SCOPE-RL: Stable and Quantitative Control of Policy Entropy in RL Post-Training

Chen Wang[1,2,*] **Zhaochun Li**[2,3,*] **Jionghao Bai**[2,4,*]
Hexuan Deng[2,5] **Guanting Dong**[6] **Ge Lan**[7,†] **Yue Wang**[2,†]

[1]College of Elite Engineers, Nankai University [2]Zhongguancun Academy
[3]Beijing Institute of Technology [4]Zhejiang University [5]Harbin Institute of Technology
[6]Renmin University of China [7]College of Software, Nankai University

Abstract

Reinforcement learning (RL) is a key paradigm for post-training large language models (LLMs), but the widely used Group Relative Policy Optimization (GRPO) often suffers from entropy collapse: exploration quickly disappears, policies converge prematurely, and sample diversity declines, ultimately harming training effectiveness. Existing remedies, including entropy bonuses and clip-based methods, rarely keep entropy within a stable exploration regime and often introduce oscillatory entropy or reward degradation. In this work, we identify a previously overlooked asymmetry in entropy dynamics: under high-temperature sampling, positive and negative samples have opposite effects on policy entropy. Specifically, high-temperature positive samples promote entropy growth, whereas negative samples suppress it. We provide a qualitative theoretical explanation for this phenomenon: when entropy decreases during policy updates, its derivative with respect to temperature is strictly positive under positive sample updates, indicating that high-temperature positive samples can counteract entropy decay, thereby slowing entropy collapse and potentially reversing it. Motivated by this insight, we propose SCOPE-RL, a stable entropy control framework through a regularization term constructed from temperature-adaptive positive samples. Extensive experiments show that SCOPE-RL consistently outperforms strong RL baselines on both Pass@1 and Pass@k. Our results provide evidence that escaping entropy collapse can improve reasoning performance, while our quantitative empirical analysis further reveals a non-monotonic entropy–performance relationship, with an optimal level of exploration for RL post-training in reasoning LLMs. Our code is available at https://github.com/597358816/SCOPE-RL.

1 Introduction

Reinforcement learning (RL) has become a central paradigm for improving the reasoning capabilities of large language models (LLMs) [1–6]. A key ingredient of RL is **exploration**, which requires the policy to generate sufficiently diverse and informative samples during training [7–10]. In reasoning LLMs, exploration is closely related to **entropy**, a measure of policy uncertainty and output diversity [11–13]. Among existing methods, Group Relative Policy Optimization (GRPO) has become a widely used backbone due to its efficiency and scalability [14–16]. However, GRPO suffers from a well-known failure mode: **entropy collapse**, where policy entropy steadily decreases, outputs become increasingly homogeneous, and exploration vanishes prematurely [17–19]. As a result, RL often fails to discover genuinely diverse reasoning strategies and may mainly sharpen behaviors already

*: Contribute equally.
† Correspondence to Ge Lan, email: lange@nankai.edu.cn.
† Correspondence to Yue Wang, email: yuewang@bza.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

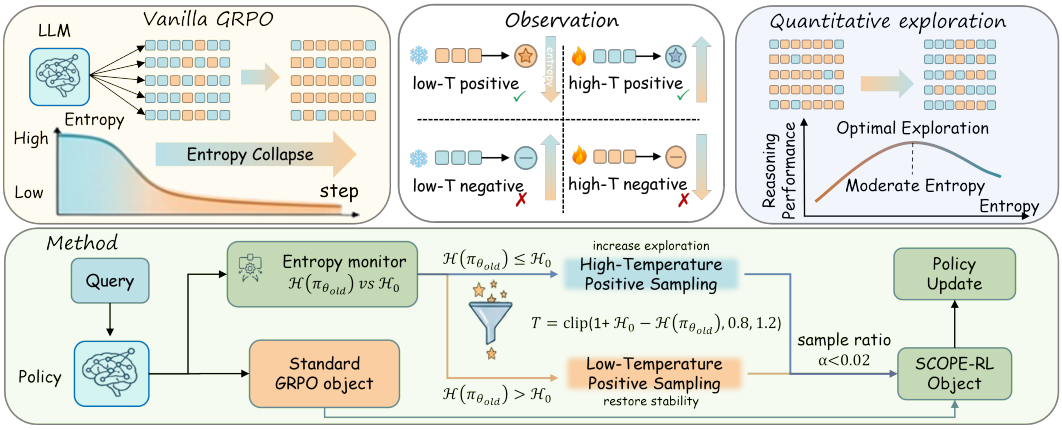


Figure 1: The overview of SCOPE-RL.

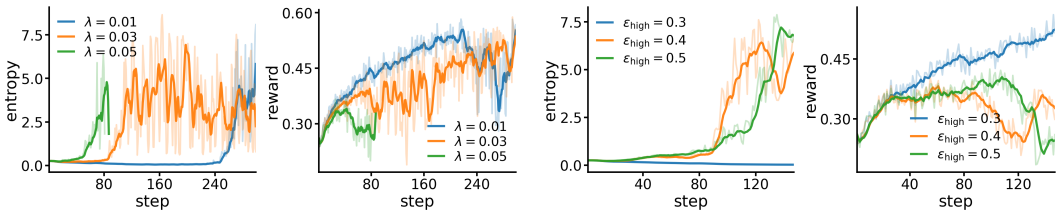
latent in the base model [20]. These findings point to a central challenge in RL post-training for reasoning LLMs: effective improvement requires restoring exploration, and restoring exploration requires overcoming entropy collapse [21].

Although entropy collapse in GRPO has been repeatedly recognized, it remains fundamentally unresolved. Existing attempts to alleviate this issue mainly follow two directions: entropy bonuses [21–24] and clip-based methods [17, 21, 25, 26]. However, both often introduce unstable training dynamics that undermine training effectiveness. In practice, entropy may oscillate between collapse and explosion, accompanied by irreversible performance degradation, rather than stabilizing at an exploration regime that is beneficial for learning. This instability often offsets the benefits brought by exploration and can even harm training. Therefore, there is still a lack of a stable entropy control mechanism that enables quantitative analysis of the entropy–performance relationship and helps identify the optimal degree of exploration in the exploration–exploitation trade-off.

To address this limitation, we revisit entropy regulation in RL post-training through the lens of sample-level entropy dynamics. In this work, we identify a previously overlooked asymmetry in entropy dynamics: positive and negative samples exert opposite effects on policy entropy, and this effect reverses across temperature regimes. Under high-temperature sampling, positive samples increase entropy, whereas negative samples accelerate entropy collapse; under low-temperature sampling, the trend reverses. We provide a qualitative theoretical explanation for this phenomenon: when entropy decreases during policy updates, the derivative of entropy with respect to temperature is strictly positive under positive-sample updates, indicating that high-temperature positive samples can counteract entropy decay, thereby slowing entropy collapse and potentially reversing it. Motivated by this insight, we propose **SCOPE-RL** (Fig. 1), a stable and quantitative entropy control framework via a regularization term built from less than 2% of temperature-adaptive positive samples. Through this mechanism, SCOPE-RL enables stable exploration control during RL post-training and maintains entropy within a controllable range, allowing us to quantitatively analyze the empirical relationship between entropy and training performance.

Our contributions are summarized as follows:

- We identify a previously overlooked temperature–sign asymmetry in sample-level entropy dynamics during RL post-training: positive and negative samples have opposite effects on policy entropy, and the direction reverses across temperature regimes.
- We provide a qualitative theoretical analysis showing that, under entropy collapse, the derivative of entropy with respect to temperature is strictly positive under positive-sample updates. Motivated by this insight, we propose SCOPE-RL, a stable entropy control framework using a temperature-adaptive positive-sample regularizer constructed from less than 2% of samples.
- We conduct extensive experiments on mathematical reasoning and knowledge-intensive benchmarks, showing that SCOPE-RL consistently outperforms strong RL baselines on both Pass@1 and Pass@k. On in-domain mathematical reasoning, it improves over GRPO by +3.23 on Qwen2.5-Math-7B, +2.10 on Qwen2.5-7B, and +1.96 on Qwen3-4B. Beyond Pass@1, it also surpasses both the base model and GRPO at Pass@1024 on AIME24 and AIME25.



(a) Entropy regularization (b) Clip higher

Figure 2: Training dynamics of Qwen2.5-Math-7B on the DAPO-17k dataset under different settings. Although both methods can induce entropy growth, the increase is often accompanied by unstable training dynamics and noticeable reward degradation.

- Our quantitative empirical analysis further shows that escaping entropy collapse improves performance, while the entropy–performance relationship is non-monotonic, with an optimal level of exploration in RL post-training.

2 Preliminary

We consider reinforcement learning post-training with verifiable rewards (RLVR) for reasoning tasks such as mathematical problem solving and code generation. Let q denote a query and $o = (o_1, \ldots, o_{|o|})$ denote a response generated autoregressively by a language model policy. We write the standard policy and its temperature-scaled variant as

$$\pi_\theta(o \mid q) = \prod_{t=1}^{|o|} \pi_\theta(o_t \mid q, o_{<t}), \qquad \pi_\theta^T(o_t \mid q, o_{<t}) = \frac{\exp(l(q, o_{<t}, o_t)/T)}{\sum_v \exp(l(q, o_{<t}, v)/T)}, \tag{1}$$

where $l(q, o_{<t}, o_t)$ is the logit of token o_t conditioned on $(q, o_{<t})$.

2.1 RL Post-Training and GRPO

In RLVR, each sampled response o is evaluated by a verifiable reward function $R(q, o)$, and the RL objective is to maximize the expected reward

$$\mathcal{J}_{\mathrm{RL}}(\theta) = \mathbb{E}_{q \sim P(Q),\, o \sim \pi_\theta(O|q)}[R(q, o)]. \tag{2}$$

Among existing methods, Group Relative Policy Optimization (GRPO) is a widely used backbone for reasoning LLMs. For each query q, it samples a group of G responses $\{o_i\}_{i=1}^G$ from $\pi_{\theta_{\mathrm{old}}}$ and computes the objective

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q),\, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\mathrm{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta) \hat{A}_i,\ \mathrm{clip}\big(r_{i,t}(\theta), 1 - \epsilon_{\mathrm{low}}, 1 + \epsilon_{\mathrm{high}}\big) \hat{A}_i \Big) \right], \tag{3}$$

where $\epsilon_{\mathrm{low}} = \epsilon_{\mathrm{high}} = 0.2$ and

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t} \mid q, o_{i,<t})}, \qquad \hat{A}_i = \frac{R(q, o_i) - \mathrm{mean}(\{R(q, o_j)\}_{j=1}^G)}{\mathrm{std}(\{R(q, o_j)\}_{j=1}^G)}. \tag{4}$$

By using group-normalized relative advantages, GRPO reduces variance and improves training stability. However, its policy entropy often decreases monotonically during training, leading to entropy collapse and premature loss of exploration.

2.2 Related Methods for Entropy

Existing remedies mainly fall into two categories: entropy regularization and clip-higher methods.

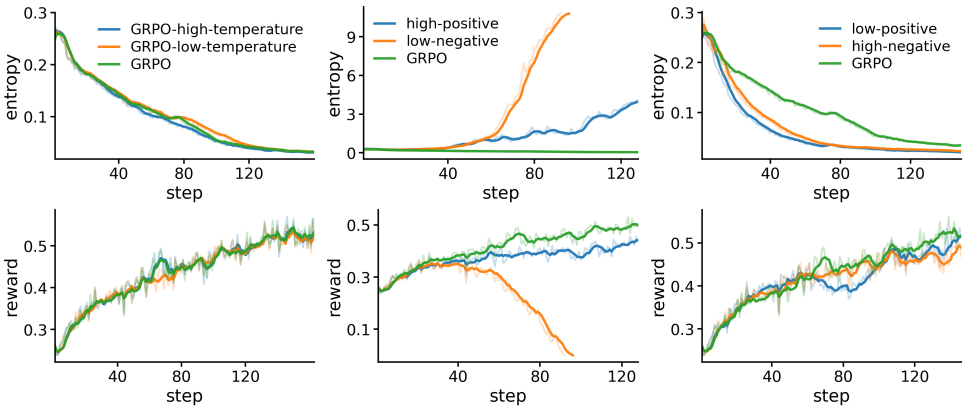


Figure 3: Observation experiments on Qwen2.5-Math-7B trained on DAPO-17k. Starting from a rollout batch size of 512×8, we additionally mix in 64 temperature-scaled samples per update. The left column varies the global sampling temperature in vanilla GRPO without separating positive and negative samples, while the middle and right columns isolate their effects under high- and low-temperature regimes. Changing temperature alone does not prevent entropy collapse; the entropy effect depends critically on both sample sign and temperature.

Entropy Regularization. The first category follows the principle of maximum-entropy reinforcement learning, which augments the reward objective with an entropy term to encourage exploration. For autoregressive language models, the token-level entropy and entropy-regularized objective are

$$\mathcal{H}(\pi_\theta) = \mathbb{E}_{q\sim P(Q),\, o\sim\pi_\theta(O|q)} \left[-\frac{1}{|o|} \sum_{t=1}^{|o|} \sum_{v} \pi_\theta(v \mid q, o_{<t}) \log \pi_\theta(v \mid q, o_{<t}) \right], \tag{5}$$
$$\mathcal{J}_{\text{MaxEnt}}(\theta) = \mathcal{J}_{\text{RL}}(\theta) + \lambda \mathcal{H}(\pi_\theta).$$

where $\lambda > 0$ controls the exploration strength. In GRPO-based training, such methods typically add entropy terms into the objective, rewards, or advantages. They are principled, but often introduce optimization bias and unstable entropy dynamics.

Clip higher. The second category modifies the symmetric clipping rule in GRPO to preserve diversity more aggressively, e.g., replacing $\text{clip}(r, 0.8, 1.2)$ with an asymmetric form

$$\text{clip}(r, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}), \qquad \epsilon_{\text{high}} > \epsilon_{\text{low}}. \tag{6}$$

For instance, DAPO enlarges the clipping range by setting $\epsilon_{\text{high}} = 0.3$ and $\epsilon_{\text{low}} = 0.2$, and empirically observes an increasing trend in entropy.

However, both Entropy Regularization and clip-higher often exhibit clear instability in practice, as shown in Fig. 2 and Table 1. As entropy increases, both methods tend to suffer from unstable training dynamics and noticeable reward drops. This suggests that both categories have difficulty delivering reliable entropy control and consistently promoting effective exploration.

3 Method

3.1 Observation and Analysis

To better understand how different sampled trajectories affect entropy during RL post-training, we conduct controlled observation experiments by varying the sampling temperature and by separately constructing auxiliary updates from positive or negative samples. The results are shown in Fig. 3. Three observations emerge.

Increasing temperature does not prevent entropy collapse. We vary the global sampling temperature in vanilla GRPO without decoupling positive and negative samples. As shown in Fig. 3, different temperatures slightly alter the entropy trajectory, but entropy still decreases monotonically under both

high- and low-temperature sampling. This shows that temperature alone is insufficient to stabilize exploration.

High-temperature positive and negative samples induce opposite entropy dynamics. Under high-temperature sampling, positive samples increase entropy while maintaining competitive reward, whereas negative samples accelerate entropy collapse. As shown in the middle column of Fig. 3, high-temperature positive samples produce a sustained entropy rise, while high-temperature negative samples drive entropy down even faster than GRPO, indicating that they reinforce rather than mitigate collapse.

The trend reverses under low-temperature sampling. Under low-temperature sampling, the entropy effect reverses: negative samples become entropy-increasing, whereas positive samples make entropy collapse more severe. However, although low-temperature negative samples can raise entropy, they also cause rapid reward collapse, making them unsuitable for stable training.

Taken together, these results reveal a temperature–sign asymmetry in sample-level entropy dynamics: the entropy effect is jointly determined by sample sign and temperature. Moreover, the failure of temperature-only control in vanilla GRPO shows that entropy regulation does not arise from temperature alone, but from its interaction with positive and negative samples. Motivated by this asymmetry, we next study a simplified bandit setting to characterize how temperature perturbs entropy drift and why high-temperature positive samples can counteract entropy collapse.

Assumption 1. *The actor policy π_θ is a tabular softmax policy; the reward is binary with $R(s,a)=1$ for positive samples; the logit gradients are orthogonal; and the update step is sufficiently small so that the one-step entropy variation admits a first-order approximation. The policy is updated via*

$$\theta_{k+1} = \theta_k + \eta \mathbb{E}_{a\sim\pi^T_{\theta_k}(\cdot|s)} \left[R(s,a) \cdot \nabla_\theta \log \pi_{\theta_k}(a|s) \right]. \tag{7}$$

Theorem 1. *Denote the function*

$$\Delta\mathcal{H}_{\theta_k}(T) = -\eta \cdot \sum_{a^*\in A^*} \pi^T_{\theta_k}(a^*|s) \times \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s)). \tag{8}$$

Then we have

$$\Delta\mathcal{H}_{\theta_k}(T) \approx \mathcal{H}(\pi_{\theta_{k+1}}) - \mathcal{H}(\pi_{\theta_k}), \tag{9}$$

and

$$\boxed{\frac{\boldsymbol{d}\Delta\mathcal{H}_{\theta_k}(T)}{\boldsymbol{d}T}|_{T=1} > 0} \quad \textit{if} \quad \Delta\mathcal{H}_{\theta_k}(T)|_{T=1} < 0. \tag{10}$$

The proof is provided in Appendix D. Theorem 1 provides a theoretical explanation for the asymmetry observed above. Specifically, when the entropy drift becomes negative, increasing the sampling temperature under positive-sample updates introduces a positive correction to the entropy dynamics, thereby counteracting entropy collapse. Motivated by this result, we propose **SCOPE-RL**, which explicitly leverages temperature-adjusted positive samples as a lightweight regularization signal to stabilize entropy and sustain exploration during RL post-training.

3.2 SCOPE-RL

Building on the above analysis, **SCOPE-RL** achieves entropy control by introducing a small auxiliary policy-gradient term constructed from temperature-adjusted positive samples. The core idea is straightforward but effective: when the policy entropy $\mathcal{H}(\pi_{\theta_{\text{old}}})$ of the previous step falls below a target threshold $\mathcal{H}_0$, SCOPE-RL draws a small number of high-temperature positive samples to promote exploration; when the entropy exceeds $\mathcal{H}_0$, it instead uses a small number of low-temperature positive samples to maintain training stability. As shown in Fig. 4, SCOPE-RL achieves entropy control by introducing a small auxiliary regularization term constructed from temperature-adjusted positive samples.

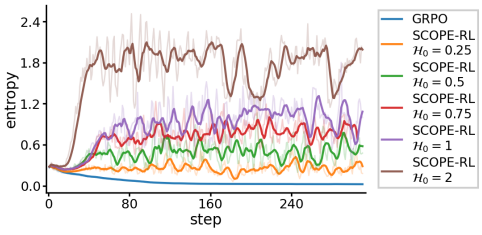


Figure 4: SCOPE-RL can not only alleviate entropy collapse, but also stably control entropy at different levels

Table 1: Main results on mathematical reasoning benchmarks. SCOPE-RL consistently improves performance across different base models.

Methods	AIME24	AIME25	AMC23	GSM8K	MATH	Minerva	Olympiad	Average
Qwen2.5-Math-7B	15.5	7.81	42.1	65.4	59.4	11.0	26.7	32.56
+GRPO	32.1	11.0	72.4	88.7	80.6	34.6	41.8	51.60
+Entropy-Reg ($\lambda = 0.01$)	34.2	10.1	73.1	88.7	79.8	35.3	40.8	51.71
+Entropy-Reg ($\lambda = 0.03$)	31.4	10.1	74.3	87.0	78.8	35.7	40.4	51.10
+Entropy-Reg ($\lambda = 0.05$)	21.6	10.8	63.6	86.5	77.0	31.9	39.2	47.23
+DAPO	33.3	13.0	73.9	87.9	79.6	34.6	40.8	51.87
+Clip higher ($\epsilon_{\text{high}} = 0.4$)	29.8	9.58	64.1	80.6	76.2	30.1	40.4	47.25
+Clip higher ($\epsilon_{\text{high}} = 0.5$)	24.0	7.92	61.8	84.9	76.4	32.7	38.9	46.66
+Entropy-Adv	31.3	11.4	72.1	87.8	78.8	37.5	42.1	51.57
+CISPO	35.0	13.8	75.1	89.7	80.2	37.1	41.6	53.21
+KL-cov	32.1	13.5	73.9	**89.9**	79.8	35.6	41.7	52.36
+SCOPE-RL	**36.4**	**18.9**	**75.3**	89.5	**82.6**	**38.2**	**43.0**	**54.84**
Δ vs. GRPO	(+4.3)	(+7.9)	(+2.9)	(+0.8)	(+2.0)	(+3.6)	(+1.2)	(+3.23)
Qwen2.5-7B	7.91	5.31	36.2	88.5	64.4	22.0	29.3	36.23
+GRPO	17.1	7.60	65.8	92.3	75.6	36.8	38.8	47.71
+DAPO	**17.7**	9.06	68.0	92.1	76.4	37.1	40.1	48.64
+Entropy-Reg ($\lambda = 0.01$)	13.6	8.85	67.4	92.3	76.8	35.5	39.1	47.65
+Entropy-Adv	14.8	10.1	67.3	91.9	76.6	**38.2**	37.5	48.06
+CISPO	15.1	8.10	66.6	92.3	77.6	37.1	38.2	47.86
+KL-cov	16.9	11.3	65.5	91.5	76.8	**38.2**	39.1	48.47
+SCOPE-RL	17.5	**12.9**	**69.3**	**92.9**	**78.0**	37.8	**40.3**	**49.81**
Δ vs. GRPO	(+0.4)	(+5.3)	(+3.5)	(+0.6)	(+2.4)	(+1.0)	(+1.5)	(+2.10)
Qwen3-4B	36.4	22.7	71.9	93.9	84.8	42.3	47.2	57.03
+GRPO	52.9	41.5	86.1	**95.2**	92.0	46.7	60.0	67.77
+DAPO	53.1	41.7	87.3	94.5	92.0	47.1	61.1	68.04
+Entropy-Reg ($\lambda = 0.01$)	52.4	42.6	88.1	94.6	91.6	46.3	60.1	67.96
+Entropy-Adv	51.6	41.7	87.0	94.2	91.2	46.0	58.4	67.16
+CISPO	52.7	42.6	89.3	94.7	91.4	47.8	60.3	68.40
+KL-cov	53.9	42.3	88.7	94.8	92.4	46.3	60.0	68.34
+SCOPE-RL	**56.5**	**43.7**	**89.7**	95.0	**92.8**	**48.9**	**61.5**	**69.73**
Δ vs. GRPO	(+3.4)	(+2.2)	(+3.6)	(-0.2)	(+0.8)	(+2.2)	(+1.5)	(+1.96)

Formally, the SCOPE-RL objective is defined as:

$$\mathcal{J}_{\text{SCOPE}}(\theta) = \mathcal{J}_{\text{GRPO}}(\theta) + \alpha\, \mathbb{E}_{q\sim P(Q),\, \{o_i\}_{i=1}^{G'} \sim \pi_{\theta_{\text{old}}}^{T}(O|q)} \left[\frac{1}{G'} \sum_{i=1}^{G'} \mathbf{1}[R(q, o_i) = 1] \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta),\ \text{clip}\big(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\big) \Big) \right]. \tag{11}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \quad \text{and} \quad T = \text{clip}(1 + \mathcal{H}_0 - \mathcal{H}(\pi_{\theta_{\text{old}}}), 0.8, 1.2). \tag{12}$$

SCOPE-RL contains two key components:

- *Positive samples as regularization*: motivated by the observations and analysis above, SCOPE-RL retains only positive samples in the auxiliary branch, as they provide a more stable signal for entropy control under temperature adjustment.
- *Temperature as regularization*: when entropy falls below the target threshold, high-temperature samples are used to encourage exploration; when entropy rises above the threshold, low-temperature samples are used to restore stability.

4 Experiments

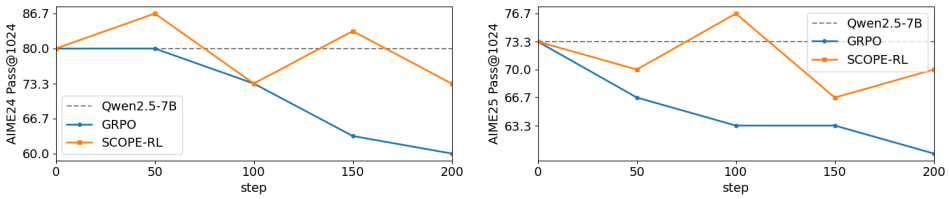
To validate SCOPE-RL, we implement it within the EasyR1 and VeRL frameworks [27, 28] and compare it against representative RL baselines, including GRPO [14], entropy-based methods such

Table 2: Generalization results on knowledge-intensive benchmarks, including ARC-Challenge, MMLU-Pro, and SuperGPQA, where SCOPE-RL demonstrates leading performance across RL methods.

Methods	Qwen2.5-Math-7B			Qwen2.5-7B			Qwen3-4B		
	ARC	MMLU	S-GPQA	ARC	MMLU	S-GPQA	ARC	MMLU	S-GPQA
Base	69.9	34.3	19.4	86.3	50.0	18.6	92.9	72.6	32.8
+GRPO	78.9	44.8	25.8	89.3	58.2	27.0	**94.6**	74.1	38.4
+DAPO	80.9	**48.6**	26.0	90.9	59.8	24.8	92.6	75.0	37.0
+Entropy-Reg	79.6	42.7	26.2	89.6	58.2	26.8	93.9	73.7	38.2
+Entropy-Adv	80.3	47.3	26.2	90.6	57.7	26.6	93.6	74.8	37.8
+CISPO	80.9	44.1	26.2	89.9	57.8	25.8	92.3	75.1	36.8
+KL-cov	79.9	48.0	26.6	90.9	59.1	26.0	93.3	75.5	37.6
+SCOPE-RL	**80.9**	48.0	**26.8**	**91.6**	**60.7**	**27.2**	94.3	**76.1**	**38.6**

Table 3: Comparison of Pass@k performance between the base model, GRPO, and SCOPE-RL across AIME24 and AIME-25. The accompanying figures show the evolution of Pass@1024 of SCOPE-RL and GRPO on AIME24 and AIME25 during training.

Benchmarks	AIME24		AIME25	
	Pass@512	Pass@1024	Pass@512	Pass@1024
Qwen2.5-7B	**71.7**	80.0	61.7	73.3
+GRPO	63.3	73.3	60.0	63.3
+SCOPE-RL	70.0	**86.7**	**66.7**	**76.7**



as Entropy-Reg [22], Entropy-Adv [24], and KL-cov [21], as well as clipping- or entropy-related approaches including DAPO [17] and CISPO [25]. Experiments are conducted on Qwen2.5-7B, Qwen2.5-Math-7B, and Qwen3-4B [29, 30], using DAPO-17K [17] for training. We evaluate both in-domain mathematical reasoning and out-of-domain generalization on seven math benchmarks, including AIME24, AIME25 [31], AMC23 [32], GSM8K [33], MATH [32], Minerva Math [34], and Olympiad [32], as well as three knowledge-intensive benchmarks, ARC-Challenge [35], MMLU-Pro[36], and SuperGPQA [37]. The detailed implementation is shown in Appendix C.

4.1 Main Results

In-domain Mathematical Reasoning. Table 1 shows that SCOPE-RL achieves the best average performance on all three base models, improving over GRPO from 51.60 to 54.84 on Qwen2.5-Math-7B, from 47.71 to 49.81 on Qwen2.5-7B, and from 67.77 to 69.73 on Qwen3-4B. It is particularly strong on challenging benchmarks such as AIME24, AIME25, and MATH, and achieves the strongest overall compared with other RL baselines.

Out-of-domain Generalization. Table 2 shows that SCOPE-RL also transfers well beyond math training data. On Qwen2.5-7B, it achieves the best results on all three knowledge benchmarks, reaching 91.6 on ARC-Challenge, 60.7 on MMLU-Pro, and 27.2 on SuperGPQA. This suggests that the benefit of entropy control is not limited to in-domain mathematical reasoning.

Pass@k Performance and Reasoning Capability. Table 3 provides a complementary view from Pass@k evaluation. On Qwen2.5-7B, GRPO underperforms the base model at Pass@1024 on both AIME24 and AIME25, while SCOPE-RL improves them to 86.7 and 76.7, respectively. The accompanying curves show the same trend dynamically: as training proceeds, Pass@1024 under GRPO degrades steadily, while SCOPE-RL maintains clearly stronger many-sample performance

Table 4: SCOPE-RL under different entropy targets $\mathcal{H}_0$, which correspond to different degrees of exploration. The overall performance is maximized at a moderate entropy level.

Benchmarks	AIME24	AIME25	AMC	GSM8K	MATH	Minerva	Olympiad	Average
Qwen2.5-math-7B	15.5	7.81	42.1	65.4	59.4	11.0	26.7	32.56
+SCOPE-RL $\mathcal{H}_0 = 0.25$	**37.9**	11.3	74.4	89.4	79.8	36.0	42.2	53.00
+SCOPE-RL $\mathcal{H}_0 = 0.50$	36.4	**18.9**	75.3	**89.5**	**82.6**	**38.2**	**43.0**	**54.84**
+SCOPE-RL $\mathcal{H}_0 = 0.75$	33.6	15.1	**77.0**	89.4	79.2	37.5	42.4	53.45
+SCOPE-RL $\mathcal{H}_0 = 1.00$	33.2	**15.6**	74.8	88.7	79.6	37.9	42.1	53.13

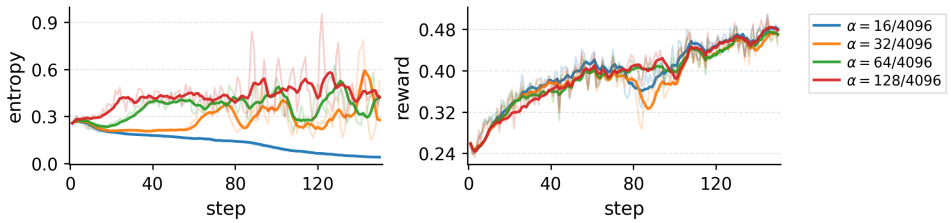


Figure 5: Entropy and reward curves of SCOPE-RL under different auxiliary sample ratios α on Qwen2.5-Math-7B, trained on DAPO-17K with a rollout batch size of 512×8 and $\mathcal{H}_0 = 0.5$.

throughout training. This is consistent with our claim that stable entropy control preserves useful exploration, which is particularly important for the reasoning capability.

Quantitative Analysis of Exploration Degree. Table 4 suggests a non-monotonic relationship between entropy level and performance in our setting. When varying the entropy target, SCOPE-RL achieves the best average performance at $\mathcal{H}_0 = 0.50$ with 54.84, outperforming $\mathcal{H}_0 = 0.25$ (53.00) and $\mathcal{H}_0 = 0.75$ (53.45). These results indicate that exploration is most beneficial when maintained within an appropriate range.

Effectiveness and Efficiency. An important consideration of SCOPE-RL is its effectiveness and efficiency trade-off. To study this, we vary the auxiliary sample ratio α, which in our implementation also controls the effective regularization strength. As shown in Fig. 5, compared with GRPO, whose entropy steadily collapses, SCOPE-RL consistently raises the entropy level and slows down collapse even with a very small auxiliary budget, indicating only marginal additional sampling cost. Meanwhile, a larger α does not necessarily improve optimization: although $\alpha = 128/4096$ yields higher entropy, it also introduces substantially larger fluctuations, suggesting reduced training stability. In contrast, moderate values such as $\alpha = 32/4096$ and $\alpha = 64/4096$ are already sufficient to prevent entropy collapse while maintaining rewards comparable to GRPO, using less than 2% additional auxiliary samples. In practice, we adopt $\alpha = 64/4096 = 1/64$ as the default setting, which provides effective entropy control with limited overhead.

Additional Results. Appendix A provides further evidence beyond the main experiments. Three-seed experiments on Llama-3.1-8B-Instruct under matched rollout budgets examine robustness across model families and training runs (Appendix A.1); expanded Pass@32 comparisons cover four base models and several anti-collapse baselines (Appendix A.2). We also report prompt-level correctness dynamics to complement the entropy analysis (Appendix A.3) and test entropy control under different clipping settings (Appendix A.4).

5 Ablations

We conduct ablation studies to understand the contribution of each design choice in SCOPE-RL and investigate the underlying mechanism behind its entropy control.

5.1 Component Ablations

We ablate the key components of SCOPE-RL in Table 5. **Temperature adjustment.** Removing temperature-adjusted sampling reduces the average score from 54.84 to 51.31 and leads to entropy collapse, showing that the positive-only auxiliary branch is insufficient by itself. **Positive-sample**

Table 5: Component ablations of SCOPE-RL on Qwen2.5-Math-7B. Removing either temperature-adjusted or positive sampling degrades performance and entropy control.

Setting	AIME24	AIME25	AMC	GSM8K	MATH	Minerva	Olympiad	Average
Entropy step	50	100	150	200	250	300	350	Entropy trends
w/o temperature	31.4	11.8	74.5	88.9	79.0	33.5	40.1	51.31 (-3.53)
	0.080	0.036	0.026	0.024	0.022	0.019	0.018	Entropy collapse
positive+negative	32.7	11.0	73.2	88.1	79.0	35.7	41.2	51.56 (-3.28)
	0.114	0.041	0.027	0.024	0.024	0.022	0.020	Entropy collapse
only negative	28.9	12.7	65.1	89.5	77.6	33.8	40.4	49.71 (-5.12)
	0.492	1.176	5.224	5.948	7.235	-	-	Entropy instability
only regularizer	-	-	-	-	-	-	-	Training failed

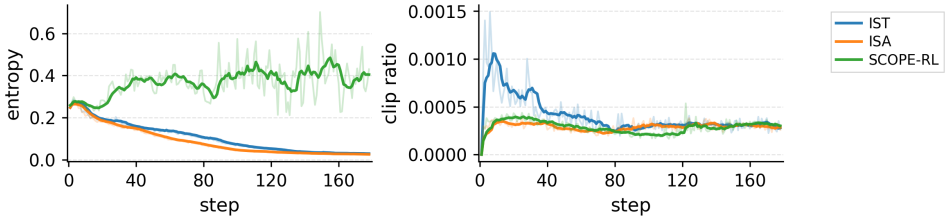


Figure 6: Importance-sampling ablations on entropy control and clipping behavior.

filtering. Including both positive and negative auxiliary samples reduces the average score to 51.56 and again leads to entropy collapse. Using only negative auxiliary samples produces unstable entropy and weaker performance, while using only the auxiliary regularizer fails to train. Together, these results support the combination of temperature adaptation and positive-sample filtering used in SCOPE-RL.

5.2 Importance-Sampling Ablations

The auxiliary samples of SCOPE-RL are drawn from $\pi^{T}_{\theta_{\text{old}}}$, but the likelihood ratio in SCOPE-RL uses $\pi_{\theta_{\text{old}}}$. **This mismatch is intentional: temperature changes which positive trajectories enter the auxiliary update, and correcting for the sampling distribution may remove the reweighting that supports entropy control.** We therefore test whether this mismatch matters empirically, while separating its effect from a possible change in clipping behavior. We consider two importance-sampling variants of the auxiliary objective: importance sampling on the likelihood ratio (IST) and importance sampling on the advantage (ISA):

$$
\begin{aligned}
r_{i,t}(\theta) &\leftarrow \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi^{T}_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, && \text{(IST)} \\
\mathbf{1}[R(q,o_i)=1] &\leftarrow \frac{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}{\pi^{T}_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})} \mathbf{1}[R(q,o_i)=1]. && \text{(ISA)}
\end{aligned}
\tag{13}
$$

The first variant clips the corrected ratio. The second retains SCOPE-RL's original ratio and clipping rule, applying the correction as a token-wise weight instead. All other components are unchanged. Fig. 6 shows that entropy declines under both importance-sampling variants while SCOPE-RL maintains higher entropy, and the clipping fractions further indicate that this degradation is not caused by altered clipping behavior, as entropy still declines when the original clipping rule is retained. **These results show that SCOPE-RL's entropy-control ability depends on the bias introduced by temperature-adjusted sampling: correcting this bias through importance weighting causes entropy control to disappear. SCOPE-RL achieves better downstream performance by confining the biased auxiliary update to a small, controlled fraction of training, so that it regulates entropy without dominating reward optimization.** Further details on the motivation, objective formulations, and interpretation of these importance-sampling ablations are provided in Appendix B.

6 Related Work

Reinforcement learning (RL) has become a core paradigm for post-training large language models (LLMs), especially for improving reasoning beyond supervised fine-tuning [38, 16, 39–41]. Among recent reasoning-oriented RL methods, Group Relative Policy Optimization (GRPO) has emerged as a widely adopted backbone due to its simplicity, scalability, and effectiveness [14–16]. However, recent evidence suggests that RL post-training often sharpens behaviors already latent in the base model rather than consistently expanding its reasoning frontier [20], motivating growing interest in exploration during RL post-training. Entropy has long been viewed as a key proxy for exploration, as it reflects policy uncertainty and output diversity and helps prevent premature convergence [11–13, 42, 43]. Accordingly, recent work has revisited entropy regulation in RL for LLMs as a means to sustain exploration and improve optimization [21–24].

Recent studies have identified entropy collapse as a major failure mode of GRPO-based training, where policy entropy decreases monotonically, sampled outputs become increasingly homogeneous, and exploration vanishes prematurely [17–19]. Existing attempts to mitigate this issue mainly fall into two directions. The first introduces entropy bonuses into rewards, advantages, or auxiliary objectives, such as Entropy Regularization (Entropy-Reg) [22], KL-Cov [21], AEnt [23], Entropy Advantage (Entropy-Adv) [24] and so on. The second relies on clipping-based or clip-higher strategies, such as DAPO [17], Clip-Cov [21], CISPO [25], CE-GPPO [26] and so on. Although both directions can alleviate entropy collapse to some extent, they often lead to unstable training dynamics, with entropy oscillating between collapse and explosion rather than remaining in a beneficial regime. In contrast, SCOPE does not directly manipulate entropy through bonuses or clipping heuristics. Instead, it exploits sample-level entropy dynamics, using a very small fraction of temperature-adaptive positive samples to achieve stable entropy control and enable quantitative analysis of entropy's effect on training.

7 Limitations

Our theoretical analysis is developed under several simplifying assumptions, including a contextual-bandit abstraction, a tabular softmax policy, binary rewards, orthogonal logit gradients, and sufficiently small updates; it also analyzes the temperature derivative only locally at $T = 1$. These assumptions are stronger than the full autoregressive setting of language modeling, but are adopted to isolate the local mechanism behind entropy control rather than to characterize all RL post-training dynamics. In particular, Theorem 1 focuses on the regime $\Delta\mathcal{H}_{\theta_k}(1) < 0$, which corresponds to the practically important setting where entropy decreases during a policy update and entropy collapse emerges. Under explicit entropy control, the opposite case $\Delta\mathcal{H}_{\theta_k}(1) > 0$ may also arise; in this regime, the key inequality in Eq. (40) no longer holds strictly, making the current proof inapplicable and a useful quantitative bound harder to derive. We therefore do not claim that the theory fully covers every controlled-entropy regime or establishes stability at arbitrary temperatures. Nevertheless, it provides a useful perspective on how temperature-adjusted positive samples reshape entropy dynamics and counteract entropy collapse. Beyond the theorem's scope, the empirical results in *Observation and Analysis* (Fig. 3) and *Effectiveness and Efficiency* (Fig. 5) show effective entropy control in the evaluated training settings.

8 Conclusion

In this work, we investigate entropy collapse in RL post-training for reasoning LLMs, and our results suggest that stable exploration is important for effective optimization. We identify a temperature–sign asymmetry in sample-level entropy dynamics and, based on this insight, propose SCOPE-RL, which regulates entropy through temperature-adaptive positive-sample regularization. Experiments show that SCOPE-RL consistently improves both Pass@1 and Pass@k over RL baselines, while enabling quantitative analysis of the exploration–performance trade-off. Overall, our results suggest that maintaining a proper level of entropy is key to improving reasoning performance in RL post-training. A promising direction for future work is to move beyond global entropy control and develop state-dependent exploration mechanisms that adapt to problem difficulty and reasoning stage, together with a more complete theory of entropy dynamics in token-level RL.

Acknowledgments

This work is supported by the Zhongguancun Academy (Grant No.s C20250203) and Natural Science Foundation of Tianjin (No.24JCQNJC02170).

References

[1] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

[2] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[4] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[5] Han Zhong, Zikang Shan, Guhao Feng, Wei Xiong, Xinle Cheng, Li Zhao, Di He, Jiang Bian, and Liwei Wang. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*, 2024.

[6] Zhichao Wang, Bin Bi, Shiva Kumar Pentyala, Kiran Ramnath, Sougata Chaudhuri, Shubham Mehrotra, Xiang-Bo Mao, Sitaram Asur, et al. A comprehensive survey of llm alignment techniques: Rlhf, rlaif, ppo, dpo and more. *arXiv preprint arXiv:2407.16216*, 2024.

[7] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[8] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

[9] Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[10] J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520, 2009.

[11] John Schulman, Xi Chen, and Pieter Abbeel. Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440*, 2017.

[12] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[13] Ofir Nachum, Mohammad Norouzi, Kelvin Xu, and Dale Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[14] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[15] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

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

[17] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[18] Gang Li, Ming Lin, Tomer Galanti, Zhengzhong Tu, and Tianbao Yang. Disco: Reinforcing large reasoning models with discriminative constrained optimization. *arXiv preprint arXiv:2505.12366*, 2025.

[19] Xingjian Zhang, Siwei Wen, Wenjun Wu, and Lei Huang. Edge-grpo: Entropy-driven grpo with guided error correction for advantage diversity. *arXiv preprint arXiv:2507.21848*, 2025.

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

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

[22] Zhenyu Hou, Xin Lv, Rui Lu, Jiajie Zhang, Yujiang Li, Zijun Yao, Juanzi Li, Jie Tang, and Yuxiao Dong. Advancing language model reasoning through reinforcement learning and inference scaling. *arXiv preprint arXiv:2501.11651*, 2025.

[23] Han Shen. On entropy control in llm-rl algorithms. *arXiv preprint arXiv:2509.03493*, 2025.

[24] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*, 2025.

[25] Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

[26] Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. Ce-gppo: Coordinating entropy via gradient-preserving clipping policy optimization in reinforcement learning. *arXiv preprint arXiv:2509.20712*, 2025.

[27] Zheng Yaowei, Lu Junting, Wang Shenzhi, Feng Zhangchi, Kuang Dongdong, and Xiong Yuwen. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

[28] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[29] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[30] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[31] HuggingFaceH4. AIME 2024 Dataset (AIME I & II). https://huggingface.co/datasets/HuggingFaceH4/aime_2024, 2025.

[32] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

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

[34] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

[35] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[36] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[37] Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, et al. Supergpqa: Scaling llm evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025.

[38] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[39] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[40] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[41] Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *arXiv preprint arXiv:2308.12067*, 2023.

[42] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[43] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[44] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

A More Results

A.1 Additional Results on Llama-3.1-8B-Instruct

To further evaluate the robustness of SCOPE-RL across different model families and training randomness, we conduct additional experiments on Llama-3.1-8B-Instruct with three independent random seeds. For a fair comparison, the rollout budgets of all baselines are strictly controlled to be identical. We report the mean and standard deviation across the three runs in Table 6.

These additional results provide further evidence that the effectiveness of SCOPE-RL transfers across model families. Compared with the main experiments on Qwen2.5-7B, Qwen2.5-Math-7B, and Qwen3-4B, the evaluation on Llama-3.1-8B-Instruct extends the empirical study to a different model family and a larger model scale. Moreover, the three independent training runs quantify the variability induced by training randomness, while controlling all methods under the same rollout budget.

Table 6: Additional results on Llama-3.1-8B-Instruct over three independent random seeds. All methods use identical rollout budgets. Results are reported as mean ± standard deviation.

Method	AIME24 HMMT25	AIME25 MATH500	AMC23 Minerva	ARC MMLU-Pro	CMIMC25 Olympiad	GSM8K SuperGPQA
SCOPE-RL	$\mathbf{11.25 \pm 1.00}$ $\mathbf{1.81 \pm 1.15}$	$\mathbf{1.53 \pm 0.33}$ $\mathbf{54.33 \pm 1.21}$	$\mathbf{28.62 \pm 1.63}$ $\mathbf{29.17 \pm 0.56}$	84.95 ± 0.00 $\mathbf{53.63 \pm 1.19}$	$\mathbf{4.87 \pm 1.34}$ $\mathbf{19.95 \pm 0.31}$	$\mathbf{88.32 \pm 0.40}$ $\mathbf{19.47 \pm 0.64}$
GRPO	7.74 ± 0.34 0.90 ± 0.24	0.24 ± 0.16 52.93 ± 1.86	26.72 ± 0.68 25.73 ± 1.33	83.72 ± 1.27 50.83 ± 1.76	3.88 ± 0.16 19.16 ± 2.23	87.09 ± 0.32 17.07 ± 0.46
DAPO	6.73 ± 0.95 0.21 ± 0.21	1.04 ± 0.21 52.20 ± 2.03	25.21 ± 1.52 24.88 ± 0.21	81.60 ± 1.74 51.49 ± 1.52	4.20 ± 0.32 18.47 ± 0.09	87.77 ± 0.50 18.27 ± 1.21
KL-Cov	8.58 ± 0.79 0.31 ± 0.32	0.45 ± 0.26 50.93 ± 2.14	22.45 ± 2.35 24.63 ± 0.64	$\mathbf{85.28 \pm 1.21}$ 52.03 ± 0.92	3.23 ± 0.56 18.82 ± 0.39	88.22 ± 0.19 17.47 ± 1.10
CISPO	Training failed					

Table 7: Expanded Pass@32 evaluation across four base models spanning three model families. SCOPE-RL achieves the highest average Pass@32 performance on all evaluated base models.

Base Model	Method	AIME24	AIME25	AMC23	CMIMC25	HMMT25	Average
Qwen2.5-Math-7B	**SCOPE-RL**	**60.00**	**40.00**	**87.50**	**20.00**	**13.30**	**44.16**
	GRPO	53.30	23.30	82.50	15.00	10.00	36.82
	DAPO	50.00	33.33	82.50	**20.00**	10.00	39.17
	KL-Cov	53.33	26.67	82.50	**20.00**	10.00	38.50
	CISPO	50.00	26.67	80.00	15.00	6.67	35.67
Qwen2.5-7B	**SCOPE-RL**	**40.00**	**36.67**	**90.00**	**32.50**	**13.33**	**42.50**
	GRPO	26.67	30.00	87.50	25.00	10.00	35.83
	DAPO	36.67	30.00	87.50	27.50	**13.33**	39.00
	KL-Cov	30.00	26.67	80.00	30.00	**13.33**	36.00
	CISPO	Training failed					
Qwen3-4B	**SCOPE-RL**	**86.67**	73.33	**100.00**	**60.00**	**56.67**	**75.33**
	GRPO	80.00	63.33	97.50	47.50	53.33	68.33
	DAPO	76.67	**76.67**	97.50	52.50	50.00	70.67
	KL-Cov	83.33	66.67	97.50	50.00	53.33	70.17
	CISPO	76.67	70.00	97.50	50.00	46.67	68.17
Llama-3.1-8B-Instruct	**SCOPE-RL**	**36.67**	**20.00**	**80.00**	15.00	**10.00**	**32.33**
	GRPO	33.33	13.33	72.50	12.50	6.67	27.67
	DAPO	33.33	13.33	75.00	15.00	**10.00**	29.33
	KL-Cov	**36.67**	10.00	72.50	**17.50**	6.67	28.67
	CISPO	Training failed					

A.2 Expanded Pass@32 Evaluation

We further expand the Pass@k evaluation to four base models spanning four models, including Qwen2.5-Math-7B, Qwen2.5-7B, Qwen3-4B, and Llama-3.1-8B-Instruct. In addition to GRPO, we compare SCOPE-RL against representative entropy- or clipping-based anti-collapse baselines, including DAPO, KL-Cov, and CISPO. The results are shown in Table 7.

SCOPE-RL achieves the highest average Pass@32 performance on all four base models. Specifically, it reaches 44.16 on Qwen2.5-Math-7B, 42.50 on Qwen2.5-7B, 75.33 on Qwen3-4B, and 32.33 on Llama-3.1-8B-Instruct, consistently outperforming GRPO and the representative anti-collapse baselines.

A.3 Whether Preserved Entropy Reflects Useful Reasoning Exploration

To examine whether the entropy preserved by SCOPE-RL corresponds to useful reasoning exploration, we analyze prompt-level training dynamics on Qwen2.5-Math-7B trained with DAPO-17K. For each epoch, prompts are categorized as *all-correct*, *mixed*, or *all-wrong* according to the correctness of their sampled responses. The results are shown in Table 8.

Table 8: Prompt-level training dynamics of GRPO and SCOPE-RL on Qwen2.5-Math-7B over four epochs.

Method	Epoch	Entropy	All-correct	Mixed	All-wrong
GRPO	1	0.192	0.105	0.484	0.411
SCOPE-RL	1	**0.310**	0.093	0.484	0.423
GRPO	2	0.151	0.163	0.442	0.395
SCOPE-RL	2	**0.418**	0.161	**0.443**	0.396
GRPO	3	0.135	0.184	0.408	0.408
SCOPE-RL	3	**0.403**	**0.225**	0.407	**0.368**
GRPO	4	0.118	0.224	0.378	0.398
SCOPE-RL	4	**0.430**	**0.265**	**0.427**	**0.318**

Under GRPO, entropy continuously decreases, while the proportion of all-correct prompts increases mainly at the expense of mixed prompts and the proportion of all-wrong prompts changes only slightly. In contrast, SCOPE-RL substantially increases the proportion of all-correct prompts while reducing all-wrong prompts, and maintains a higher proportion of mixed prompts at the end of training. These results suggest that the preserved entropy is associated with continued discovery of successful reasoning trajectories rather than merely increased output randomness.

A.4 Robustness to Clipping Settings

We further study whether SCOPE-RL remains effective under different clipping settings. In addition to the default symmetric clipping $(\epsilon_{\text{low}}, \epsilon_{\text{high}}) = (0.2, 0.2)$, we conduct controlled experiments with two asymmetric variants, $(0.2, 0.3)$ and $(0.3, 0.2)$, both combined with SCOPE-RL. Fig. 7 reports the corresponding entropy trajectories.

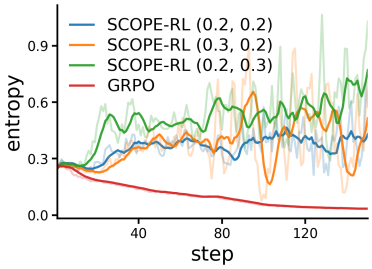


Figure 7: Entropy trajectories of SCOPE-RL under different clipping settings on Qwen2.5-Math-7B.

GRPO exhibits clear entropy collapse, with entropy decreasing from 0.2368 to 0.0395. In contrast, SCOPE-RL maintains substantially higher entropy under all three clipping configurations. In the final interval, the two asymmetric variants retain entropy values of 0.5028 and 0.5795, respectively.

These results show that changing clipping bounds affects the absolute entropy trajectory, but does not undermine the entropy control effect of SCOPE-RL. Together with the broader experiments across models, benchmarks, random seeds, and identical rollout budgets, this provides further evidence that SCOPE-RL is empirically robust to clipping settings.

B Details of Importance-Sampling Ablations

The auxiliary samples in SCOPE-RL are drawn from the temperature-adjusted policy $\pi^{T}_{\theta_{\text{old}}}$, while the likelihood ratio is computed using the original policy $\pi_{\theta_{\text{old}}}$. This intentional distribution mismatch introduces a temperature-dependent bias into the auxiliary update, which may be essential for entropy control. To investigate whether this bias contributes to the effectiveness of SCOPE-RL, we introduce two importance-sampling ablations, IST and ISA. In particular, IST corrects the likelihood-ratio denominator to match the sampling distribution, but simultaneously changes the clipping behavior.

ISA is therefore introduced to examine the effect of temperature-dependent reweighting while preserving the original likelihood ratio and clipping rule.

For clarity, we present the complete objectives of GRPO, SCOPE-RL, IST, and ISA below.

GRPO. The standard GRPO objective is:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{q\sim P(Q) \\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}\hat{A}_i, \text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}} \right)\hat{A}_i \right) \right].$$

SCOPE-RL. SCOPE-RL adds a temperature-adjusted positive-sample regularizer to the GRPO objective:

$$J_{\text{SCOPE}}(\theta) = J_{\text{GRPO}}(\theta) + \alpha\,\mathbb{E}_{\substack{q\sim P(Q) \\ \{o_i\}_{i=1}^{G'}\sim\pi_{\theta_{\text{old}}}^{T}}} \left[\frac{1}{G'}\sum_{i=1}^{G'}\mathbf{1}[R(q,o_i)=1]\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, 1-\epsilon, 1+\epsilon \right) \right) \right].$$

IST: Importance Sampling on the Likelihood Ratio. IST replaces the denominator of the auxiliary likelihood ratio with the temperature-adjusted sampling policy:

$$J_{\text{IST}}(\theta) = J_{\text{GRPO}}(\theta) + \alpha\,\mathbb{E}_{\substack{q\sim P(Q) \\ \{o_i\}_{i=1}^{G'}\sim\pi_{\theta_{\text{old}}}^{T}}} \left[\frac{1}{G'}\sum_{i=1}^{G'}\mathbf{1}[R(q,o_i)=1]\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}^{T}(o_{i,t}\mid q,o_{i,<t})}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}^{T}(o_{i,t}\mid q,o_{i,<t})}, 1-\epsilon, 1+\epsilon \right) \right) \right].$$

ISA: Importance Sampling on the Advantage. ISA instead applies a temperature-dependent weight to the positive-sample advantage while retaining the original SCOPE-RL likelihood ratio and clipping rule:

$$J_{\text{ISA}}(\theta) = J_{\text{GRPO}}(\theta) + \alpha\,\mathbb{E}_{\substack{q\sim P(Q) \\ \{o_i\}_{i=1}^{G'}\sim\pi_{\theta_{\text{old}}}^{T}}} \left[\frac{1}{G'}\sum_{i=1}^{G'}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\frac{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}^{T}(o_{i,t}\mid q,o_{i,<t})}\mathbf{1}[R(q,o_i)=1] \min\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, \text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, 1-\epsilon, 1+\epsilon \right) \right) \right].$$

The motivation for introducing ISA is to disentangle the effect of temperature-dependent sample weighting from changes in clipping behavior. Unlike IST, which modifies both the likelihood ratio and the clipping region, ISA leaves the original ratio and clipping operator unchanged and modifies only the token-wise advantage weight. Thus, ISA provides a complementary comparison for assessing whether the entropy-control degradation observed under IST can be attributed solely to altered clipping behavior.

As shown in Fig. 6, both IST and ISA exhibit declining policy entropy, whereas SCOPE-RL maintains a substantially higher entropy level. In particular, the entropy decline under ISA, despite retaining the original clipping rule, indicates that changes in clipping behavior alone cannot explain the degradation. These results support the conclusion that the deliberate temperature-induced weighting scheme in SCOPE-RL plays an important role in entropy regulation. By restricting this biased auxiliary update to a small fraction of training, SCOPE-RL achieves effective entropy control without dominating the original reward optimization.

Table 9: Implementation settings. Unless stated otherwise, the RL and SCOPE-RL settings apply to the three Qwen main experiments.

Setting	Value
Qwen main experiments	
Training data	DAPO-17K
Hardware	$8\times$ A800 GPUs (40 GB each)
Optimizer; learning rate	AdamW; 1×10^{-6}
Training seed	1
Training steps	340 (Qwen2.5-Math-7B); 204 (Qwen2.5-7B); 170 (Qwen3-4B)
Prompt batch size; mini-batch size	512; 128
Rollouts per prompt	8 ($512 \times 8 = 4096$ standard rollouts per batch)
Rollout temperature; maximum response length	1.0; 8192 tokens
Clipping bounds $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$	$(0.2, 0.2)$
Reward	Binary, verifiable reward
Llama-3.1-8B-Instruct extension	
Training data and duration	Math12K; four epochs (91 updates)
Training seeds	1, 2, and 3
Rollout-budget control	Standard rollouts reduced by the number of auxiliary rollouts; equal total rollout count across methods
SCOPE-RL auxiliary branch	
Target entropy $\mathcal{H}_0$	0.5
Temperature range	$[0.8, 1.2]$
Nominal coefficient α	$64/(512 \times 8) = 1/64$
Target positive trajectories	64 retained per update; use all available if fewer are found
Auxiliary candidate count	Estimated from the preceding update's pass rate
Feedback statistic	Mean token entropy over valid generated tokens in the preceding update
Evaluation	
Maximum response length; top-p	8192 tokens; 0.95
Sampling temperature	0.1 for Pass@1; 1.0 for Pass@k

C Detailed implementation

C.1 Availability of Positive Samples in the Auxiliary Branch

The auxiliary branch of SCOPE-RL relies on positive samples. In practice, we handle their availability as follows:

- **Requirement on training difficulty.** RLVR requires the training data to provide a sufficient number of positive reward signals. In practice, the dataset should not be excessively difficult, and we generally maintain an initial pass rate above approximately 30%. Otherwise, insufficient positive feedback becomes a general limitation of RLVR training rather than a limitation specific to SCOPE-RL.
- **Dynamic collection of auxiliary positive samples.** In our default setting, the standard rollout batch contains $512 \times 8 = 4096$ samples, while the auxiliary branch targets 64 positive samples. We use the pass rate from the previous training step to estimate the number of auxiliary prompts required to obtain approximately 64 positives and then perform the rollout. If more than 64 positive samples are obtained, we retain 64; if fewer than 64 are obtained, we use all available positive samples without additional sampling.
- **Robustness to insufficient positives.** SCOPE-RL does not require an exact number of positive samples at every training step. As shown in Fig. 5, even 32 auxiliary samples are sufficient to maintain effective entropy control. In our experiments, cases with fewer than 16 positive auxiliary samples are rare. Even if the entropy-control strength is temporarily weaker at an

individual step, SCOPE-RL can recover at subsequent steps because such low-positive-rate events empirically do not persist across consecutive updates.

C.2 Parameter settings

We follow the default EasyR1 setup for all experiments and run all models on 8 A800 GPUs. The full training configuration is listed in Table 9.

D Proof of Theorem 1

D.1 Preliminary

We consider the problem of fine-tuning Large Language Models using Reinforcement Learning within a mathematically rigorous framework. The LLM generation process is modeled as a finite-horizon Markov Decision Process.

Notations Let Q denote the query space, and let $P(Q)$ be the distribution over queries. Let $\mathcal{V}$ denote the discrete token space (vocabulary). Given a query $q \sim P(Q)$, the model generates a response trajectory $o = (o_1, o_2, \ldots, o_{|o|})$, where each token $o_t \in \mathcal{A}$ and the sequence length $|o|$ is bounded by a maximum length L. To ensure a unified mathematical formulation, we assume all trajectories have a fixed length L. For any response shorter than L, we append a special padding token o_{pad} to its suffix until the length reaches L. At time step t, the state s_t is defined as the concatenation of the query and the partial response generated so far, i.e., $s_t = (q, o_{<t})$. The policy $\pi_\theta(\cdot|s_t)$, parameterized by θ, maps the current state s_t to a probability distribution over $\mathcal{A}$. The probability of generating a complete trajectory o given q is given by the chain rule:

$$\pi_\theta(o|q) = \prod_{t=1}^{L} \pi_\theta(o_t|s_t).$$

Throughout this section, we adopt the same definitions and notation used in the main text. We formally define entropy at two different levels of granularity.

First, we define the token-level entropy at time step t given state s_t as the Shannon entropy of the policy's output distribution:

$$\mathcal{H}_t(\pi_\theta; s_t) \coloneqq -\sum_{o_t \in \mathcal{A}} \pi_\theta(o_t|s_t) \log \pi_\theta(o_t|s_t) = -\mathbb{E}_{o_t \sim \pi_\theta(\cdot|s_t)} \left[\log \pi_\theta(a|s_t)\right]. \tag{14}$$

Based on this, we define the cumulative expected entropy $\mathcal{H}(\pi_\theta; q)$ for a query q as the expectation of the sum of token-level entropy $\mathcal{H}_t$ over the generated trajectory:

$$\mathcal{H}(\pi_\theta; q) \coloneqq \mathbb{E}_{o \sim \pi_\theta(O|q)} \left[\sum_{t=1}^{L} \mathcal{H}_t(\pi_\theta; s_t)\right]. \tag{15}$$

Second, we define the trajectory-level entropy $\mathcal{H}'(\pi_\theta; q)$, which measures the uncertainty of the entire response sequence o:

$$\mathcal{H}'(\pi_\theta; q) \coloneqq -\sum_{o} \pi_\theta(o|q) \cdot \log \pi_\theta(o|q) = -\mathbb{E}_{o \sim \pi_\theta(O|q)} \left[\log \pi_\theta(o|q)\right]. \tag{16}$$

The following lemma establishes the equivalence between these two formulations $\mathcal{H}(\pi_\theta; q)$ and $\mathcal{H}'(\pi_\theta; q)$, providing the theoretical basis for optimizing sequence diversity via step-wise regularization.

Lemma 1. *The cumulative expected token-level entropy $\mathcal{H}(\pi_\theta; q)$ is equivalent to the trajectory-level entropy $\mathcal{H}'(\pi_\theta; q)$. That is, for any policy π_θ and query q, it holds that*

$$\mathcal{H}(\pi_\theta; q) = \sum_{t=1}^{L} \mathbb{E}_{s_t \sim \pi_\theta(\cdot|q)}[\mathcal{H}_t(\pi_\theta; s_t)] = \mathcal{H}'(\pi_\theta; q).$$

Proof. We begin by considering the cumulative expected token-level entropy. According to (14) and (15), it holds that

$$\mathcal{H}(\pi_\theta; q) = \mathbb{E}_{o\sim\pi_\theta(O|q)}\left[\sum_{t=1}^{L}\mathcal{H}_t(\pi_\theta; s_t)\right] = \sum_{t=1}^{L}\mathbb{E}_{o\sim\pi_\theta(O|q)}[\mathcal{H}_t(\pi_\theta; s_t)] = \sum_{t=1}^{L}\mathbb{E}_{s_t\sim\pi_\theta(\cdot|q)}[\mathcal{H}_t(\pi_\theta; s_t)] \tag{17}$$

Next, we turn to the analysis of the trajectory-level entropy, as defined in (16).

$$\begin{aligned}\mathcal{H}'(\pi_\theta; q) &= -\mathbb{E}_{o\sim\pi_\theta(O|q)}\left[\log\pi_\theta(o|q)\right] = -\mathbb{E}_{o\sim\pi_\theta(O|q)}\left[\sum_{t=1}^{L}\log\pi_\theta(o_t|s_t)\right] \\ &= -\sum_{t=1}^{L}\mathbb{E}_{o\sim\pi_\theta(O|q)}\left[\log\pi_\theta(o_t|s_t)\right] = -\sum_{t=1}^{L}\mathbb{E}_{s_{t+1}\sim\pi_\theta(\cdot|q)}\left[\log\pi_\theta(o_t|s_t)\right] \\ &= \sum_{t=1}^{L}\mathbb{E}_{s_t\sim\pi_\theta(\cdot|q)}[\mathcal{H}_t(\pi_\theta; s_t)].\end{aligned} \tag{18}$$

Combining (17) and (18), we prove the conclusion. □

Remark 1. *Lemma 1 justifies the practical implementation of entropy regularization in auto-regressive models. The llama establishes an equality between trajectory entropy and the expected sum of conditional token entropies under the fixed-length formulation. In practice, we monitor the mean entropy over valid generated tokens to reduce sensitivity to response length. This length-normalized statistic is an empirical control signal; when response lengths vary, it is not generally equal to trajectory entropy.*

D.2 Problem Simplified

Exploration in LLM reasoning is intuitively reflected by the diversity of the generated response trajectories. Following Lemma 1, this global diversity is mathematically equivalent to the cumulative token-level entropy $\mathcal{H}(\pi_\theta; q)$. We therefore use mean token entropy as a practical proxy for policy uncertainty during training, while the equivalence in Lemma 1 applies to expected cumulative token entropy under its stated assumptions.

In this work, we analyze the relationship between entropy dynamics and parameter updates. While the reinforcement learning objective targets sequence-level attributes, the training process physically operates on the shared parameters θ at each specific token generation step. Based on the linearity of differentiation, the gradient of the sequence entropy is fundamentally driven by the superposition of gradients from individual time steps:

$$\begin{aligned}\nabla_\theta\mathcal{H}(\pi_\theta; q) &= \nabla_\theta\mathbb{E}_{o\sim\pi_\theta(O|q)}\left[\sum_{t=1}^{L}\mathcal{H}_t(\pi_\theta; s_t)\right] \\ &= \mathbb{E}_{o\sim\pi_\theta(O|q)}\left[\sum_{t=1}^{L}\left(\nabla_\theta\mathcal{H}_t(\pi_\theta; q) + \mathcal{H}_t(\pi_\theta; q)\cdot\nabla_\theta\log\pi_\theta(o|q)\right)\right] \\ &\approx \mathbb{E}_{o\sim\pi_\theta(O|q)}\left[\sum_{t=1}^{L}\nabla_\theta\mathcal{H}_t(\pi_\theta; s_t)\right].\end{aligned} \tag{19}$$

Remark 2. *As explicitly derived in (19), the full gradient of the sequence entropy comprises two distinct components: (1) the instantaneous gradient term $\nabla_\theta\mathcal{H}_t$, which reflects how the parameter update directly alters the shape of the policy distribution at the current state; and (2) the trajectory-shift term (the positive-sample regularizer weighted by cumulative entropy), which accounts for how policy changes alter the distribution of states likely to be visited in future steps. In the subsequent theoretical analysis, we explicitly focus on the instantaneous gradient component. This theoretical simplification is justified because our primary goal is to uncover the micro-mechanism of how temperature modulation re-allocates the gradient contribution among tokens (particularly by dampening the excessive reinforcement of dominant tokens) at any given decision step, thereby locally and directly counteracting the tendency of the policy to collapse toward determinism. While trajectory shift is relevant for long-term value propagation, the modification of the local distributional shape constitutes the dominant effect regarding the prevention of entropy collapse. Therefore, isolating this component allows us to mathematically decouple and rigorously prove the direct control authority of temperature T over the optimization dynamics.*

D.3 Detailed Proof

In this section, we provide the rigorous proof for the temperature-modulated entropy control mechanism. As justified in Section D.2, we focus on a contextual bandit setting.

D.3.1 Problem Setup and Definition

Consider a generic state s, we denote the action space as $\mathcal{A}$. The policy $\pi_\theta(\cdot|s)$ is parameterized as a Softmax distribution over the action space $\mathcal{A}$:

$$\pi_\theta(a|s) = \frac{\exp(z(s,a))}{\sum_{a' \in \mathcal{A}} \exp(z(s,a'))},$$

where $z(s,a)$ is the logit for the state-action pair (s,a) under parameter θ. Furthermore, given a temperature T, we define:

$$\pi_\theta^T(a|s) = \frac{\exp(\frac{z(s,a)}{T})}{\sum_{a' \in \mathcal{A}} \exp(\frac{z(s,a')}{T})},$$

we write $z_{\theta_k}(s,a)$ simply as $z_k(s,a)$, and we write $\pi_\theta^T(a|s)$ simply as $\pi_\theta(a|s)$ when $T=1$.

The standard expected-reward objective for this contextual bandit problem is

$$J(\theta) = \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[R(s,a) \right]. \tag{20}$$

where $R(s,a)$ is the reward for token a given positive context s.

To isolate the effect of temperature-modulated sampling, at iteration k we consider the following auxiliary surrogate objective with the sampling distribution $\pi_{\theta_k}^T$ held fixed:

$$J_{\text{aux}}^{(k)}(\theta; T) = \mathbb{E}_{a \sim \pi_{\theta_k}^T(\cdot|s)} \left[R(s,a) \log \pi_\theta(a \mid s) \right]. \tag{21}$$

The corresponding local update is

$$\theta_{k+1} = \theta_k + \eta \nabla_\theta J_{\text{aux}}^{(k)}(\theta; T)\Big|_{\theta=\theta_k} = \theta_k + \eta \mathbb{E}_{a \sim \pi_{\theta_k}^T(\cdot|s)} \left[R(s,a) \nabla_\theta \log \pi_{\theta_k}(a \mid s) \right]. \tag{22}$$

where η is the learning rate.

Remark 3. *Note that, unlike the standard RL policy gradient, (22) estimates the policy gradient using samples drawn from the fixed temperature-modulated policy $\pi_{\theta_k}^T$. This mismatch introduces a deliberate bias into the gradient estimate. SCOPE-RL is designed to exploit this temperature-induced bias, and the following analysis characterizes how it reshapes the policy update to regulate policy entropy.*

Within this contextual bandit framework, we define the policy entropy for parameter θ at a generic state s a:

$$\mathcal{H}(\pi_\theta) = -\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[\log \pi_\theta(a|s)] \tag{23}$$

D.3.2 Theoretical Assumptions

To render the analysis tractable and reveal the core mechanism, we introduce the following standard assumptions.

Assumption 2 (Binary Reward)**.** *For a given initial state s, a corresponding action space $\mathcal{A}$, the reward for (s,a) is binary:*

$$R(s,a) = \begin{cases} 1 & \text{if } a \in \mathcal{A}^*; \\ 0 & \text{otherwise}, \end{cases}$$

where $\mathcal{A}^*$ denotes the set of reference actions for state s.

Assumption 3 (Orthogonal Gradients)**.** *Following standard analyses in deep learning theory (e.g., NTK regime), we assume the gradients of logits for different tokens are orthogonal, i.e.,*

$$\langle \nabla_\theta z_k(s,a), \nabla_\theta z_k(s,b) \rangle = c_{a,b} \cdot \delta_{a,b}, \quad \forall a, b \in \mathcal{A},$$

where $c_{a,b} \in \mathbb{R}$ and $\delta_{a,b} = \boldsymbol{1}[a=b]$ is the Kronecker delta. In the following, we set $c_{a,b} \equiv 1$.

Remark 4. *We acknowledge that Assumption 3 is an idealization given the shared parameters in LLMs. However, assuming gradient orthogonality is a standard practice in deep learning theory (e.g., the Neural Tangent Kernel regime [44]) to model the capacity of over-parameterized networks to adjust outputs independently. Crucially, this assumption decouples the complex interactions between tokens, allowing us to mathematically isolate the mechanistic effect of temperature modulation on the gradient weights, thereby rendering the qualitative analysis tractable.*

D.3.3 Formal statement of Theorem 1

Theorem 2. *Let the sequence of parameters $\{\theta_k\}$ be generated by (22). Define*

$$\Delta\mathcal{H}_k(T) = -\eta \sum_{a^*\in\mathcal{A}^*} \pi_{\theta_k}^T(a^* \mid s)\,\mathrm{Cov}_{\pi_{\theta_k}}\left(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a' \mid s)\right). \tag{24}$$

Then, under the first-order approximation,

$$\Delta\mathcal{H}_k(T) \approx \mathcal{H}(\pi_{\theta_{k+1}}) - \mathcal{H}(\pi_{\theta_k}). \tag{25}$$

Moreover, if

$$\Delta\mathcal{H}_k(1) < 0,$$

then

$$\left.\frac{\mathrm{d}\Delta\mathcal{H}_k(T)}{\mathrm{d}T}\right|_{T=1} > 0. \tag{26}$$

Proof. For a given parameter θ_k and a relatively small learning rate η, leveraging Taylor'expansion under first-order approximation, we have

$$\mathcal{H}(\pi_{\theta_{k+1}}) - \mathcal{H}(\pi_{\theta_k}) \approx \langle \nabla_\theta \mathcal{H}(\pi_{\theta_k}), \theta_{k+1} - \theta_k \rangle.$$

We then turn to derive what $\nabla_\theta \mathcal{H}(\pi_{\theta_k})$ is, according to the definition of $\mathcal{H}$, we have

$$\begin{aligned}
\nabla_\theta \mathcal{H}(\pi_{\theta_k}) &= -\nabla_\theta \mathbb{E}_{a\sim\pi_{\theta_k}(\cdot|s)}[\log \pi_{\theta_k}(a|s)] \\
&= -\nabla_\theta \sum_{a\in\mathcal{A}} [\pi_{\theta_k}(a|s) \cdot \log \pi_{\theta_k}(a|s)] \\
&= -\mathbb{E}_{a\sim\pi_{\theta_k}}[\nabla_\theta \log \pi_{\theta_k}(a|s) \cdot \log \pi_{\theta_k}(a|s) + \nabla_\theta \log \pi_{\theta_k}(a|s)] \\
&= -\mathbb{E}_{a\sim\pi_{\theta_k}}[\nabla_\theta \log \pi_{\theta_k}(a|s) \cdot \log \pi_{\theta_k}(a|s)].
\end{aligned}$$

the last equality follows from $\sum_{a\in\mathcal{A}} \nabla_\theta \pi_{\theta_k}(a|s) = \nabla_\theta \sum_{a\in\mathcal{A}} \pi_{\theta_k}(a|s) = \nabla_\theta 1 = 0$.

Then we have

$$\begin{aligned}
\mathcal{H}(\pi_{\theta_{k+1}}) - \mathcal{H}(\pi_{\theta_k}) &\approx -\left\langle \mathbb{E}_{a\sim\pi_{\theta_k}(\cdot|s)}[\nabla_\theta \log \pi_{\theta_k}(a|s) \cdot \log \pi_{\theta_k}(a|s)], \theta_{k+1} - \theta_k \right\rangle \\
&= -\mathbb{E}_{a\sim\pi_{\theta_k}(\cdot|s)}\left[\log \pi_{\theta_k}(a|s) \cdot \langle \nabla_\theta \log \pi_{\theta_k}(a|s), \theta_{k+1} - \theta_k \rangle\right] \\
&\approx -\mathbb{E}_{a\sim\pi_{\theta_k}(\cdot|s)}\left[\log \pi_{\theta_k}(a|s) \cdot (\log \pi_{\theta_{k+1}}(a|s) - \log \pi_{\theta_k}(a|s))\right] \\
&\approx -\mathbb{E}_{a\sim\pi_{\theta_k}(\cdot|s)}\left[\log \pi_{\theta_k}(a|s) \cdot \sum_{a'\in\mathcal{A}} \frac{\mathbf{d}\log \pi_{\theta_k}(a|s)}{\mathbf{d}z_k(s,a')} \cdot \Delta z_k(s,a')\right].
\end{aligned}$$

We now evaluate $\frac{\mathbf{d}\log \pi_{\theta_k}(a|s)}{\mathbf{d}z_k(s,a')}$ and $\Delta z_k(s,a')$ in turn.

$$\frac{\mathbf{d}\log \pi_{\theta_k}(a|s)}{\mathbf{d}z_k(s,a')} = \delta_{a,a'} - \pi_{\theta_k}(a'|s);$$

According to the parameter update rule in (22), it holds that

$$\begin{aligned}
\Delta z_k(s,a') &\approx \left\langle \frac{\mathbf{d}z_k(s,a')}{\mathbf{d}\theta}, \theta_{k+1} - \theta_k \right\rangle \\
&= \left\langle \nabla_\theta z_k(s,a'), \eta \cdot \mathbb{E}_{a\sim\pi_{\theta_k}^T(\cdot|s)}[\mathbf{1}[a\in\mathcal{A}^*] \cdot \nabla_\theta \log \pi_{\theta_k}(a|s)] \right\rangle \\
&= \eta \cdot \sum_{a^*\in\mathcal{A}^*} \pi_{\theta_k}^T(a^*|s) \cdot \left\langle \nabla_\theta z_k(s,a'), \nabla_\theta z_k(s,a^*) - \mathbb{E}_{a\sim\pi_{\theta_k}}[\nabla_\theta z_k(s,a)] \right\rangle \\
&= \eta \cdot \sum_{a^*\in\mathcal{A}^*} \left[\pi_{\theta_k}^T(a^*|s) \cdot (\delta_{a',a^*} - \pi_{\theta_k}(a'|s))\right],
\end{aligned} \tag{27}$$

where the first equality in (27) follows from

$$\begin{aligned}\nabla_\theta \log \pi_{\theta_k}(a|s) &= \nabla_\theta \left[z_k(s,a) - \log \sum_{a' \in \mathcal{A}} \exp(z_k(s,a')) \right] \\ &= \nabla_\theta z_k(s,a) - \frac{1}{\sum_{a' \in \mathcal{A}} \exp(z_k(s,a'))} \left(\sum_{a' \in \mathcal{A}} \exp(z_k(s,a')) \cdot \nabla_\theta z_k(s,a') \right) \\ &= \nabla_\theta z_k(s,a) - \mathbb{E}_{a \sim \pi_{\theta_k}} [\nabla_\theta z_k(s,a)],\end{aligned}$$

and the last equality in (27) follows from Assumption 3. Then, it holds that

$$\begin{aligned}\mathcal{H}(\pi_{\theta_{k+1}}) - \mathcal{H}(\pi_{\theta_k}) &\approx - \sum_{a' \in \mathcal{A}} \Delta z_k(s,a') \cdot \mathbb{E}_{a \sim \pi_{\theta_k}(\cdot|s)} \left[\log \pi_{\theta_k}(a|s) \cdot \frac{\mathbf{d} \log \pi_{\theta_k}(a|s)}{\mathbf{d} z_k(s,a')} \right] \\ &= - \sum_{a' \in \mathcal{A}} \Delta z_k(s,a') \cdot \mathbb{E}_{a \sim \pi_{\theta_k}(\cdot|s)} \left[\log \pi_{\theta_k}(a|s) \cdot (\delta_{a,a'} - \pi_{\theta_k}(a'|s)) \right] \\ &= - \sum_{a' \in \mathcal{A}} \Delta z_k(s,a') \cdot \left[\pi_{\theta_k}(a'|s) \cdot \log \pi_{\theta_k}(a'|s) - \pi_{\theta_k}(a'|s) \cdot \mathbb{E}_{a \sim \pi_{\theta_k}(\cdot|s)} [\log \pi_{\theta_k}(a|s)] \right] \\ &= - \left(\mathbb{E}_{a \sim \pi_{\theta_k}(\cdot|s)} [\Delta z_k(s,a) \cdot \log \pi_{\theta_k}(a|s)] - \mathbb{E}_{a \sim \pi_{\theta_k}(\cdot|s)} [\Delta z_k(s,a)] \cdot \mathbb{E}_{a \sim \pi_{\theta_k}(\cdot|s)} [\log \pi_{\theta_k}(a|s)] \right) \\ &= -\mathrm{Cov}_{\pi_{\theta_k}} (\Delta z_k(s,a), \log \pi_{\theta_k}(a|s)) \\ &= -\mathrm{Cov}_{\pi_{\theta_k}} (\Delta z_k(s,a), z_k(a|s)),\end{aligned} \tag{28}$$

take (27) into (28), we have

$$\mathcal{H}(\pi_{\theta_{k+1}}) - \mathcal{H}(\pi_{\theta_k}) \approx -\eta \cdot \sum_{a^* \in \mathcal{A}^*} \pi_{\theta_k}^T(a^*|s) \cdot \mathrm{Cov}_{\pi_{\theta_k}} (z_k(s,a), \delta_{a,a^*} - \pi_{\theta_k}(a|s)) \equiv \Delta \mathcal{H}_k(T).$$

We now proceed to the proof of the second part of the lemma. We first partition the set $\mathcal{A}^*$ into three subsets: $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$, where

$$\mathcal{A}_1 = \left\{ a | a \in \mathcal{A}^*, z_k(s,a) < \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s)} [z(s,a')] \right\}; \tag{29}$$

$$\mathcal{A}_2 = \left\{ a | a \in \mathcal{A}^*, z_k(s,a) \geq \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s)} [z(s,a')], \mathrm{Cov}_{a' \sim \pi_{\theta_k}} (z_k(s,a'), \delta_{a',a}) < C_{\pi_k} \right\}; \tag{30}$$

$$\mathcal{A}_3 = \left\{ a | a \in \mathcal{A}^*, z_k(s,a) \geq \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s)} [z(s,a')], \mathrm{Cov}_{a' \sim \pi_{\theta_k}} (z_k(s,a'), \delta_{a',a}) \geq C_{\pi_k} \right\}; \tag{31}$$

where $C_{\pi_k} = \mathrm{Cov}_{a' \sim \pi_{\theta_k}} (z_k(s,a'), \pi_{\theta_k}(a'|s))$. According to the condition $\Delta \mathcal{H}_k(1) < 0$, we have

$$\begin{aligned}0 &< \sum_{a^* \in \mathcal{A}^*} \pi_{\theta_k}(a^*|s) \cdot \mathrm{Cov}_{\pi_{\theta_k}} (z_k(s,a), \delta_{a,a^*} - \pi_{\theta_k}(a|s)) \\ &= \sum_{i=1}^{3} \sum_{a^* \in \mathcal{A}_i} \pi_{\theta_k}(a^*|s) \cdot \mathrm{Cov}_{\pi_{\theta_k}} (z_k(s,a), \delta_{a,a^*} - \pi_{\theta_k}(a|s)) \\ &\leq \sum_{a^* \in \mathcal{A}_2 \bigcup \mathcal{A}_3} \pi_{\theta_k}(a^*|s) \cdot \mathrm{Cov}_{\pi_{\theta_k}} (z_k(s,a), \delta_{a,a^*} - \pi_{\theta_k}(a|s)).\end{aligned} \tag{32}$$

We now turn to the analysis of $\frac{\mathbf{d} \Delta \mathcal{H}_k(T)}{\mathbf{d} T}$. According to the definition of π_θ, it holds that

$$\frac{\mathbf{d} \pi_{\theta_k}^T(a|s)}{\mathbf{d} T} = -T^{-2} \cdot \pi_{\theta_k}^T(a|s) \cdot \left[z_k(s,a) - \mathbb{E}_{\pi_{\theta_k}^T} [z_k(s,a')] \right].$$

We evaluate the temperature derivative at the reference temperature $T = 1$. At $T = 1$, we have

$$\frac{\mathbf{d} \pi_{\theta_k}^T(a|s)}{\mathbf{d} T}|_{T=1} = -\pi_{\theta_k}(a|s) \cdot \left[z_k(s,a) - \mathbb{E}_{\pi_{\theta_k}} [z_k(s,a')] \right] = -\mathrm{Cov}_{\pi_{\theta_k}} (z_k(s,a'), \delta_{a,a'}).$$

Thus it holds that

$$\begin{aligned}\frac{\mathbf{d}\Delta\mathcal{H}(T)}{\mathbf{d}T}|_{T=1} &= -\eta\cdot\sum_{a^*\in\mathcal{A}^*}\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))\cdot\frac{\mathbf{d}\pi^T_{\theta_k}(a^*|s)}{\mathbf{d}T}\\ &= -\eta\cdot\sum_{a^*\in\mathcal{A}^*}\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))\cdot(-\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}))\\ &= \eta\cdot\left[\sum_{a^*\in\mathcal{A}_1\bigcup\mathcal{A}_2\bigcup\mathcal{A}_3}\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))\cdot\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*})\right].\end{aligned}$$

We now proceed with a case-by-case analysis based on different a^*.

For $a^*\in\mathcal{A}_1$, we have $z_k(s,a)<\mathbb{E}_{a'\sim\pi_{\theta_k}(\cdot|s)}[z(s,a')]$, and then $\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*})<0$. According to Lemma 2, it holds that $C_{\pi_k}\geq 0$, then we have

$$\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))=\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*})-C_{\pi_k}<0, \tag{33}$$

therefore

$$\sum_{a^*\in\mathcal{A}_1}\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))\cdot\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*})\geq 0, \tag{34}$$

We first consider the boundary case that $\mathcal{A}_2=\varnothing$, by combining (32) and (33), we have

$$\sum_{a^*\in\mathcal{A}_3}\pi_{\theta_k}(a^*\mid s)\,\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a),\delta_{a,a^*}-\pi_{\theta_k}(a\mid s))>0. \tag{35}$$

Therefore, there exists at least one $a^*\in\mathcal{A}_3$ such that

$$\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a),\delta_{a,a^*}-\pi_{\theta_k}(a\mid s))>0. \tag{36}$$

For the particular $a^*\in\mathcal{A}_3$ whose existence is guaranteed by (35), we further have

$$\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*})=\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'\mid s))+C_{\pi_k}>C_{\pi_k}\geq 0. \tag{37}$$

Therefore, the product corresponding to this a^* is strictly positive, while all other terms in $\mathcal{A}_3$ are non-negative. Hence,

$$\sum_{a^*\in\mathcal{A}_3}\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'\mid s))\cdot\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*})>0. \tag{38}$$

Hence the contribution from $\mathcal{A}_3$ to $\frac{d\Delta\mathcal{H}_k(T)}{dT}\Big|_{T=1}$ is strictly positive, while the contribution from $\mathcal{A}_1$ is non-negative, which is calculate in (34). Therefore,

$$\frac{d\Delta\mathcal{H}_k(T)}{dT}\Big|_{T=1}>0. \tag{39}$$

In the case that $\mathcal{A}_2\neq\varnothing$. For $a^*\in\mathcal{A}_2$, we have $z_k(s,a^*)\geq\mathbb{E}_{a'\sim\pi_{\theta_k}(\cdot|s)}[z(s,a')]$ and $\mathrm{Cov}_{a'\sim\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*})<C_{\pi_k}$. Let $u=\max_{a^*\in\mathcal{A}_2}\{z_k(s,a^*)\}$, therefore it holds that

$$\begin{aligned}&\sum_{a^*\in\mathcal{A}_2}\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))\cdot\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*})\\ =&\sum_{a^*\in\mathcal{A}_2}\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))\cdot\pi_{\theta_k}(a^*|s)\cdot\left[z_k(s,a^*)-\mathbb{E}_{\pi_{\theta_k}}[z_k(s,a')]\right]\\ =&\sum_{a^*\in\mathcal{A}_2}\underbrace{\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))}_{<0}\cdot\pi_{\theta_k}(a^*|s)\cdot\left[\underbrace{z_k(s,a^*)-u}_{\leq 0}+u-\mathbb{E}_{\pi_{\theta_k}}[z_k(s,a')]\right]\\ \geq&\sum_{a^*\in\mathcal{A}_2}\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'),\delta_{a',a^*}-\pi_{\theta_k}(a'|s))\cdot\pi_{\theta_k}(a^*|s)\cdot\left[u-\mathbb{E}_{\pi_{\theta_k}}[z_k(s,a')]\right].\end{aligned}$$

For $a^* \in \mathcal{A}_3$, we have

$$\begin{aligned}
&\sum_{a^*\in\mathcal{A}_3} \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s)) \cdot \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*}) \\
=&\sum_{a^*\in\mathcal{A}_3} \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s)) \cdot \pi_{\theta_k}(a^*|s) \cdot [z_k(s,a^*) - \mathbb{E}_{\pi_{\theta_k}}[z_k(s,a')]] \\
=&\sum_{a^*\in\mathcal{A}_3} \underbrace{\mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s))}_{\geq 0} \cdot \pi_{\theta_k}(a^*|s) \cdot [\underbrace{z_k(s,a^*) - u}_{>0} + u - \mathbb{E}_{\pi_{\theta_k}}[z_k(s,a')]] \\
>&\sum_{a^*\in\mathcal{A}_3} \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s)) \cdot \pi_{\theta_k}(a^*|s) \cdot [u - \mathbb{E}_{\pi_{\theta_k}}[z_k(s,a')]],
\end{aligned} \tag{40}$$

where the inequality holds according to Lemma 3, which formulate that for any $a \in \mathcal{A}_2$ and $b \in \mathcal{A}_3$, it holds that

$$z_k(s,a) < z_k(s,b). \tag{41}$$

We now combining the analyses from the three subsets $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3$. The total derivative of the entropy change with respect to temperature is the sum of contributions from all $a^* \in \mathcal{A}^*$. Notice that for the dominant subset $\mathcal{A}_2 \cup \mathcal{A}_3$ (where $z(s,a^*)$ is large), the term $[u - \mathbb{E}[z]]$ is positive according to the definition of u and $\mathcal{A}_2 \cup \mathcal{A}_3$. The covariance terms are positive by construction in these subsets. For the subset $\mathcal{A}_1$ (tail actions), the contribution is also non-negative as proven above.

Rigorously, by combining the results from the three cases discussed above, we arrive at the following conclusion:

$$\begin{aligned}
\frac{\mathbf{d}\Delta\mathcal{H}_k(T)}{\mathbf{d}T}|_{T=1} =& \eta \cdot \left[\sum_{a^*\in\mathcal{A}_1\cup\mathcal{A}_2\cup\mathcal{A}_3} \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s)) \cdot \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*}) \right] \\
\geq& \eta \cdot \left[\sum_{a^*\in\mathcal{A}_2\cup\mathcal{A}_3} \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s)) \cdot \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*}) \right] \\
>& \eta \cdot \left[\sum_{a^*\in\mathcal{A}_2\cup\mathcal{A}_3} \pi_{\theta_k}(a^*|s) \cdot \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s)) \cdot \left(u - \mathbb{E}_{\pi_{\theta_k}}[z_k(s,a')]\right) \right] \\
=& \eta \cdot \left(u - \mathbb{E}_{\pi_{\theta_k}}[z_k(s,a')]\right) \cdot \left[\sum_{a^*\in\mathcal{A}_2\cup\mathcal{A}_3} \pi_{\theta_k}(a^*|s) \cdot \mathrm{Cov}_{\pi_{\theta_k}}(z_k(s,a'), \delta_{a',a^*} - \pi_{\theta_k}(a'|s)) \right] \\
\geq& 0.
\end{aligned}$$

This confirms that in the face of collapse, increasing temperature systematically introduces a relative entropy-increasing component to the update. □

D.3.4 Lemmas used in the proof of Theorem 2

We provide below the proofs of two lemmas used in the proof of Theorem 2.

Lemma 2. *Let*

$$C_{\pi_k} = \mathrm{Cov}_{a'\sim\pi_{\theta_k}(\cdot|s)}\left(z_k(s,a'), \pi_{\theta_k}(a' \mid s)\right). \tag{42}$$

Then $C_{\pi_k} \geq 0$. Moreover, if $\pi_{\theta_k}(\cdot \mid s)$ is non-uniform, then $C_{\pi_k} > 0$.

Proof. By the pairwise representation of covariance,

$$2C_{\pi_k} = \sum_{a,b\in\mathcal{A}} \pi_{\theta_k}(a \mid s)\pi_{\theta_k}(b \mid s)\left(z_k(s,a) - z_k(s,b)\right) \cdot \left(\pi_{\theta_k}(a \mid s) - \pi_{\theta_k}(b \mid s)\right). \tag{43}$$

Since $\pi_{\theta_k}(\cdot \mid s)$ is obtained by applying softmax to the logits,

$$z_k(s,a) > z_k(s,b) \quad \Longleftrightarrow \quad \pi_{\theta_k}(a \mid s) > \pi_{\theta_k}(b \mid s). \tag{44}$$

Therefore, every summand above is non-negative, and hence $C_{\pi_k} \geq 0$.

If the policy is non-uniform, there exists at least one pair (a, b) with $z_k(s, a) \neq z_k(s, b)$, for which the corresponding summand is strictly positive. Thus $C_{\pi_k} > 0$. □

Lemma 3. *For any $a \in \mathcal{A}_2$ and $b \in \mathcal{A}_3$,*

$$z_k(s, a) < z_k(s, b). \tag{45}$$

Proof. Define

$$g_k(a) \triangleq \mathrm{Cov}_{a' \sim \pi_{\theta_k}(\cdot|s)} \left(z_k(s, a'), \delta_{a',a} \right). \tag{46}$$

Direct calculation gives

$$g_k(a) = \pi_{\theta_k}(a \mid s) \left(z_k(s, a) - \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s)}[z_k(s, a')] \right). \tag{47}$$

Let

$$\mu_k \triangleq \mathbb{E}_{a' \sim \pi_{\theta_k}(\cdot|s)}[z_k(s, a')]. \tag{48}$$

By the definitions of $\mathcal{A}_2$ and $\mathcal{A}_3$,

$$z_k(s, a) \geq \mu_k, \qquad g_k(a) < C_{\pi_k} \leq g_k(b). \tag{49}$$

By Lemma 2, $C_{\pi_k} \geq 0$, so $g_k(b) \geq 0$, which implies $z_k(s, b) \geq \mu_k$.

Suppose, for contradiction, that $z_k(s, a) \geq z_k(s, b)$. Softmax monotonicity gives

$$\pi_{\theta_k}(a \mid s) \geq \pi_{\theta_k}(b \mid s). \tag{50}$$

Together with

$$z_k(s, a) - \mu_k \geq z_k(s, b) - \mu_k \geq 0, \tag{51}$$

we obtain

$$\begin{aligned} g_k(a) &= \pi_{\theta_k}(a \mid s) \left(z_k(s, a) - \mu_k \right) \\ &\geq \pi_{\theta_k}(b \mid s) \left(z_k(s, b) - \mu_k \right) = g_k(b), \end{aligned} \tag{52}$$

contradicting $g_k(a) < C_{\pi_k} \leq g_k(b)$. Therefore,

$$z_k(s, a) < z_k(s, b). \tag{53}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have mentioned in the abstract.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed it in Section Limitation.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The detailed assumptions and proofs of the theoretical results are presented in the appendix, while here we provide a brief introduction in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We use open-sourced models, which are easy to reproduce.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide codes in our supplemental material. The data and models we used are all open-sourced.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide in Detailed implementation in appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Experiments on LLMs are too expensive to run for many times.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide in our appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We follow NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate

Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We follow the license and terms of use in our experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.

- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.