A}(\tau)$ a token-level or trajectory-level quantity? (iv) how should the actual optimizer update be distinguished from the local logit-coordinate signal? and (v) what is the exact relationship between the $p_v(1-p_v)$ leading term and token-specific corrections? This appendix resolves each of these points by developing a fixed-context, conditional analysis in logit coordinates, and arriving at the exact non-asymptotic identity stated in Proposition 1.

A.2 Setup and Notation

Definition 1 (Basic objects)**.** *We fix the following:*

- *Input prompt x;*
- *Completion trajectory $\tau = (y_1, \ldots, y_T)$ sampled from $\pi_\theta(\cdot \mid x)$;*
- *Context at step t: $c_t = (x, y_{<t})$ where $y_{<t} = (y_1, \ldots, y_{t-1})$;*
- *Vocabulary $\mathcal{V}$;*
- *Pre-softmax logit vector $z_t = (z_{t,u})_{u\in\mathcal{V}}$.*

The softmax policy is given by

$$\pi_\theta(u \mid c_t) \;=\; \frac{e^{z_{t,u}}}{\sum_{k\in\mathcal{V}} e^{z_{t,k}}}, \qquad u \in \mathcal{V}. \tag{16}$$

Remark 1 (Strict positivity)**.** *By (16), $e^{z_{t,u}} > 0$ for every u, so*

$$0 < \pi_\theta(u \mid c_t) < 1, \qquad \forall\, u \in \mathcal{V}. \tag{17}$$

This fact is used below to eliminate absolute values: $|1-p_v| = 1-p_v$ and $|0-p_v| = p_v$.

We fix a step t and token $v \in \mathcal{V}$ throughout and abbreviate $p_u := \pi_\theta(u \mid c_t)$ and $p_v := \pi_\theta(v \mid c_t)$.

A.3 Global vs. Conditional Expectation

This is the most critical step of the setup. Since π_θ samples an entire trajectory, the global expectation $\mathbb{E}_{\tau\sim\pi_\theta(\cdot|x)}[\,\cdot\,]$ treats the context c_t as random, which makes $p_v = \pi_\theta(v \mid c_t)$ a random variable. However, our analysis concerns how local softmax geometry and token-level redistribution structure emerge *given* a specific context.

Definition 2 (Conditional expectation notation)**.** *We write $\mathbb{E}_t[\,\cdot\,] := \mathbb{E}[\,\cdot \mid c_t\,]$. Under this conditioning c_t is fixed, so p_v is a deterministic constant.*

Key distinction. Under the global expectation $\mathbb{E}_\tau$ the context c_t and hence p_v are random; under the conditional expectation $\mathbb{E}_t$ the context is fixed and p_v is a constant. Every core identity below is a *conditional* statement in the sense of Definition 2.

A.4 Minimal Assumptions

Assumption 1 (Conditional integrability)**.** *The advantage magnitude is integrable conditional on context: $\mathbb{E}[|\hat{A}(\tau)| \mid c_t] < \infty$, and moreover $\mathbb{E}[|\hat{A}(\tau)| \mid c_t, y_t = u] < \infty$ for every $u \in \mathcal{V}$.*

Assumption 2 (Strict probability range)**.** $0 < p_v < 1$ *for all $v \in \mathcal{V}$ (Remark 1).*

All statements in this appendix are made under Assumptions 1–2.

A.5 GRPO Advantage is a Trajectory-Level Scalar

Given G completions $\tau_1, \ldots, \tau_G$ sampled from the same prompt x with rewards $r(\tau_i)$, GRPO computes the normalized advantage

$$\hat{A}(\tau_i) \;=\; \frac{r(\tau_i) - \mu_r}{\sigma_r}, \quad \mu_r = \tfrac{1}{G}\sum_{i=1}^{G} r(\tau_i), \tag{18}$$

with σ_r the within-group standard deviation of rewards. The *same* scalar is broadcast to every step t of τ_i: $\hat{A}_t = \hat{A}(\tau)$.

Remark 2 (No independence from y_t)**.** *$\hat{A}(\tau)$ is a function of the* entire *completion, not of a single token, and typically $\hat{A}(\tau) \not\perp y_t$. The derivation below makes no independence assumption.*

A.6 Local On-Policy Score Form

The local contribution at step t is

$$\nabla_\theta \mathcal{J}_t(\theta) \;=\; \mathbb{E}_{\tau\sim\pi_\theta(\cdot|x)}\big[\,\hat{A}(\tau)\,\nabla_\theta \log \pi_\theta(y_t \mid c_t)\,\big]. \tag{19}$$

Remark 3 (Scope of the analysis)**.** *Equation* (19) *is the* unclipped, on-policy, token-time local score form. *Importance ratios, clipping, KL penalties, optimizer preconditioning, and minibatch aggregation are intentionally excluded; every quantity defined below is an ascent-direction signal associated with this local score, and is distinct from the realized parameter update produced by an actual optimizer.*

A.7 Softmax Log-Derivative Identity

Proposition 2 (Softmax log-derivative)**.** *For the softmax policy* (16)*,*

$$\frac{\partial}{\partial z_{t,v}} \log \pi_\theta(y_t \mid c_t) \;=\; \mathbf{1}[y_t = v] - p_v. \tag{20}$$

Proof. From the softmax definition,

$$\log \pi_\theta(y_t \mid c_t) \;=\; z_{t,y_t} - \log \sum_{k\in\mathcal{V}} e^{z_{t,k}}.$$

Differentiating with respect to $z_{t,v}$:

First term. $\partial z_{t,y_t}/\partial z_{t,v} = \mathbf{1}[y_t = v]$: it equals 1 if $y_t = v$ (since then $z_{t,y_t} = z_{t,v}$) and 0 otherwise (since z_{t,y_t} does not depend on $z_{t,v}$).

Second term. By the chain rule,

$$\frac{\partial}{\partial z_{t,v}} \log \sum_k e^{z_{t,k}} \;=\; \frac{e^{z_{t,v}}}{\sum_k e^{z_{t,k}}} \;=\; p_v.$$

Combining, $\partial \log \pi_\theta / \partial z_{t,v} = \mathbf{1}[y_t = v] - p_v$. □

Remark 4. *When $y_t = v$ the derivative equals $1 - p_v > 0$; when $y_t \neq v$ it equals $-p_v < 0$. That is, the log-probability is sensitive in the positive direction to the logit of the sampled token, and in the negative direction to logits of non-sampled tokens.*

A.8 Exact Conditional Logit-Coordinate Score Identity

Definition 3 (Conditional logit-coordinate score)**.**

$$g_{t,v}(c_t) \;:=\; \mathbb{E}\left[\hat{A}(\tau)\,\frac{\partial \log \pi_\theta(y_t \mid c_t)}{\partial z_{t,v}} \,\middle|\, c_t\right]. \tag{21}$$

Proposition 3 (Exact conditional score identity)**.**

$$g_{t,v}(c_t) \;=\; \mathbb{E}_t\big[\,\hat{A}(\tau)\,(\mathbf{1}[y_t = v] - p_v)\,\big]. \tag{22}$$

Proof. Substitute (20) into (21); the result follows by definition of $\mathbb{E}_t$. □

Note that $g_{t,v}(c_t)$ is *not* the parameter-space gradient of $\mathcal{J}$; it is the exact conditional score experienced by the logit-coordinate direction at a fixed context.

A.9 Realized Local Ascent Contribution

Definition 4 (Realized local ascent contribution)**.** *For a context c_t and realized sample y_t, the local ascent-direction signal at logit coordinate v is*

$$\Delta^{\mathrm{loc}}_{t,v} \;:=\; \eta\,\hat{A}(\tau)\,(\mathbf{1}[y_t = v] - p_v), \tag{23}$$

where $\eta > 0$ is a step-size-like scalar.

Remark 5 (Interpretation)**.** *$\Delta^{\mathrm{loc}}_{t,v}$ is not the exact post-update logit difference: Adam preconditioning, clipping, KL regularization, minibatch averaging, and nonlocal coupling through shared parameters are all excluded. More precisely, it is the* realized local ascent contribution in logit coordinates *at a fixed context.*

A.10 Interpretation of Signed Local Dynamics

Case $\hat{A}(\tau) > 0$ (local one-step sharpening). For $v = y_t$, $\Delta^{\mathrm{loc}}_{t,v} = \eta\hat{A}(\tau)(1 - p_v) > 0$; for $v \neq y_t$, $\Delta^{\mathrm{loc}}_{t,v} = -\eta\hat{A}(\tau)p_v < 0$. The sampled token is reinforced, all alternatives are suppressed, and the local update exerts sharpening pressure.

Case $\hat{A}(\tau) < 0$ (local flattening). For $v = y_t$, $\Delta^{\mathrm{loc}}_{t,v} < 0$; for $v \neq y_t$, $\Delta^{\mathrm{loc}}_{t,v} > 0$. Probability mass flows away from the sampled token and toward alternatives, yielding flattening pressure.

Local redistribution magnitude. From Definition 4,

$$\big|\Delta^{\mathrm{loc}}_{t,v}\big| \;=\; \eta\,|\hat{A}(\tau)|\,\big|\mathbf{1}[y_t = v] - p_v\big|. \tag{24}$$

Definition 5 (Expected absolute local redistribution magnitude)**.** $S_v := \mathbb{E}_t\big[\,|\Delta^{\mathrm{loc}}_{t,v}|\,\big]$.

Remark 6 (What S_v is and is not)**.** *S_v is not the entropy itself, nor a direct entropy decrement, nor a signed expected update. It is the exact conditional magnitude of local logit-coordinate redistribution induced on token v under fixed context.*

A.11 Exact Conditional Decomposition of S_v

Event partition and conditional means. Let $E_v := \{y_t = v\}$ and $E_v^c := \{y_t \neq v\}$, and define the token-conditioned advantage means

$$\begin{aligned} \alpha_v &:= \mathbb{E}[\,|\hat{A}(\tau)| \mid c_t, y_t = v\,], \\ \beta_v &:= \mathbb{E}[\,|\hat{A}(\tau)| \mid c_t, y_t \neq v\,], \end{aligned} \tag{25}$$

both finite under Assumption 1.

Total-expectation decomposition. From Definition 5 and (24),

$$\begin{aligned} S_v &= \eta\, \mathbb{E}_t\big[\, |\hat{A}(\tau)|\, \big|\mathbf{1}[y_t = v] - p_v\big|\,\big] \\ &= \eta\, \Pr(E_v \mid c_t) \\ &\quad \cdot \mathbb{E}\big[\, |\hat{A}(\tau)|\, |1 - p_v|\,\big|\, c_t, E_v\big] \\ &\quad + \eta\, \Pr(E_v^c \mid c_t) \\ &\quad \cdot \mathbb{E}\big[\, |\hat{A}(\tau)|\, |0 - p_v|\,\big|\, c_t, E_v^c\big]. \end{aligned} \tag{26}$$

By Assumption 2, $|1-p_v| = 1-p_v$ and $|0-p_v| = p_v$, and $\Pr(E_v \mid c_t) = p_v,\ \Pr(E_v^c \mid c_t) = 1 - p_v$. Since p_v is a constant under conditioning, we factor it out of the inner expectations. Substituting gives

$$\begin{aligned} S_v &= \eta\big[\, p_v(1-p_v)\alpha_v + (1-p_v)p_v\beta_v\,\big] \\ &= \eta\, p_v(1-p_v)(\alpha_v + \beta_v). \end{aligned} \tag{27}$$

Proposition 4 (First exact identity)**.** $S_v = \eta\, p_v(1-p_v)(\alpha_v + \beta_v)$.

A.12 Exact Expansion of β_v

Since $\{y_t \neq v\}$ is the disjoint union $\bigsqcup_{u \neq v}\{y_t = u\}$, the law of total expectation gives

$$\beta_v = \sum_{u \neq v} \Pr(y_t = u \mid c_t, y_t \neq v)\, \alpha_u, \tag{28}$$

where for all $u \in \mathcal{V}$, $\alpha_u := \mathbb{E}[\, |\hat{A}(\tau)| \mid c_t, y_t = u\,]$. By Bayes' rule, $\Pr(y_t = u \mid c_t, y_t \neq v) = p_u/(1-p_v)$ for $u \neq v$, so

$$\boxed{\beta_v = \sum_{u \neq v} \frac{p_u}{1-p_v}\, \alpha_u} \tag{29}$$

A.13 Global Conditional Scale and Deviation Variables

Definition 6 (Global conditional scale)**.** $A_{\text{scale}} := \mathbb{E}[\, |\hat{A}(\tau)| \mid c_t\,] = \mathbb{E}_t[\, |\hat{A}(\tau)|\,]$. *Applying the law of total expectation over y_t:*

$$A_{\text{scale}} = \sum_{u \in \mathcal{V}} p_u\, \alpha_u. \tag{30}$$

Definition 7 (Deviation variables)**.** *For each $u \in \mathcal{V}$, let $\delta_u := \alpha_u - A_{\text{scale}}$; so that, $\alpha_u = A_{\text{scale}} + \delta_u$.*

Lemma 1 (Weighted zero-mean property)**.** $\sum_{u \in \mathcal{V}} p_u\, \delta_u = 0$.

Proof. $\sum_u p_u \delta_u = \sum_u p_u(\alpha_u - A_{\text{scale}}) = \sum_u p_u \alpha_u - A_{\text{scale}} \sum_u p_u = A_{\text{scale}} - A_{\text{scale}} \cdot 1 = 0$, using (30) and $\sum_u p_u = 1$. □

A.14 Exact Deviation Decomposition of β_v

Substituting $\alpha_u = A_{\text{scale}} + \delta_u$ into (29):

$$\beta_v = A_{\text{scale}} \sum_{u \neq v} \frac{p_u}{1-p_v} + \sum_{u \neq v} \frac{p_u}{1-p_v}\, \delta_u. \tag{31}$$

The first coefficient equals 1: $\sum_{u \neq v} p_u/(1-p_v) = (1-p_v)/(1-p_v) = 1$, using $\sum_{u \neq v} p_u = 1 - p_v$. Defining the residual

$$\tilde{\delta}_v := \sum_{u \neq v} \frac{p_u}{1-p_v}\, \delta_u, \tag{32}$$

we obtain the deviation decomposition

$$\boxed{\beta_v = A_{\text{scale}} + \tilde{\delta}_v} \tag{33}$$

A.15 Exact Simplification of the Residual

We now express $\tilde{\delta}_v$ purely in terms of δ_v, which produces the compact form used in the main result.

Lemma 2 (Simplification of $\tilde{\delta}_v$)**.** $\tilde{\delta}_v = -\,p_v\, \delta_v/(1-p_v)$.

Proof. From Lemma 1, $p_v\delta_v + \sum_{u \neq v} p_u \delta_u = 0$, hence $\sum_{u \neq v} p_u \delta_u = -p_v\delta_v$. Substituting into (32) gives $\tilde{\delta}_v = -p_v\delta_v/(1-p_v)$. □

Corollary 1 (Combined deviation)**.**

$$\delta_v + \tilde{\delta}_v = \delta_v\, \frac{1-2p_v}{1-p_v}. \tag{34}$$

Proof. $\delta_v + \tilde{\delta}_v = \delta_v - \frac{p_v}{1-p_v}\delta_v = \delta_v\, \frac{(1-p_v)-p_v}{1-p_v} = \delta_v\, \frac{1-2p_v}{1-p_v}$. □

A.16 Exact Non-Asymptotic Identity

We now combine the preceding results. From Proposition 4, $S_v = \eta\, p_v(1-p_v)(\alpha_v + \beta_v)$. Substituting $\alpha_v = A_{\text{scale}} + \delta_v$ and (33) gives

$$\alpha_v + \beta_v = 2A_{\text{scale}} + \delta_v + \tilde{\delta}_v, \tag{35}$$

so that

$$\begin{aligned} S_v &= \eta\, p_v(1-p_v)\big[\, 2A_{\text{scale}} + (\delta_v + \tilde{\delta}_v)\,\big] \\ &= 2\eta\, A_{\text{scale}}\, p_v(1-p_v) \\ &\quad + \eta\, p_v(1-p_v)(\delta_v + \tilde{\delta}_v). \end{aligned} \tag{36}$$

Applying Corollary 1, $p_v(1-p_v)(\delta_v + \tilde{\delta}_v) = p_v(1-p_v) \cdot \frac{1-2p_v}{1-p_v}\, \delta_v = p_v(1-2p_v)\, \delta_v$, so

$$\boxed{\begin{aligned} S_v &= 2\eta\, A_{\text{scale}}\, p_v(1-p_v) \\ &\quad + \eta\, p_v(1-2p_v)\, \delta_v. \end{aligned}} \tag{37}$$

This is an exact non-asymptotic identity for S_v at a fixed context c_t; no approximation has been used.

A.17 Covariance Interpretation

We reinterpret δ_v as a conditional covariance. Let $I_v := \mathbf{1}[y_t = v]$; then

$$\begin{aligned}\mathbb{E}_t[\,|\hat{A}(\tau)|\,I_v\,] &= \Pr(y_t = v \mid c_t) \\ &\quad \cdot \mathbb{E}[\,|\hat{A}(\tau)| \mid c_t, y_t = v\,] \\ &= p_v\,\alpha_v.\end{aligned} \tag{38}$$

Therefore

$$\begin{aligned}\mathrm{Cov}_t(|\hat{A}(\tau)|, I_v) &= \mathbb{E}_t[\,|\hat{A}(\tau)|\,I_v\,] \\ &\quad - \mathbb{E}_t[\,|\hat{A}(\tau)|\,]\,\mathbb{E}_t[I_v] \\ &= p_v\,\alpha_v - A_{\text{scale}}\,p_v \\ &= p_v\,(\alpha_v - A_{\text{scale}}) \\ &= p_v\,\delta_v.\end{aligned} \tag{39}$$

For $p_v > 0$,

$$\boxed{\delta_v \;=\; \frac{\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v])}{p_v}} \tag{40}$$

Hence δ_v measures the *conditional coupling* between the sampling event $\{y_t = v\}$ and the trajectory-level advantage magnitude.

Alternative covariance identity. From (30) and the decomposition $A_{\text{scale}} = p_v\alpha_v + (1-p_v)\beta_v$,

$$\begin{aligned}\alpha_v - A_{\text{scale}} &= \alpha_v - p_v\alpha_v - (1-p_v)\beta_v \\ &= (1-p_v)(\alpha_v - \beta_v),\end{aligned} \tag{41}$$

so

$$\boxed{\begin{aligned}&\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v]) \\ &\quad = p_v(1-p_v)(\alpha_v - \beta_v)\end{aligned}} \tag{42}$$

The residual is large only when *both* $p_v(1-p_v)$ and $|\alpha_v - \beta_v|$ are non-negligible—i.e., intermediate probability *and* a meaningful difference in expected advantage magnitude between sampling v and not sampling it.

A.18 Exact Covariance Form

Substituting (40) into (37):

$$\begin{aligned}&\eta\,p_v(1-2p_v)\,\delta_v \\ &\quad = \eta\,p_v(1-2p_v)\,\frac{\mathrm{Cov}_t(|\hat{A}(\tau)|, I_v)}{p_v} \\ &\quad = \eta(1-2p_v)\,\mathrm{Cov}_t(|\hat{A}(\tau)|, I_v).\end{aligned} \tag{43}$$

yielding the exact covariance form

$$\boxed{\begin{aligned}S_v \;&=\; 2\eta\,A_{\text{scale}}\,p_v(1-p_v) \\ &\quad +\; \eta(1-2p_v)\,\mathrm{Cov}_t(|\hat{A}(\tau)|, I_v).\end{aligned}} \tag{44}$$

A.19 Main Proposition

Definition 8 (Leading term and residual)**.**

$$L_v \;:=\; 2\eta\,A_{\text{scale}}\,p_v(1-p_v), \tag{45}$$

$$R_v \;:=\; \eta(1-2p_v)\,\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v]). \tag{46}$$

Proposition 5 (Main proposition; restatement of Proposition 1)**.** *Under Assumptions 1–2, for every fixed context c_t and every token $v \in \mathcal{V}$,*

$$S_v \;=\; L_v + R_v, \tag{47}$$

where $S_v = \mathbb{E}_t[\,|\Delta^{\text{loc}}_{t,v}|\,]$ is the exact conditional expected absolute local ascent-contribution magnitude, L_v is the smooth leading geometry term, and R_v is the token–outcome coupling residual.

Proof. Summarizing the chain above:

1. $S_v = \eta\,p_v(1-p_v)(\alpha_v + \beta_v)$ (Proposition 4).
2. $\beta_v = A_{\text{scale}} + \tilde{\delta}_v$ ((33)).
3. $\tilde{\delta}_v = -\,p_v\,\delta_v/(1-p_v)$ (Lemma 2).
4. $\delta_v + \tilde{\delta}_v = \delta_v\,(1-2p_v)/(1-p_v)$ (Corollary 1).
5. $S_v = 2\eta\,A_{\text{scale}}\,p_v(1-p_v) + \eta\,p_v(1-2p_v)\,\delta_v$ ((37)).
6. $\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v]) = p_v\,\delta_v$ ((40)).
7. $\eta\,p_v(1-2p_v)\,\delta_v = \eta(1-2p_v)\,\mathrm{Cov}_t(\cdots)$.

□

A.20 Structural Interpretation: Ordinary vs. Branch-Defining Tokens

The decomposition $S_v = L_v + R_v$ admits a clean structural reading (which is an interpretation, not a theorem).

Ordinary tokens. Syntactic continuations and follow-ups within an already determined reasoning path do not appreciably change the final outcome whether sampled or not, so $\alpha_v \approx \beta_v$. By (42), $\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v]) \approx 0$, hence $R_v \approx 0$ and $S_v \approx L_v$. The leading geometry alone captures their local redistribution behavior.

Branch-defining tokens. At positions where sampling v vs. not sampling v materially alters the downstream trajectory, $|\alpha_v - \beta_v|$ is non-negligible, and so is R_v. These are precisely the positions at which the conditional covariance of advantage magnitude with the sampling indicator is structurally non-vanishing.

Caveat. This distinction is a structural interpretation based on Proposition 5, not an exact theorem about any particular class of tokens.

A.21 From the Exact Identity to the Practical Surrogate

Leading-term surrogate as a modeling choice. Directly computing S_v during RL training is impractical because α_v, β_v, and $\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v])$ are all unknown. SCOPE therefore adopts $S_v \approx L_v$ as a *practical modeling choice*; this is neither a rigorous approximation theorem nor an assertion that $R_v = 0$.

Instance-level practical surrogate score. For a sampled completion τ_i at token position t with generated token $y_{i,t}$, policy probability $p_{i,t} := \pi_\theta(y_{i,t} \mid c_{i,t})$, and trajectory-level advantage $\hat{A}_i$, we define

$$\begin{aligned} s_{i,t} &:= p_{i,t}(1 - p_{i,t})\,[\hat{A}_i]_+, \\ [\hat{A}_i]_+ &:= \max(\hat{A}_i, 0). \end{aligned} \tag{48}$$

This score simultaneously captures three properties: (i) the leading redistribution geometry $p_{i,t}(1-p_{i,t})$; (ii) restriction to the positive-advantage sharpening side only; and (iii) computability at the level of individual sampled-token instances.

Why $[\hat{A}]_+$ instead of $|\hat{A}|$? The exact target S_v is sign-agnostic in $\hat{A}$ through $|\hat{A}(\tau)|$. However, collapse-aware control specifically targets the *sharpening* side: $\hat{A}(\tau) > 0$ events reinforce the sampled token and suppress alternatives, directly producing local one-step sharpening and thus driving entropy decay. Using $[\hat{A}_i]_+$ rather than $|\hat{A}_i|$ therefore isolates the regime that we want SCOPE to regulate.

$s_{i,t}$ is not an exact Monte Carlo estimator. Equation (48) does not estimate α_v, β_v, does not recover the covariance correction R_v, and does not integrate over the full conditional distribution. It is a principled surrogate that preserves the leading redistribution geometry and the positive-advantage sharpening signal.

Top-$q\%$ masking. Let $\mathcal{B}$ denote the set of valid generated token instances in the minibatch, and let τ_q be the top-$q\%$ threshold of $\{s_{i,t}\}_{(i,t)\in\mathcal{B}}$. The selection mask is

$$m_{i,t} := \mathbf{1}[\, s_{i,t} \geq \tau_q \,]. \tag{49}$$

Selective KL regularization. A per-token KL penalty is applied only on selected instances:

$$\mathcal{L}_{\text{KL-top}} = \lambda \frac{\sum_{(i,t)\in\mathcal{B}} m_{i,t}\, D^{\text{ref}}_{i,t}}{\sum_{(i,t)\in\mathcal{B}} m_{i,t} + \varepsilon}, \tag{50}$$

where $D^{\text{ref}}_{i,t} := D_{\text{KL}}\big(\pi_\theta(\cdot \mid c_{i,t}) \,\|\, \pi_{\text{ref}}(\cdot \mid c_{i,t})\big)$, $\lambda > 0$ is the regularization strength, and $\varepsilon > 0$ is a denominator-stabilization constant.

A.22 Summary: Exact Results, Interpretation, Design Choices

Exact statements (under fixed context c_t). The following hold without approximation:

$$\begin{aligned} g_{t,v}(c_t) &= \mathbb{E}_t\big[\hat{A}(\tau)(\mathbf{1}[y_t{=}v] - p_v)\big], \\ |\Delta^{\text{loc}}_{t,v}| &= \eta\,|\hat{A}(\tau)|\,|\mathbf{1}[y_t{=}v] - p_v|, \\ S_v &= \eta\, p_v(1-p_v)(\alpha_v + \beta_v), \\ S_v &= 2\eta\, A_{\text{scale}}\, p_v(1-p_v) \\ &\quad + \eta\, p_v(1-2p_v)\,\delta_v, \\ S_v &= 2\eta\, A_{\text{scale}}\, p_v(1-p_v) \\ &\quad + \eta(1-2p_v)\,\mathrm{Cov}_t(\cdots), \\ \mathrm{Cov}_t &= p_v(1-p_v)(\alpha_v - \beta_v), \end{aligned} \tag{51}$$

where Cov_t abbreviates $\mathrm{Cov}_t(|\hat{A}(\tau)|, \mathbf{1}[y_t = v])$.

Interpretive statements (not exact theorems). Positive-advantage events are the direct source of local one-step sharpening; S_v is a local redistribution magnitude rather than a direct entropy decrement; L_v dominates at ordinary-continuation tokens; R_v is relatively important at branch-defining tokens.

Practical design choices (not derived as theorems). Using $S_v \approx L_v$ as a leading surrogate; defining $s_{i,t} = p_{i,t}(1 - p_{i,t})[\hat{A}_i]_+$ as an instance-level score; selecting collapse-aware candidates via top-$q\%$ masking; and applying selective KL regularization only to the selected tokens.

A.23 Tightness of the Leading-Term Surrogate

Appendix A.21 justified the modeling choice $S_v \approx L_v$ by computational necessity, but did not characterize when this approximation is tight and when it is loose. This subsection provides such a characterization. The result is that the surrogate is automatically tight in precisely the probability regime that SCOPE's score $s_{k,t}$ ranks highest, so the discarded residual R_v tends to be small at the very token instances SCOPE selects.

Ratio $|R_v|/L_v$. From Proposition 5 and Definition 8 together with the covariance identity (42),

$$\frac{|R_v|}{L_v} = \frac{\eta\,|1-2p_v|\cdot p_v(1-p_v)\cdot|\alpha_v-\beta_v|}{2\eta\,A_{\text{scale}}\,p_v(1-p_v)} = \frac{|1-2p_v|\cdot|\alpha_v-\beta_v|}{2\,A_{\text{scale}}}. \tag{52}$$

The $p_v(1-p_v)$ factor cancels, so the tightness of $S_v \approx L_v$ is governed by two independent factors: the deviation coefficient $|1-2p_v|$ and the normalized coupling strength $|\alpha_v-\beta_v|/A_{\text{scale}}$.

Three regimes. We analyze (52) under the three token categories introduced in Appendix A.20.

Ordinary continuation tokens. By definition, sampling v versus not sampling it does not materially alter the downstream trajectory, so $\alpha_v \approx \beta_v$ and $|\alpha_v-\beta_v|/A_{\text{scale}} \ll 1$. Hence $|R_v|/L_v \ll 1$ regardless of p_v. The leading-term surrogate is tight.

Branch-defining tokens with extreme p_v. When a branch-defining token has p_v close to 0 or 1, $|1-2p_v|$ is close to 1 and $|\alpha_v-\beta_v|/A_{\text{scale}}$ is of order 1. Hence $|R_v|/L_v$ is of order 1 and the surrogate is loose. However, for such tokens the geometry factor $p_v(1-p_v)$ is itself small, and consequently both L_v and $s_{k,t}$ are small: these tokens are unlikely to be selected by the top-$q\%$ mask, so the looseness of the surrogate does not affect SCOPE's actual interventions.

Branch-defining tokens with intermediate p_v. This is the regime SCOPE targets: p_v away from $\{0,1\}$, and $|\alpha_v-\beta_v|/A_{\text{scale}}$ of order 1. Here the factor $|1-2p_v|$ becomes the decisive quantity.

Automatic suppression at $p_v = 1/2$. At $p_v = 1/2$, (52) vanishes identically:

$$p_v = \tfrac{1}{2} \implies |R_v|/L_v = 0. \tag{53}$$

More generally, for p_v in a neighborhood of $1/2$, $|1-2p_v|$ is small, so the surrogate is tight even when $|\alpha_v-\beta_v|$ is large. This is the crux of the tightness argument: SCOPE's geometry factor $p_v(1-p_v)$ attains its maximum at $p_v = 1/2$, and the same point is a zero of the residual coefficient $(1-2p_v)$. The two functions are aligned:

p_v	$p_v(1-p_v)$	$\lvert1-2p_v\rvert$	$\lvert R_v\rvert/L_v$
$\to 0$	$\to 0$	$\to 1$	loose ($L_v \to 0$)
0.3	0.21	0.4	moderate
0.5	0.25	0	0
0.7	0.21	0.4	moderate
$\to 1$	$\to 0$	$\to 1$	loose ($L_v \to 0$)

Because the top-$q\%$ mask ranks by $p_v(1-p_v)\,[\hat{A}]_+$, the selected instances concentrate near $p_v \approx 1/2$, which is precisely where $|R_v|/L_v$ is minimized among the branch-defining candidates.

Residual failure mode. The analysis above does not claim $|R_v|/L_v = 0$ at every selected instance; it claims that the expected $|R_v|/L_v$ over selected instances is smaller than over an unrestricted population of branch-defining tokens. The residual failure mode remains: a branch-defining token with p_v moderately displaced from $1/2$ (e.g., $p_v \in \{0.3, 0.7\}$) and a large $|\alpha_v-\beta_v|$ can have $|R_v|/L_v$ of order 0.4, which is non-negligible. SCOPE's surrogate is not uniformly tight; it is tight in the regime the surrogate itself selects into. Recovering R_v at these moderate-displacement branch-defining tokens is the future-work direction noted in the Limitations section.

B Compute Budget and Runtime Analysis

This section reports the compute budget, runtime measurements, and overhead analysis for all experiments.

Experimental Setup. All experiments are conducted on **4× NVIDIA H200 NVL GPUs** under identical training configurations. Specifically, we use bf16 precision, gradient accumulation steps of 32, gradient checkpointing, per-device batch size of 4, 8 generations per prompt, maximum completion length of 3072, and 117 optimization steps. These settings are shared across all methods, including GRPO, INTUITOR, and SCOPE.

Runtime Comparison. Table 4 summarizes the runtime statistics. Under both RLVR and RLIF settings, SCOPE exhibits a modest increase in wall-clock time compared to the corresponding baselines. This difference closely follows the increase in mean completion length. As rollout generation dominates the computational cost in GRPO-style RL training, the observed runtime differences are attributable to longer generated sequences rather than additional algorithmic overhead.

Overall Computational Overhead. Across all runs, the number of optimization steps, batch size, rollout count, GPU configuration, numerical precision, and optimization settings are strictly identical. SCOPE does not introduce additional forward or backward passes; its entropy-aware masking and selective KL regularization are integrated within the

existing computation graph. Under matched rollout lengths, training throughput remains comparable. The observed runtime differences are consistent with the increase in generated tokens per step.

Longer Rollouts as Preserved Reasoning. The source of the 2–3% runtime difference also admits a behavioral reading. Under both paradigms, SCOPE rollouts are consistently longer than the corresponding baselines in mean completion length (RLVR: 772.79 $\rightarrow$ 829.83 tokens, $+7.4\%$; RLIF: 909.70 $\rightarrow$ 943.19 tokens, $+3.7\%$). Because the only difference between each pair of runs is the presence of selective KL regularization at the top-5% of redistribution-ranked tokens, this length increase is itself a consequence of SCOPE's entropy-preservation mechanism: a policy that is prevented from prematurely collapsing at collapse-critical decision points retains the option to extend its reasoning trajectory rather than commit to an early answer. Conversely, collapsed policies tend toward shortcut continuations and premature termination, which shortens mean completion length under the same decoding budget. Read this way, the runtime overhead is not a cost of the SCOPE algorithm *per se*—the computation graph is unchanged (previous paragraph)—but the computational expression of the very behavior SCOPE is designed to preserve: longer, more deliberate reasoning chains made possible by avoiding premature entropy collapse.

C Additional Pass@k Results and Extended Analysis

This appendix provides additional Pass@k results that complement the 7B analysis in Section 4.5. We report (i) per-scale improvement curves (ΔPass@k = SCOPE $-$ Baseline) across all model scales (Figure 4), and (ii) full Pass@k curves for the 3B and 1.5B scales (Figures 5 and 6). We then expand on two observations from Section 4.5 whose implications warrant unpacking beyond the space available in the main text (Appendices C.1 and C.2).

Figure 4 directly plots the per-scale improvement of SCOPE over the corresponding baseline across all model scales. On saturating benchmarks (MATH500), gains diminish at large k as both methods approach ceiling coverage. On harder benchmarks (AMC, AIME24/25), gains *grow* with k, confirming that SCOPE preserves reasoning diversity that becomes increasingly valuable under larger sampling budgets.

Figures 5 and 6 report Pass@k curves for the 3B and 1.5B scales under the same setup as Figure 3. The qualitative patterns are consistent across scales, with one key difference: the base-model crossover observed at 7B under RLIF (where the untrained model surpasses trained methods at large k) progressively weakens at 3B and vanishes at 1.5B. This confirms that the crossover is driven by the base model's latent knowledge—at 7B, sufficient knowledge exists to benefit from uninhibited diversity, whereas at 1.5B, the knowledge bottleneck limits what diversity can achieve, making SCOPE's directed entropy preservation more valuable than raw diversity at all sampling budgets.

C.1 Simultaneous Improvement of Pass@1 and Pass@k

A central tension of RL post-training is the exploration–exploitation trade-off discussed in Section 1: methods that raise single-sample accuracy (Pass@1) typically do so by sharpening the policy, which erodes the distributional breadth required for Pass@k at larger k (Yue et al., 2025; Chen et al., 2025a). This trade-off is directly visible in our own baselines. Under RLIF on the 7B model, INTUITOR lifts Pass@1 on AIME24 relative to the base model but is overtaken by the base model's Pass@k at sufficiently large k (Figure 3, bottom right); the same pattern appears on AIME25. GRPO exhibits a similar, if less pronounced, pattern under RLVR. In both cases, the baseline buys Pass@1 gains by compressing distributional support, and pays for it at larger sampling budgets.

SCOPE departs from this pattern. Across all model scales and both paradigms, SCOPE simultaneously improves Pass@1 *and* preserves or increases Pass@k relative to its direct baseline. On AIME24 at the 7B scale under RLIF, SCOPE improves Pass@1 from 0.054 to 0.115 (a $2.1\times$ relative gain) while also improving Pass@128 relative to INTUITOR, closing most of the base-model gap that INTUITOR had opened. Analogous behavior holds on AMC and AIME25, and under RLVR the gap against the base model is closed entirely.

We attribute this departure to the mechanism by which SCOPE intervenes. Uniform or dense entropy penalties reduce confidence *indiscriminately*, preventing collapse at the cost of also suppressing the beneficial sharpening that drives Pass@1 gains. By contrast, SCOPE regularizes only the $\sim 5\%$ of tokens where the update rule itself predicts collapse-inducing redistribution. At the remaining $\sim 95\%$ of positions, the policy is free to

Table 4: Compute budget and runtime comparison under identical training configurations.

Setting	Method	Steps	Mean Length	Wall-clock	Rel. Time
RLVR	GRPO	117	772.79	8h 51m 46s	1.00×
	SCOPE	117	829.83	9h 08m 04s	1.03×
RLIF	INTUITOR	117	909.70	9h 16m 10s	1.00×
	SCOPE	117	943.19	9h 27m 12s	1.02×

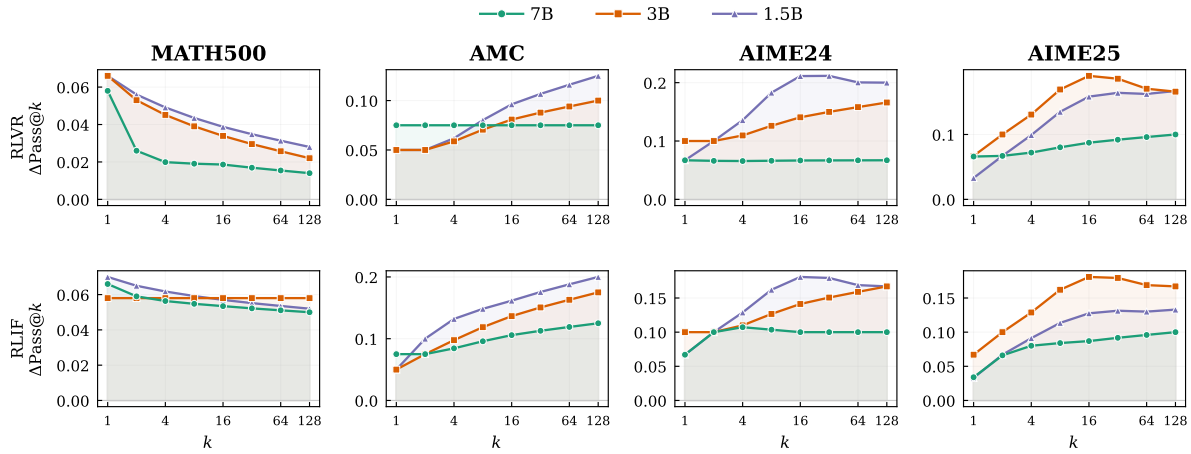

Figure 4: **Pass@k improvement (ΔPass@k = SCOPE − Baseline) across model scales.**

sharpen normally, preserving the Pass@1 benefit of RL optimization; at the collapse-critical minority, KL regularization prevents the premature overconfidence that would otherwise destroy Pass@k coverage. In this sense SCOPE resolves the exploration–exploitation trade-off not by interpolating more carefully along its Pareto frontier, but by *decomposing* the two pressures onto disjoint subsets of tokens—sharpening where it is locally safe, preserving uncertainty where it is structurally critical.

C.2 Scale-Dependent Crossover and the Value of RL Post-Training

The most striking qualitative pattern across the Pass@k curves—Figure 3 for the 7B scale together with the 3B and 1.5B results in Figures 5 and 6—is the scale-dependent fate of the base model under RLIF at large k. At 7B, the untrained base model surpasses both INTUITOR and SCOPE on AIME24/25 as k approaches 128. At 3B, the base model approaches trained methods at large k but does not overtake them. At 1.5B, the base model never catches up; SCOPE dominates at every k up to 128, including on the hardest benchmarks where Pass@1 is near zero.

This pattern has a clean mechanistic reading consistent with prior findings that RL redistributes probability mass within the base model's existing support rather than creating new capabilities (Yue et al., 2025). At 7B, the base policy already places non-negligible mass on solution-relevant trajectories for AIME-level problems; with a sufficiently large sampling budget, raw distributional breadth is enough to surface these trajectories, and RL's sharpening—even when moderated by SCOPE—becomes a liability rather than a benefit beyond a threshold k. At 1.5B, the base policy's support over solution-relevant trajectories is too thin for breadth alone to recover them; sampling 128 times from a distribution that barely covers the solution manifold yields near-zero Pass@128, and directed guidance via RL is required to move mass toward the right region of output space in the first place.

Two broader implications follow. First, the question "does RL post-training help?" is underspecified without jointly fixing a model scale and a sampling budget. For a fixed method, the sign of the RL-versus-base comparison can flip as k varies, and the k at which it flips shrinks as model scale grows. Evaluations that report only Pass@1—or only Pass@k at a single k—risk reaching conclusions that reverse under other sampling budgets, particularly for larger models where the crossover falls within the range typically used in test-time scaling experiments.

Second, the value of entropy-control methods such as SCOPE is itself scale-conditioned, but in a direction that favors their necessity rather than

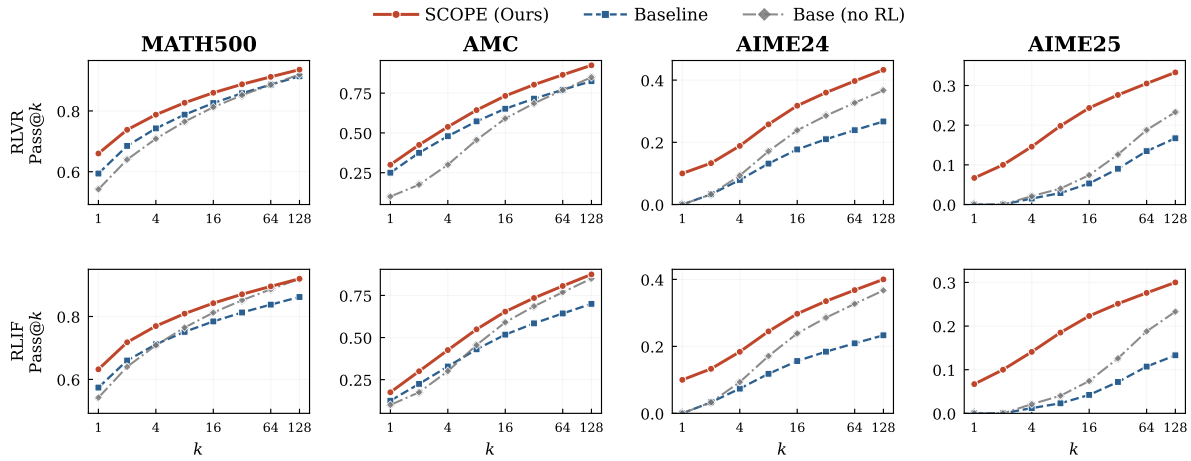

Figure 5: **Pass@k scaling on Qwen2.5-3B.** Same setup as Figure 3. SCOPE maintains a consistent advantage across all k values. The base model begins to approach trained methods at large k on AIME24/25 under RLIF, foreshadowing the crossover observed at 7B, but does not yet surpass SCOPE.

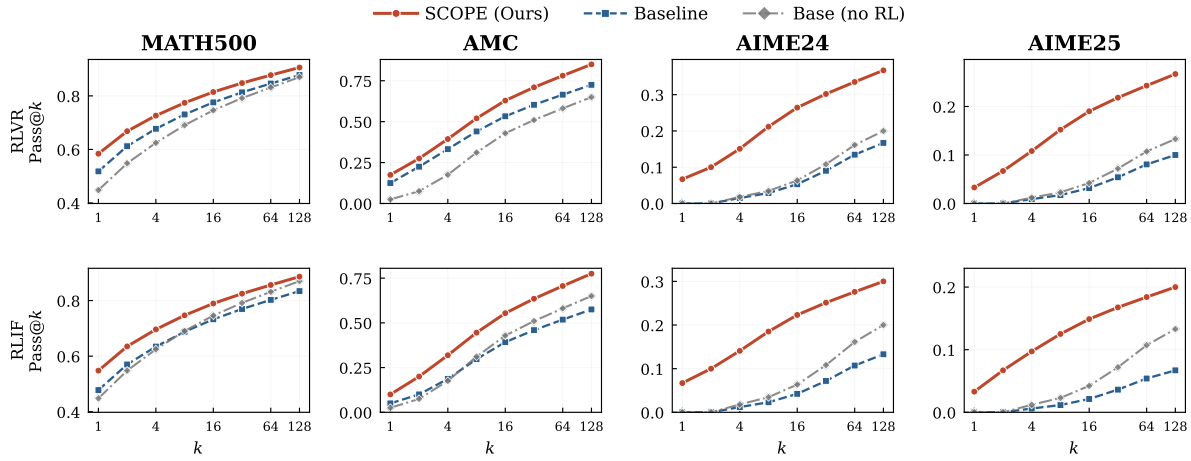

Figure 6: **Pass@k scaling on Qwen2.5-1.5B.** Same setup as Figure 3. At this scale, the base model's limited knowledge prevents diversity from translating into coverage: SCOPE outperforms all methods at every k, including Pass@128. On AIME24/25, the baseline achieves near-zero Pass@1 but gradually recovers at large k, while SCOPE sustains meaningful Pass@k growth across the entire range.

diminishing it. At small scales, where base capability is the binding constraint, SCOPE's targeted entropy preservation contributes at every k because the alternative—a collapsed trained policy—is the only thing worse than the already-weak base. At large scales, where base capability is abundant and the real challenge is preventing RL from narrowing the distribution below the level needed to exploit that capability, SCOPE's selective regularization is what permits simultaneous Pass@1 and Pass@k gains (Appendix C.1). The crossover observed between the base model and INTUITOR at 7B is, on this reading, a direct visualization of collapse severity: methods that collapse the policy too aggressively lose to the base model at large k; methods that regularize collapse—*where it matters*—do not.

D Extended Analysis of Cross-Architecture Generalization

This appendix unpacks two observations from Section 4.6 whose mechanistic readings connect the cross-architecture results directly to the theoretical framing of SCOPE.

D.1 Collapse Headroom: Why Gains on Qwen2.5-Math-7B Are Smaller

Section 4.6 attributes the reduced gains on Qwen2.5-Math-7B partly to hyperparameters tuned for Qwen2.5-7B-Base. A more structural account is also available. Qwen2.5-Math-7B is obtained by extensive math-focused continued pre-training on top of Qwen2.5-7B-Base—over 1T tokens of math-specific data drawn from web sources, books,

code, and model-synthesized problems (Yang et al., 2024)—rather than additional SFT or RLHF. Although this procedure is pre-training rather than post-training, its distributional effect is similar in the direction relevant to SCOPE: by heavily over-representing mathematical text, continued pre-training concentrates the initial policy on mathematical reasoning trajectories before RL begins. In the terminology of Proposition 1, for most math-relevant tokens the leading redistribution geometry $p_v(1 - p_v)$ is already small because p_v is already close to 0 or 1 at initialization.

This yields a clean principle: *SCOPE's gain is proportional to the remaining collapse headroom of the initial policy.* Because SCOPE mitigates collapse that *would have occurred* during RL, a policy whose math-relevant distribution is already concentrated offers little headroom for SCOPE to preserve—regardless of the absolute quality or task-competence of the model. The gain pattern on Qwen2.5-Math-7B is consistent with this view: directionally identical to Qwen2.5-7B-Base, with harder benchmarks still exhibiting larger relative gains than easier ones, but uniformly compressed in magnitude. The one benchmark where SCOPE regresses slightly (AIME24 RLIF: $0.144 \rightarrow 0.141$) is also the one where the Qwen2.5-Math-7B baseline is strongest relative to SCOPE on Qwen2.5-7B-Base ($0.144 > 0.115$), suggesting that this particular RLIF run did not induce the kind of collapse SCOPE is designed to prevent—there was simply no collapse for SCOPE to counteract.

This framing is distinct from the hyperparameter explanation in that it predicts *when* SCOPE will provide diminishing returns in future applications: domain-specialized base models produced by heavy continued pre-training on the target domain, models that have undergone multiple rounds of SFT or RLHF, and more broadly any setting in which the initial policy is already concentrated on the target distribution before RL begins. It also clarifies that "smaller gains on Qwen2.5-Math-7B" is not evidence against SCOPE's utility but rather a consequence of its intended mechanism: SCOPE preserves entropy that *would otherwise collapse*, so when that collapse does not occur in the first place, neither does the gain.

D.2 Support Redistribution: Why LLaMA Gains Vanish on Hard Benchmarks

The LLaMA3.1-8B-Instruct result—modest gains on easier benchmarks (GSM8K, MATH500), null gains on AIME24/25—admits a sharper reading than "LLaMA is unfit for RL." A central claim of Yue et al. (2025) is that RL post-training does not implant new capabilities but redistributes probability mass within the base policy's support: if a solution trajectory is absent from the base policy, no amount of entropy control, reward shaping, or gradient modification will make it samplable.

Our cross-architecture results constitute a direct empirical confirmation of this support-redistribution view. On GSM8K and MATH500, where LLaMA's baseline is nontrivial (0.756, 0.533), the base support evidently contains solution trajectories, and SCOPE improves over baseline much as on Qwen. On AIME24/25, where the baseline is at or near zero (AIME24: $0.028 \rightarrow 0.028$; AIME25: $0.000 \rightarrow 0.000$), the base support does not contain solution trajectories, and SCOPE has nothing to redistribute toward. The same SCOPE optimizer, applied to the same model on the same task, produces positive gains on some benchmarks and null gains on others, with the transition aligning precisely with where the baseline signals nonzero support.

Two implications follow. First, the LLaMA result should not be read as a limitation of SCOPE but as a visible consequence of its *intended* scope: SCOPE targets collapse-induced loss of diversity, not insufficiency of initial support. These are distinct failure modes—the first is created by RL optimization and removable by entropy control, the second is a property of the pretrained model that no post-training method can repair—and conflating them obscures what post-training methods can and cannot deliver. Second, cross-architecture evaluations of any RL post-training method must span multiple difficulty levels simultaneously. An evaluation restricted to AIME24/25 across architectures would conclude "SCOPE does not generalize to LLaMA"; one restricted to GSM8K would conclude the opposite. Both would be partial truths, and only the joint pattern reveals that SCOPE's generalization is gated by base-support coverage—a property independent of SCOPE itself.

This closes a loop in the paper's narrative. Section 2 cites Yue et al. (2025) as theoretical motivation for entropy-aware RL, and the cross-architecture experiment in Section 4.6 now independently confirms that the same support-redistribution principle bounds the generalization of our own method.