DaGRPO: Rectifying Gradient Conflict in Reasoning via Distinctiveness-Aware Group Relative Policy Optimization

Xuan Xie,[1] Xuan Wang,[2*] Wenjie Wang,[2] Shuai Chen,[2] Wei Lin[2]

[1]Tsinghua University
[2]Meituan

Abstract

The evolution of Large Language Models (LLMs) has catalyzed a paradigm shift from superficial instruction following to rigorous long-horizon reasoning. While Group Relative Policy Optimization (GRPO) has emerged as a pivotal mechanism for eliciting such post-training reasoning capabilities due to its exceptional performance, it remains plagued by significant training instability and poor sample efficiency. We theoretically identify the root cause of these issues as the lack of distinctiveness within on-policy rollouts: for routine queries, highly homogeneous samples induce destructive gradient conflicts; whereas for hard queries, the scarcity of valid positive samples results in ineffective optimization. To bridge this gap, we propose **D**istinctiveness-**a**ware **G**roup **R**elative **P**olicy **O**ptimization (**DaGRPO**). DaGRPO incorporates two core mechanisms: (1) *Sequence-level Gradient Rectification*, which utilizes fine-grained scoring to dynamically mask sample pairs with low distinctiveness, thereby eradicating gradient conflicts at the source; and (2) *Off-policy Data Augmentation*, which introduces high-quality anchors to recover training signals for challenging tasks. Extensive experiments across 9 mathematical reasoning and out-of-distribution (OOD) generalization benchmarks demonstrate that DaGRPO significantly surpasses existing SFT, GRPO, and hybrid baselines, achieving new state-of-the-art performance (e.g., a +4.7% average accuracy gain on math benchmarks). Furthermore, in-depth analysis confirms that DaGRPO effectively mitigates gradient explosion and accelerates the emergence of long-chain reasoning capabilities.

1 Introduction

The quest for Artificial General Intelligence (AGI) demands that Large Language Models (LLMs) transcend simple pattern matching to master complex reasoning. Unlike standard generation tasks, mathematical and logical reasoning requires models to navigate vast solution spaces and maintain rigorous logical consistency over long horizons—capabilities often referred to as "System 2" thinking (Li et al. 2025). Consequently, the post-training paradigm has shifted toward the elicitation of these complex Chain-of-Thought (CoT) capabilities (Jaech et al. 2024; Guo et al. 2025; Team et al. 2025). While Supervised Fine-Tuning (SFT) provides a foundation, it often fails to generalize beyond the training distribution or correct self-generated errors. This necessitates the adoption of Reinforcement Learning (RL) with Verifiable Rewards to facilitate robust self-improvement. In this landscape, Group Relative Policy Optimization (GRPO) (Shao et al. 2024) has anchored itself as a pivotal technique. By introducing Group Relative Advantage, GRPO obviates the dependency on a computationally expensive Value Function, significantly reducing memory footprint while effectively driving performance breakthroughs in tasks requiring multi-step logical deduction, such as mathematics and coding.

However, as an on-policy algorithm, GRPO inherently suffers from training instability—a challenge where off-policy counterparts like DPO (Rafailov et al. 2023) demonstrate superior robustness. Prior studies (Trott et al. 2019; Simoni et al. 2025) attribute this instability primarily to two factors:

1. *Reward Sparsity*: For challenging queries requiring long-horizon reasoning, the vast majority of on-policy rollouts yield low rewards, resulting in zero policy gradients that contribute negligibly to policy updates.
2. *Gradient Conflict*: As illustrated in Figure 1, training Qwen2.5-0.5B-Instruct with a group size of $G = 8$ reveals that for an identical prompt, distinct responses within the group exhibit divergent or even diametrically opposed gradient directions on specific parameters during back-propagation. This conflict causes valid gradient signals to cancel each other out or deviate from the optimal trajectory, thereby severely compromising both training stability and sample efficiency.

However, we argue that these phenomena are merely superficial symptoms. The root cause of GRPO's instability lies in the *lack of distinctiveness* among on-policy rollouts—a deficiency inextricably linked to GRPO's inherent optimization mechanism.. Prior literature (Wu et al. 2025) suggests that GRPO implicitly operates under a Contrastive Learning hypothesis: by normalizing intra-group rewards, the algorithm implicitly categorizes responses into positive (positive advantage) and negative (negative advantage) samples, aiming to boost the probability of the former while suppressing the latter. Crucially, the efficacy of such contrastive optimization hinges on the semantic distinctiveness of the sampled trajectories: effective gradient signals emerge only when high-quality positive anchors are contrasted against sufficiently distinct negative references.

*Corresponding author: wangxuan39@meituan.com
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

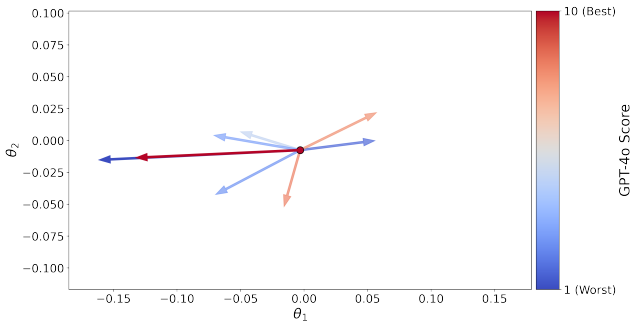

Figure 1: **Visualization of Gradient Conflict.** Using Qwen2.5-0.5B-Instruct with group size $G = 8$, we visualize the gradient directions of distinct responses for an identical prompt. The color spectrum (Red=High Score, Blue=Low Score) reflects GPT-4o evaluations.

In practice, however, datasets often provide only input prompts and final ground truths, lacking fine-grained supervision for intermediate deductive steps. Consequently, the on-policy sampling process encounters a twofold dilemma:

- *"Homogenized Conflict" in Easy Tasks*: For routine queries, the majority of on-policy reasoning trajectories generated by the model exhibit high homogeneity. This results in insufficient logical distinctiveness between positive and negative samples. Consequently, critical reasoning tokens shared across these samples simultaneously receive contradictory signals—reinforcing and suppressing—triggering severe gradient conflicts. As validated by GPT-4o scoring in Figure 1, these low-distinctiveness pairs act as noise rather than valid feedback for logical refinement.
- *"Benchmark Absence" in Hard Tasks*: For queries exceeding current capabilities, the model typically generates uniformly poor rollouts with rewards approaching zero. The absence of a high-quality positive anchor within the group renders the entire batch ineffective. In summary, homogenization in easy tasks and quality degradation in hard tasks collectively result in a severe lack of distinctiveness in the reasoning space.

To address these challenges, we propose **DaGRPO**, a novel framework that explicitly enhances sample distinctiveness to rectify gradient update directions. Specifically, DaGRPO incorporates two core mechanisms: 1. *Sequence-level Gradient Rectification*: We introduce a fine-grained scoring mechanism (via LLM-as-a-Judge) to quantify intra-group distinctiveness, dynamically masking "ambiguous" pairs that fall below a critical threshold. This eradicates gradient conflicts stemming from low-quality contrastive pairs at the source, ensuring that policy updates are driven exclusively by signals with clear discriminative value. 2. *Off-policy Data Augmentation*: To mitigate benchmark absence in hard tasks, we employ a hybrid sampling strategy that injects off-policy demonstrations from a stronger model (e.g., DeepSeek-R1) into the on-policy group as high-quality positive anchors. This mechanism not only recovers the gradient signals for challenging queries but also provides explicit logical guidance.

Extensive evaluations across 6 competition-level mathematical reasoning datasets and 3 out-of-distribution (OOD) benchmarks demonstrate the superiority of our approach. DaGRPO achieves state-of-the-art performance across all evaluated benchmarks. Specifically, on in-distribution math tasks, DaGRPO attains an average accuracy of 48.8%—a substantial +4.7% gain over the standard GRPO baseline—significantly outperforming strong competitors like Oat-Zero (Liu et al. 2025) and LUFFY (Yan et al. 2025). Furthermore, DaGRPO exhibits superior generalization capabilities and training stability on OOD tasks.
Our main contributions are summarized as follows:

- We provide a comprehensive analysis of gradient conflict and sample inefficiency in GRPO through the lens of contrastive learning. We identify the "*Lack of Distinctiveness*" within on-policy rollouts as the fundamental pathology driving optimization instability.
- We propose DaGRPO, a framework that synergizes two core mechanisms: *Sequence-level Gradient Rectification* to dynamically filter low-distinctiveness samples for conflict elimination, and *Off-policy Data Augmentation* to introduce high-quality anchors, thereby recovering training signals for challenging tasks.
- We conduct extensive experiments across 9 benchmarks. Empirical results confirm that DaGRPO effectively mitigates gradient conflicts while significantly pushing the performance boundaries and OOD generalization capabilities of LLMs in complex reasoning tasks.

2 Related Work

Reinforcement Learning for LLM Post-training. Reinforcement Learning (RL) has established itself as the cornerstone paradigm for post-training Large Language Models (LLMs), evolving from preference alignment to complex reasoning elicitation. Foundational work by Christiano et al. (2017) and Ouyang et al. (2022) (InstructGPT) solidified PPO (Schulman et al. 2017) as the standard for aligning models with human intent. While PPO remains effective, its dependency on Value Functions and Reward Models incurs substantial memory overhead, a bottleneck particularly acute in reasoning tasks that require extensive exploration. To mitigate these costs, Direct Preference Optimization (DPO) (Rafailov et al. 2023) emerged as a reward-free, off-policy alternative, optimizing directly on static preference data. Subsequent variants like IPO (Azar et al. 2024) and KTO (Ethayarajh et al. 2024) further refined this approach to address overfitting and data constraints. On-policy and off-policy methods present inherent trade-offs: On-policy methods (e.g., PPO) ensure consistency between sampling and policy distributions, fostering better exploration, yet suffer from poor sample efficiency. Conversely, off-policy methods like DPO optimize data utilization but remain susceptible to distribution shift.
Optimization and Analysis of GRPO. As the core algorithm of DeepSeek-R1 (Guo et al. 2025), Group Relative Policy Optimization (GRPO) replaces the traditional value function with group-wise relative advantage estimation, demonstrating a superior performance-compute ratio in mathematical and code reasoning tasks. However, its widespread adoption

has exposed inherent limitations, catalyzing a surge of targeted research. Addressing training instability, DAPO (Yu et al. 2025) identifies the risk of entropy collapse and proposes clip-higher and dynamic sampling strategies, GSPO (Zheng et al. 2025) improves the stability of MoE model training by elevating the optimization granularity from the token level to the sequence level. GTPO (Simoni et al. 2025) attempts to mitigate gradient conflict by selectively masking the gradient contributions of conflicting tokens. However, this strategy typically only identifies continuous conflicting tokens at the sequence boundaries (i.e., beginning and end), effectively ignoring the conflicts triggered by insufficient logical distinctiveness deep within the reasoning chains. Concurrently, Wu et al. (2025) reformulate the GRPO objective from a contrastive learning perspective, revealing its intrinsic mathematical connection with DPO. To further tackle the sample inefficiency of exclusive on-policy sampling, recent works have explored hybrid paradigms incorporating off-policy data. For instance, CHORD (Zhang et al. 2025) utilizes a global coefficient to dynamically balance the gradient weights between SFT data (off-policy) and GRPO samples (on-policy), while LUFFY (Yan et al. 2025) integrates offline trajectories directly into online groups. Different from these approaches, we identify the lack of sample distinctiveness as the root cause of optimization instability and propose a distinctiveness-aware mechanism that dynamically filters homogeneous conflicts and introduces high-quality contrastive references.

3 Preliminaries

Unlike PPO, which relies on a learned Value Function for advantage estimation, GRPO employs a group-based relative estimation strategy that significantly reduces computational overhead. Specifically, for each input prompt q, GRPO samples a group of G responses $\{o_i\}_{i=1}^G$ from the old policy $\pi_{\theta_{\text{old}}}$, where a reward model assigns a scalar reward $R(o_i)$ to each response. The advantage A_i for the i-th response is then derived via intra-group normalization:

$$A_i = \frac{R(o_i) - \text{mean}\left(\{R(o_i) \mid i = 1, \ldots, G\}\right)}{\text{std}\left(\{R(o_i) \mid i = 1, \ldots, G\}\right)} \tag{1}$$

Building on this, GRPO adopts a clipped surrogate objective similar to PPO, incorporating a KL divergence penalty term as a constraint:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{q\sim\mathcal{Q} \\ o_i\sim\pi_{\theta_{\text{old}}}}} \Bigg[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \Bigg(\\ \min\left(r_{i,t}(\theta)A_i, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)A_i\right) \\ - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \Bigg) \Bigg]. \end{aligned} \tag{2}$$

where, $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q, o_{i,<t})}$. Assuming the clipping operation is inactive and omitting the KL divergence penalty, the GRPO optimization objective reduces to:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathop{\mathbb{E}}_{\substack{q\sim\mathcal{Q} \\ o_i\sim\pi_{\theta_{\text{old}}}}} \left[\frac{1}{G}\sum_{i=1}^{G} \frac{1}{|o_i|}\sum_{t=1}^{|o_i|} r_{i,t}(\theta) \cdot A_i \right], \tag{3}$$

Consequently, the simplified policy gradient is derived as:

$$\nabla_\theta \tilde{\mathcal{J}}_{\text{GRPO}} = \frac{1}{G}\sum_{i=1}^{G} \frac{A_i}{|o_i|}\sum_{t=1}^{|o_i|} r_{i,t}(\theta) \cdot \nabla_\theta \log \pi_\theta(o_{i,t}|q, o_{i,<t}) \tag{4}$$

Notably, the advantage A_i is computed at the sequence-level (based on the final reward of the complete response o_i), yet during gradient updates, it is uniformly broadcasted as a scalar weight to every token within the sequence. This implies that every token receives an identical reinforcement signal, regardless of its actual contribution to the final reasoning outcome. As discussed subsequently, when sample distinctiveness is low, this "Granularity Mismatch" induces severe gradient conflicts, thereby compromising training stability. Since π_θ is initialized from $\pi_{\theta_{\text{old}}}$, the probability ratio $r_{i,t}(\theta) \approx 1$ during the initial phase of updates. To facilitate the theoretical analysis of gradient directions in subsequent sections, we adopt the approximation $r_{i,t}(\theta) \approx 1$, focusing primarily on directional conflicts arising from sign differences.

4 Analysis of Gradient Conflict in GRPO

Building upon the "granularity mismatch" identified in Section 3, we provide a rigorous analysis of how this mismatch precipitates the gradient conflict problem. The core issue stems from the fact that GRPO assigns sequence-level scalar rewards to fine-grained token-level updates, creating ambiguity in credit assignment.

Consider a group of G responses $\{o_i\}_{i=1}^G$ generated for a specific prompt q. Based on the sign of the estimated advantage A_i, we partition the responses into a positive set $G^+ = \{o_i \mid A_i > 0\}$ and a negative set $G^- = \{o_i \mid A_i < 0\}$. The gradient of the GRPO objective can be formally decomposed as:

$$\nabla_\theta \tilde{\mathcal{J}}_{\text{GRPO}} \approx \underbrace{\frac{1}{G}\sum_{o_j\in G^+} \frac{A_j}{|o_j|}\sum_{t=1}^{|o_j|} g_{j,t}}_{\text{Positive Component } \mathbf{v}^+} + \underbrace{\frac{1}{G}\sum_{o_k\in G^-} \frac{A_k}{|o_k|}\sum_{t=1}^{|o_k|} g_{k,t}}_{\text{Negative Component } \mathbf{v}^-} \tag{5}$$

where $g_{i,t} = \nabla_\theta \log \pi_\theta(o_{i,t}|q, o_{i,<t})$ represents the token-level log-probability gradient.

In reasoning tasks, the sampled responses within a group often exhibit significant structural homogeneity, as the policy π_θ tends to reuse established reasoning patterns. Let $\mathcal{S}_{\text{shared}}$ denote the set of state-action pairs (h, w)—where w is the token and h is the history context—that appear simultaneously in both positive and negative trajectories within the same group. For such a shared instance appearing in $o_j \in G^+$ and $o_k \in G^-$, the aggregated gradient update for this specific pattern is:

$$\nabla_\theta \tilde{\mathcal{J}}_{\text{shared}} \propto \left(\frac{A_j}{|o_j|} + \frac{A_k}{|o_k|} \right) \nabla_\theta \log \pi_\theta(w|h) \tag{6}$$

Since $A_j > 0$ and $A_k < 0$, these coefficients impose opposing optimization forces on the same parameters. This phenomenon constitutes a form of *destructive interference*

(Yu et al. 2020), which inevitably leads to severe training instability.

Specifically, when the magnitudes of the advantages are comparable (i.e., $|A_j|/|o_j| \approx |A_k|/|o_k|$), the effective gradient $\nabla_\theta \tilde{\mathcal{J}}_{\text{shared}}$ approaches zero. This neutralizes the learning signal for essential structural knowledge (e.g., reasoning templates or grammatical correctness), forcing the model to "unlearn" or ignore valid patterns simply because they co-occurred with incorrect reasoning steps in negative samples. On the other hand, these counteracting updates create a "tug-of-war" dynamic, causing the optimization trajectory to oscillate violently in the parameter space (zig-zagging) rather than following a consistent descent direction. This prevents the policy from settling into a stable optimum and significantly hinders convergence efficiency.

5 Method

As demonstrated in Section 4, the gradient conflict in GRPO fundamentally arises when shared tokens—representing identical logical steps—simultaneously receive conflicting high-magnitude advantage signals ($A_i > 0$, $A_k < 0$) despite lacking semantic distinctiveness. To mitigate this, we propose a Sequence-level Gradient Rectification mechanism. The core idea is to impose a "distinctiveness margin" on the policy update: we explicitly mask out sample pairs that fail to exhibit a sufficient quality gap, ensuring that the policy optimization is driven solely by samples with clear distinguishability.

Fine-grained Quality Assessment. While GRPO relies on binary rule-based rewards, these coarse signals fail to capture the nuance of reasoning quality. For each generated response o_i, we employ a LLM-as-a-Judge to compute a fine-grained scalar score $S(o_i) \in [1, 10]$. To ensure robust evaluation, we retrieve a high-quality demonstration o_{demo} (generated by DeepSeek-R1 and verified by math-verify) for the same prompt q as a reference. We prompt GPT-4o to score o_i against o_{demo} based on three dimensions: correctness of the final answer, validity of the reasoning process, and completeness of the logical steps:

$$S(o_i) = F_{\text{judge}}\left(q, o_i, o_{\text{demo}}\right), \quad i = 1, \ldots, G \tag{7}$$

Distinctiveness-aware Masking. To filter out ambiguous samples, we establish a dynamic gating mechanism based on the score margin between the positive set G^+ and the negative set G^-. Let $S_{\text{max}-} = \max\{S(o_k) \mid o_k \in G^-\}$ be the score of the "best" negative sample, and $S_{\text{min}+} = \min\{S(o_j) \mid o_j \in G^+\}$ be the score of the "worst" positive sample. We define rectification masks $\lambda^+(\cdot)$ and $\lambda^-(\cdot)$ using a distinctiveness threshold δ:

1. **Positive Mask $\lambda^+(j)$:** A positive sample o_j is retained only if its quality strictly exceeds the best negative sample by the margin δ.

$$\lambda^+(j) = \mathbb{I}\big[S(o_j) - S_{\text{max}-} \geqslant \delta\big] \tag{8}$$

2. **Negative Mask $\lambda^-(k)$:** A negative sample o_k is retained only if its quality is strictly lower than the worst positive sample by the margin δ.

$$\lambda^-(k) = \mathbb{I}\big[S_{\text{min}+} - S(o_k) \geqslant \delta\big] \tag{9}$$

where $\mathbb{I}[\cdot]$ is the indicator function. This mechanism effectively shields the gradient from "borderline" samples that cause confusion.

Incorporating these masks into the GRPO framework, the rectified policy gradient becomes:

$$\nabla_\theta \tilde{\mathcal{J}}_{\text{Rect}} = \frac{1}{G}\Bigg[\sum_{o_j \in G^+} \frac{\lambda^+(j) A_j}{|o_j|} \sum_{t=1}^{|o_j|} g_{j,t} + \sum_{o_k \in G^-} \frac{\lambda^-(k) A_k}{|o_k|} \sum_{t=1}^{|o_k|} g_{k,t}\Bigg] \tag{10}$$

By dynamically masking low-distinctiveness samples, we significantly alleviate gradient conflicts throughout the optimization process.

Off-policy Data Augmentation. For complex reasoning tasks, models often suffer from exploration collapse. Due to capability limitations, pure on-policy sampling may fail to generate valid reasoning paths, resulting in near-zero rewards for the entire group (i.e., $\forall i, R(o_i) \approx 0$). Consequently, relative advantages vanish ($A_i \approx 0$), rendering the gradient signal uninformative. Furthermore, even if the model occasionally hits the correct final answer via spurious correlations, the reasoning process might be flawed. Reinforcing such low-quality "lucky guesses" leads to overfitting and performance degradation.

Therefore, it is imperative to integrate high-quality off-policy demonstrations (generated by an external strong policy, e.g., DeepSeek-R1) to serve as explicit reasoning anchors. These anchors provide a clear logical reference, guiding the optimization when on-policy exploration fails. We employ a hybrid strategy similar to LUFFY (Yan et al. 2025) to construct a mixed sample group:

$$G_{\text{mix}} = G_{\text{on}} \cup G_{\text{off}} \tag{11}$$

Within this mixed group, we re-calculate the standardized advantages. Crucially, off-policy samples typically yield higher rewards, naturally serving as high-quality positive anchors to drive policy updates, while failed on-policy rollouts serve as negatives. The advantage is computed over the union:

$$A_i = \frac{R(o_i) - \text{mean}(\{R(o_i) \mid o_i \in G_{\text{mix}}\})}{\text{std}(\{R(o_i) \mid o_i \in G_{\text{mix}}\})} \tag{12}$$

As training progresses and the model acquires the capability to generate correct responses independently, on-policy samples naturally gain dominance in the positive set, thereby encouraging autonomous exploration.

Combining the Distinctiveness Masking with the Hybrid Update, the final gradient for DaGRPO aggregates contributions from the entire mixed batch, applying masks dynamically based on the unified score distribution:

$$\nabla_\theta \tilde{\mathcal{J}}_{\text{DaGRPO}} = \frac{1}{|G_{\text{mix}}|}\Bigg[\sum_{o_j \in G^+_{\text{mix}}} \frac{\lambda^+(j) A_j}{|o_j|} \sum_{t=1}^{|o_j|} g_{j,t} + \sum_{o_k \in G^-_{\text{mix}}} \frac{\lambda^-(k) A_k}{|o_k|} \sum_{t=1}^{|o_k|} g_{k,t}\Bigg] \tag{13}$$

Crucially, the distinctiveness-aware masking described in Eq. (13) applies to the *mixed* group G_{mix}. This provides a safety mechanism against noisy or low-quality off-policy data. If an injected anchor o_{off} is factually incorrect (low score) or only marginally better than the current policy's best response (score gap $< \delta$), the mask λ will automatically zero out its contribution. Consequently, DaGRPO is robust to the quality variance of external demonstrators, ensuring that the policy only learns from anchors that provide a strictly superior and unambiguous reasoning signal.

6 Experimental Setups

Training Datasets We utilize OpenR1-Math-46k-8192 (Yan et al. 2025) as our training dataset, a subset of OpenR1-Math-220k (Face 2025) comprising approximately 46,000 prompt-response pairs, with reference responses generated by the expert model DeepSeek-R1 (Guo et al. 2025). Before training, the dataset has been filtered to remove responses exceeding 8k tokens and responses verified as incorrect by math-verify. Within the DaGRPO framework, these demonstration responses serve a dual purpose: acting as reference answers for the LLM-as-a-Judge mechanism and as high-quality reasoning anchors for Off-policy Data Augmentation.

Benchmarks We evaluate our method on a comprehensive suite of 9 benchmarks, categorized into in-distribution mathematical reasoning and out-of-distribution (OOD) generalization tasks:

- **Mathematical Reasoning (6 Datasets):** To rigorously assess mathematical capabilities, we test on AIME 2024 (Zhang and Math-AI 2024), AIME 2025 (Zhang and Math-AI 2025), AMC (Li et al. 2024), MATH–500 (Hendrycks et al. 2021), Minerva (Lewkowycz et al. 2022), and OlympiadBench (He et al. 2024). Following standard practices for datasets with limited sample sizes (AIME and AMC), we report the avg@32 metric. For the remaining larger benchmarks, we report pass@1.
- **OOD Generalization (3 Datasets):** To verify that our method improves general reasoning beyond math without catastrophic forgetting, we evaluate on ARC–c (Clark et al. 2018), GPQA–Diamond (Rein et al. 2024) (expert–level biology, physics, and chemistry), and MMLU–Pro (Wang et al. 2024) (interdisciplinary complex reasoning across 57+ subjects).
- **Evaluation Settings:** For all evaluations, we set the sampling temperature to 0.6, top–p to 1.0, and the maximum response length to 8,192 tokens.

Baselines To demonstrate the superiority of DaGRPO, we used Qwen2.5-Math-7B (Yang et al. 2024) as the base model and compared it with various strong baseline methods. The baselines are classified as follows: **Supervised Fine-Tuning (SFT):** Standard supervised fine-tuning using prompts and expert responses in the OpenR1-Math-46k dataset. **On-policy RL methods:** Original GRPO training based on binary rule rewards. **Hybrid SFT+RL methods**, including: SFT-GRPO: A two-stage method, first performing SFT fine-tuning, and then GRPO reinforcement learning optimization; LUFFY (Yan et al. 2025): A hybrid policy optimization method that mixes On-policy and Off-policy data, which we reproduced under the same settings. **Previous RLVR methods:** We also compared previous Reinforcement Learning with Verifiable Rewards (RLVR) methods, including SimpleRL-Zero (Zeng et al. 2025), OpenReasoner-Zero (Hu et al. 2025), PRIME-Zero (Cui et al. 2025), and Oat-Zero (Liu et al. 2025).

Implementation Details Following the configuration of Dr.GRPO (Liu et al. 2025), we removed the length normalization and standard deviation normalization terms in advantage estimation, and removed the KL divergence constraint to encourage the model to fully explore. For each prompt, we sample $G = 8$ rollouts, with sampling temperature set to 1.0, top-p set to 1.0, and maximum length of 8,192 tokens. We use a global rollout batch size of 128 and an update batch size of 64. The learning rate is set to 1×10^{-6}. All experiments are conducted for 500 training steps. We adopt a strict binary reward mechanism: the reward is 1 if the final answer is correct, otherwise 0. No additional format rewards or length rewards are used. On this basis, we prompt GPT-4o to perform fine-grained scoring for each rollout, and based on experience, we set the distinctiveness threshold δ to 3.

7 Experimental Results

7.1 Main Results

Superior Performance on In–Distribution Reasoning. As presented in Table 1, DaGRPO establishes a new state–of–the–art performance among the compared methods. On the aggregate of 6 mathematical reasoning benchmarks, DaGRPO achieves the highest average accuracy of **48.8%**, significantly outperforming the standard GRPO baseline (44.1%) by a substantial margin of **+4.7%**. Furthermore, DaGRPO consistently surpasses competitive RLVR baselines, such as Oat–Zero (43.7%), and outperforms the hybrid strategy LUFFY (47.1%), demonstrating the effectiveness of our distinctiveness–aware optimization framework over simple data mixing strategies.

Efficacy of Gradient Rectification (Mechanism 1). The results of the DaGRPO (w/o Off-policy) variant provide strong empirical support for our gradient rectification mechanism. Even without the injection of off–policy anchors, simply filtering out low–distinctiveness samples yields an average ID score of **45.9%**, which already surpasses both the SFT (44.1%) and GRPO (44.1%) baselines. This validates our theoretical analysis in Section 4: mitigating gradient conflicts from ambiguous samples is sufficient to stabilize training and unlock performance gains inherent in on–policy exploration.

Critical Role of Anchors in Hard Tasks (Mechanism 2). The contribution of the Off–policy Anchor Injection becomes evident in challenging tasks. Comparing the full DaGRPO with the variant without off-policy data, we observe significant improvements on difficult benchmarks such as **AIME 2024** (+6.3% gain from 23.2% to 29.5%) and **OlympiadBench** (+4.6% gain from 51.2% to 55.8%). This confirms that in scenarios where reward sparsity limits on–policy exploration, the high–quality anchors successfully bridge the gap, guiding the model to master complex reasoning paths that are otherwise inaccessible.

Table 1: Main results on in-distribution mathematical reasoning and out-of-distribution generalization benchmarks. We compare DaGRPO with strong baselines including standard SFT, On-policy RL (GRPO), Hybrid methods (SFT-GRPO, LUFFY), and recent RLVR approaches (SimpleRL-Zero, Oat-Zero, etc.). "DaGRPO (w/o Off-policy)" denotes our variant equipped solely with the distinctiveness-aware gradient rectification mechanism. The best and second-best results are highlighted in **bold** and underlined, respectively. * means the results are taken from the corresponding paper.

Model	In-Distribution Performance						Out-of-Distribution Performance			
	AIME 24/25	AMC	MATH-500	Minerva	Olympiad	Avg	ARC-c	GPQA	MMLU-Pro	Avg
Qwen2.5-Math	10.4/3.3	29.8	47.6	9.1	16.1	19.4	10.0	12.6	14.4	12.3
Qwen2.5-Math-Instruct	12.5/10.2	48.5	80.4	32.7	41.0	37.6	70.3	24.7	34.1	43.0
SimpleRL-Zero*	27.0/6.8	54.9	76.0	25.0	34.7	37.4	30.2	23.2	34.5	29.3
OpenReasoner-Zero*	16.5/15.0	52.1	82.4	33.1	47.1	41.0	66.2	29.8	**58.7**	51.6
PRIME-Zero*	17.0/12.8	54.0	81.4	39.0	40.3	40.7	73.3	18.2	32.7	41.4
Oat-Zero*	**33.4**/11.9	61.2	78.0	34.6	43.4	43.7	70.1	23.7	41.7	45.2
SFT	22.2/**22.3**	52.8	82.6	**40.8**	43.7	44.1	75.2	24.7	42.7	47.5
GRPO	22.7/14.8	58.5	81.8	40.0	46.7	44.1	82.2	40.4	49.2	57.2
SFT → GRPO	25.7/21.6	62.2	84.6	38.2	46.8	46.5	67.7	30.8	40.5	49.3
LUFFY	23.0/20.8	**63.4**	85.4	35.6	54.3	47.1	81.2	38.8	52.8	57.6
DaGRPO (w/o Off-policy)	23.2/16.1	63.1	85.0	36.7	51.2	45.9	81.3	**45.9**	52.4	**59.9**
DaGRPO	29.5/20.9	63.0	**86.0**	37.5	**55.8**	**48.8**	**83.1**	40.9	53.0	59.0

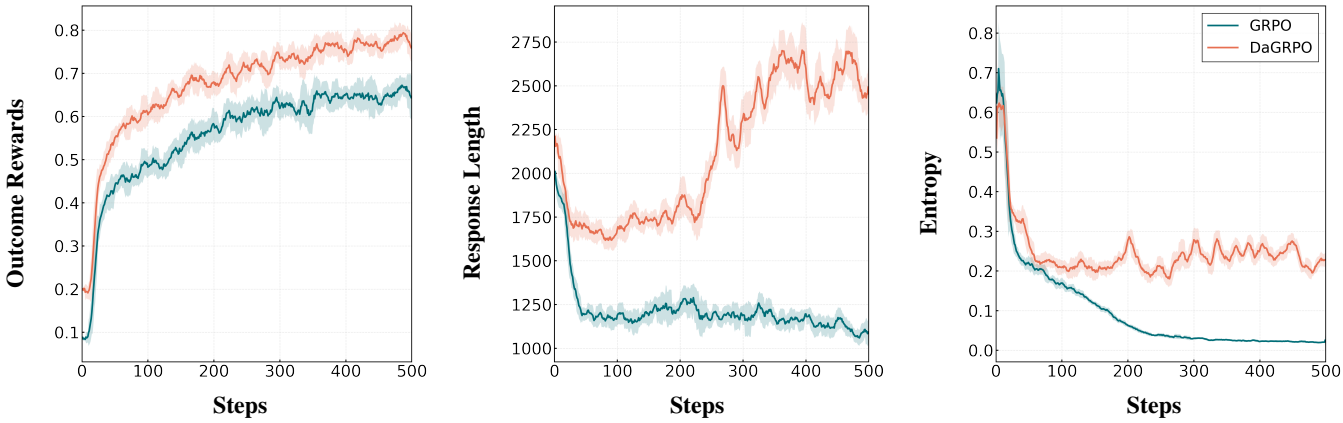

Figure 2: **Training Dynamics.** Comparison of training curves between DaGRPO and baseline GRPO over 500 steps. **Left:** Average outcome rewards per step. **Middle:** Average response length (tokens). **Right:** Policy entropy.

Robust Out–of–Distribution Generalization. A key observation is that DaGRPO significantly outperforms baselines on OOD tasks (Avg 59.9% vs. GRPO 57.2%). We attribute this to the preservation of policy entropy shown in Figure 2 (Right). Standard GRPO tends to collapse into narrow, repetitive solution templates (low entropy) to maximize rewards on in-distribution math problems, which harms generalization. By masking out ambiguous samples that encourage "safe" but mediocre responses, DaGRPO maintains a higher exploration entropy, preventing overfitting to specific prompt-response patterns and preserving the model's broad reasoning capabilities for unseen domains like biology (GPQA) and science (ARC-c).

7.2 Training Dynamics Analysis

To delve deeper into the optimization stability and emergent behaviors of DaGRPO, we visualize the training dynamics in Figure 2. The comparison reveals three critical advantages of our method:

Enhanced Sample Efficiency and Convergence. As shown in the outcome reward curve (Figure 2 Left), DaGRPO exhibits a significantly faster convergence rate and a higher asymptotic performance compared to GRPO. The initial "jump–start" (Step 0–50) can be attributed to the injection of high–quality off–policy anchors, which provide immediate, valid gradient signals. Throughout the training, DaGRPO maintains a consistent performance gap over GRPO, validating that our gradient rectification mechanism effectively filters out noisy gradients, allowing the model to focus on learning from high–distinctiveness signals.

Emergence of Long–Chain Reasoning. The response length evolution (Figure 2 Middle) presents the most striking difference. While GRPO suffers from "length collapse"—quickly converging to shorter, potentially shortcut–based responses (∼1200 tokens)— DaGRPO exhibits a distinct "U–shaped" trajectory. After an initial adjustment, DaGRPO undergoes a

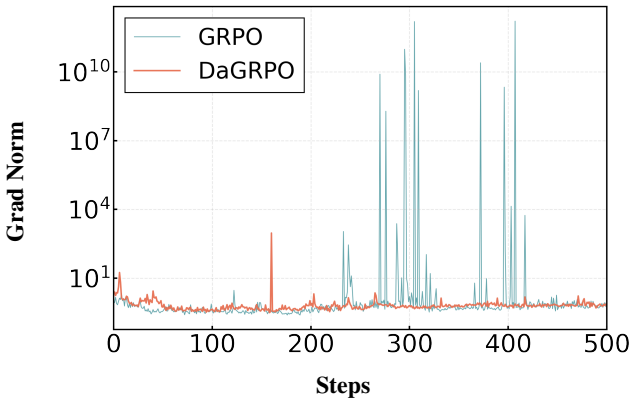

Figure 3: Evolution of Gradient Norms throughout the training process. The y-axis is plotted on a logarithmic scale.

phase transition around Step 200, where the response length surges to over 2500 tokens. Suggesting that DaGRPO encourages the model to engage in deeper thinking processes. Instead of taking heuristic shortcuts, the model learns to expand its reasoning steps to ensure logical consistency.

Mitigation of Mode Collapse. The entropy plot (Figure 2 Right) explains the mechanism behind the aforementioned behaviors. GRPO's entropy rapidly collapses to near–zero (< 0.05), indicating severe **mode collapse** where the policy becomes deterministic and ceases exploration. In contrast, DaGRPO maintains a healthy level of entropy ($\sim$0.25) throughout training. By masking out ambiguous samples that often drive the policy towards safe but suboptimal modes, DaGRPO sustains a "Buffer Zone" for exploration, enabling the model to continuously discover and refine diverse reasoning paths.

Optimization Stability. Figure 3 offers visual proof of DaGRPO's robustness via gradient norm analysis. Standard GRPO suffers from severe numerical instability, frequently exhibiting gradient explosions of magnitude 10^{10}, which directly corroborates the destructive impact of gradient conflicts. Conversely, DaGRPO maintains a smooth and bounded gradient flow by filtering out low-distinctiveness samples. This "denoising" mechanism effectively eliminates destructive interference, ensuring a highly stable optimization process.

8 Conclusion

In this paper, we addressed the challenges of gradient conflict and low sample efficiency in the post-training of LLM reasoning capabilities by proposing Distinctiveness-aware Group Relative Policy Optimization (DaGRPO). Through rigorous theoretical analysis, we identified that the root cause of instability in GRPO lies in the lack of semantic distinctiveness within self-sampled reasoning trajectories, which leads to destructive gradient interference. By introducing a sequence-level gradient rectification mechanism to filter low-distinctiveness samples and integrating off-policy data augmentation, DaGRPO successfully smoothes the optimization trajectory and recalls training signals for hard tasks. Empirical results demonstrate that DaGRPO achieves state-of-the-art performance across multiple mathematical reasoning and out-of-distribution benchmarks, exhibiting superior training stability and the emergence of long-chain reasoning.

Despite its promising performance, DaGRPO has certain limitations. First, the gradient rectification mechanism relies on an LLM-as-a-Judge for fine-grained scoring, which introduces additional computational overhead and latency during the training phase. However, this is not a permanent bottleneck; future work can distill the judge's scoring capability into a lightweight, regression-based Reward Model (RM) to replace the online API calls, significantly reducing training latency while maintaining distinctiveness sensitivity. Second, the off-policy augmentation component depends on high-quality external expert data (e.g., from DeepSeek-R1) acting as anchors. In scenarios where strong prior models are unavailable, generating high-distinctiveness sample pairs solely through self-bootstrapping remains a challenging open problem. We leave these explorations for future work to further advance efficient and autonomous reasoning alignment.

References

Azar, M. G.; Guo, Z. D.; Piot, B.; Munos, R.; Rowland, M.; Valko, M.; and Calandriello, D. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, 4447–4455. PMLR.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cui, G.; Yuan, L.; Wang, Z.; Wang, H.; Zhang, Y.; Chen, J.; Li, W.; He, B.; Fan, Y.; Yu, T.; et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Ethayarajh, K.; Xu, W.; Muennighoff, N.; Jurafsky, D.; and Kiela, D. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Face, H. 2025. Open r1: A fully open reproduction of deepseek-r1.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; et al. 2024. 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)*, 3828–3850.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring

mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Hu, J.; Zhang, Y.; Han, Q.; Jiang, D.; Zhang, X.; and Shum, H.-Y. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; et al. 2022. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35: 3843–3857.

Li, J.; Beeching, E.; Tunstall, L.; Lipkin, B.; Soletskyi, R.; Huang, S.; Rasul, K.; Yu, L.; Jiang, A. Q.; Shen, Z.; et al. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9): 9.

Li, Z.-Z.; Zhang, D.; Zhang, M.-L.; Zhang, J.; Liu, Z.; Yao, Y.; Xu, H.; Zheng, J.; Wang, P.-J.; Chen, X.; et al. 2025. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Liu, Z.; Chen, C.; Li, W.; Qi, P.; Pang, T.; Du, C.; Lee, W. S.; and Lin, M. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741.

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Simoni, M.; Fontana, A.; Rossolini, G.; and Saracino, A. 2025. Gtpo: Trajectory-based policy optimization in large language models. *arXiv preprint arXiv:2508.03772*.

Team, K.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Trott, A.; Zheng, S.; Xiong, C.; and Socher, R. 2019. Keeping your distance: Solving sparse reward tasks using self-balancing shaped rewards. *Advances in Neural Information Processing Systems*, 32.

Wang, Y.; Ma, X.; Zhang, G.; Ni, Y.; Chandra, A.; Guo, S.; Ren, W.; Arulraj, A.; He, X.; Jiang, Z.; et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290.

Wu, Y.; Ma, L.; Ding, L.; Li, M.; Wang, X.; Chen, K.; Su, Z.; Zhang, Z.; Huang, C.; Zhang, Y.; et al. 2025. It Takes Two: Your GRPO Is Secretly DPO. *arXiv preprint arXiv:2510.00977*.

Yan, J.; Li, Y.; Hu, Z.; Wang, Z.; Cui, G.; Qu, X.; Cheng, Y.; and Zhang, Y. 2025. Learning to reason under off-policy guidance. *arXiv preprint arXiv:2504.14945*.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; et al. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yu, T.; Kumar, S.; Gupta, A.; Levine, S.; Hausman, K.; and Finn, C. 2020. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836.

Zeng, W.; Huang, Y.; Liu, Q.; Liu, W.; He, K.; Ma, Z.; and He, J. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Zhang, W.; Xie, Y.; Sun, Y.; Chen, Y.; Wang, G.; Li, Y.; Ding, B.; and Zhou, J. 2025. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *arXiv preprint arXiv:2508.11408*.

Zhang, Y.; and Math-AI, T. 2024. American Invitational Mathematics Examination (AIME) 2024.

Zhang, Y.; and Math-AI, T. 2025. American Invitational Mathematics Examination (AIME) 2025.

Zheng, C.; Liu, S.; Li, M.; Chen, X.-H.; Yu, B.; Gao, C.; Dang, K.; Liu, Y.; Men, R.; Yang, A.; et al. 2025. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*.