DIAGNOSING COLLAPSE IN MIXED-POLICY RL: THE UNDERLYING MECHANISM AND A UNIFIED SOLUTION FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Compared to pure on-policy reinforcement learning (RL), mixed-policy RL integrates the exploration of online rollouts with the exploitation of off-policy external traces, effectively surpassing the inherent capability boundaries of the base large language model. Nevertheless, directly applying standard importance sampling to integrate external data induces performance degradation. Although prior methods circumvent this instability by redesigning the surrogate objective, the fundamental cause of the collapse and the unifying theoretical mechanism underlying these disparate solutions remain poorly understood. In this paper, we reveal the fundamental cause of training collapse in mixed-policy RL. Our analysis demonstrates that the instability is not explained by the variance of importance ratios, but driven by severe gradient conflicts induced by low-probability external tokens against the rest of the batch. Motivated by this insight, we propose a unified gradient-suppression framework that reveals existing solutions, despite their different formulations, are fundamentally united by an implicit mechanism that suppresses gradient weights of low-probability tokens. To demonstrate the validity and utility of this framework, we derive a direct probability-aware reweighting method as a concrete instantiation. Extensive evaluations across multiple backbones and benchmarks demonstrate the superior performance of this method, while simultaneously validating the theoretical soundness of our framework and its predictive capacity for method design.

1 INTRODUCTION

Reinforcement learning has emerged as a scalable post-training paradigm for improving the reasoning capabilities of large language models (e.g., OpenAI-o1 (Jaech et al., 2024) and DeepSeek-R1 (Guo et al., 2025)). Specifically, Reinforcement Learning with Verifiable Reward (RLVR) methods based on relative policy optimization (e.g., GRPO (Shao et al., 2024) and RLOO (Ahmadian et al., 2024)) bypass the reliance on human annotation and have proven highly effective in enhancing model performance, particularly in solving complex mathematical and coding tasks. Despite its empirical success, recent analyses (Shao et al., 2024; Yue et al., 2025) point out that RLVR struggles to surpass the reasoning boundaries of the base model, relying on utilizing reasoning patterns already in the base model. Moreover, on challenging problems where the model fails to sample any correct solution, all group-relative advantages reduce to zero, leading to a complete stall in policy updates.

To overcome these limitations, recent methods augment the rollout group with external high-quality trajectories (Yan et al., 2025; Dong et al., 2025), providing supervisory guidance on harder instances that the model fails to explore autonomously. However, for these offline data, the true behavior policy is virtually inaccessible. Straightforwardly employing the model's previous policy, $\pi_{\theta_{\text{old}}}$, as a proxy for this unknown behavior policy—without any correction—leads to significant performance degradation. LUFFY (Yan et al., 2025) and RL-PLUS (Dong et al., 2025) have successfully circumvented this degradation by redesigning the off-policy objective with importance sampling. These approaches effectively assimilate knowledge from external trajectories, allowing them to surpass the performance of purely on-policy methods. However, the underlying mechanisms driving their empirical success remain poorly understood. This lack of clarity raises two fundamental questions:

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

(i) What exactly drives the failure when directly employing $\pi_{\theta_{old}}$ for importance sampling? (ii) What is the unifying theoretical mechanism underlying these seemingly disparate solutions?

In this paper, we systematically identify the root cause of training collapse in mixed-policy RL and propose a unified theoretical framework to overcome it. By analyzing the magnitude and direction of the gradients, we reveal that the training failure is not caused by gradient fluctuations stemming from the excessive variance of importance sampling ratios, but rather by severe gradient conflicts between low-probability external tokens and other tokens within the batch. Motivated by this fundamental insight, we introduce a **unified gradient-suppression framework** that serves as our core contribution. This framework provides a principled lens to demystify existing successful but heuristically designed approaches: we demonstrate that despite their disparate formulations, methods like LUFFY and RL-PLUS are fundamentally united by an implicit mechanism that suppresses the gradient weights of low-probability tokens. Consequently, our framework elevates the understanding of mixed-policy alignment from disparate empirical tricks to a unified, principled design space.

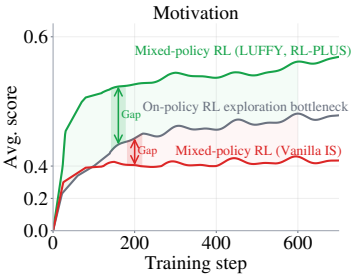


Figure 1: The panel presents the motivating empirical observation that gives rise to the two core questions posed in the Introduction: why does Vanilla IS fail on off-policy data, and what shared mechanism underlies the success of otherwise disparate solutions.

To empirically validate the predictive power and utility of our framework, we instantiate it into a simple yet explicit **Probability-aware Reweighting** method. We conduct extensive experiments across multiple model backbones (spanning the Qwen and LLaMA families from 0.5B to 8B parameters) on both in-domain mathematical reasoning and out-of-domain scientific benchmarks. Our results demonstrate that this direct instantiation significantly outperforms existing complex mixed-policy solutions. Furthermore, through systematic interval swap ablations and the construction of novel off-policy objectives, we validate the necessity of targeted token suppression and deeply showcase how our unified framework can rigorously guide future algorithmic innovations.

In summary, we make the following contributions:

1. **A new diagnostic perspective on training collapse in Mixed-Policy RL.** We provide empirical and theoretical evidence that training collapse in mixed-policy optimization is primarily driven by gradient conflict induced by low-probability external tokens—not by the large variance of importance sampling ratios as is commonly assumed (Section 3).
2. **A unified gradient-suppression framework.** We propose a theoretical framework that unifies seemingly disparate methods (LUFFY, RL-PLUS, and our approach) under a single mechanism: implicit gradient suppression for low-probability tokens. This framework not only demystifies their empirical success but also provides a principled design space for future off-policy objectives (Sections 4 and 5).
3. **Superior performance across diverse models and benchmarks.** Guided by our framework, we design a simple probability-aware reweighting method that achieves superior performance compared to existing specialized solutions across five model backbones (0.5B–8B parameters, spanning both Qwen and LLaMA families), six in-domain mathematical reasoning benchmarks, and three out-of-domain scientific benchmarks (Section 5).

2 Preliminaries

Reinforcement Learning with Verifiable Rewards (RLVR). In reinforcement learning with verifiable rewards (RLVR), a policy π_θ generates trajectories in response to a prompt $q \sim \mathcal{Q}$, and a deterministic verifier assigns a reward $R(\tau)$ based on the correctness of the final answer. To circumvent the need for a learned value network, group-relative methods such as GRPO (Shao et al., 2024) estimate advantages by normalizing rewards within a rollout group.

Mixed Policy Gradient Optimization. Building upon the standard GRPO formulation, LUFFY introduces a mixed-policy paradigm by augmenting the online rollout batch with external off-policy

samples. The relative advantage $\hat{A}_i$ for each trajectory τ_i is computed as:

$$\hat{A}_i = \frac{R(\tau_i) - \text{mean}(\mathcal{G}_{\text{on}} \cup \mathcal{G}_{\text{off}})}{\text{std}(\mathcal{G}_{\text{on}} \cup \mathcal{G}_{\text{off}})}, \tag{1}$$

where $\mathcal{G}_{\text{on}} = \{R(\tau_i) \mid \tau_i \sim \pi_{\theta_{\text{old}}}(\tau), i = 1, 2, \ldots, N_{\text{on}}\}$ and $\mathcal{G}_{\text{off}} = \{R(\tau_j) \mid \tau_j \sim \pi_\beta(\tau), j = 1, 2, \ldots, N_{\text{off}}\}$.

Extending from the GRPO objective, the generalized mixed objective is thus structured as:

$$\mathcal{J}(\theta) = \frac{1}{Z}\Bigg[\underbrace{\sum_{j=1}^{N_{\text{off}}} \sum_{t=1}^{|\tau_j|} \hat{r}_{j,t}(\theta, \beta)\hat{A}_j}_{\text{Off-policy Objective: } \mathcal{J}_{\text{OFF}}(\theta)} + \underbrace{\sum_{i=1}^{N_{\text{on}}} \sum_{t=1}^{|\tau_i|} \min\big(r_{i,t}(\theta)\hat{A}_i,\ \text{clip}\big(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_i\big)}_{\text{On-policy Objective: } \mathcal{J}_{\text{ON}}(\theta)} \Bigg], \tag{2}$$

where $Z = \sum_{j=1}^{N_{\text{off}}} |\tau_j| + \sum_{i=1}^{N_{\text{on}}} |\tau_i|$, $\hat{r}_{j,t}(\theta, \beta) = \pi_\theta(\tau_{j,t}|q, \tau_{j,<t})/\pi_\beta(\tau_{j,t}|q, \tau_{j,<t})$ and $r_{i,t}(\theta) = \pi_\theta(\tau_{i,t}|q, \tau_{i,<t})/\pi_{\theta_{\text{old}}}(\tau_{i,t}|q, \tau_{i,<t})$. For brevity in subsequent sections, we drop trajectory arguments and write $\hat{r}(\theta, \beta) \triangleq \hat{r}_{j,t}(\theta, \beta)$, $\pi_\theta \triangleq \pi_\theta(\tau_{j,t} \mid q, \tau_{j,<t})$, and $\pi_{\theta_{\text{old}}} \triangleq \pi_{\theta_{\text{old}}}(\tau_{j,t} \mid q, \tau_{j,<t})$.

Inaccessibility of the behavior policy. Equation 2 introduces the importance sampling ratio $\hat{r}(\theta, \beta)$ to correct the sampling bias introduced by external data. However, the true behavior policy π_β is practically inaccessible because the external data is often generated by models with incompatible vocabularies, sampled from unknown model mixtures, or authored by humans. Thus, practical off-policy optimization typically relies on behavior policy estimators $\hat{\pi}^\beta$, data support set approximations, or substituting π_β with a known reference policy (Levine et al., 2020). We summarize the three surrogate ratios analyzed in this paper below.

Surrogate ratios. Vanilla importance sampling (Vanilla IS) directly adopts the old training policy $\pi_{\theta_{\text{old}}}$ as the reference policy for estimating π_β; **LUFFY** (Yan et al., 2025) stabilizes the importance ratio by setting $\pi_\beta = 1$ and applying a bounding function $f(x) = x/(x + 0.1)$ to the ratio; **RL-PLUS** (Dong et al., 2025) focuses on mitigating distributional mismatch with Multiple Importance Sampling (MIS) combined with an exploration bonus. The three surrogate ratios are

$$\hat{r}^{\text{Vanilla}}(\theta) = \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}, \quad \hat{r}^{\text{LUFFY}}(\theta) = \frac{\pi_\theta}{\pi_\theta + 0.1}, \quad \hat{r}^{\text{RL-PLUS}}(\theta) = \frac{4\pi_\theta}{3\pi_{\theta_{\text{old}}} + 1}\big(1 - sg(\pi_\theta)\big)^{0.5}, \tag{3}$$

where $sg(\cdot)$ denotes the stop-gradient operator.

3 Empirical Analysis

In this section, we conduct an in-depth empirical investigation to understand the underlying causes of training collapse when directly applying vanilla importance sampling to external trajectories. As our point of departure, we analyze the gradients—the immediate mechanism behind model updates—focusing on their magnitude and direction. Specifically, we first analyze the ratio variance that induces fluctuations in gradient magnitude (§3.1). We then turn to gradient direction, investigating which tokens dominate the batch-level updates and how their gradients interact with the updates of other tokens (§3.2).

3.1 Vanilla IS Fails Irrespective of Ratio Variance

A standard explanation in prior RL literature is that optimization fails when the importance ratio fluctuates too violently, which is also the intuition behind PPO-style clipping (Schulman et al., 2015; 2017). If this explanation applies to our setting, then mitigating ratio variance should suffice to restore stable training under Vanilla IS. We therefore test this hypothesis by conducting a comparative experiment that only modifies $\mathcal{J}_{\text{OFF}}(\theta)$, using three off-policy strategies (mentioned in Section 2) with two additional variance-control diagnostics. The first is clipped Vanilla IS, which truncates the off-policy ratio to $[0.8, 1.2]$ and directly suppresses large ratio fluctuations. The second is an on-policy setting[1], where samples are drawn from $\pi_{\theta_{\text{old}}}$ itself, so the effective ratio stays at 1 and ratio variance is eliminated altogether. The full training setup is given in Section 5.1.

[1]Here, the training data still include external data, and "on-policy setting" refers specifically to setting `ppo_epochs` = 1 and `train_batch_size` = `mini_batch_size`, ensuring that $\pi_\theta = \pi_{\theta_{\text{old}}}$ during the update.

As summarized in Figure 2, reducing the variance from 0.021 to 0.011 and ultimately to 0 does not improve performance. By contrast, LUFFY and RL-PLUS achieve the highest average scores even though their surrogate weights exhibit larger variances, 0.035 and 0.037, respectively. These observations jointly lead to a negative conclusion: reducing ratio variance does not rescue Vanilla IS.

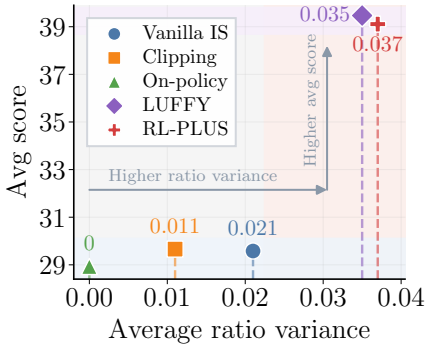


Figure 2: Average ratio variance during training versus average performance across methods. Lower ratio variance does not improve average performance.

3.2 Investigating Gradient Dynamics in Mixed-Policy Training

Since any parameter updates decompose into a scalar weight and a gradient direction, and Section 3.1 ruled out the former (ratio variance), this section shifts its focus to the latter: the gradient dynamics. Compared to pure on-policy RL, mixed-policy training introduces external data into the optimization process.

This raises a critical question: *could the parameter update directions driven by certain external trajectories conflict with those of other samples in the batch, thereby hindering the model's overall optimization?*

3.2.1 Dominance and Conflict of Low-Probability Tokens

To verify whether such potential conflicts actually occur and identify which samples dominate the update direction, we partition all batch tokens into ten equal-width intervals according to their policy probability π_θ and examine both the token count and the aggregated gradient magnitude within each interval.

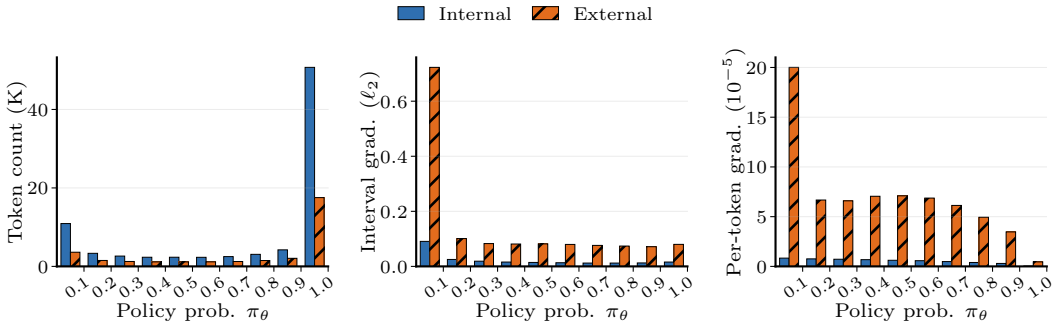


Figure 3: Gradient magnitudes and token counts across different probability ranges. From left to right: token count in each probability interval, the total ℓ_2 gradient norm contributed by tokens in that interval, and the average ℓ_2 gradient norm per token.

Token distribution and interval-wise gradient magnitude. As shown in Figure 3, the token count across probability intervals exhibits a pronounced U-shaped pattern. Notably, although the external trajectories are not sampled by the current policy, the high-probability intervals ($\pi_\theta \geq 0.9$) still harbor the largest number of tokens. Despite this population imbalance, the maximum gradient magnitude emerges unambiguously in the lowest-probability external interval ($\pi_\theta \in [0, 0.1)$). Moreover, the per-token gradient magnitude appears loosely inversely correlated with the token probability. Together, these observations suggest that a numerically minor subset of low-probability external tokens exerts a disproportionate influence on the overall batch gradient.

Gradient conflict between low-probability external tokens and the rest. To determine whether this disproportionate influence is constructive or destructive, we partition all batch tokens into two mutually exclusive subsets: the external low-probability group $\mathcal{B}^{\text{ext}}_{\text{low}}$ ($\pi_\theta \in [0, 0.1)$) and $\mathcal{B}_{\text{rest}}$ containing all remaining tokens. We then compute the aggregated gradient of each subset and track their interaction across consecutive

training steps (Figure 4). Two findings jointly characterize a severe gradient conflict. First, the cosine similarity between the two aggregated gradients, $\cos(g_{\text{low}}^{\text{ext}}, g_{\text{rest}})$, remains persistently negative across all observed steps, indicating that the two subsets push the parameters in systematically opposing directions. Second, the ℓ_2 norms of the two gradients are structurally comparable ($\|g_{\text{low}}^{\text{ext}}\|_2 \approx \|g_{\text{rest}}\|_2$), confirming that the external low-probability tokens are not a negligible perturbation but rather an equally powerful opposing force. This rules out the possibility that the conflict could be absorbed as minor noise by the dominant gradient component.

Low-probability external updates hinder optimization on $\mathcal{B}_{\text{rest}}$. At each of five consecutive training steps, we form separate updates using only $\mathcal{B}_{\text{rest}}$ or only $\mathcal{B}_{\text{low}}^{\text{ext}}$, and measure the resulting loss change on $\mathcal{B}_{\text{rest}}$ relative to the preceding step (Table 10). Averaged across the five steps, the $\mathcal{B}_{\text{rest}}$-only update reduces this loss by 0.1123, whereas the $\mathcal{B}_{\text{low}}^{\text{ext}}$-only update increases it by 0.0516. The external low-probability update therefore hinders optimization of $\mathcal{B}_{\text{rest}}$ in this local comparison.

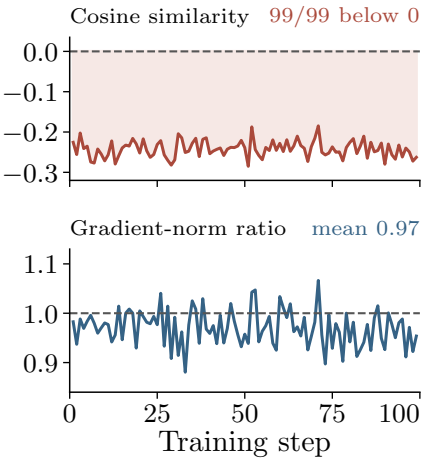


Figure 4: Persistent gradient conflict during training: Cosine similarity and gradient norm ratio between $\mathcal{B}_{\text{low}}^{\text{ext}}$ and $\mathcal{B}_{\text{rest}}$.

3.2.2 Token-wise Gradient Scaling: General Form and Vanilla Specialization

Figure 3 demonstrates a negative correlation between token-gradient magnitudes and their probabilities. To further investigate, this section provides a formal derivation of the intrinsic relationship between token probabilities and token-wise gradient magnitudes. For brevity, π_θ denotes the current probability of the sampled token a_i.

Gradient scaling for generalized surrogate objectives. Consider a single sampled token a_i with an advantage signal $\hat{A}$. For any differentiable surrogate ratio $\hat{r}(\theta)$, the single-token objective is $\mathcal{J}(\theta) = \hat{r}(\theta) \cdot \hat{A}$. By applying the chain rule through the softmax logits and bounding the resulting Jacobian, we show in Appendix G.1 that the ℓ_2-norm of the gradient at any layer ℓ with parameters $\theta^{(\ell)}$ is structurally governed by an *effective gradient factor* $\rho(\theta)$:

$$\left\|\nabla_{\theta^{(\ell)}} \mathcal{J}\right\|_2 \propto |\hat{A}\, \rho(\theta)|, \quad \text{where} \quad \rho(\theta) \triangleq \pi_\theta \frac{\partial \hat{r}(\theta)}{\partial \pi_\theta}(1 - \pi_\theta). \tag{4}$$

Specialization to Vanilla IS. For the Vanilla IS, the effective gradient factor evaluates to $\rho(\theta) = \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}(1 - \pi_\theta)$. The importance ratio tightly centers around 1 ($\pi_\theta / \pi_{\theta_{\text{old}}} \approx 1$) during typical training updates. Under this condition, the general gradient relationship reduces to:

$$\left\|\nabla_{\theta^{(\ell)}} \mathcal{J}(\theta)\right\| \;\propto\; |\hat{A}| \cdot (1 - \pi_\theta). \tag{5}$$

This analytically demonstrates that, under the standard ratio, the gradient coefficient gradually decreases as the token probability increases.

4 Method

Mitigating gradient conflicts primarily involves two dimensions: gradient direction and magnitude. However, achieving gradient direction alignment via gradient projection incurs prohibitive computational overhead. Therefore, focusing on gradient magnitude, we propose a unified gradient-suppression framework: stable mixed-policy optimization should attenuate the update gradients induced by low-probability external tokens.

4.1 PROBABILITY-AWARE REWEIGHTING

Guided by this framework, we down-weight external tokens whose policy probability π_θ is below a threshold c in the off-policy maximization objective (with $0 \le w_1 \ll 1$ and $w_2 = 1$):

$$\mathcal{J}_{\text{OFF}}(\theta) = \sum_{j=1}^{N_{\text{off}}} \sum_{t=1}^{|\tau_j|} \omega(\pi_\theta) \cdot \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}} \cdot \hat{A}_j, \quad \text{where} \quad \omega(\pi_\theta) = \begin{cases} w_1, & \text{if } \pi_\theta \le c, \\ w_2, & \text{otherwise}, \end{cases} \tag{6}$$

4.2 A UNIFIED PERSPECTIVE ON SOLUTIONS

In this section, we aim to demonstrate that seemingly distinct approaches—such as LUFFY and RL-PLUS—can fundamentally be unified under our gradient-suppression framework.

Theoretical Analysis. As shown in Table 1, we derive the effective gradient factor $\rho(\theta)$ for all objectives using Eq. equation 4. Within the low-probability regime ($\pi_\theta < 0.1$), both LUFFY and RL-PLUS yield smaller gradient weight coefficients, effectively scaling down the update magnitude compared with Vanilla IS (Appendix G.2).

Table 1: Unified view of the surrogate ratio $\hat{r}$, effective gradient weight $\rho(\theta)$, and low-probability token suppression mechanism across mixed-policy objectives.

Method	Surrogate Ratio $\hat{r}$	Gradient Weight $\rho(\theta)$	Suppression Mechanism
Vanilla IS	$\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}$	$\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}(1-\pi_\theta)$	None
LUFFY	$\frac{\pi_\theta}{\pi_\theta+0.1}$	$\frac{0.1\pi_\theta}{(\pi_\theta+0.1)^2}(1-\pi_\theta)$	Implicit: $\hat{r}(\theta)$ redesigning
RL-PLUS	$\frac{4\pi_\theta}{3\pi_{\theta_{\text{old}}}+1}\left(1-sg(\pi_\theta)\right)^{0.5}$	$\frac{4\pi_\theta}{3\pi_{\theta_{\text{old}}}+1}(1-\pi_\theta)^{1.5}$	Implicit: $\hat{r}(\theta)$ redesigning
Ours	$\omega(\pi_\theta)\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}$	$\omega(\pi_\theta)\frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}(1-\pi_\theta)$	Explicit: $\omega(\pi_\theta)$ reweighting

Empirical Observations. Figure 5 further supports this gradient-suppression effect during training. In the left and right panels, both LUFFY and RL-PLUS exhibit smaller effective weight coefficients and smaller average per-token gradients than Vanilla IS across the relevant probability range. The middle panel shows that the total gradient mass in low-probability intervals is also markedly reduced relative to Vanilla IS. Notably, the left- and right-panel curves follow the same probability-dependent trend, indicating that smaller coefficients are consistently associated with weaker empirical gradients. This agreement validates the theoretical characterization in Eq. equation 4.

Conclusion. Despite different approaches (redesigning $\hat{r}(\theta)$ or explicit reweighting), all three methods ultimately conform to the unified gradient-suppression framework by suppressing the gradient updates for low-probability external tokens.

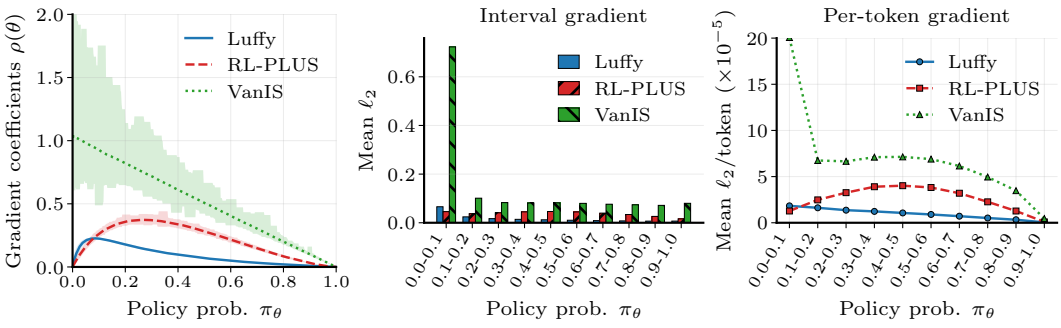


Figure 5: Effective gradient weights and probability-bucket gradients induced by different off-policy objectives. **Left:** Empirical $\rho(\theta)$ values computed on external tokens during training. **Middle:** Total external-token gradient magnitude in each probability interval. **Right:** Average gradient magnitude per external token in each interval.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUPS

We follow the mixed-policy RLVR protocol adopted by recent works (Yan et al., 2025; Dong et al., 2025), using Elliott/Openr1-Math-46k-8192 as the training dataset. We evaluate in-distribution mathematical reasoning on six benchmarks: AIME 2024, AIME 2025, AMC, MATH-500, Minerva (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). For out-of-domain evaluation, we use ARC-C (Clark et al., 2018), GPQA-diamond (Rein et al., 2023), and MMLU-Pro (Wang et al., 2024) (see Appendix E.1 for details). We study two model families across scales:

- **Qwen Family:** Qwen2.5-0.5B (Qwen et al., 2025), Qwen2.5-Math-1.5B/7B (Yang et al., 2024);
- **LLaMA Family:** LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024).

Baselines. We compare with three groups of baselines. **Pure RL methods:** SimpleRL-Zero (Zeng et al., 2025), GRPO (Shao et al., 2024), OpenReasoner-Zero (Hu et al., 2025), PRIME-Zero (Cui et al., 2025), and Oat-Zero (Liu et al., 2025). These methods represent standard on-policy RLVR training and its stronger variants with different implementations of policy optimization. **RL + SFT hybrids:** GRPO+SFT loss, HPT (Lv et al., 2025), and SRFT (Fu et al., 2025). These baselines integrate verifier-based RL with supervised imitation, either through joint optimization or staged training. **Mixed-policy RL methods:** LUFFY and RL-PLUS.

Table 2: Main results on in-distribution and out-of-distribution benchmarks for the Qwen2.5-Math-7B setting. Best results in each column are shown in bold, and second-best results are underlined.

Model	In-Distribution Performance					Out-of-Distribution Performance			
	AIME 24/25	AMC	MATH-500	Minerva	Olympiad	ARC-c	GPQA*	MMLU-Pro	**Avg.**
Qwen2.5-Math-7B	11.5/4.9	31.3	43.6	7.4	15.6	18.2	11.1	16.9	17.8
Pure RL Methods									
SimpleRL-Zero	27.0/6.8	54.9	76.0	25.0	34.7	30.2	23.2	34.5	34.7
GRPO	23.3/9.2	60.2	81.2	36.4	44.2	77.2	30.8	48.8	45.7
OpenReasoner-Zero	16.5/15.0	52.1	82.4	33.1	47.1	66.2	29.8	**58.7**	44.5
PRIME-Zero	17.0/12.8	54.0	81.4	39.0	40.3	73.3	18.2	32.7	41.0
Oat-Zero	**33.4**/11.9	61.2	78.0	34.6	43.4	70.1	23.7	41.7	44.2
RL+SFT Hybrid Methods									
RL w/ SFT Loss	19.5/16.4	49.7	80.4	34.9	39.4	71.2	23.7	43.2	42.0
HPT	27.1/22.1	64.2	86.8	36.0	55.6	81.6	36.9	52.1	51.4
SRFT	18.4/15.5	55.9	83.8	**42.6**	48.9	80.5	36.8	**58.7**	49.0
Mixed-policy RL Methods									
LUFFY	27.5/20.0	**68.7**	86.8	32.7	51.9	83.1	40.9	52.3	51.5
RL-PLUS	30.0/**22.5**	65.4	88.8	39.0	57.8	**83.9**	40.9	54.1	53.6
Ours	29.6/21.7	63.6	**89.0**	40.8	**58.7**	82.9	**43.4**	54.3	**53.8**

5.2 MAIN RESULTS

Table 2 presents the main comparison on the Qwen2.5-Math-7B backbone. Three observations stand out. First, despite relying on a straightforward probability-aware reweighting mechanism, our method achieves performance on par with—and in several metrics surpasses—the two concurrent mixed-policy approaches. Second, all three mixed-policy RL methods consistently outperform pure RL baselines, confirming that once the gradient conflict induced by low-probability external tokens is properly mitigated, mixed-policy training can more effectively harness external knowledge to improve reasoning capability. Third, the gains are not confined to the in-distribution mathematical benchmarks: all mixed-policy methods also yield

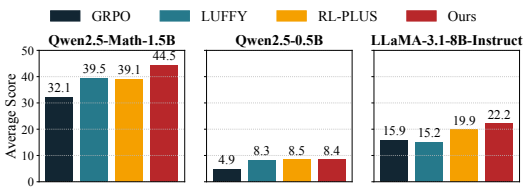


Figure 6: Cross-backbone comparison on the six mathematical benchmarks. Full per-benchmark results are deferred to Appendix Table 11.

substantial improvements on the out-of-distribution evaluation suite, indicating that the reasoning abilities acquired from external mathematical data generalize well to unseen domains.

Figure 6 summarizes the cross-backbone average results. Across all settings, our method consistently matches or surpasses both LUFFY and RL-PLUS, highlighting the *versatility* of the proposed approach—it remains effective regardless of model family, parameter count, or the initial reasoning ability of the backbone.

5.3 Unified Framework Validation

To further examine the validity of the unified framework in Section 4.2, we conduct two targeted validation studies under the same mixed-policy training protocol as the main experiments, using Qwen2.5-Math-1.5B as the backbone and modifying only the surrogate ratio used for off-policy updates.

5.3.1 Interval Swap Ablation

Figure 5 shows that LUFFY and RL-PLUS suppress gradient coefficients below the vanilla IS curve across both low- and high-probability regimes. Therefore, we perform interval-swap ablations on the surrogate ratios to determine if the observed effect originates specifically from the low-probability regime.

We design a hybrid ratio $\hat{r}_{\text{hybrid}}(\theta)$ that applies different surrogate ratios in different intervals:

$$\hat{r}_{\text{hybrid}}(\theta) = \begin{cases} \hat{r}_{\text{low}}(\theta), & \text{if } \pi_\theta < p, \\ \hat{r}_{\text{high}}(\theta), & \text{if } \pi_\theta \geq p, \end{cases} \quad \text{where } \hat{r}_{\text{low}}, \hat{r}_{\text{high}} \in \{\hat{r}^{\text{LUFFY}}(\theta), \hat{r}^{\text{RL-PLUS}}(\theta), \hat{r}^{\text{Vanilla}}(\theta)\} \quad (7)$$

Table 3: Interval-swap ablation. The low-probability surrogate ratio determines training stability among the tested combinations.

Low-probability ratio $\hat{r}_{\text{low}}$	High-probability ratio $\hat{r}_{\text{high}}$	Avg(6math) ↑	Outcome
Low-probability interval: suppressing ratio			
$\hat{r}^{\text{LUFFY}}$	$\hat{r}^{\text{LUFFY}}$	39.47	**Work**
	$\hat{r}^{\text{Vanilla}}$	38.68	
$\hat{r}^{\text{RL-PLUS}}$	$\hat{r}^{\text{RL-PLUS}}$	39.11	
	$\hat{r}^{\text{Vanilla}}$	39.59	
Low-probability interval: no suppression			
$\hat{r}^{\text{Vanilla}}$	$\hat{r}^{\text{LUFFY}}$	27.89	**Failed**
	$\hat{r}^{\text{RL-PLUS}}$	28.15	
	$\hat{r}^{\text{Vanilla}}$	29.58	

As shown in Table 3, the outcome depends entirely on $\hat{r}_{\text{low}}$ within the tested combinations: replacing $\hat{r}_{\text{low}}$ with the designed ratio yields stable training regardless of $\hat{r}_{\text{high}}$, whereas keeping $\hat{r}_{\text{low}} = \hat{r}^{\text{Vanilla}}$ leads to collapse regardless of $\hat{r}_{\text{high}}$. Thus, low-probability suppression determines stability in this experimental setting.

5.3.2 Additional Objective Validation

We construct Additional off-policy ratio variants to verify whether their empirical training behavior aligns with the unified framework's predictions. As shown in Table 4 and Figure 8, the first two variants—which suppress low-probability tokens—achieved successful training results, whereas the third variant did not. This demonstrates a complete agreement between theoretical predictions and empirical observations. Consequently, these results indicate that the framework is not only descriptive of existing methods but also predictive for newly designed off-policy objectives.

Table 4: Additional objective validation. The behaviors of these objectives predicted by the unified framework align consistently with the experimental results.

Variant	Surrogate Ratio $\hat{r}$	Gradient Weight $\rho(\theta)$	Low-prob. suppression	Prediction	Avg score
1	$\frac{6\pi_\theta}{5\pi_{\theta_{\text{old}}}+1}(1-sg(\pi_\theta))$	$\frac{6\pi_\theta}{5\pi_{\theta_{\text{old}}}+1}(1-\pi_\theta)^2$	✓	Work	38.62
2	π_θ	$\pi_\theta(1-\pi_\theta)$	✓	Work	39.21
3	$\frac{\pi_\theta+0.1\log\pi_\theta}{\pi_{\theta_{\text{old}}}+0.1}$	$\frac{\pi_\theta+0.1}{\pi_{\theta_{\text{old}}}+0.1}(1-\pi_\theta)$	✗	Failed	28.45

6 Additional Analysis and Discussion

Due to space constraints, we present three supporting analyses in the appendix. Appendix B examines the generality of gradient conflict between low-probability external tokens and the rest of the batch. Appendix C discusses what information these tokens carry and why suppressing their gradients can still benefit learning. Appendix D reports hyperparameter sensitivity experiments for the probability threshold and low-probability token weight.

7 Related Work

Mixed-Policy RL for Language Models. Recent works have demonstrated that purely on-policy methods (e.g., GRPO) often struggle to overcome the reasoning bottlenecks of base models, especially when they fail to sample correct trajectories autonomously (Shao et al., 2024; Ahmadian et al., 2024; Yue et al., 2025). To address this, several mixed-policy methods have been proposed to incorporate high-quality external data. However, performing off-policy evaluation on external data is notoriously challenging in the LLM context. The true behavior policy π_β is typically inaccessible due to several practical factors: vocabulary mismatches between source and target models, the unknown mixture of models generating the dataset, and the inherent lack of a formal underlying generation probability distribution for human-annotated data.

Surrogate Objective Designs. To circumvent the unavailability of π_β, various heuristic approximations have been proposed. When adopting the naive approach of using the current policy $\pi_{\theta_{\text{old}}}$ as a proxy, the optimization quickly succumbs to severe instabilities. To mitigate this, **LUFFY** (Yan et al., 2025) treats the expert policy as deterministic ($\pi_\beta \to 1$), yielding an initial importance ratio $\hat{r} = \pi_\theta/\pi_\beta = \pi_\theta$. Upon observing severe entropy collapse during training with this ratio, they apply a bounding function $f(x) = x/(x+c)$ to stabilize the updates. Similarly, **RL-PLUS** (Dong et al., 2025) focuses on mitigating distributional mismatch with Multiple Importance Sampling (MIS). They model the behavior policy as a mixture of the current policy and an unknown expert policy, approximating the expert with a uniform distribution $\mathcal{U} \to 1$. Furthermore, they apply an exploration penalty $(1-sg(\pi_\theta))^{0.5}$ to promote learning on low-probability tokens, ultimately arriving at their final surrogate objective. While these methods empirically circumvent the failure of Vanilla IS, their derivations rely on different theoretical frameworks, independent heuristics. In contrast, our analysis transcends these disparate starting points by demonstrating that their success is fundamentally governed by a shared, implicit mechanism at the gradient level: the suppression of optimization updates from low-probability external tokens.

8 Conclusion

This work offers a new lens on mixed-policy RL: performance degradation is not primarily caused by fluctuations in importance-sampling ratios, but instead arises from gradient conflicts introduced by low-probability external tokens. The proposed gradient-suppression framework unifies several existing methods under a common view and provides a principled design space for future off-policy objectives. Based on this framework, we introduce Probability-aware Reweighting, a simple yet effective method that delivers strong performance across multiple models on both in-domain and out-of-domain datasets. Ultimately, our work provides a concrete foundation for future research on incorporating high-quality external data to push the capability boundaries of models.

AI USE STATEMENT

In this work, we used generative AI tools to assist in drafting portions of the manuscript and editing its language for clarity. We also used Qwen3-4B and Qwen3-0.6B to generate external reasoning trajectories from prompts in `POLARIS-Project/Polaris-Dataset-53K`, forming two datasets for the additional gradient-conflict analysis. We reviewed and revised all AI-assisted text and take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies an optimization method for mixed-policy reinforcement learning on mathematical reasoning tasks. It does not involve human subjects or crowdsourcing, and we are not aware of specific ethical concerns arising from the research presented here.

REPRODUCIBILITY STATEMENT

We provide runnable source code in the Supplementary Material. To reproduce our experiments, install the dependencies listed in `requirements.txt` and run the scripts in `exp_scripts/`. Section 5.1 describes the models, training data, mixed-policy protocol, baselines, benchmarks, and evaluation metrics used in our experiments. Appendix E.1 provides additional training details, and the remaining experimental appendix reports further results. The assumptions and derivations supporting the theoretical analysis are given in the main text and the theoretical appendix.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

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.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, et al. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*, 2025.

Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, et al. Rl-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. *arXiv preprint arXiv:2508.00222*, 2025.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850, 2024.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24158–24188, 2026.

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

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

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.

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

Tao Liu, Taiqiang Wu, Runming Yang, Shaoning Sun, Junjie Wang, and Yujiu Yang. Profit: Leveraging high-value signals in sft via probability-guided token selection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15383–15401, 2026.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, et al. Towards a unified view of large language model post-training. *arXiv preprint arXiv:2509.04419*, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

John Schulman, Sergey Levine, Philipp Moritz, Michael I Jordan, and Pieter Abbeel. Trust region policy optimization. In *International Conference on Machine Learning*, pp. 1889–1897, 2015.

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

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

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 115452–115486. Curran Associates, Inc., 2025. doi: 10.52202/085713-3850. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/a797c2d2e0c1fdabf4d1ab8cd0b465c6-Paper-Conference.pdf.

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.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*, 2025.

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

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.

Weihao Zeng, Yuzhen Huang, Wei Liu, Keqing He, Qian Liu, Zejun Ma, and Junxian He. 7b model and 8k examples: Emerging reasoning with reinforcement learning is both effective and efficient, 2025.

APPENDIX CONTENTS

A Limitations and Future Work **14**

B Generality of Gradient Conflict Between Low-Probability External Tokens and the Rest of the Batch **14**

B.1 Shared-Parameter Interaction Between Tokens 14

B.2 Directional Interaction at the LM Head 14

B.3 Conflict Across Data Sources and Trajectory Generators 15

B.4 Why Magnitude Suppression Mitigates Directional Conflict 15

C Discussion of Low-Probability External Tokens **16**

D Hyperparameter Sensitivity Ablations **17**

E Additional Experimental Details **17**

E.1 Training and Evaluation Details 17

E.2 Gradient-Source Intervention Details 18

E.3 Experimental Results Details 18

F Additional Analysis of Low-Probability External Tokens **18**

F.1 Diagnostic Setup and External Data 18

F.2 Frequent Token Strings in the Suppressed Subset 19

G Additional Theoretical Derivations **20**

G.1 Gradient Scaling Derivation 20

G.2 Proof of Gradient Weight Inequalities 21

A Limitations and Future Work

While our approach demonstrates promising results, several limitations remain. First, constrained by available GPU resources, we did not evaluate our framework on larger models (i.e. , more than 8B parameters). Scaling our method to larger architectures is a key direction for future work. Second, the current application of our mixed-policy is limited to the RLVR domain. Generalizing this method to other domains requires further investigation, particularly regarding how to optimally design and balance offline and on-policy signals in broader contexts.

B Generality of Gradient Conflict Between Low-Probability External Tokens and the Rest of the Batch

This section examines why low-probability external tokens can oppose the rest of a mixed-policy batch, whether the observed conflict persists across data sources, and why suppressing the corresponding gradient can improve optimization. The theoretical conditions below are sufficient conditions for conflict.

B.1 Shared-Parameter Interaction Between Tokens

Consider two token instances A and B with prefixes x_A, x_B and target token IDs y_A, y_B. All backbones evaluated in this work are dense autoregressive models. Although different prefixes produce different activations, their token losses backpropagate through the same attention, MLP, normalization, and LM-head parameters. The input embedding table is also shared, even though different token IDs initially access different rows. Parameter sharing permits cross-token interference but does not determine its sign. We can characterize that sign explicitly at the shared LM head.

B.2 Directional Interaction at the LM Head

Let $p_i \in \mathbb{R}^V$ be the full vocabulary distribution at token instance i, e_{y_i} the one-hot target vector, h_i the final hidden state, and $z_i = W_{\text{out}} h_i$ the LM-head logits. For the token objective $\mathcal{J}_i = \hat{r}_i(\theta)\hat{A}_i$, define $\gamma_i = \hat{A}_i \pi_i \, \partial \hat{r}_i / \partial \pi_i$ and the softmax residual $\delta_i = e_{y_i} - p_i$. The LM-head gradient and the inner product of two such gradients factorize as

$$\nabla_{W_{\text{out}}} \mathcal{J}_i = \gamma_i \delta_i h_i^\top, \tag{8}$$

$$\langle \nabla_{W_{\text{out}}} \mathcal{J}_A, \nabla_{W_{\text{out}}} \mathcal{J}_B \rangle_F = \gamma_A \gamma_B (\delta_A^\top \delta_B)(h_A^\top h_B). \tag{9}$$

The three factors capture the signs and weights induced by the advantages and surrogate objective, competition between target distributions in logit space, and similarity of contextual representations, respectively. In particular, two positive-advantage external tokens have $\gamma_A \gamma_B > 0$ when their surrogate derivatives have the same sign.

Expanding the residual inner product yields

$$\delta_A^\top \delta_B = \mathbb{I}[y_A = y_B] - p_A(y_B) - p_B(y_A) + p_A^\top p_B. \tag{10}$$

For the same target $y_A = y_B = y$, this equals $(1 - p_A(y))(1 - p_B(y)) + \sum_{k \neq y} p_A(k) p_B(k) \geq 0$. For distinct targets, it becomes $p_A^\top p_B - p_A(y_B) - p_B(y_A)$, which is negative whenever $p_A(y_B) + p_B(y_A) > p_A^\top p_B$.

This condition has a target–competitor interpretation. A low-probability external target y_A may compete with a token y_B that the model favors under prefix x_A. The gradient for A raises the logit of y_A while suppressing y_B. If y_B is the correct target at another token instance B, its gradient instead raises y_B. Thus, a sufficient condition for LM-head conflict is

$$\gamma_A \gamma_B > 0, \qquad h_A^\top h_B > 0, \qquad p_A(y_B) + p_B(y_A) > p_A^\top p_B, \tag{11}$$

under which the LM-head inner product in Eq. equation 9 is negative.

The empirical claim concerns aggregated gradients, not every low–high token pair. Let $\mathcal{B}_{\text{low}}^{\text{ext}}$ contain low-probability external tokens and $\mathcal{B}_{\text{rest}}$ all remaining tokens. Their exact LM-head interaction is

$$\left\langle G_{\text{low}}^W, G_{\text{rest}}^W \right\rangle_F = \sum_{A \in \mathcal{B}_{\text{low}}^{\text{ext}}} \sum_{B \in \mathcal{B}_{\text{rest}}} \gamma_A \gamma_B (\delta_A^\top \delta_B)(h_A^\top h_B). \tag{12}$$

Table 5: Gradient interaction between low-probability external tokens and the rest of the batch across two external-trajectory generators.

Dataset	$\|G_{\text{low}}\|_2/\|G_{\text{rest}}\|_2$	Average cosine
Polaris–Qwen3-4B	1.14	-0.38
Polaris–Qwen3-0.6B	1.30	-0.45

The aggregate gradients conflict when negative target–competitor contributions dominate positive contributions. For the full model, the inner product also includes all other disjoint parameter blocks:

$$\langle G_{\text{low}}, G_{\text{rest}} \rangle = \left\langle G^W_{\text{low}}, G^W_{\text{rest}} \right\rangle_F + \sum_{\ell \neq W} \left\langle G^{(\ell)}_{\text{low}}, G^{(\ell)}_{\text{rest}} \right\rangle. \tag{13}$$

If the LM-head contribution is $-D < 0$ and

$$D > \sum_{\ell \neq W} \left\| G^{(\ell)}_{\text{low}} \right\|_2 \left\| G^{(\ell)}_{\text{rest}} \right\|_2, \tag{14}$$

then the full-model inner product is negative by the Cauchy–Schwarz inequality. This is a sufficient condition, not an assertion that the LM head always dominates.

B.3 Conflict Across Data Sources and Trajectory Generators

To test whether the conflict in Figure 4 is specific to OpenR1, we constructed two additional datasets from `POLARIS-Project/Polaris-Dataset-53K`, using Qwen3-4B and Qwen3-0.6B to generate external trajectories. On Qwen2.5-Math-1.5B, we partitioned each batch into low-probability external tokens and the remaining tokens and measured their aggregated gradients. Both datasets exhibit negative average cosine similarity, while the low-probability gradient has a norm comparable to or larger than that of the rest of the batch (Table 5).

These results extend the directional and magnitude observations to different external data sources and trajectory generators. They establish persistence in the tested settings, without implying that every model or dataset exhibits the same conflict.

B.4 Why Magnitude Suppression Mitigates Directional Conflict

Probability-aware reweighting scales gradients according to token probability rather than computing each token's alignment. Let G_{low} and G_{rest} denote the two aggregated gradients, and let $w_1 \in [0, 1]$ be the weight on the low-probability subset. The resulting ascent direction is $G(w_1) = G_{\text{rest}} + w_1 G_{\text{low}}$. For a sufficiently small step $\theta^+ = \theta + \eta G(w_1)$, a first-order expansion of the remaining-token objective gives

$$\mathcal{J}_{\text{rest}}(\theta^+) - \mathcal{J}_{\text{rest}}(\theta) = \eta\big(\|G_{\text{rest}}\|_2^2 + w_1 \langle G_{\text{rest}}, G_{\text{low}} \rangle\big) + O(\eta^2). \tag{15}$$

When $\langle G_{\text{rest}}, G_{\text{low}} \rangle < 0$, the first-order improvement decreases as w_1 increases. Suppressing the conflicting component therefore improves first-order progress on the remaining-token objective. The effective gradient factor $\rho(\theta)$ in Eq. equation 4 explains the magnitude of this component; the cosine analysis identifies its opposing direction.

An individual low-probability token can still be positively aligned with G_{rest}, in which case down-weighting removes a locally helpful contribution. Token probability is a practical group-level proxy for the observed aggregate conflict, and $w_1 > 0$ can retain part of this signal. The interval-swap ablation in Table 3 provides a complementary intervention: in the evaluated settings, suppressing the low-probability interval permits stable training irrespective of the surrogate used in the high-probability interval, whereas retaining Vanilla IS in the low-probability interval leads to failure. Together with Eq. equation 15, this supports a causal optimization role for suppressing the conflicting subset within the evaluated setting.

C Discussion of Low-Probability External Tokens

High-quality external trajectories may contain useful information that the current policy is unlikely to sample, making gradient suppression on such tokens seem to discard precisely the novelty that mixed-policy RL aims to exploit. This apparent tension raises a question: *If low-probability external tokens may contain useful new reasoning, how can suppressing their gradients improve performance?* We investigate this question through experiments with Qwen2.5-Math-1.5B (see Appendix F.1 for details), complemented by further analyses and insights from prior work:

External tokens are only *relatively* lower-probability than internal tokens, rather than predominantly low-probability in an absolute sense. As shown in Table 6, external tokens have lower mean probabilities than internal ones (0.78 vs. 0.94 on Polaris; 0.80 vs. 0.95 on OpenR1), **yet most of them still lie in the $[0.9, 1]$ interval (65% and 71%)**.

Table 6: Mean token probability and the percentage of tokens in each probability interval under the current policy, averaged over the first 50 training steps on the two datasets.

Dataset	Tokens	Mean π_θ	$\pi_\theta < 0.1$	$\pi_\theta \in [0.1, 0.9)$	$\pi_\theta \in [0.9, 1]$
Polaris	External	0.78	9.8%	25.2%	65%
	Internal	0.94	1.9%	9.1%	89%
OpenR1	External	0.80	12.0%	17.0%	71%
	Internal	0.95	1.6%	6.4%	92%

On the one hand, the external trajectories are high-quality, correct responses with strong logical and semantic coherence, so many tokens are consistent with patterns already well captured by the model and therefore receive high generation probabilities. **On the other hand**, their generation by a different policy introduces a distributional mismatch that makes their probabilities relatively lower than those of internal tokens.

Fitting Relatively Lower-Probability External Tokens Accompanies Accuracy Gains. Figure 7 presents the model's sampling probabilities on external trajectories when tokens with $\pi_\theta < 0.05$ are clipped during training. As shown, these probabilities increase overall as training progresses. This suggests that the model acquires reasoning knowledge from external trajectories, alongside an overall improvement in training accuracy. This aligns with the "Mixed Learn" paradigm of mixed-policy RL, where external trajectories support exploitation and internal trajectories drive exploration. Specifically, the model effectively fits these relatively low-probability tokens, causing their probabilities to rise gradually, rather than being restricted to the clipped low-probability segment, which accounts for only a small fraction of external tokens.

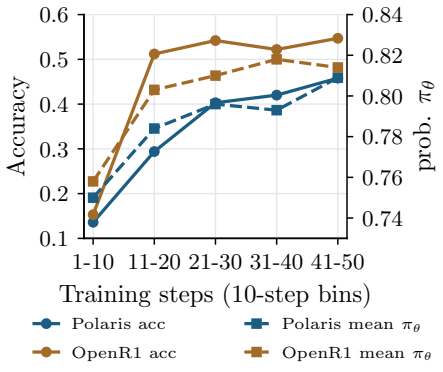


Figure 7: Training dynamics of mean external-token probability under the current policy and training accuracy.

Absolute low-probability tokens lack core reasoning We further examine the most frequent token strings of low-probability subset ($\pi_\theta < 0.05$) As shown in Appendix Table 12, 86.0% of their occurrences are function words, discourse markers, hedges, connectives tokens, including `Wait`, `maybe`, `Alternatively`, `However`, and so on. Many of these tokens are semantically replaceable and do not carry core reasoning content. Such fork-like tokens can be important in pure RL, where low-probability branches support internal exploration and help preserve diverse reasoning paths, as emphasized by *Beyond the 80/20 Rule* (Wang et al., 2025) and *Lp-Reg* (Huang et al., 2026).

However, the objective differs when exploiting external trajectories, including in mixed-policy RL. The primary goal is to learn the knowledge and reasoning structures encoded in high-quality external trajectories. Consistent with *ProFit* (Liu et al., 2026), such low-probability regions tend to contain

more semantically replaceable, non-core expressions, while higher-probability tokens more often carry the core semantic and logical structure. Suppressing gradients on absolute low-probability external tokens can therefore reduce less useful updates without discarding most of the reasoning signal.

D Hyperparameter Sensitivity Ablations

We examine the two hyperparameters of probability-aware reweighting in Eq. 6: the low-probability threshold c and its weight w_1, with $w_2 = 1$ throughout. Each sweep changes one parameter while holding the other fixed. We report the average score on the six mathematical reasoning benchmarks used in Section 5; these scores should be compared within each sweep. Tables 7–9 report the weight sweep on OpenR1 and threshold sweeps on OpenR1 and Polaris, respectively.

Table 7: Sensitivity to the low-probability weight w_1 on OpenR1 with Qwen2.5-Math-7B and fixed $c = 0.1$. Scores are averages over six mathematical benchmarks.

w_1	0	0.05	0.1	0.3	0.5
Avg. score	48.39	48.51	**48.97**	46.24	45.17

Table 8: Sensitivity to threshold c on OpenR1 with Qwen2.5-Math-7B and fixed $w_1 = 0$. Scores are averages over six mathematical benchmarks.

c	0.02	0.05	0.1	0.2	0.3
Avg. score	48.65	**50.65**	48.39	48.12	47.76

Table 9: Sensitivity to threshold c on Polaris with Qwen2.5-Math-1.5B and fixed $w_1 = 0$. Scores are averages over six mathematical benchmarks.

c	0.02	0.05	0.07	0.1
Avg. score	36.14	37.26	**39.58**	36.66

Across these sweeps, two patterns emerge. First, performance is approximately inverted-U-shaped as suppression strength changes: suppressing too few low-probability external tokens may leave harmful interference, whereas suppressing too many may remove useful signal. Second, the best tested threshold differs across configurations: $c = 0.05$ for Qwen2.5-Math-7B on OpenR1 and $c = 0.07$ for Qwen2.5-Math-1.5B on Polaris. Thus, the preferred hyperparameters should be tuned for the model and data rather than treated as fixed constants.

E Additional Experimental Details

E.1 Training and Evaluation Details

Training. In Elliott/Openr1-Math-46k-8192, each prompt is paired with one verified correct external reasoning trace shorter than 8192 tokens. For each prompt, we use a group size of 8 trajectories: pure on-policy methods use 8 on-policy rollouts, while mixed-policy methods use 7 on-policy rollouts and 1 off-policy trace. For our probability-aware reweighting method, we set $w_1 = 0$, $w_2 = 1$, and $c = 0.05$. This setup is consistent with the broader use of curated public math corpora such as NuminaMath (Li et al., 2024) in reasoning post-training. All methods are trained under the same RL pipeline for fair comparison. Following Dr.GRPO-style practice, we remove the standard-deviation normalization in group advantage computation and the token-length normalization in the policy loss, and we do not use an additional KL penalty. The maximum response length is set to 8192 tokens, the

Table 10: Step-wise change in loss on $\mathcal{B}_{\text{rest}}$ under updates from either token subset. Both conditions start from Qwen2.5-Math-1.5B and differ only in the tokens used for each update.

Step	$\mathcal{B}_{\text{rest}}$ only	$\mathcal{B}_{\text{low}}^{\text{ext}}$ only
1	-0.1158	$+0.0533$
2	-0.1137	$+0.0565$
3	-0.1152	$+0.0517$
4	-0.1050	$+0.0462$
5	-0.1119	$+0.0502$
Mean	$\mathbf{-0.1123}$	$\mathbf{+0.0516}$

rollout temperature is 1.0, and rewards are computed only by Math-Verify without auxiliary format or length bonuses.

Evaluation. For all evaluations, the decoding temperature is set to 0.6. Following LUFFY, we report avg@32 on AIME 2024, AIME 2025, and AMC, and pass@1 on all other benchmarks.

E.2 Gradient-Source Intervention Details

Both conditions start from Qwen2.5-Math-1.5B and use the same training data, optimizer, learning rate, and other settings. They differ only in which tokens contribute to the parameter update: either the remaining tokens $\mathcal{B}_{\text{rest}}$ or the low-probability external tokens $\mathcal{B}_{\text{low}}^{\text{ext}}$ ($\pi_\theta \in [0, 0.1)$). We evaluate the loss on $\mathcal{B}_{\text{rest}}$ after each of five consecutive steps. For update source S, each table entry is the change relative to the preceding step in that condition, $\Delta L_{\text{rest},t}^{S} = L_{\text{rest}}(\theta_t^S) - L_{\text{rest}}(\theta_{t-1}^S)$; negative values indicate improvement. The mean is computed over the five step-wise changes.

Every $\mathcal{B}_{\text{low}}^{\text{ext}}$-only update increases the loss on $\mathcal{B}_{\text{rest}}$, while every $\mathcal{B}_{\text{rest}}$-only update decreases it (Table 10). Thus, the conflicting low-probability external-token gradient directly worsens the remaining-token loss in this intervention.

E.3 Experimental Results Details

Table 11: Full cross-model comparison on six mathematical reasoning benchmarks. For each backbone other than Qwen2.5-Math-7B, we report four RL methods: GRPO, LUFFY, RL-PLUS, and ours.

Backbone	Method	AIME 24	AIME 25	AMC	MATH-500	Minerva	Olympiad	Avg.
Qwen2.5-Math-1.5B	GRPO	12.0	8.1	42.0	66.4	30.0	33.7	32.1
	LUFFY	17.1	15.8	48.5	80.4	30.1	44.9	39.5
	RL-PLUS	18.8	15.8	50.0	76.6	28.3	45.2	39.1
	Ours	**23.8**	**20.8**	**57.8**	**82.4**	**33.8**	**48.6**	**44.5**
Qwen2.5-0.5B	GRPO	0.0	0.0	5.1	17.0	3.3	4.2	4.9
	LUFFY	0.0	**0.4**	9.3	29.0	**5.5**	5.6	8.3
	RL-PLUS	0.0	0.0	**10.8**	**29.4**	4.4	6.5	**8.5**
	Ours	**0.4**	0.0	9.9	29.2	3.7	**7.4**	8.4
LLaMA-3.1-8B-Instruct	GRPO	2.9	0.0	19.3	39.4	19.9	13.6	15.9
	LUFFY	2.9	0.8	11.8	42.2	19.9	13.6	15.2
	RL-PLUS	**7.9**	0.8	20.8	48.6	22.4	19.0	19.9
	Ours	5.4	**3.8**	**24.7**	**53.2**	**23.5**	**22.5**	**22.2**

F Additional Analysis of Low-Probability External Tokens

F.1 Diagnostic Setup and External Data

The probability-distribution and training-dynamics diagnostics use Qwen2.5-Math-1.5B over the first 50 training steps with $c = 0.05$ and $w_1 = 0$. Alongside the OpenR1 data used in the main

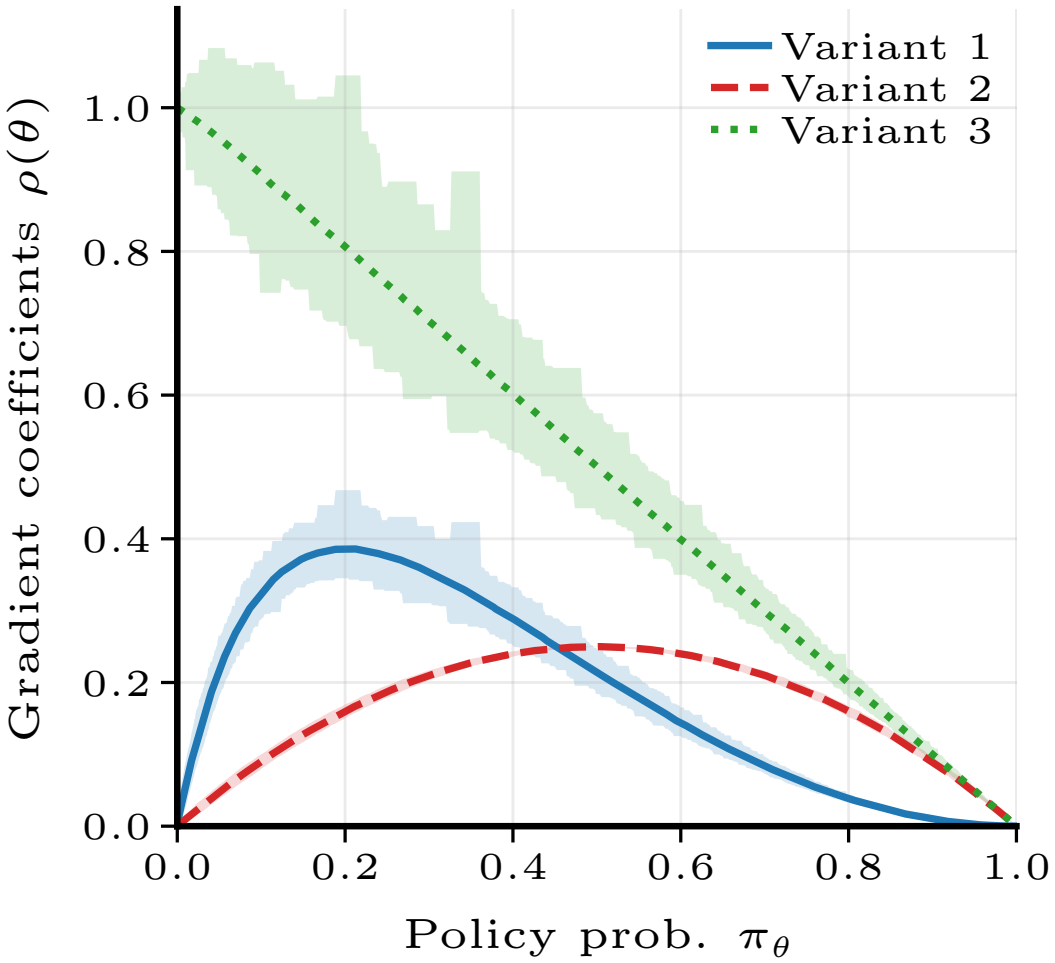


Figure 8: Training dynamics of the three variants. Variants 1 and 2 exhibit stable training and improvement, while Variant 3 fails.

experiments, we analyze 16K correct external trajectories derived from the Polaris math corpus[2]. For Polaris, eight candidates per prompt were generated with Qwen3-0.6B, then deduplicated against OpenR1 and filtered for correctness and the 8192-token length limit.

F.2 Frequent Token Strings in the Suppressed Subset

Table 12 records the 100 most frequent token strings among OpenR1 external tokens with $p < 0.05$. Bold token strings mark function words, discourse markers, hedges, connectives, punctuation, whitespace, and formatting markers identifiable from their surface form (80 of 100 strings, accounting for 86.0% of occurrences within the top 100).

Table 12: Most frequent token strings among OpenR1 external tokens with $p < 0.05$. Bold identifies linguistic expressions, punctuation, whitespace, and formatting markers by surface form. Counts and frequencies are reported from the token diagnostic.

Rank	Token string	Count	Frequency	Rank	Token string	Count	Frequency
1	**' ('**	52476	1.495%	51	**' However'**	12451	0.355%
2	**' the'**	50233	1.431%	52	**' all'**	12418	0.354%
3	**' '**	49140	1.400%	53	**' For'**	12415	0.354%
4	**' in'**	44028	1.254%	54	**' But'**	12228	0.348%
5	**' Wait'**	43737	1.246%	55	' compute'	12183	0.347%

[2] https://huggingface.co/datasets/POLARIS-Project/Polaris-Dataset-53K

Table 12 (continued)

Rank	Token string	Count	Frequency	Rank	Token string	Count	Frequency
6	','	36817	1.049%	56	' so'	11758	0.335%
7	' that'	35671	1.016%	57	'Therefore'	11622	0.331%
8	'.'	32401	0.923%	58	' x'	11503	0.328%
9	' maybe'	30927	0.881%	59	' here'	11327	0.323%
10	'Wait'	28860	0.822%	60	' no'	11318	0.322%
11	' is'	26963	0.768%	61	' -'	10702	0.305%
12	'1'	26593	0.758%	62	' there'	10430	0.297%
13	'.\n \n '	24864	0.708%	63	'0'	10388	0.296%
14	':'	24581	0.700%	64	' even'	10232	0.292%
15	' but'	24404	0.695%	65	' If'	10152	0.289%
16	'Alternatively'	24370	0.694%	66	' Because'	10142	0.289%
17	' a'	24176	0.689%	67	' two'	10132	0.289%
18	'?'	22493	0.641%	68	'3'	9915	0.283%
19	' if'	22332	0.636%	69	' actually'	9604	0.274%
20	':\n \n '	21433	0.611%	70	'4'	9559	0.272%
21	' Hmm'	20989	0.598%	71	' let'	9473	0.270%
22	' perhaps'	19926	0.568%	72	'First'	9367	0.267%
23	' when'	19716	0.562%	73	' it'	9249	0.264%
24	' for'	19031	0.542%	74	' total'	9178	0.262%
25	' Let'	18975	0.541%	75	' another'	8968	0.256%
26	' would'	18889	0.538%	76	' yes'	8843	0.252%
27	' with'	18686	0.532%	77	' where'	8806	0.251%
28	' Therefore'	18681	0.532%	78	' That'	8687	0.248%
29	' this'	18243	0.520%	79	' +'	8684	0.247%
30	' Then'	17726	0.505%	80	' at'	8665	0.247%
31	' each'	17322	0.494%	81	' see'	8525	0.243%
32	' and'	16698	0.476%	82	' Alternatively'	8425	0.240%
33	' since'	16043	0.457%	83	' both'	8314	0.237%
34	' which'	15765	0.449%	84	' first'	8303	0.237%
35	' The'	15627	0.445%	85	' not'	8177	0.233%
36	' as'	15133	0.431%	86	' just'	8130	0.232%
37	' we'	14849	0.423%	87	'5'	8047	0.229%
38	' from'	14478	0.413%	88	' ='	7838	0.223%
39	' can'	14388	0.410%	89	' because'	7829	0.223%
40	'But'	14238	0.406%	90	' possible'	7807	0.222%
41	' Since'	13881	0.396%	91	' Maybe'	7799	0.222%
42	' me'	13515	0.385%	92	' of'	7655	0.218%
43	' to'	13443	0.383%	93	' ['	7490	0.213%
44	'2'	13379	0.381%	94	' such'	7443	0.212%
45	' So'	13099	0.373%	95	'So'	7263	0.207%
46	'**'	12997	0.370%	96	' only'	7169	0.204%
47	'='	12870	0.367%	97	'Hmm'	7167	0.204%
48	'<th'	12800	0.365%	98	' how'	7155	0.204%
49	' I'	12700	0.362%	99	' then'	7081	0.202%
50	'Okay'	12529	0.357%	100	' might'	6982	0.199%

G Additional Theoretical Derivations

G.1 Gradient Scaling Derivation

This appendix provides the full mathematical derivation for the gradient scaling bounds presented in Section 3.2.2.

General gradient for $\mathcal{J} = \hat{r}(\theta) \cdot \hat{A}$. Consider a single sampled token a_i with autoregressive context s_i, let $z_i \in \mathbb{R}^V$ denote the logits over a vocabulary of size V, and let $p_\theta = \mathrm{softmax}(z_i)$ with n-th component $p_{\theta,n}$. Let $\hat{A} \in \mathbb{R}$ be the advantage signal (treated as constant with respect to the current parameters), and let $\hat{r}(\theta)$ be any differentiable surrogate ratio induced by π_θ. The single-token objective is $\mathcal{J}(\theta) = \hat{r}(\theta) \cdot \hat{A}$.

The standard softmax derivative gives, for any logit index n:

$$\frac{\partial \pi_\theta}{\partial z_{i,n}} = \pi_\theta\big(\mathbf{1}_{v_n = a_i} - p_{\theta,n}\big). \tag{16}$$

Applying the chain rule to $\mathcal{J} = \hat{r}(\theta) \cdot \hat{A}$, we obtain:

$$\frac{\partial \mathcal{J}}{\partial z_{i,n}} = \hat{A}\, \frac{\partial \hat{r}(\theta)}{\partial \pi_\theta}\, \pi_\theta\big(\mathbf{1}_{v_n = a_i} - p_{\theta,n}\big). \tag{17}$$

Writing this in vector form with $e_{a_i} \in \mathbb{R}^V$ (the one-hot vector for a_i):

$$\nabla_{z_i} \mathcal{J} = \hat{A}\, \pi_\theta\, \frac{\partial \hat{r}(\theta)}{\partial \pi_\theta}\, \big(e_{a_i} - p_\theta\big). \tag{18}$$

Defining the *gradient scaling function*

$$s(\theta) \triangleq \pi_\theta \frac{\partial \hat{r}(\theta)}{\partial \pi_\theta},$$

we can write $\nabla_{z_i} \mathcal{J} = \hat{A}\, s(\theta)\, \big(e_{a_i} - p_\theta\big)$.

Extension to any layer via the chain rule. Let $\theta^{(\ell)}$ denote the parameters of layer ℓ, and define the Jacobian $J_i^{(\ell)} \triangleq \frac{\partial z_i}{\partial \theta^{(\ell)}}$. The gradient at layer ℓ is:

$$\nabla_{\theta^{(\ell)}} \mathcal{J} = \hat{A}\, s(\theta)\, \big(J_i^{(\ell)}\big)^\top \big(e_{a_i} - p_\theta\big). \tag{19}$$

Norm scaling: the role of $|\hat{A}\, s(\theta)| \cdot (1 - \pi_\theta)$. The vector $\big(e_{a_i} - p_\theta\big)$ has component $(1 - \pi_\theta)$ at index a_i and $-p_{\theta,k}$ elsewhere. Its norm satisfies:

$$\big\|e_{a_i} - p_\theta\big\|_2^2 = \big(1 - \pi_\theta\big)^2 + \sum_{k \neq a_i} p_{\theta,k}^2. \tag{20}$$

Since $\sum_{k \neq a_i} p_{\theta,k} = 1 - \pi_\theta$, the Cauchy–Schwarz inequality yields:

$$\big(1 - \pi_\theta\big) \;\le\; \big\|e_{a_i} - p_\theta\big\|_2 \;\le\; \sqrt{2}\,\big(1 - \pi_\theta\big). \tag{21}$$

Define the *effective gradient factor*

$$\rho(\theta) \triangleq s(\theta)(1 - \pi_\theta).$$

Let $\sigma_{\min}(J_i^{(\ell)})$ and $\sigma_{\max}(J_i^{(\ell)})$ denote the minimum and maximum singular values of the Jacobian $J_i^{(\ell)}$. By applying the properties of matrix-vector products ($\sigma_{\min}(M)\|x\|_2 \le \|Mx\|_2 \le \sigma_{\max}(M)\|x\|_2$) to Eq. equation 19 alongside Eq. equation 21, we obtain the mathematically rigorous layer-wise bound:

$$\boxed{\sigma_{\min}(J_i^{(\ell)})\, |\hat{A}\, \rho(\theta)| \;\le\; \big\|\nabla_{\theta^{(\ell)}} \mathcal{J}\big\|_2 \;\le\; \sqrt{2}\, \sigma_{\max}(J_i^{(\ell)})\, |\hat{A}\, \rho(\theta)|.} \tag{22}$$

G.2 Proof of Gradient Weight Inequalities

In Section 4.2, we claimed that the effective gradient weights $\rho(\theta)$ of RL-PLUS and LUFFY are systematically smaller than that of Vanilla IS for low-probability tokens. We provide the formal proofs below. Recall the effective gradient weights derived in Table 1:

$$\rho_{\text{Vanilla}}(\theta) = \frac{\pi_\theta}{\pi_{\theta_{\text{old}}}}(1 - \pi_\theta),$$
$$\rho_{\text{LUFFY}}(\theta) = \frac{0.1\pi_\theta}{(\pi_\theta + 0.1)^2}(1 - \pi_\theta),$$
$$\rho_{\text{RL-PLUS}}(\theta) = \frac{4\pi_\theta}{3\pi_{\theta_{\text{old}}} + 1}(1 - \pi_\theta)^{1.5}.$$

We assume $\pi_\theta \in (0, 1)$ and $\pi_{\theta_{\text{old}}} \in (0, 1)$ throughout the analysis.

Proof for RL-PLUS. We want to prove $\rho_{\text{RL-PLUS}}(\theta) < \rho_{\text{Vanilla}}(\theta)$. Comparing the two expressions and dividing by the positive term $\pi_\theta(1 - \pi_\theta) > 0$, we get:

$$\frac{4}{3\pi_{\theta_{\text{old}}} + 1}(1 - \pi_\theta)^{0.5} < \frac{1}{\pi_{\theta_{\text{old}}}}$$
$$\iff \quad 4\pi_{\theta_{\text{old}}}\sqrt{1 - \pi_\theta} < 3\pi_{\theta_{\text{old}}} + 1.$$

Since $\pi_\theta > 0$, we have $\sqrt{1 - \pi_\theta} < 1$. Therefore, it strictly holds that:

$$4\pi_{\theta_{\text{old}}}\sqrt{1 - \pi_\theta} < 4\pi_{\theta_{\text{old}}}. \tag{23}$$

Furthermore, because $\pi_{\theta_{\text{old}}} < 1$, it is naturally true that:

$$4\pi_{\theta_{\text{old}}} = 3\pi_{\theta_{\text{old}}} + \pi_{\theta_{\text{old}}} < 3\pi_{\theta_{\text{old}}} + 1. \tag{24}$$

Combining these inequalities yields $4\pi_{\theta_{\text{old}}}\sqrt{1-\pi_\theta} < 3\pi_{\theta_{\text{old}}} + 1$. This concludes the proof that RL-PLUS strictly reduces the effective gradient weight globally compared to Vanilla IS.

Proof for LUFFY. We analyze the condition under which $\rho_{\text{LUFFY}}(\theta) < \rho_{\text{Vanilla}}(\theta)$. Comparing the two expressions and again dividing by $\pi_\theta(1-\pi_\theta) > 0$:

$$\frac{0.1}{(\pi_\theta + 0.1)^2} < \frac{1}{\pi_{\theta_{\text{old}}}}$$
$$\iff \quad \pi_{\theta_{\text{old}}} < 10(\pi_\theta + 0.1)^2.$$

This defines the exact mathematical condition under which LUFFY produces smaller gradient weights. For low-probability tokens ($\pi_\theta < 0.1$), the tightest upper bound is achieved as $\pi_\theta \to 0$, which yields:

$$10(\pi_\theta + 0.1)^2 > 10(0.1)^2 = 0.1. \tag{25}$$

As established in Section 3.2, severe gradient conflicts are driven by external tokens originating from the lowest probability interval ($\pi_{\theta_{\text{old}}} \in [0, 0.1)$). For these tokens, we naturally have $\pi_{\theta_{\text{old}}} < 0.1$. Therefore, it strictly holds that:

$$\pi_{\theta_{\text{old}}} < 0.1 < 10(\pi_\theta + 0.1)^2. \tag{26}$$

Thus, for the problematic low-probability external tokens, LUFFY consistently suppresses the gradient weights relative to Vanilla IS.