Too Correct to Learn: Reinforcement Learning on Saturated Reasoning Data

Zhenwen Liang[1,†], **Yujun Zhou**[1,2,†*], **Sidi Lu**[1], **Xiangliang Zhang**[2], **Haitao Mi**[1], **Dong Yu**[1]

[1]Tencent AI Lab, [2]University of Notre Dame,

[†] Equal contribution

Correspondence to: zhenwzliang@global.tencent.com

Abstract

Reinforcement Learning (RL) enhances LLM reasoning, yet a paradox emerges as models scale: strong base models saturate standard benchmarks (e.g., MATH), yielding correct but homogeneous solutions. In such environments, the lack of failure cases causes the advantage signal in group-relative algorithms (e.g., GRPO) to vanish, driving policies into mode collapse. To address this, we propose Constrained Uniform Top-K Sampling (CUTS), a parameter-free decoding strategy enforcing structure-preserving exploration. Unlike standard sampling that follows model biases, CUTS flattens the local optimization landscape by sampling uniformly from constrained high-confidence candidates. We integrate this into Mixed-CUTS, a training framework synergizing exploitative and exploratory rollouts to amplify intra-group advantage variance. Experiments on Qwen3 models demonstrate that our approach prevents policy degeneration and significantly boosts out-of-domain generalization. Notably, Mixed-CUTS improves Pass@1 accuracy on the challenging AIME25 benchmark by up to 15.1% over standard GRPO, validating that maintaining diversity within the high-probability region of the model distribution is critical for rigorous reasoning.

1 Introduction

RL is central to aligning Large Language Models (LLMs) with complex reasoning tasks (Jaech et al., 2024; Guo et al., 2025; Yang et al., 2025). Leveraging outcome-based supervision, algorithms like Group Relative Policy Optimization (GRPO) (Shao et al., 2024) transform models from pattern matchers into rigorous reasoners (Yu et al., 2025a; Zheng et al., 2025; Huang et al., 2025; Liang et al., 2025a; Liu et al., 2025b; Zhao et al., 2025).

However, RL's efficacy increasingly hinges on data difficulty. High-quality reasoning datasets are scarce and rapidly absorbed into community training pipelines, rendering benchmarks *saturated*—where strong base models already solve most instances (Liu et al., 2025a; Yang et al., 2025). This growing prevalence of **saturated reasoning data** fundamentally alters the learning dynamics of RL for LLMs.

For strong base models (e.g., Qwen3), standard datasets like MATH have saturated, yielding high baseline success rates (Yang et al., 2025). This poses a critical challenge for group-relative learning: when a model generates homogeneous correct solutions, intra-group reward variance collapses toward zero. Lacking failure cases or contrast, the relative advantage signal vanishes (Zhu et al., 2025). Consequently, the model succumbs to **saturation-induced mode collapse**—not because it is incorrect, but because it is *too correct to learn*. The policy becomes trapped in local optima of "easy successes," ceasing to explore generalizable strategies (Zhou et al., 2025). Standard entropy regularization fails here by indiscriminately penalizing confidence, disrupting coherent reasoning rather than restoring learning signals (Cui et al., 2025).

This reveals a structural limitation: on saturated data, correctness alone provides insufficient training signals. To address this, we argue that effective RL requires explicitly reintroducing diversity to reignite the advantage signal. Rather than altering objectives, we focus on the decoding process as a controllable intervention point.

We introduce **Constrained Uniform Top-K Sampling (CUTS)**, an inference-time operator designed to break the "rich-get-richer" dynamics of standard sampling. Instead of adhering to skewed distributions that favor dominant paths, CUTS flattens the local landscape by sampling uniformly from a constrained set of high-confidence (Top-K) candidates. This decouples generation probability from historical preference, compelling the model to explore semantically valid but underestimated

[*]Work done during Yujun's Internship at Tencent AI Lab.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 205–215

tokens. By restricting uniformity to this confidence-filtered window, CUTS ensures structural coherence while enabling controlled exploration.

We integrate this operator into **Mixed-CUTS**, a framework leveraging a dual-stream rollout strategy to amplify GRPO's intra-group variance. For each prompt, we generate a mixture of *exploitative* (standard sampling) and *exploratory* (CUTS) trajectories. This hybrid design anchors the baseline while injecting necessary diversity. Crucially, even when all solutions are correct, the structural contrast between standard and CUTS-induced paths restores informative gradient signals, preventing convergence stagnation on saturated benchmarks.

Contributions. (1) We diagnose and formalize saturation-induced collapse: a failure mode in group-relative RL where high baseline correctness on easy datasets causes the advantage signal to vanish. (2) We propose Mixed Constrained Uniform Top-K Sampling (Mixed-CUTS), a parameter-free decoding operator that enforces structure-preserving exploration to counteract this stagnation. (3) Empirically, we demonstrate significant generalization gains on various benchmarks, validating that maintaining diversity is essential when correctness alone provides insufficient signal.

2 Method

2.1 Preliminaries

We build our training framework upon GRPO (Shao et al., 2024). Unlike PPO (Schulman et al., 2017) that requires a parametric value network, GRPO eliminates the critic to reduce memory overhead, instead using group statistics as a baseline. Formally, given query $\mathbf{q}$, the reference policy $\pi_{\theta_{\text{old}}}$ samples G outputs $\{\mathbf{o}_1, \ldots, \mathbf{o}_G\}$, yielding rewards $\{r_1, \ldots, r_G\}$. Advantages are computed by standardizing rewards within the group:

$$\hat{A}_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G) + \epsilon} \tag{1}$$

Crucially, this *trajectory-level* advantage is applied uniformly to every token t, i.e., $\hat{A}_{i,t} = \hat{A}_i$. The policy θ is then updated by maximizing a clipped surrogate objective that ensures stability within a trust region:

$$\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\left\{\frac{\pi_\theta(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}\hat{A}_{i,t},\right.$$
$$\left.\text{clip}\left(\frac{\pi_\theta(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \mathbf{q}, \mathbf{o}_{i,<t})}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}\right)\hat{A}_{i,t}\right\} \tag{2}$$

The Vanishing-Advantage Problem. A critical limitation of Eq. 1 arises on *saturated* datasets. If the model succeeds on all G trajectories (i.e., $r_i = 1, \forall i$), the standard deviation $\text{std}(r)$ becomes zero, causing the standardized advantage $\hat{A}_i$ to vanish or depend solely on the stabilizer ϵ. Consequently, optimization stalls despite high accuracy, as the lack of contrast eliminates the learning signal. This necessitates a mechanism to explicitly guarantee non-zero intra-group variance.

2.2 Constrained Uniform Top-K Sampling (CUTS)

Standard autoregressive decoding samples x_t from $P_\theta(x_t \mid \mathbf{q}, \mathbf{x}_{<t})$. On saturated data, however, this proportional nature induces **mode collapse**: distributions become excessively peaked around dominant paths, suppressing valid alternatives. To counteract this, we propose **Constrained Uniform Top-K Sampling (CUTS)**, a parameter-free operator that constructs a locally flattened proposal distribution $Q(x_t \mid \mathbf{q}, \mathbf{x}_{<t})$ via a three-stage process: *Select, Filter, and Equalize*. CUTS introduces no additional trainable parameters and operates purely at inference time, with a small set of decoding hyperparameters.

Selection and Filtering. At step t, we extract the top-K tokens $\mathcal{V}_{\text{top-}K}$. To preclude incoherent tail generations, we apply a probability threshold δ to define the valid candidate set:

$$\mathcal{S}_t = \left\{v \in \mathcal{V}_{\text{top-}K} \mid P_\theta(v \mid \mathbf{q}, \mathbf{x}_{<t}) \geq \delta\right\}. \tag{3}$$

This constraint restricts exploration to a high-confidence token neighborhood, serving as a proxy for local plausibility. We handle edge cases explicitly: if $\mathcal{S}_t = \varnothing$, we fallback to $\mathcal{V}_{\text{top-}K}$; if $|\mathcal{S}_t| = 1$, the operator degenerates to a deterministic choice.

Uniform Equalization. To decouple generation probability from the model's bias within $\mathcal{S}_t$, we define the proposal distribution as a uniform prior:

$$Q(x_t = v \mid \mathbf{q}, \mathbf{x}_{<t}) = \begin{cases} \frac{1}{|\mathcal{S}_t|} & \text{if } v \in \mathcal{S}_t, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

This equalization enforces local "width-first" exploration, enabling the traversal of plausible reasoning paths that are underrepresented in the standard distribution.

Prefix Protection. Given the sensitivity of reasoning tasks to early decisions, we employ a warm-up mechanism to preserve initial stability. We apply CUTS only after a prefix of T_{warm} tokens; prior to this, standard sampling is used to establish a coherent problem setup.

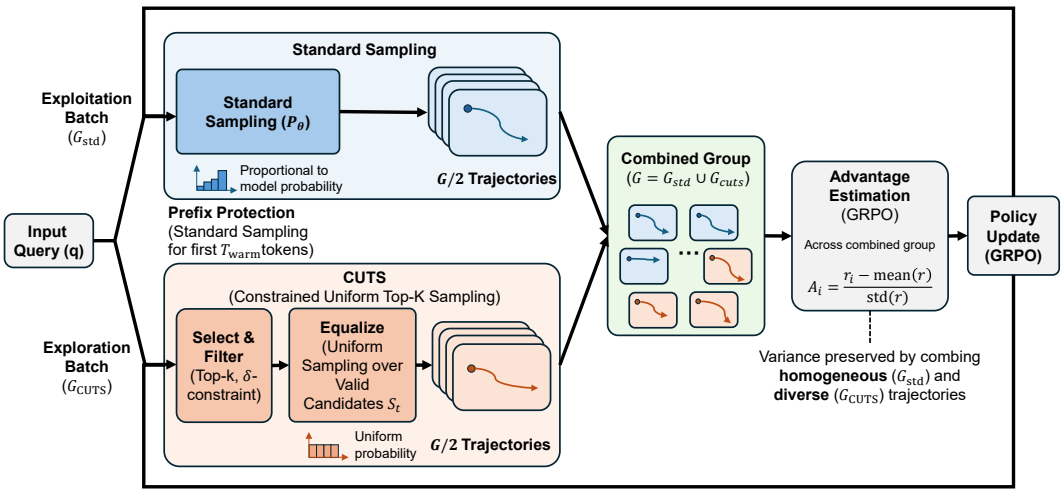

Figure 1: **The Mixed-CUTS Framework.** The framework combines exploitative rollouts ($\mathcal{G}_{\text{std}}$) and exploratory rollouts ($\mathcal{G}_{\text{CUTS}}$) to preserve advantage variance under saturated training conditions. The CUTS operator enforces uniform sampling within a constrained Top-K candidate set, decoupling generation from model bias.

2.3 The Mixed-CUTS Training Framework

To balance exploration with policy stability, we introduce the **Mixed-CUTS** framework within the GRPO paradigm in Figure 1. For each query $\mathbf{q}$, we generate a hybrid group of G responses, partitioned into two subsets with equal size:

- **Exploitation Batch ($\mathcal{G}_{\text{std}}$):** Trajectories generated via standard sampling, which anchor the baseline to the policy's current bias.
- **Exploration Batch ($\mathcal{G}_{\text{CUTS}}$):** Trajectories generated via CUTS, which inject diversity by uncovering plausible but under-explored paths.

Advantages are computed over the combined group $\mathcal{G} = \mathcal{G}_{\text{std}} \cup \mathcal{G}_{\text{CUTS}}$. In saturated regimes where $\mathcal{G}_{\text{std}}$ collapses to uniform high rewards, $\mathcal{G}_{\text{CUTS}}$ introduces necessary outcome variability. This explicitly restores intra-group variance, generating informative relative signals for optimization.

On-policy vs. Behavior Policy. Although Mixed-CUTS induces a mixed behavior policy μ, we retain the standard clipped objective with $\pi_{\theta_{\text{old}}}$, and the resulting off-policy bias is tightly bounded by three compounding restrictions. *First*, CUTS redistributes probability mass only within the Top-K candidate set, so any token emitted by μ is already assigned non-trivial probability under the model's own distribution. *Second*, the minimum-probability threshold δ prunes low-confidence tail tokens, keeping the proposal distribution inside a local trust region of semantically plausible continuations. *Third*, the PPO clipping on the importance ratio $\pi_\theta/\pi_{\theta_{\text{old}}}$ further limits the per-step policy update, so even when a CUTS token has a low probability under $\pi_{\theta_{\text{old}}}$, the gradient contribution is clipped to a bounded range. Together these constraints ensure that Mixed-CUTS injects exploratory variance without severe divergence from the current policy, which is consistent with the empirically stable training curves observed across all of our runs.

Why Mixed-CUTS Restores the Advantage Signal. We now formalize why mixing an exploratory sub-group with the standard sub-group is guaranteed to keep the intra-group variance away from zero in saturated regimes. Consider a single prompt with a combined group $\mathcal{G}_{\text{mixed}} = \mathcal{G}_{\text{std}} \cup \mathcal{G}_{\text{CUTS}}$ of size G, split into two equal sub-groups of size $G/2$. Let $(\mu_{\text{std}}, \sigma^2_{\text{std}})$ and $(\mu_{\text{CUTS}}, \sigma^2_{\text{CUTS}})$ denote the sample mean and variance of the rewards within each sub-group. By the law of total variance applied to the combined group,

$$\sigma^2_{\text{mixed}} = \tfrac{1}{2}(\sigma^2_{\text{std}} + \sigma^2_{\text{CUTS}}) + \tfrac{1}{4}(\mu_{\text{std}} - \mu_{\text{CUTS}})^2. \quad (5)$$

The first "within-group" term captures the sampling noise inside each sub-group; the second "between-group" term is a non-negative penalty activated whenever the two sub-groups have different expected rewards. In a *saturated* regime, standard GRPO corresponds to $\mathcal{G} = \mathcal{G}_{\text{std}}$ with $\sigma^2_{\text{std}} \to 0$, so the advantage in Eq. 1 collapses. Because CUTS equalizes probabilities over the constrained Top-K subset and therefore deviates from the greedy mode of $\pi_{\theta_{\text{old}}}$, it changes the expected per-prompt reward of the exploratory sub-group, i.e. $\mu_{\text{CUTS}} \neq \mu_{\text{std}}$: on "too easy" prompts ($\mu_{\text{std}} \to 1$) CUTS occasionally stumbles onto suboptimal branches, lowering μ_{CUTS}; on "too hard" prompts ($\mu_{\text{std}} \to 0$) CUTS occasionally hits a correct alternative branch

the greedy policy systematically misses, raising μ_{CUTS}. Substituting either extreme into Eq. 5 gives $\sigma^2_{\text{mixed}} \gtrsim \frac{1}{2}\sigma^2_{\text{CUTS}} + \frac{1}{4}(\mu_{\text{std}} - \mu_{\text{CUTS}})^2 > 0$, so the intra-group variance is structurally prevented from collapsing as long as the exploratory sub-group behaves differently from the exploitative one. A complete case-by-case derivation of the two saturated extremes is provided in Appendix D.

3 Experiments

3.1 Experimental Setup

We train on the **MATH Training Set** (Hendrycks et al., 2021) on Qwen3-1.7B and 4B (non-thinking mode). Details are in Appendix B.

3.2 Main Results

Table 1 compares MIXED-CUTS with GRPO and "Thinking Mode" baselines. Results highlight four key insights:

Surpassing Intrinsic "Thinking" Capabilities. MIXED-CUTS enables the 1.7B model to outperform its "Thinking Mode" (28.1% vs 24.9% on AIME25) without extended inference overhead. This suggests successful distillation of System-2-like reasoning into a standard, efficient policy.

Breaking the Saturation Trap. MIXED-CUTS dominates out-of-domain. On Qwen3-4B, it beats GRPO by **+15.1%** (AIME25) and **+13.5%** (AIME24). This validates our saturation hypothesis: while GRPO collapses on easy data (MATH) due to vanishing gradients, our structured exploration sustains the advantage signal, driving robust generalization.

Robustness Over Randomness. Pass@1 gains significantly outweigh Pass@16 gains (+4.4% vs +0.7% on 4B). This proves our method does not merely rely on random coverage but fundamentally shifts probability mass toward correct paths, improving policy reliability.

Scalability with Model Size. Benefits amplify with scale (AIME25 gain: +5.3% on 1.7B $\rightarrow$ +15.1% on 4B). Larger models benefit more from CUTS's "width-first" exploration, which unlocks latent reasoning branches that standard greedy sampling prematurely prunes.

3.3 Generalization beyond Mathematics

The gains of MIXED-CUTS are not limited to mathematical benchmarks. We evaluate our *MATH-trained* Qwen3-4B checkpoints, without any further task-specific fine-tuning, on two comprehensive general-reasoning benchmarks outside the mathematical domain: **MMLU-Pro**, a multi-domain knowledge benchmark covering biology, law, literature, and more, and **SuperGPQA**, a graduate-level multi-discipline QA benchmark. As shown in Table 2, MIXED-CUTS consistently outperforms the standard GRPO baseline on both non-mathematical distributions (**+1.06%** on MMLU-Pro and **+1.25%** on SuperGPQA), despite the RL optimization being restricted to mathematical data. By preventing mode collapse on one axis (math), Mixed-CUTS enhances the model's broader structural exploration capabilities, transferring to diverse language tasks and confirming that our method improves *fundamental* reasoning capabilities rather than overfitting to a single domain.

3.4 Analysis of Training Dynamics

Figure 2 illustrates Qwen3-4B training dynamics, evidencing the efficacy of MIXED-CUTS against saturation.

Further Introducing Uncertainty. The middle plot validates our mechanism. Standard GRPO (Grey) shows stagnant entropy ($\approx$0.20-0.25), confirming rapid convergence to "safe" patterns due to vanishing gradients. In contrast, MIXED-CUTS (Orange) sustains steady growth, acting as a variance-preservation mechanism that prevents premature convergence and keeps the optimization landscape active.

Emergence of Deeper Reasoning. Increased entropy signals deeper reasoning, not noise. While GRPO plateaus at $\approx$1200 tokens (Left), MIXED-CUTS drives trajectory lengths to peak over 1800. By forcing exploration of "second-best" tokens, CUTS unlocks latent "System 2" behaviors—such as self-correction—that are otherwise suppressed by greedy baselines.

From Exploration to Generalization. Extended exploration translates to robust generalization. On the harder AIME25 (Middle-Right), GRPO stagnates ($\approx$0.25), whereas MIXED-CUTS diverges sharply at step 30 to reach **0.40**. This correlation validates our premise: breaking the saturation trap on easy data can generalize LLMs to complex tasks.

Consistency Gains Are a Stable Optimization Outcome. The rightmost plot tracks the AIME25 maj@16 consistency metric over training. The $+23.2\%$ improvement reported in Table 6 is not a late-stage artifact or random fluctuation: the MIXED-CUTS maj@16 curve pulls away from the GRPO curve early in training and the gap is

Table 1: **Main results comparing MIXED-CUTS and GRPO on MATH.** We report Pass@1 and Pass@16 (%) across five benchmarks, including "Thinking Mode" baselines for reference. Δ denotes the gain over GRPO.

Model	MATH		AIME24		AIME25		AMC		GPQA	
	Pass@1	Pass@16	Pass@1	Pass@16	Pass@1	Pass@16	Pass@1	Pass@16	Pass@1	Pass@16
Qwen3-1.7B(Non-thinking)										
Base Model	70.2	90.5	12.9	36.5	11.7	27.8	39.8	72.9	32.1	80.3
GRPO	83.6	93.9	29.5	60.7	22.8	44.5	59.8	82.9	34.2	80.1
MIXED-CUTS	85.1	95.3	32.3	62.0	28.1	52.5	62.7	88.0	36.0	79.4
Δ	+1.5	+1.4	+2.8	+1.3	+5.3	+8.0	+2.9	+5.1	+1.8	-0.7
Base Model (Thinking)	82.6	93.2	28.9	62.8	24.9	44.5	57.5	82.0	34.9	75.3
Qwen3-4B (Non-thinking)										
Base Model	82.5	94.4	24.2	54.0	21.5	48.0	61.3	82.9	45.3	83.9
GRPO	86.4	95.9	32.5	63.8	26.6	57.9	68.9	88.9	48.1	84.6
MIXED-CUTS	90.8	96.6	46.0	73.5	41.7	71.9	76.7	91.9	50.1	84.5
Δ	+4.4	+0.7	+13.5	+9.7	+15.1	+14.0	+7.8	+3.0	+2.0	-0.1
Base Model (Thinking)	89.9	95.7	54.1	75.5	42.1	62.3	73.6	88.3	52.0	84.4

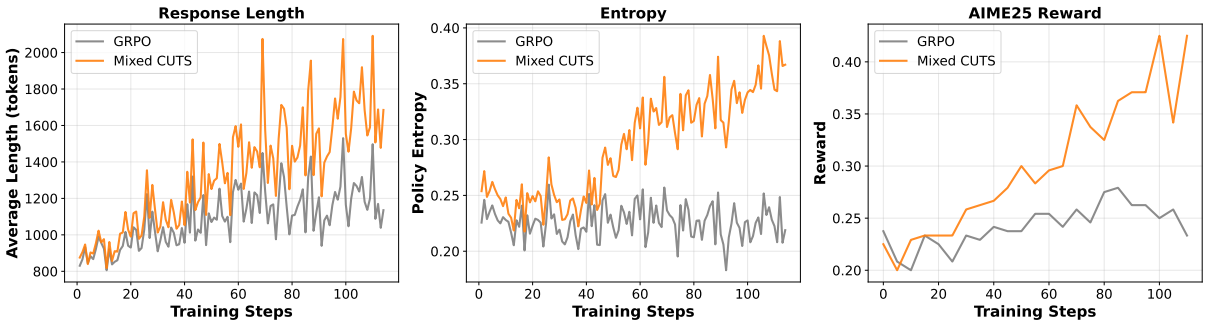

Figure 2: **Training Dynamics (Qwen3-4B).** Evolution of (Left) Response Length, (Middle-Left) Policy Entropy, (Middle-Right) AIME25 Reward, and (Right) AIME25 maj@16 consistency. Unlike GRPO (Grey), MIXED-CUTS (Orange) breaks the saturation trap by sustaining high entropy and inducing longer reasoning chains, driving both superior out-of-domain generalization and substantially stronger majority-vote consistency.

Table 2: Zero-shot cross-domain accuracy of Qwen3-4B trained solely on MATH, evaluated on non-mathematical reasoning benchmarks.

Training Method	MMLU-Pro	SuperGPQA
Base Model	63.80%	33.05%
Standard GRPO	68.59%	40.03%
MIXED-CUTS (Ours)	**69.65%**	**41.28%**

robustly maintained throughout the later optimization phase, confirming that the consistency gain is a stable, continuously compounding outcome of the Mixed-CUTS objective rather than an artifact of the particular evaluation checkpoint.

4 Conclusion

We identified a critical bottleneck in LLM-RL: on saturated datasets, standard sampling induces vanishing gradients and mode collapse. To address this, we introduced **CUTS**, a lightweight operator enforcing local uniformity within the high-probability region of the model distribution. Integrated into the **Mixed-CUTS** framework, this approach explicitly restores informative advantage signals even in saturated regimes. Empirically, we confirm that such parameter-free diversification yields substantial gains in reasoning robustness and out-of-domain generalization. As models scale, strategies like CUTS that unlock latent capabilities beyond static data ceilings will be essential for continuous self-improvement. Future work will extend this uniform-prior exploration to code generation and agentic planning.

Limitation

While Mixed-CUTS is motivated by the vanishing-advantage phenomenon in group-relative policy optimization, we do not yet provide a formal convergence or optimality analysis characterizing how decoding-time uniformization interacts with the GRPO objective. In particular, the mixed behavior policy induced by CUTS introduces a controlled

deviation from strict on-policy sampling, and although this deviation is empirically stable under clipping, its long-horizon effect on policy improvement is not theoretically quantified. Moreover, our notion of diversity is operationalized through local distribution flattening within Top-K candidates, which serves as a practical proxy rather than a formally grounded exploration criterion. Developing a principled theoretical framework that connects decoding-level interventions, advantage variance preservation, and convergence guarantees in saturated regimes remains an open direction for future research.

References

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

Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, and 1 others. 2025. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*.

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

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

Chengsong Huang, Wenhao Yu, Xiaoyang Wang, Hongming Zhang, Zongxia Li, Ruosen Li, Jiaxin Huang, Haitao Mi, and Dong Yu. 2025. R-zero: Self-evolving reasoning llm from zero data. *arXiv preprint arXiv:2508.05004*.

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

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

Zongxia Li, Wenhao Yu, Chengsong Huang, Rui Liu, Zhenwen Liang, Fuxiao Liu, Jingxi Che, Dian Yu, Jordan Boyd-Graber, Haitao Mi, and 1 others. 2025. Self-rewarding vision-language model via reasoning decomposition. *arXiv preprint arXiv:2508.19652*.

Zhenwen Liang, Ruosen Li, Yujun Zhou, Linfeng Song, Dian Yu, Xinya Du, Haitao Mi, and Dong Yu. 2025a. Clue: Non-parametric verification from experience via hidden-state clustering. *arXiv preprint arXiv:2510.01591*.

Zhenwen Liang, Sidi Lu, Wenhao Yu, Kishan Panaganti, Yujun Zhou, Haitao Mi, and Dong Yu. 2025b. Can llms guide their own exploration? gradient-guided reinforcement learning for llm reasoning. *arXiv preprint arXiv:2512.15687*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025a. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Rui Liu, Dian Yu, Tong Zheng, Runpeng Dai, Zongxia Li, Wenhao Yu, Zhenwen Liang, Linfeng Song, Haitao Mi, Pratap Tokekar, and 1 others. 2025b. Vogue: Guiding exploration with visual uncertainty improves multimodal reasoning. *arXiv preprint arXiv:2510.01444*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

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

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

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

Xiangqi Wang, Yue Huang, Yujun Zhou, Xiaonan Luo, Kehan Guo, and Xiangliang Zhang. 2025b. Causally-enhanced reinforcement policy optimization. *arXiv preprint arXiv:2509.23095*.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, and 1 others. 2025. Rag-gym: Optimizing reasoning and search agents with process supervision. *arXiv preprint arXiv:2502.13957*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025a. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhaoning Yu, Will Su, Leitian Tao, Haozhu Wang, Aashu Singh, Hanchao Yu, Jianyu Wang, Hongyang Gao, Weizhe Yuan, Jason Weston, and 1 others. 2025b. Restrain: From spurious votes to signals–self-driven rl with self-penalization. *arXiv preprint arXiv:2510.02172*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Yulai Zhao, Haolin Liu, Dian Yu, SY Kung, Haitao Mi, and Dong Yu. 2025. One token to fool llm-as-a-judge. *arXiv preprint arXiv:2507.08794*.

Tong Zheng, Hongming Zhang, Wenhao Yu, Xiaoyang Wang, Xinyu Yang, Runpeng Dai, Rui Liu, Huiwen Bao, Chengsong Huang, Heng Huang, and 1 others. 2025. Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. 2025. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*.

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

A Related Works

RL for LLM Reasoning. RL is the standard paradigm for enhancing LLM reasoning in objective domains like mathematics and coding (Liu et al., 2025a; Li et al., 2025; Wang et al., 2025b; Yu et al., 2025a; Xiong et al., 2025; Dai et al., 2025). Recent advancements rely on GRPO and its variants (Guo et al., 2025; Yang et al., 2025; Yu et al., 2025a; Wang et al., 2025a), which efficiently estimate baselines via group averages without separate value networks. Despite this efficiency, optimization instability remains a challenge. Recent works observe that while RL elicits long chain-of-thought (CoT) reasoning, policies often rapidly converge to homogeneous generation patterns, stalling further improvement (Zhou et al., 2025; Liang et al., 2025b; Yu et al., 2025b).

The Challenge of Saturation and Mode Collapse. A critical bottleneck is the exploration-exploitation trade-off (Wang et al., 2025a; Zhou et al., 2025). Traditionally, mode collapse was attributed to sparse rewards (Cui et al., 2025). However, we identify a distinct **saturation-induced collapse** in strong models on standard benchmarks. In these "easy-task" regimes, high baseline success rates cause intra-group reward variance to vanish (Yu et al., 2025a). Lacking negative contrast, the relative advantage signal disappears, disincentivizing the exploration of superior strategies (Zhu et al., 2025). While entropy bonuses (Cui et al., 2025) attempt to mitigate this, they often encourage nonsensical diversity: because entropy regularization indiscriminately penalizes confidence, it can disrupt coherent reasoning chains and lead to diversity that is incoherent rather than meaningful. In contrast, our decoding-time intervention (CUTS) performs *structure-preserving* exploration: by restricting its uniform sampling strictly to the high-confidence Top-K subset, CUTS maintains local semantic validity while effectively breaking the mode collapse, offering a more controlled and stable exploration mechanism than global entropy bonuses.

B Detailed Experimental Setup

B.1 Datasets

We conduct large-scale training using the canonical **MATH dataset** (Hendrycks et al., 2021). To rigorously evaluate our models, we employ a comprehensive suite of five benchmarks designed to measure both in-domain retention and out-of-domain generalization: **MATH-500**, **AIME24**, **AIME25**,

AMC (Li et al., 2024), and **GPQA-Diamond (GPQA)** (Rein et al., 2024). Implementation details are provided in Appendix B.

B.2 Models and Configurations

We utilize the **Qwen3** series (Yang et al., 2025) as our backbone models, specifically the 1.7B and 4B parameter variants, with non-thinking mode. To fully accommodate the extensive reasoning chains required for complex mathematical problem solving, we scale the generation capacity according to the model size. We configure the maximum generation length to **5,000 tokens** for the 1.7B model and extend it to **12,000 tokens** for the 4B model. This extended context window is critical for avoiding truncation during the exploration of deep reasoning paths in the rollout phase.

B.3 System Prompt

For all experiments, we used the following system prompt to guide the model's generation format, ensuring that it produces a step-by-step reasoning process and a clearly marked final answer (Zeng et al., 2025):

System Prompt

```
Please reason step by step, and put
your final answer within \boxed{}.
```

B.4 Answer and Reasoning Extraction

To implement the scoring criteria described in the main text, we apply the following extraction procedure for each generated response o_i:

- **Final Answer Extraction (for Validity):** We parse the response to find the content within the final occurrence of the `\boxed{·}` command. A response is deemed "valid" only if this command is present and its content contains at least one numeric digit. This extracted numeric string is used for the majority vote.

B.5 Hyperparameter Settings

This section summarizes the key hyperparameters used in MIXED-CUTS training and decoding. Unless otherwise specified, all parameters are fixed across experiments and model scales.

CUTS Decoding Hyperparameters. The CUTS operator is fully parameter-free with respect to model training and introduces only a small number of decoding-time hyperparameters. At each decoding step, we retrieve the Top-K candidate tokens based on the model's original distribution, with $K = 5$ in all experiments. To ensure semantic plausibility, we apply a probability threshold $\delta = 0.03$ to filter out low-confidence candidates. The remaining tokens are assigned a uniform probability, enforcing local width-first exploration. To preserve early reasoning stability, CUTS is disabled for the first $T_{\text{warmup}} = 5$ tokens, during which standard decoding is applied.

Table 3: Key hyperparameters used in MIXED-CUTS.

Category	Value
Top-K (K)	5
Probability Threshold (δ)	0.03
Warm-up Tokens (T_{warmup})	5
Group Size (G)	16
Exploitation / Exploration Split	8 / 8
Training Batch Size	128
PPO Mini-batch Size	32
KL Loss Type	Low-variance KL
KL Coefficient	1×10^{-3}
Rollout Temperature	1.0
Rollout Top-p	1.0
Validation Top-p / Top-k	0.8 / 20
Validation Temperature	1.0

Mixed-CUTS Sampling Strategy. Within the GRPO framework, we generate a group of $G = 16$ trajectories per prompt. The group is evenly split into an exploitation batch ($G_{\text{std}} = 8$), generated via standard sampling, and an exploration batch ($G_{\text{cuts}} = 8$), generated using CUTS. Advantages are computed jointly over the combined group.

Training and Optimization Settings. We use a global training batch size of 128, with PPO mini-batches of size 32. KL regularization is enabled using a low-variance KL estimator with coefficient 1×10^{-3}. During rollout, we use temperature sampling with $T = 1.0$ and disable nucleus truncation (top-$p = 1.0$) to avoid confounding exploration effects. Validation rollouts follow standard decoding settings.

Table 3 provides a concise summary of the main hyperparameters.

B.6 Hyperparameter Sensitivity Analysis

Beyond the default configuration in Table 3, we independently probe the sensitivity of Mixed-CUTS to its two CUTS-specific decoding hyperparameters on AIME25, varying $\delta \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$ at fixed $K = 5$ and

$K \in \{3, 5, 7, 9\}$ at fixed $\delta = 0.03$, for both Qwen3-1.7B and Qwen3-4B. Results are reported in Tables 4 and 5.

Table 4: Impact of the minimum-probability threshold δ on AIME25 (fixed $K = 5$). Pass@1 / Pass@16 (%).

δ	Qwen3-1.7B	Qwen3-4B
0.01	22.0 / 44.1	35.0 / 64.2
0.02	26.5 / 50.2	40.0 / 69.5
0.03 (Default)	**28.1 / 52.5**	**41.7 / 71.9**
0.04	27.2 / 51.4	40.6 / 70.8
0.05	26.0 / 49.8	39.8 / 70.1

Table 5: Impact of the Top-K candidate-set size on AIME25 (fixed $\delta = 0.03$). Pass@1 / Pass@16 (%).

Top-K	Qwen3-1.7B	Qwen3-4B
$K = 3$	25.5 / 48.9	38.9 / 68.5
$K = 5$ (Default)	**28.1 / 52.5**	**41.7** / 71.9
$K = 7$	26.8 / 53.4	40.2 / **72.8**
$K = 9$	21.5 / 48.2	35.5 / 68.2

Robustness across reasonable ranges. Performance remains highly stable across $\delta \in [0.02, 0.05]$ and $K \in [3, 7]$, and every configuration in these ranges consistently outperforms the standard GRPO baseline (26.6% on Qwen3-4B AIME25), indicating that Mixed-CUTS does not require careful per-model tuning.

The necessity of the δ filter. Setting δ too low (e.g., $\delta = 0.01$) causes a sharp drop (22.0% on 1.7B, 35.0% on 4B), because the weak filter allows low-quality tail tokens into the candidate set, and those tokens occasionally corrupt the reasoning chain. This empirically validates the design choice of filtering Top-K by a minimum probability.

Exploration–noise tradeoff in K. $K = 7$ achieves a slightly higher Pass@16 than the default $K = 5$ on both model sizes, because a marginally wider search space uncovers more diverse correct paths when multiple samples are drawn. However, excessive K ($K = 9$) introduces noise, causing Pass@1 to drop much more severely (e.g., -6.2% on 4B) than Pass@16 (-3.7%), matching the theoretical expectation that sampling noise heavily impacts single-shot accuracy but is partially mitigated by multi-sample evaluation.

C Additional Experiments

C.1 Robustness Analysis: Majority Vote Performance

In addition to Pass@1 and Pass@16, we evaluate the models using Majority Vote (maj@16), a metric that reflects the model's internal consistency and confidence. Unlike Pass@N, which measures the existence of a correct solution in the sample space, maj@16 measures whether the correct solution dominates the probability distribution. The results are detailed in Table 6.

Significant Gains in Solution Consistency. MIXED-CUTS demonstrates remarkable improvements in consistency compared to the GRPO baseline. On the Qwen3-4B model, we observe a massive +23.2% improvement on AIME25 (maj@16 increases from 31.9% to 55.1%). This indicates that our method does not simply "stumble upon" the correct answer through random exploration; rather, it fundamentally reshapes the policy to assign high probability mass to correct reasoning paths. Standard GRPO, by contrast, often struggles to achieve consensus on hard tasks due to optimization instability, leading to lower majority vote scores despite decent Pass@N performance.

Beating "Thinking Mode" Consistency. It is particularly noteworthy that MIXED-CUTS (operating in standard mode) achieves higher consistency than the base model's native "Thinking Mode" on several key benchmarks. For instance, on the 4B scale, MIXED-CUTS achieves a maj@16 of 55.1% on AIME25, surpassing the Thinking Mode's 54.0%. Similarly, on the 1.7B scale, our method outperforms Thinking Mode on AIME25 (36.9% vs 31.3%) and AMC (74.3% vs 68.6%). This result reinforces our claim that MIXED-CUTS effectively distills the benefits of extensive search into a robust, low-latency policy that yields reliable, consistent solutions without the computational overhead of recursive thinking.

C.2 Performance with Abundant Hard Data

We further verify that Mixed-CUTS remains effective when trained directly on high-quality hard data, showing that its benefits are orthogonal to data difficulty rather than a substitute for hard data. We train Qwen3-4B directly on the **DAPO-17K** dataset, a large-scale collection of high-quality reasoning prompts analogous in difficulty to AIME-level training distributions, and evaluate on the

Table 6: Comparison of Majority Vote performance (maj@16) on the MATH dataset and out-of-domain benchmarks. Results represent the accuracy when selecting the most consistent answer from 16 sampled paths. Δ values indicate the improvement of Mixed-CUTS over the GRPO baseline.

Model	MATH	AIME24	AIME25	AMC	GPQA
Qwen3-1.7B					
Base Model (Non-thinking)	77.6	21.1	16.2	50.9	34.3
GRPO	89.3	45.7	29.7	71.0	36.5
MIXED-CUTS	90.4	49.2	36.9	74.3	38.0
Δ	+1.1	+3.5	+7.2	+3.3	+1.5
Base Model (Thinking)	88.9	45.4	31.3	68.6	38.7
Qwen3-4B					
Base Model (Non-thinking)	88.3	33.1	24.8	70.5	48.1
GRPO	90.1	43.0	31.9	78.1	51.4
MIXED-CUTS	94.0	54.9	55.1	83.0	53.1
Δ	+3.9	+11.9	+23.2	+4.9	+1.7
Base Model (Thinking)	92.4	67.5	54.0	81.2	54.8

same five benchmarks used in the main text (Table 7).

Consistent gains on hard data. Training on harder data raises the performance floor: standard GRPO on DAPO reaches 54.1% on AIME25 Pass@1, compared to 26.6% on MATH. On top of this stronger baseline, MIXED-CUTS still yields consistent absolute gains across all five benchmarks (e.g., $+1.9\%$ on AIME25, $+3.1\%$ on AMC, $+1.7\%$ on GPQA for Pass@1), with even larger margins on Pass@16 (e.g., $+4.9\%$ on AIME25, $+7.0\%$ on GPQA). The "vanishing advantage" phenomenon re-emerges once the model begins to saturate even on harder data, and Mixed-CUTS continues to break this new saturation bottleneck by systematically exploring valid alternative semantic branches.

Beyond the "data wall". The absolute algorithmic gain is larger on MATH than on DAPO ($+15.1\%$ vs. $+1.9\%$ on AIME25 Pass@1), and this pattern highlights where Mixed-CUTS contributes most. When abundant, ultra-hard labeled data is available, the inherent difficulty of the prompts already forces the model into high-variance exploration, so the vanishing-advantage collapse is less severe. The more interesting regime is the opposite one: curating increasingly difficult high-quality reasoning datasets becomes unsustainably expensive as model capabilities scale. Mixed-CUTS shows that *easy*, easily-saturated data still carries exploitable learning signal—by structurally enforcing exploration on simple datasets like MATH, the model acquires generalized reasoning skills (the $+15.1\%$ gain on AIME25) without needing an endless supply of DAPO-level prompts. Even when hard data *is* abundant, Mixed-CUTS still provides orthogonal gains on top of it.

D Variance Preservation: Full Derivation

This appendix provides the complete case-by-case derivation of the variance-preservation argument sketched in Section 2 (the discussion following Eq. 5). The goal is to show that, in the two saturated extremes that kill the GRPO advantage signal, the intra-group variance σ^2_{mixed} of a Mixed-CUTS group is strictly bounded away from zero.

Setup. For a single prompt $\mathbf{q}$, let $\mathcal{G}_{\text{std}}$ and $\mathcal{G}_{\text{CUTS}}$ be the standard and CUTS sub-groups of size $G/2$, with binary rewards $r_i \in \{0, 1\}$ (correct / incorrect). Let $(\mu_{\text{std}}, \sigma^2_{\text{std}})$ and $(\mu_{\text{CUTS}}, \sigma^2_{\text{CUTS}})$ be their sample means and variances. The combined group has size G, mean $\mu_{\text{mixed}} = \frac{1}{2}(\mu_{\text{std}} + \mu_{\text{CUTS}})$, and variance

$$\sigma^2_{\text{mixed}} = \tfrac{1}{2}(\sigma^2_{\text{std}} + \sigma^2_{\text{CUTS}}) + \tfrac{1}{4}(\mu_{\text{std}} - \mu_{\text{CUTS}})^2,$$

by the law of total variance for two equal sub-groups.

Table 7: Performance of Qwen3-4B trained directly on the hard **DAPO-17K** dataset. Pass@1 / Pass@16 (%) across five reasoning benchmarks.

Model (Qwen3-4B, trained on DAPO-17K)	AIME24	MATH-500	AIME25	AMC	GPQA
Standard GRPO	65.6 / 82.6	92.6 / 96.3	54.1 / 68.7	84.6 / 95.6	54.9 / 81.9
MIXED-CUTS (Ours)	**67.5 / 84.1**	**93.8 / 97.8**	**56.0 / 73.6**	**87.7 / 97.7**	**56.6 / 88.9**

Key behavioral assumption. By construction, CUTS does not follow the model's greedy mode: within the Top-K subset filtered by the probability threshold δ, it replaces the model's skewed distribution with a uniform one. Whenever $|\mathcal{S}_t| \geq 2$ and the Top-K mass is non-trivially concentrated on the greedy token, the exploratory sub-group therefore produces trajectories that differ semantically from the greedy trajectories generated by $\pi_{\theta_{\text{old}}}$. In expectation over prompts, this behavioral gap implies $\mu_{\text{CUTS}} \neq \mu_{\text{std}}$ on any prompt where the model is not already deterministic.

Case A: "Too easy" saturated prompt ($\mu_{\text{std}} \to 1$, $\sigma^2_{\text{std}} \to 0$). All standard samples succeed, so the greedy policy is essentially deterministic on this prompt and the within-group variance of $\mathcal{G}_{\text{std}}$ vanishes. CUTS, by decoupling sampling from the peaked distribution, occasionally selects a semantically valid but sub-optimal branch that does not lead to the canonical solution; some of these branches fail, pushing μ_{CUTS} below 1. Substituting $\mu_{\text{std}} \to 1$, $\sigma^2_{\text{std}} \to 0$, $\mu_{\text{CUTS}} < 1$ into the decomposition yields

$$\sigma^2_{\text{mixed}} \approx \tfrac{1}{2}\sigma^2_{\text{CUTS}} + \tfrac{1}{4}(1 - \mu_{\text{CUTS}})^2 > 0.$$

A non-zero advantage signal is therefore preserved for exactly the "all-correct" prompts where standard GRPO breaks down.

Case B: "Too hard" saturated prompt ($\mu_{\text{std}} \to 0$, $\sigma^2_{\text{std}} \to 0$). All standard samples fail because the greedy policy repeatedly commits to the same incorrect reasoning path. CUTS forces the model to uniformly consider its Top-K alternatives, giving it a non-trivial probability of stumbling onto a correct intermediate step that the greedy mode systematically misses; some of these exploratory trajectories succeed, pushing μ_{CUTS} above 0. Substituting $\mu_{\text{std}} \to 0$, $\sigma^2_{\text{std}} \to 0$, $\mu_{\text{CUTS}} > 0$ into the decomposition yields

$$\sigma^2_{\text{mixed}} \approx \tfrac{1}{2}\sigma^2_{\text{CUTS}} + \tfrac{1}{4}\mu^2_{\text{CUTS}} > 0.$$

Even on prompts where standard GRPO sees only failures, Mixed-CUTS recovers a positive variance and, importantly, a positive advantage $\hat{A}_i$ for the rare CUTS trajectories that happen to succeed—exactly the learning signal required to escape the "too-hard" failure mode.

Conclusion. In both saturated extremes, σ^2_{mixed} is strictly lower-bounded by a non-zero quantity driven by the between-group difference $(\mu_{\text{std}} - \mu_{\text{CUTS}})^2$. The standardized advantage in Eq. 1 is therefore kept away from the degenerate $0/\epsilon$ regime, and the policy gradient remains informative. This is the formal statement of the "structural variance preservation" claim made in the main text: Mixed-CUTS does not rely on noise to restore contrast—it relies on the structural behavioral gap between greedy and Top-K-uniform decoding, and this gap is precisely the second term of Eq. 5.

E AI Writing Assistance Declaration

We utilized generative AI models solely to improve the readability and clarity of the manuscript. The scope of assistance was limited to grammatical correction and stylistic polishing of the content originally written by the authors.