LEARNING FROM CORRECTNESS TRENDS: SAMPLE-WISE HISTORICAL ADVANTAGE SHAPING FOR MATHEMATICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning is becoming an important approach for improving the reasoning ability of large language models. Group Relative Policy Optimization (GRPO) avoids the training cost of an additional value network through group-wise relative rewards and has been widely used for reasoning model training. However,the standard GRPO constructs its optimization signal primarily from the current candidate group and does not retain how the same training sample has evolved across previous visits. Consequently, candidate groups with similar current relative rewards receive similar update strengths even when their sample-wise learning trajectories differ. In this paper, we propose TrendGRPO, a sample-wise historical advantage shaping method that enhances within-group comparison with cross-iteration information. TrendGRPO maintains a correctness history for each sample across repeated visits and uses the deviation between the current group accuracy and historical accuracy to asymmetrically reshape positive and negative normalized advantages. We evaluate TrendGRPO on multiple base models and mathematical reasoning benchmarks. Ablations and training diagnostics support historical sample-wise conditioning in advantage shaping.

1 INTRODUCTION

The performance of large language models(LLMs) in complex reasoning and mathematical reasoning tasks has become one of the key indicators to evaluate the capability of artificial intelligence systems Wang et al. (2026b). Chain-of-thought (CoT) reasoning Barua et al. (2026), as a pioneering approach, enables LLMs to decompose complex problems into intermediate reasoning steps and has significantly improved their performance on arithmetic and symbolic reasoning benchmarks. However, prompt engineering alone is insufficient to systematically improve the reasoning limits of a base model Hu et al. (2025). Therefore, supervised fine-tuning (SFT) has been widely used to build basic reasoning abilities. Nevertheless, SFT Zelikman et al. (2022) mainly learns from static question-answer pairs and lacks the ability to explore and correct reasoning paths through dynamic trial and error, making it difficult to handle complex reasoning challenges in open-domain settings Uesato et al. (2022); Lightman et al. (2024). To address this limitation, reinforcement learning from human feedback has been introduced into language model training, allowing models to optimize their generated outputs through reward signals rather than simply imitating static data distributions Ouyang et al. (2022). Among these methods, Proximal Policy Optimization (PPO) has been widely adopted because it limits the difference between old and new policies by clipping policy updates, helping to improve training stability Schulman et al. (2017). However, PPO requires an additional value network, leading to extra computational cost and possible training instability. More recently, Group Relative Policy Optimization (GRPO) Shao et al. (2024); Mroueh et al. (2026) samples multiple responses for the same prompt and performs group-wise reward normalization, removing the need for an additional value network. As a result, GRPO has become an efficient approach for improving the reasoning ability of LLMs.

Although GRPO has shown strong performance in reasoning model training Shao et al. (2024); Guo et al. (2025), its advantage estimation is based on relative rewards within the current candidate group Shao et al. (2024); Mroueh et al. (2026). Therefore, it does not explicitly maintain a persistent state of how the same training sample has evolved over previous visits. Consequently, two

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.

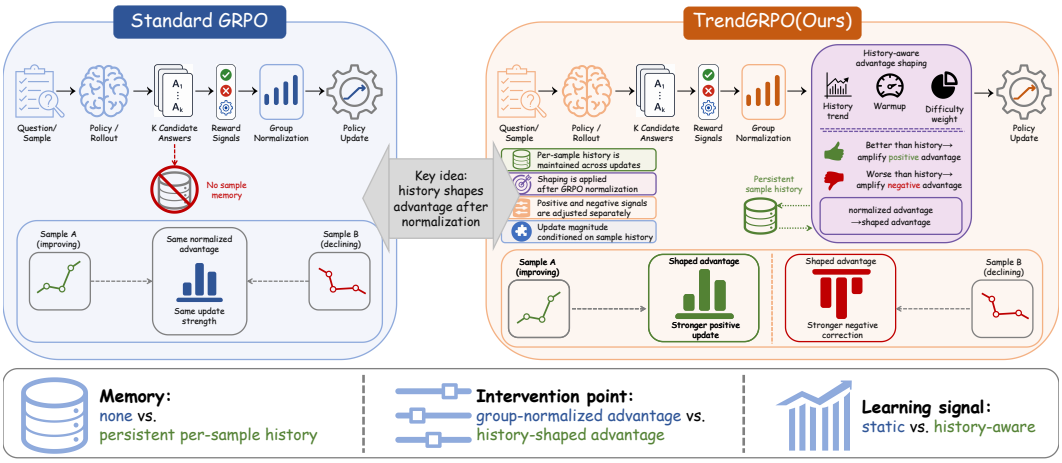


Figure 1: Core difference between standard GRPO and TrendGRPO. TrendGRPO augments within-group comparison with persistent sample-wise temporal comparison and asymmetrically conditions the magnitudes of positive and negative advantages.

samples with identical current-group reward patterns receive the same normalized advantages, even if one is improving relative to its history, while the other is deteriorating. In this sense, the optimization signal of standard GRPO is temporally myopic at the sample level. We focus on optimization regimes with repeated sample visits, where the accumulated correctness history can serve as an additional conditioning signal for advantage construction. To this end, we propose TrendGRPO, which maintains persistent sample-wise correctness statistics and compares the current candidate-group performance with the historical performance of the same sample. The resulting current-to-historical deviation is applied after GRPO normalization to asymmetrically reshape positive and negative advantages. This design introduces the temporal comparison in the sample-wise way as an additional conditioning axis for the GRPO advantages. Because history-conditioned shaping is performed after group-wise normalization, it leaves rollout generation and the clipped policy objective unchanged and does not require no additional learned reward model. Its low integration overhead is therefore a practical consequence of the post-normalization design.

Our key contributions are as follows:

1. We identify sample-wise temporal comparison as a complementary optimization signal for GRPO. Unlike the standard within-group comparison, it captures how the same training sample evolves over repeated visits.
2. We propose TrendGRPO, which maintains persistent sample-level correctness states and performs asymmetric post-normalization shaping of positive and negative advantages according to current-versus-historical performance. This design preserves the candidate ordering established by GRPO while making the update magnitude history conditioned.
3. We evaluate TrendGRPO on three base models and multiple mathematical reasoning benchmarks. Controlled comparisons, ablations, training dynamics, sensitivity analysis, and failure-streak diagnostics support the effectiveness and activation behavior of the proposed historical signal.

2 Related Work

2.1 GRPO-Based Advantage Optimization

Group Relative Policy Optimization (GRPO) Shao et al. (2024) is a critic-free variant of Proximal Policy Optimization that estimates advantages by standardizing rewards within a group of responses sampled for the same prompt Mroueh et al. (2026). This group-relative formulation has become widely used in reinforcement learning for mathematical reasoning, but its optimization sig-

nal is constructed primarily from information available in the current rollout. Recent studies have refined GRPO-based optimization from several perspectives. DAPO Yu et al. (2026) adopts dynamic sampling to reduce training on candidate groups with insufficient reward variation. OAR Li et al. (2026) improves credit assignment through outcome-grounded token-level advantage reshaping, while DVAO Jiang et al. (2026) dynamically balances optimization objectives in multi-reward settings. AVSPO He et al. (2026) diagnoses and mitigates advantage collapse, and SA-MRPO Wang et al. (2026c) adjusts multi-reward weights according to objective saturation. These methods primarily refine the optimization signal using the current rollout, reward structure, sampling behavior, or token-level credit assignment. TrendGRPO introduces a complementary conditioning axis by maintaining the persistent correctness history of each training sample across optimization steps. It compares the current candidate-group accuracy with the historical accuracy of the same sample and applies the resulting current-to-historical deviation after GRPO normalization to asymmetrically reshape positive and negative advantages. Accordingly, TrendGRPO augments GRPO's within-group comparison with sample-wise temporal comparison.

2.2 Conditioning Signals for GRPO Optimization

Existing GRPO extensions also use training information from different sources to adjust the optimization signal. GPRS Zhu et al. (2025) replaces exact numerical rewards with within-group preference comparisons, providing more robust relative reward signals. GRPO-LEAD Zhang & Zuo (2025) reweights the advantage according to estimated problem difficulty to place greater emphasis on hard samples. PASS Wang et al. (2026a) introduces process-level supervision and reorganizes advantages from different sources, while GiGPO Feng et al. (2026) modifies the sampling structure to achieve finer-grained credit assignment. Recent studies have also begun to use historical information accumulated during training. HA-DW Yang et al. (2026) adaptively reweights group-relative advantages based on continuously updated difficulty anchors and training dynamics to reduce estimation bias related to problem difficulty. ExGRPO Zhan et al. (2026), in contrast, stores and reuses valuable historical rollouts through experience management and uses a hybrid policy objective to balance exploration and experience reuse. TrendGRPO uses neither global difficulty anchors nor stored or replayed historical trajectories. Instead, it maintains compact and persistent correctness statistics for each training sample. By comparing the accuracy of the current candidate group with the historical accuracy of the same sample, TrendGRPO constructs a current-to-history deviation and asymmetrically reshapes positive and negative advantages after GRPO normalization. This design preserves the relative ranking of candidates within the current group while allowing the update magnitude to respond to the learning trajectory of each sample across iterations.

3 Method

We propose TrendGRPO, a sample-wise history-conditioned advantage shaping method for mathematical reasoning. The overall framework is shown in Fig. 2. Standard GRPO constructs an advantage by comparing candidates within the current group. TrendGRPO retains this within-group comparison and introduces an additional temporal comparison between the current group accuracy and the historical accuracy of the same sample. The resulting signal conditions the magnitude of positive and negative normalized advantages before policy optimization.

The placement of the historical signal after GRPO normalization is central to the method. It preserves the relative candidate ordering established by the current group while allowing the update strength to depend on the sample's cross-iteration learning trajectory. Thus, TrendGRPO changes how the final optimization advantage is constructed without replacing the native GRPO normalization or clipped policy objective.

3.1 Candidate Generation and Base Reward

Given a training sample i, the policy model π_θ generates K candidate responses $\{o_{i,1}, o_{i,2}, \ldots, o_{i,K}\}$. For the j-th candidate response, the task adapter first computes a binary correctness score $y_{i,j} \in \{0, 1\}$, where $y_{i,j} = 1$ indicates that the candidate response is correct, and $y_{i,j} = 0$ otherwise. For mathematical reasoning tasks, an incorrect answer may still have a numerical magnitude close to the ground-truth answer. Therefore, we further introduce a magnitude-based

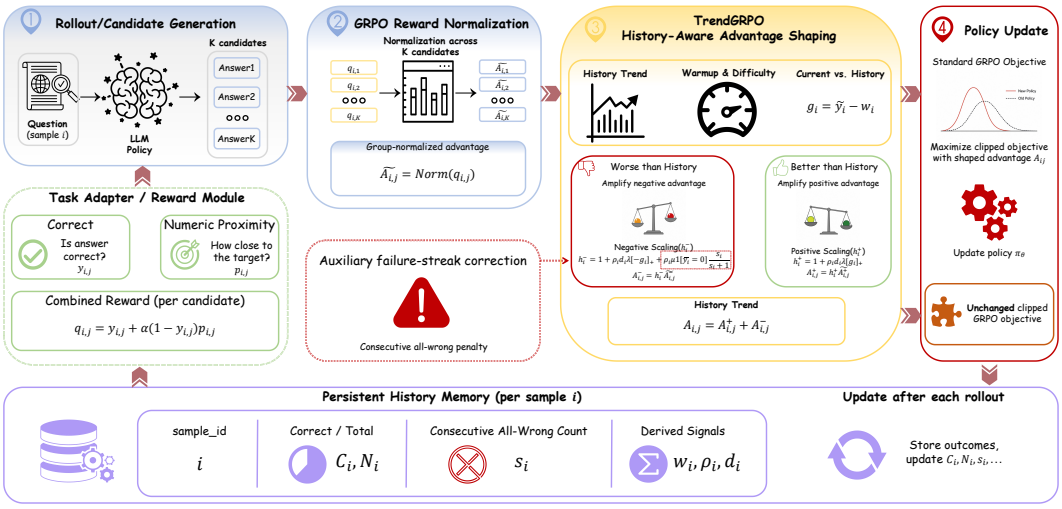


Figure 2: Overview of **TrendGRPO**. TrendGRPO preserves the standard GRPO reward normalization process and introduces sample-level historical accuracy trends after advantage standardization to dynamically shape positive and negative advantages.

proximity score $p_{i,j} \in [0, 1]$. If both the model output and the ground-truth answer can be parsed as numerical values, we denote the predicted answer as $\hat{a}_{i,j}$ and the ground-truth answer as a_i. The numerical proximity is then defined as:

$$p_{i,j} = \exp\left(-\left|\log(1 + |\hat{a}_{i,j}|) - \log(1 + |a_i|)\right|\right) \tag{1}$$

If a valid numerical answer cannot be parsed, we set $p_{i,j} = 0$.

Finally, the base reward for each candidate response is defined as:

$$q_{i,j} = y_{i,j} + \alpha(1 - y_{i,j})p_{i,j}, \tag{2}$$

where α controls the weight of the numerical proximity reward. This design ensures that correct responses receive the highest reward, while providing fine-grained training signals for incorrect answers with similar numerical magnitudes.

3.2 GRPO: Group-wise Reward Normalization

After obtaining the candidate-level rewards, GRPO performs group-wise normalization over the K candidate responses of the same sample. Let the rewards in this group be $\{q_{i,1}, \ldots, q_{i,K}\}$. The standardized advantage is defined as:

$$\widetilde{A}_{i,j} = \frac{q_{i,j} - \bar{q}_i}{\sigma(q_i) + \epsilon}, \tag{3}$$

where $\bar{q}_i$ and $\sigma(q_i)$ denote the mean and standard deviation of the candidate rewards for sample i, respectively, and ϵ is a numerical stability term. This step preserves the relative preference learning mechanism of GRPO, in which the model learns the relative differences between better and worse candidates for the same problem. TrendGRPO does not modify this normalization process, but instead introduces sample-level historical trends after $\widetilde{A}_{i,j}$ is obtained.

3.3 Sample-level Historical Memory

TrendGRPO maintains a compact cumulative history for each training sample. For sample i, the history contains three key quantities: C_i, N_i, s_i, where C_i denotes the cumulative number of correct candidates generated for the sample, N_i denotes the total number of candidates generated so far, and s_i denotes the number of consecutive rollouts in which all candidates are incorrect. These cumulative

statistics provide a long-term reference, and the historical accuracy of sample i is defined as:

$$w_i = \begin{cases} \frac{C_i}{N_i}, & N_i > 0, \\ 0, & N_i = 0. \end{cases} \tag{4}$$

In the current rollout, the average accuracy of the candidate group for sample i is defined as:

$$\bar{y}_i = \frac{1}{K} \sum_{j=1}^{K} y_{i,j} \tag{5}$$

Accordingly, the sample-level accuracy trend is defined as $g_i = \bar{y}_i - w_i$. When $g_i > 0$, the current candidate group performs better than the historical level of the sample; when $g_i < 0$, the current performance is below its historical level; and when g_i is close to 0, the current rollout brings no significant change.

3.4 Warmup and Difficulty Adjustment

To avoid relying too heavily on unstable historical statistics when a sample has been visited only a few times, TrendGRPO introduces a historical warmup coefficient:

$$\rho_i = \text{clip}\left(\frac{N_i/K}{v_{\text{warm}}}, 0, 1\right), \tag{6}$$

where v_{warm} denotes the number of prior visits required for the historical trend to take full effect. Before sufficient history is available, ρ_i limits its influence; with more visits, ρ_i approaches 1.

In addition, to place greater emphasis on samples with moderate difficulty, we define a difficulty weight:

$$d_i = d_{\min} + (1 - d_{\min}) 4 w_i (1 - w_i), \tag{7}$$

where $d_{\min}$ denotes the minimum difficulty weight. When the historical accuracy w_i is close to 0.5, the sample has higher uncertainty and thus a larger d_i; when the sample is almost always correct or almost always incorrect, d_i becomes smaller.

3.5 Sample-Wise History-Conditioned Advantage Shaping

Given the group-normalized advantage $\widetilde{A}_{i,j}$, TrendGRPO first decomposes it into positive and negative components:

$$\begin{aligned} \widetilde{A}^{+}_{i,j} &= \max(\widetilde{A}_{i,j}, 0), \\ \widetilde{A}^{-}_{i,j} &= \min(\widetilde{A}_{i,j}, 0). \end{aligned} \tag{8}$$

Here, $\widetilde{A}^{+}_{i,j}$ represents the positive learning signal from candidates that perform better than the group baseline, while $\widetilde{A}^{-}_{i,j}$ represents the negative learning signal assigned to relatively worse candidates. TrendGRPO preserves the signs and relative ordering established by GRPO, but conditions the magnitudes of these two components separately on the historical state of the corresponding training sample.

Based on the sample-level historical trend, TrendGRPO constructs the positive scaling weight as:

$$h^{+}_i = 1 + \rho_i d_i \lambda [g_i]_+, \tag{9}$$

where $[g_i]_+ = \max(g_i, 0)$. When the current candidate group performs better than the historical level of the sample, the positive advantage is amplified, reinforcing candidates associated with the observed improvement.

For the negative component, TrendGRPO constructs the scaling weight as:

$$h^{-}_i = 1 + \rho_i d_i \lambda [-g_i]_+ + \rho_i \mu \mathbf{1}[\bar{y}_i = 0] \frac{s_i}{s_i + 1}, \tag{10}$$

where $[-g_i]_+ = \max(-g_i, 0)$ and $\mathbf{1}[\cdot]$ denotes the indicator function. The trend-dependent term strengthens the negative advantage when the current performance falls below the historical level.

The final term provides an auxiliary persistent-failure correction: for all-wrong groups with consecutive previous failures, it amplifies existing negative advantages but does not create a signal from zero group-normalized advantages.

Accordingly, the positive, negative, and final shaped advantages are:

$$\begin{aligned} A^{+}_{i,j} &= h^{+}_{i}\,\widetilde{A}^{+}_{i,j}, \\ A^{-}_{i,j} &= h^{-}_{i}\,\widetilde{A}^{-}_{i,j}, \\ A_{i,j} &= A^{+}_{i,j} + A^{-}_{i,j}. \end{aligned} \tag{11}$$

Since $\widetilde{A}^{-}_{i,j} \leq 0$, increasing h^{-}_{i} makes the negative advantage more negative, thereby strengthening the correction applied to relatively worse candidates. By separately scaling and recombining the positive and negative components, TrendGRPO transforms the history-agnostic normalized advantage into a sample-wise history-conditioned optimization signal. Appendix A.6 provides a detailed optimization analysis of this advantage reconstruction.

3.6 History Update and Policy Optimization

After each rollout, TrendGRPO updates the sample-level history according to the correctness of the current candidate group:

$$\begin{aligned} C_i &\leftarrow C_i + \sum_{j=1}^{K} y_{i,j}, \\ N_i &\leftarrow N_i + K. \end{aligned} \tag{12}$$

The consecutive failure count s_i is reset when the current candidate group contains a correct response; otherwise, it is incremented:

$$s_i \leftarrow \begin{cases} 0, & \sum_{j=1}^{K} y_{i,j} > 0, \\ s_i + 1, & \sum_{j=1}^{K} y_{i,j} = 0. \end{cases} \tag{13}$$

Finally, TrendGRPO performs a GRPO-style policy update using the shaped advantage $A_{i,j}$:

$$\mathcal{L}(\theta) = -\mathbb{E}_{i,j}\left[\min\left(r_{i,j}(\theta)A_{i,j}, \mathrm{clip}\left(r_{i,j}(\theta), 1-\delta, 1+\delta\right)A_{i,j}\right)\right] \tag{14}$$

where $r_{i,j}(\theta)$ is the probability ratio between new and old policies, and δ is the clipping coefficient.

TrendGRPO retains the model architecture, candidate-generation procedure, group-wise normalization, and clipped GRPO objective, while replacing the history-agnostic update magnitude with a sample-wise history-conditioned one. The complete procedure is summarized in Appendix A.3.

4 Experiments

4.1 Experimental Setup

We evaluate TrendGRPO on several text-based mathematical reasoning benchmarks, including GSM8K Cobbe et al. (2021), MATH500 Lightman et al. (2024), AIME2024, Minerva Math Lewkowycz et al. (2022), GaoKao 2023 En Zhang et al. (2023), Olympiad Bench He et al. (2024), and College Math. All datasets use a unified evaluation prompt template that requires the model to provide the necessary reasoning and place the final answer in `\boxed{}`. We select Qwen3-0.6B, Qwen3-8B, and OLMo-3-7B-Instruct as the base models, and train both standard GRPO and TrendGRPO with sample-level historical advantage shaping for each model. For our controlled comparisons, GRPO and TrendGRPO use the same training data, training schedule, prompt template, and evaluation settings.

4.2 Model Training

All reinforcement learning experiments are implemented using the veRL framework, with training data from DAPO-Math-17K Yu et al. (2026). To control the training cost, we sample 2,048 examples

Table 1: Main results on textual mathematical reasoning benchmarks. Results of external closed-source, open-source, and reasoning models are collected from their corresponding public reports, while GRPO and TrendGRPO results are evaluated under our unified setting. See Appendix A.1 for details of external result sources. "–" denotes unavailable results.

Method	GSM8K	MATH	MATH500	AIME24	Minerva Math	GaoKao 2023 En	Olympiad Bench	College Math	Avg.
Closed-Source Models									
GPT-4o Hurst et al. (2024)	92.9	81.1	–	–	36.8	67.5	43.3	48.5	61.7
Claude-3.5-Sonnet Anthropic (2024)	96.1	71.1	78.3	16.0	–	–	–	–	–
Open-Source Models									
Llama-3.1-8B-Instruct Grattafiori et al. (2024)	76.6	47.2	–	–	21.7	38.4	15.4	33.8	38.9
Llama-3.1-70B-Instruct Grattafiori et al. (2024)	94.1	65.7	–	–	34.2	54.0	27.7	42.5	53.0
InternLM2-Math-Plus-20B Ying et al. (2024)	87.9	56.5	–	–	20.2	51.9	23.1	37.5	46.2
InternLM2-Math-Plus-Mixtral8x7B Ying et al. (2024)	92.1	59.4	–	–	26.8	49.6	25.0	37.5	48.4
Mathstral-7B-v0.1 Jiang et al. (2023)	84.9	56.6	–	–	16.2	46.0	21.5	33.7	43.2
NuminaMath-72B-CoT Beeching et al. (2024)	90.8	66.7	–	3.3	25.0	58.4	32.6	39.7	45.2
DeepSeekMath-7B-RL Shao et al. (2024)	88.2	52.4	–	–	20.6	43.6	19.0	37.5	43.6
Qwen2.5-Math-7B-Instruct Yang et al. (2024)	95.2	83.6	–	–	37.1	66.8	41.6	46.8	61.9
Reasoning Models									
Bespoke-Stratos-7B Labs (2025)	–	–	82.0	20.0	–	–	–	–	–
Open-Reasoner-Zero-7B Hu et al. (2026)	–	–	81.4	17.9	–	–	–	–	–
DeepSeek-R1-Distill-Qwen-1.5B Guo et al. (2025)	–	–	83.9	28.9	–	–	–	–	–
Eurus-2-7B-PRIME Cui et al. (2025)	–	–	79.2	26.7	38.6	–	40.3	–	–
Qwen-2.5-Math-7B+PURE-PRM+VR Cheng et al. (2026)	–	–	82.6	20.0	37.1	–	44.1	–	–
Our Models									
Qwen3-0.6B + GRPO	75.36	–	51.20	3.33	16.09	45.17	20.74	42.65	36.36
Qwen3-0.6B + TrendGRPO	**75.71**	–	**53.20**	3.33	**19.54**	**47.26**	**22.67**	**43.87**	**37.94**
OLMo-3-7B-Instruct + GRPO	94.16	–	77.00	6.67	56.70	55.87	33.93	50.16	53.50
OLMo-3-7B-Instruct + TrendGRPO	**95.45**	–	**77.60**	6.67	**57.47**	**57.18**	**34.67**	**50.40**	**54.21**
Qwen3-8B + GRPO	94.09	–	83.80	30.00	43.30	68.93	37.78	72.03	61.42
Qwen3-8B + TrendGRPO	**96.21**	–	**85.00**	30.00	**46.36**	**70.23**	**41.63**	**73.66**	**63.30**

from the training set using a fixed random seed and use the same subset for all models and ablation experiments. For each problem, the policy model samples $K = 8$ candidate responses. We train for 6 epochs with a batch size of 8, a PPO mini-batch size of 8, and a maximum generation length of 2048. By default, TrendGRPO uses $\lambda = 1.0$, $\mu = 1.0$, and $\alpha = 0.2$. Further details are provided in Appendix A.2. We will release the training scripts, evaluation scripts, data indices, and complete configurations after publication.

4.3 Main Result

Table 1 presents the performance of TrendGRPO on multiple mathematical reasoning benchmarks. TrendGRPO achieves higher average accuracy across all three base models. Specifically, the average accuracy increases from 36.36 to 37.94 on Qwen3-0.6B, from 53.50 to 54.21 on OLMo-3-7B, and from 61.42 to 63.30 on Qwen3-8B, while AIME2024 accuracy remains unchanged across all three backbones. These results show that the benefit of sample-wise historical shaping is observed across the three tested backbones, spanning two model families and different model scales.

TrendGRPO also reaches or surpasses the publicly reported results of several strong models on some benchmarks. Taking Qwen3-8B as an example, TrendGRPO achieves 85.00 on MATH500, exceeding the reported results of several public reasoning models on this benchmark. It reaches 96.21 on GSM8K, which is comparable to closed-source models such as GPT-4o and Claude-3.5-Sonnet, and obtains 41.63 on Olympiad Bench, outperforming most open-source models and several reasoning models listed in the table. The external results are included as contextual references, as their training data, inference budgets, and evaluation settings differ across sources.

4.4 Sensitivity to the Trend Coefficient

Figure 3 shows the effect of the trend coefficient λ on TrendGRPO. Across the three datasets, performance generally follows a roughly unimodal pattern as λ changes. When $\lambda = 0$, i.e., when the trend-dependent scaling component is disabled while the proximity reward and auxiliary failure-streak correction remain unchanged, the overall performance is lower than that under the default setting, indicating that sample-level historical information provides an additional effective optimization signal during training. As λ increases, performance on the three datasets first improves and reaches the best or near-best results around $\lambda = 1.0$, suggesting that a moderate level of historical trend shaping can better adjust the normalized advantage of GRPO. When λ is further increased to 1.5 or 2.0, performance does not continue to improve and even shows a slight decrease on some datasets. This suggests that the historical trend signal is more suitable as an adjustment term for the advantage, rather than an overly strong factor that dominates policy updates.

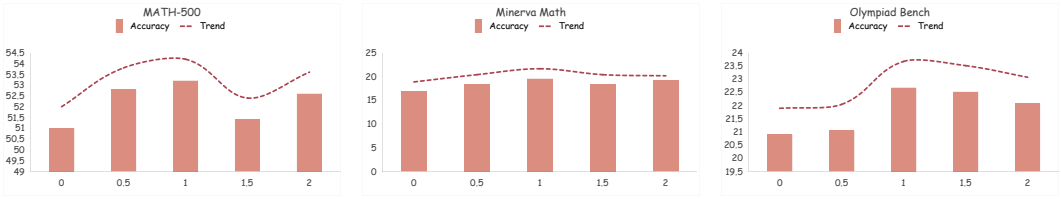


Figure 3: Sensitivity analysis of the trend coefficient λ on Qwen3-0.6B.

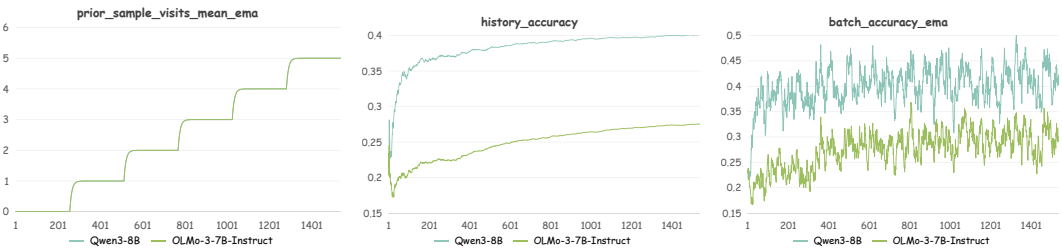


Figure 4: Dynamics of sample-wise history accumulation and rollout accuracy.

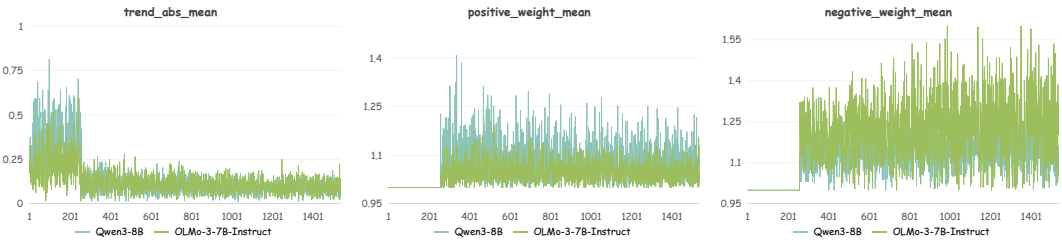


Figure 5: Dynamics of temporal deviations and history-conditioned advantage weights.

4.5 Dynamics of Sample-Wise Historical States

Figure 4 presents the evolution of sample-wise historical states and rollout accuracy during training. The mean number of prior visits increases stepwise across training epochs, confirming that the correctness history of each sample is persistently retained and updated over repeated visits. Because the two models use the same sampling schedule, their prior-visit trajectories overlap. For both Qwen3-8B and OLMo-3-7B, historical accuracy increases smoothly as training progresses, reflecting the cumulative evolution of sample-level correctness states. Batch accuracy also improves rapidly during the early stage and subsequently fluctuates within a relatively stable range, consistent with the stochastic nature of rollout-based reinforcement learning. These dynamics establish that the historical state used by TrendGRPO is accumulated across optimization steps rather than derived solely from the current candidate group.

4.6 History-Conditioned Advantage Shaping

Figure 5 connects the sample-wise temporal deviation with the resulting positive and negative shaping weights. During the initial pass over the training set, the shaping weights remain at 1, even though the raw trend magnitude fluctuates, because no prior observations are yet available and the history warm-up suppresses their influence. Once samples begin to be revisited, as shown in Figure 4, both weights start to deviate from 1. The mean absolute trend subsequently decreases and stabilizes, indicating that current candidate-group performance becomes closer to the historical level of the corresponding samples. Nevertheless, the positive and negative weights continue to fluctuate above 1, showing that remaining sample-specific deviations asymmetrically modulate the magnitudes of normalized advantages. These dynamics verify that TrendGRPO operates through sample-wise history-conditioned advantage shaping after GRPO normalization.

Table 2: Ablation of sample-wise historical conditioning on Qwen3-0.6B. "Same-sample" indicates whether the historical state belongs to the current training sample; "Asym. Shaping" indicates separate positive/negative shaping.

Method	Same-sample	Asym. Shaping	MATH500	Minerva Math	Olympiad Bench	Avg.
GRPO	–	–	51.20	16.09	20.74	29.34
Global History	×	✓	49.60	18.77	21.48	29.95
Shuffled History	×	✓	51.40	17.62	19.70	29.57
Symmetric History Shaping	✓	×	52.80	18.39	20.89	30.69
TrendGRPO	✓	✓	**53.20**	**19.54**	**22.67**	**31.80**

Table 3: Ablation study of TrendGRPO on Qwen3-0.6B.

Dataset	GRPO	w/o Trend	w/o Prox.	w/o Streak Corr.	TrendGRPO
MATH500	51.20	51.00	50.40	53.00	**53.20**
Minerva Math	16.09	16.86	17.24	17.24	**19.54**
Olympiad Bench	20.74	20.89	21.93	20.89	**22.67**

4.7 Ablation of historical conditioning

Table 2 examines whether the gains of TrendGRPO come from general historical statistics or from the correspondence between each sample and its own historical state. Compared with GRPO, Global History brings only limited and unstable improvements. When the historical shaping mechanism is retained but the sample-history correspondence is broken, Shuffled History performs close to GRPO on average, indicating that arbitrary historical signals or training-stage information alone cannot explain the performance gains. After restoring the historical state of the same sample, Symmetric History Shaping increases the average accuracy to 30.69. Further applying asymmetric shaping to positive and negative advantages, TrendGRPO achieves improvements on all three datasets, with the average accuracy reaching 31.80. Accurate sample-history correspondence is the basis of the method, while asymmetric advantage shaping further strengthens the optimization signal.

4.8 Component ablation

Table 3 shows the contribution of each component in TrendGRPO. Removing trend scaling leads to clearly lower performance than the full method on all three datasets, confirming that the deviation of current performance from the historical state of each sample serves as the core conditioning signal in TrendGRPO. Removing the numerical proximity reward also reduces overall performance, indicating that it provides complementary candidate-level information beyond binary correctness. The contribution of persistent-failure correction is more task-dependent: its effect on MATH500 is relatively small, while it brings clearer gains on the more challenging Minerva Math and Olympiad Bench, which is consistent with its design for targeted correction of persistently failed samples. The full TrendGRPO achieves the best performance on all three datasets, showing that trend scaling plays the central role, while the numerical proximity reward and persistent-failure correction provide complementary gains.

5 Conclusion

This paper introduces TrendGRPO, which extends GRPO from purely within-group comparison to sample-wise temporal comparison across repeated visits. TrendGRPO maintains a persistent correctness state for each training sample and uses the deviation between current and historical performance to asymmetrically shape positive and negative normalized advantages. Experiments across three tested backbones and multiple mathematical reasoning benchmarks show higher average accuracy than the controlled GRPO baseline, with gains on most evaluated benchmarks. Ablations and training diagnostics further support sample-wise historical conditioning as the core optimization signal, with numerical proximity and failure-streak correction serving as complementary components. Because the shaping is applied after GRPO normalization, the method retains the original rollout and clipped policy optimization procedures with little additional training overhead.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing, improving clarity, and refining the presentation of the manuscript. We have not used generative AI tools to generate the core research idea, design the method, produce experimental results, or perform automatic evaluation for reported benchmarks, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for minor auxiliary writing suggestions, including wording refinements. We have reviewed all AI-assisted work. All AI-assisted text was manually checked and revised by the authors, and all experimental results reported in the paper were obtained from our own implementations and evaluations. We 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 focuses on improving reinforcement learning algorithms for mathematical reasoning in large language models. We do not involve human subjects, private user data, or sensitive personal information. All experiments are conducted on public or publicly available mathematical reasoning datasets. We do not release new datasets containing personal or sensitive content. While the underlying principle of history-conditioned advantage shaping may be applicable to other domains with appropriate reward signals, our empirical study is limited to mathematical reasoning benchmarks. We do not identify additional ethical risks specific to TrendGRPO beyond those generally associated with reinforcement learning and large language model training and deployment.

REPRODUCIBILITY STATEMENT

We provide the main algorithmic details of TrendGRPO in Section 3, including the definition of sample-level historical statistics, trend-based advantage shaping, and persistent failure penalty. The experimental setup, training data, model backbones, evaluation benchmarks, and key hyperparameters are described in Section 4 and Appendix A.2. Appendix A.1 further reports the sources of external benchmark results used for comparison. We will release the source code, training scripts, evaluation scripts, and configuration files to facilitate reproduction of the reported GRPO and TrendGRPO experiments.

REFERENCES

Anthropic. Claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Josh Barua, Seun Eisape, Kayo Yin, and Alane Suhr. Long chain-of-thought reasoning across languages. In *International Conference on Learning Representations*, volume 2026, pp. 77834–77857, 2026.

Edward Beeching, Shengyi Costa Huang, Albert Jiang, Jia Li, Benjamin Lipkin, Zihan Qina, Kashif Rasul, Ziju Shen, Roman Soletskyi, and Lewis Tunstall. Numinamath 72b cot. https://huggingface.co/AI-MO/NuminaMath-72B-CoT, 2024.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. *Advances in Neural Information Processing Systems*, 38:131646–131671, 2026.

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

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.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026.

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.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. *arXiv preprint arXiv:2605.21125*, 2026.

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. *Advances in Neural Information Processing Systems*, 38:162239–162262, 2026.

Ke Hu, Zhehuai Chen, Chao-Han Huck Yang, Piotr Żelasko, Oleksii Hrinchuk, Vitaly Lavrukhin, Jagadeesh Balam, and Boris Ginsburg. Chain-of-thought prompting for speech translation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *ArXiv*, abs/2310.06825, 2023. URL https://api.semanticscholar.org/CorpusID:263830494.

Guochao Jiang, Jingyi Song, Guofeng Quan, Chuzhan Hao, Guohua Liu, and Yuewei Zhang. Dvao: Dynamic variance-adaptive advantage optimization for multi-reward reinforcement learning. *arXiv preprint arXiv:2605.25604*, 2026.

Bespoke Labs. Bespoke-stratos: The unreasonable effectiveness of reasoning distillation. www.bespokelabs.ai/blog/bespoke-stratos-the-unreasonable-effectiveness-of-reasoning-distillation, 2025. Accessed: 2025-01-22.

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.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24681–24693, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, Jiri Navratil, Jarret Ross, and Jesus Rios. Revisiting group relative policy optimization: Insights into on-policy and off-policy training. In *International Conference on Learning Representations*, volume 2026, pp. 110991–111007, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

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.

Zayne Sprague, Fangcong Yin, Juan Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 94118–94162, 2025.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Chao Wang, Hongtao Tian, Tao Yang, Yunsheng Shi, Ting Yao, and Wenbo Ding. Process advantage signal shaping: A paradigm-agnostic middleware for process-supervised rl in llm reasoners. *arXiv preprint arXiv:2606.29296*, 2026a.

Haozhe Wang, Qixin Xu, Che Liu, Junhong Wu, Fangzhen Lin, and Wenhu Chen. Emergent hierarchical reasoning in llms through reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 74519–74543, 2026b.

Yixuan Wang, Yifei Chen, Haichao Zhang, Haozheng Luo, Xander Wu, Jie Ni, Yun Fu, Nuno Vasconcelos, and Yijiang Li. Learn what's left, not what's mastered: Saturation aware advantage reweighting for multi-reward policy optimization. *arXiv preprint arXiv:2608.16072*, 2026c.

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.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, et al. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*, 2026.

Huaiyuan Ying, Shuo Zhang, Linyang Li, Zhejian Zhou, Yunfan Shao, Zhaoye Fei, Yichuan Ma, Jiawei Hong, Kuikun Liu, Ziyi Wang, et al. Internlm-math: Open math large language models toward verifiable reasoning. *arXiv preprint arXiv:2402.06332*, 2024.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. Star: Self-taught reasoner. In *Proceedings of the NIPS*, volume 22, pp. 3, 2022.

Runzhe Zhan, Yafu Li, Zhi Wang, Xiaoye Qu, Dongrui Liu, Jing Shao, Derek Wong, and Yu Cheng. Exgrpo: Learning to reason from experience. In *International Conference on Learning Representations*, volume 2026, pp. 12810–12846, 2026.

Jixiao Zhang and Chunsheng Zuo. Grpo-lead: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5642–5665, 2025.

Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*, 2023.

Huaisheng Zhu, Siyuan Xu, Hangfan Zhang, Teng Xiao, Zhimeng Guo, Shijie Zhou, Shuyue Hu, and Vasant G Honavar. Reinforcement learning for large language models via group preference reward shaping. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21398–21411, 2025.

Table 4: Sources of external benchmark results.

Model / Group	Source
GPT-4o	Qwen2.5-Math Technical Report Yang et al. (2024)
Claude-3.5-Sonnet (GSM8K)	To CoT or not to CoT? Sprague et al. (2025)
Claude-3.5-Sonnet (MATH)	The Llama 3 Herd of Models Grattafiori et al. (2024)
Claude-3.5-Sonnet (MATH500, AIME2024)	DeepSeek-R1 official evaluation table DeepSeek-AI (2025)
Open-source models	Qwen2.5-Math Technical Report Yang et al. (2024)
Bespoke-Stratos-7B	Bespoke-Stratos Labs (2025)
Open-Reasoner-Zero-7B	Open-Reasoner-Zero Hu et al. (2026)
DeepSeek-R1-Distill-Qwen-1.5B	Deepseek-r1 Guo et al. (2025)
Eurus-2-7B-PRIME	Process reinforcement through implicit rewards Cui et al. (2025)
Qwen-2.5-Math-7B+PURE-PRM+VR	PURE Cheng et al. (2026)

Table 5: Training hyperparameters for GRPO and TrendGRPO.

Hyperparameter	Value
Training dataset	DAPO-Math-17K
Training subset size	2,048 samples
Sampling seed	42
Base models	Qwen3-0.6B, Qwen3-8B, OLMo-3-7B-Instruct
RL algorithm	GRPO
Rollout responses per prompt K	8
Training epochs	6
Train batch size	8
PPO mini-batch size	8
Learning rate	1×10^{-6}
Maximum prompt length	1024
Maximum response length	2048
KL coefficient	0
Precision	bfloat16
Number of GPUs	2
Checkpoint interval	500 steps
Validation interval	100 steps
Validation samples	128
Logger	SwanLab
Trend coefficient λ	1.0
Proximity coefficient α	0.2
Stagnation penalty coefficient μ	1.0
History warmup visits	1
Difficulty floor $d_{\min}$	0.25
Numerical stability ϵ	10^{-12}

A Appendix

A.1 Sources of External Benchmark Results

To provide a more complete reference for the main experimental results, Table 1 reports not only our reproduced results for GRPO and TrendGRPO, but also the performance of several representative closed-source, open-source, and reasoning models. Since some models do not release their full weights, or their inference services, prompts, and decoding settings are difficult to reproduce exactly, we use the results reported in public technical reports or official release pages as references. Table 4 summarizes the specific sources of the external model results included in the main results table.

A.2 Training Hyperparameters

Table 5 summarizes the training hyperparameters used for GRPO and TrendGRPO. To ensure a fair comparison, all methods use the same training set, random seed, number of rollout responses,

Algorithm 1 TrendGRPO: Sample-wise Historical Advantage Shaping

Require: Policy model π_θ, training samples $\mathcal{D}$, number of rollouts K, coefficients λ, α, μ, warmup visits v_{warm}, difficulty floor $d_{\min}$

Initialize sample history $C_i \leftarrow 0$, $N_i \leftarrow 0$, $s_i \leftarrow 0$ for each sample i
for each training iteration **do**
  Sample a mini-batch of prompts $\mathcal{B}$
  for each sample $i \in \mathcal{B}$ **do**
    Generate K candidate responses $\{o_{i,j}\}_{j=1}^{K}$ using π_θ
    Compute correctness $y_{i,j}$ and numerical proximity $p_{i,j}$
    Compute candidate reward $q_{i,j} = y_{i,j} + \alpha(1 - y_{i,j})p_{i,j}$
    Compute group-normalized advantage $\widetilde{A}_{i,j} = \mathrm{Norm}(\{q_{i,j}\}_{j=1}^{K})$
    Compute historical accuracy $w_i = C_i / \max(N_i, 1)$
    Compute current group accuracy $\bar{y}_i = \frac{1}{K}\sum_{j=1}^{K} y_{i,j}$
    Compute trend signal $g_i = \bar{y}_i - w_i$
    Compute warmup coefficient ρ_i and difficulty weight d_i
    Split advantage $\widetilde{A}^{+}_{i,j} = \max(\widetilde{A}_{i,j}, 0)$ and $\widetilde{A}^{-}_{i,j} = \min(\widetilde{A}_{i,j}, 0)$
    Compute shaping weights h_i^{+} and h_i^{-} using g_i, s_i, ρ_i, and d_i
    Obtain shaped advantage $A_{i,j} = h_i^{+}\widetilde{A}^{+}_{i,j} + h_i^{-}\widetilde{A}^{-}_{i,j}$
    Update history $C_i \leftarrow C_i + \sum_{j=1}^{K} y_{i,j}$, $N_i \leftarrow N_i + K$
    Update $s_i \leftarrow 0$ if $\sum_{j=1}^{K} y_{i,j} > 0$, otherwise $s_i \leftarrow s_i + 1$
  end for
  Update policy π_θ with the GRPO clipped objective using shaped advantages $A_{i,j}$
end for

Table 6: Diagnostic of the auxiliary failure-streak term on Qwen3-0.6B. Samples are grouped by consecutive all-wrong count s_i; columns report visit fraction, historical accuracy, negative shaping weight, zero-information fraction, and recovery rates within the next 1 or 3 visits.

Failure bucket	Sample frac.	Avg. hist. acc.	Avg. neg. weight	Zero-info	Recovery@next	Recovery@3 visits
$s_i = 0$	0.144	0.108	1.013	0.006	–	–
$s_i = 1$	0.206	0.028	1.017	0.016	0.152	0.277
$s_i = 2$	0.164	0.011	1.505	0.008	0.097	0.195
$s_i \geq 3$	0.487	0.002	1.757	0.008	0.055	0.099

batch size, and other training hyperparameters. The only difference is that TrendGRPO applies our proposed history-conditioned advantage shaping.

A.3 Pseudocode of TrendGRPO

Algorithm 1 summarizes the training procedure of TrendGRPO. Its main difference from standard GRPO lies in the advantage shaping step after normalization.

A.4 Diagnostics of the Auxiliary Failure-Streak Term

As a diagnostic of the auxiliary failure-streak term, Table 6 groups training visits by their consecutive all-wrong count s_i and examines how the negative shaping weight changes when this term is activated. For each group, we report its visit ratio, historical accuracy, average negative shaping weight, zero-information ratio, and recovery rates from the all-wrong state within the next one and three visits. Longer failure streaks are associated with lower historical accuracy and lower subsequent recovery rates: both Recovery@next and Recovery@3 visits decrease as s_i increases. This suggests that the failure-streak state tracks samples that are increasingly difficult for the current policy to recover. Consistent with its design, the average negative shaping weight increases from 1.013 to 1.757, indicating that the auxiliary term applies progressively stronger negative scaling as repeated failures accumulate. Meanwhile, the zero-information ratio remains below 2% across all groups, showing that the term is usually applied when the underlying group-normalized advantages retain non-zero

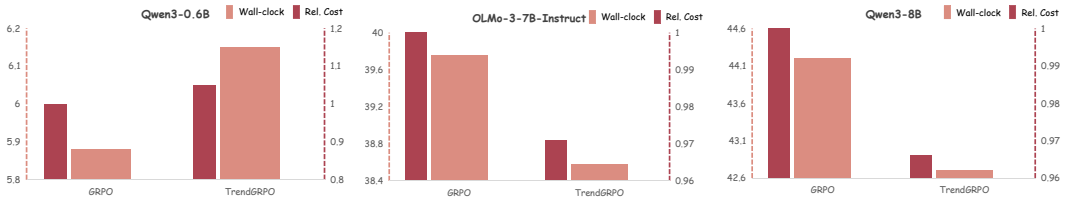


Figure 6: Training cost comparison between GRPO and TrendGRPO across different base models.

relative signals. It therefore reshapes existing negative advantages rather than creating a learning signal for zero-information groups. These observations show that the auxiliary failure-streak term behaves as intended: its strongest corrections are applied to samples with longer failure streaks, which are also associated with lower historical accuracy and lower subsequent recovery rates.

A.5 Training Cost Analysis

Figure 6 compares the training time of GRPO and TrendGRPO across different base models. The relative training costs of TrendGRPO are 1.05, 0.971, and 0.966, respectively, remaining close to those of standard GRPO. This indicates that sample-level historical statistics and post-normalization advantage shaping introduce little additional training overhead. Overall, the additional scalar history updates and advantage-scaling operations account for only a small fraction of the total training cost.

A.6 Optimization Interpretation of TrendGRPO

This section characterizes how TrendGRPO modifies the optimization signal of GRPO. We show that the proposed reconstruction induces a bounded, history-conditioned, and sign-specific reweighting of the clipped GRPO surrogate. We further establish its group-mean effect, sign and ordering preservation, and zero-signal invariance. These results characterize the optimization structure induced by TrendGRPO.

Unified scaling form. For a group-normalized GRPO advantage $\widetilde{A}_{i,j}$, define

$$m_{i,j} = h_i^+ \mathbf{1}[\widetilde{A}_{i,j} \geq 0] + h_i^- \mathbf{1}[\widetilde{A}_{i,j} < 0]. \tag{15}$$

The reconstructed advantage can then be written as

$$A_{i,j} = m_{i,j} \widetilde{A}_{i,j}. \tag{16}$$

The multiplier $m_{i,j}$ is computed from the fixed rollout batch and the pre-update history of sample i. It is therefore treated as constant with respect to the policy parameters during the corresponding optimization update.

Exact reweighting of the clipped surrogate. For notational simplicity, let $r_{i,j}(\theta)$ denote the importance-ratio term appearing in the GRPO surrogate, with its token index suppressed. The following argument applies elementwise when the implementation uses token-level importance ratios.

Let

$$\ell_\delta(r, a) = \min\left(ra,\ \mathrm{clip}(r, 1-\delta, 1+\delta)a\right) \tag{17}$$

denote the clipped surrogate contribution to be maximized. For any non-negative scalar m,

$$\ell_\delta(r, ma) = m\ell_\delta(r, a). \tag{18}$$

This follows directly from the positive homogeneity of both arguments of the minimum. Since $m_{i,j} \geq 0$, we therefore have

$$\ell_\delta\left(r_{i,j}(\theta), A_{i,j}\right) = m_{i,j}\ell_\delta\left(r_{i,j}(\theta), \widetilde{A}_{i,j}\right). \tag{19}$$

Conditional on the same fixed rollout batch, the two empirical surrogate gradients satisfy

$$\begin{aligned} \nabla_\theta \mathcal{J}_{\mathrm{Trend}} &= \mathbb{E}\left[m_{i,j}\nabla_\theta \ell_\delta\left(r_{i,j}(\theta), \widetilde{A}_{i,j}\right)\right], \\ \nabla_\theta \mathcal{J}_{\mathrm{GRPO}} &= \mathbb{E}\left[\nabla_\theta \ell_\delta\left(r_{i,j}(\theta), \widetilde{A}_{i,j}\right)\right]. \end{aligned} \tag{20}$$

Consequently,

$$\nabla_\theta \mathcal{J}_{\text{Trend}} - \nabla_\theta \mathcal{J}_{\text{GRPO}} = \mathbb{E}\left[(m_{i,j} - 1)\nabla_\theta \ell_\delta\left(r_{i,j}(\theta), \widetilde{A}_{i,j}\right)\right]. \tag{21}$$

Thus, TrendGRPO preserves the functional form of the clipped GRPO surrogate while inducing a history-conditioned reweighting of its candidate-wise contributions. It does not, however, optimize the same empirical surrogate as GRPO.

Symmetric scaling and asymmetric reconstruction. Define

$$\bar{h}_i = \frac{h_i^+ + h_i^-}{2}, \qquad \Delta h_i = \frac{h_i^+ - h_i^-}{2}. \tag{22}$$

Using

$$\widetilde{A}_{i,j}^+ + \widetilde{A}_{i,j}^- = \widetilde{A}_{i,j}, \qquad \widetilde{A}_{i,j}^+ - \widetilde{A}_{i,j}^- = |\widetilde{A}_{i,j}|, \tag{23}$$

the reconstructed advantage admits the decomposition

$$\boxed{A_{i,j} = \bar{h}_i \widetilde{A}_{i,j} + \Delta h_i |\widetilde{A}_{i,j}|}. \tag{24}$$

The first term corresponds to sign-symmetric sample-level scaling, whereas the second term changes the balance between positive and negative advantages.

For compactness, let

$$\zeta_i = \rho_i \mu \mathbf{1}[\bar{y}_i = 0] \frac{s_i}{s_i + 1} \tag{25}$$

denote the auxiliary failure-streak contribution. From the definitions of h_i^+ and h_i^-, we obtain

$$\begin{aligned} \bar{h}_i &= 1 + \frac{\rho_i d_i \lambda}{2} |g_i| + \frac{\zeta_i}{2}, \\ \Delta h_i &= \frac{\rho_i d_i \lambda}{2} g_i - \frac{\zeta_i}{2}. \end{aligned} \tag{26}$$

Therefore, the temporal deviation g_i controls the direction of the asymmetric component. The auxiliary streak term contributes an additional negative adjustment only when its activation condition is satisfied.

Because GRPO normalization gives

$$\sum_{j=1}^{K} \widetilde{A}_{i,j} = 0, \tag{27}$$

Equation 24 implies

$$\frac{1}{K} \sum_{j=1}^{K} A_{i,j} = \Delta h_i \frac{1}{K} \sum_{j=1}^{K} |\widetilde{A}_{i,j}|. \tag{28}$$

A common sample-level multiplier would only change $\bar{h}_i$ and preserve the zero group mean. TrendGRPO additionally modifies Δh_i, producing a history-conditioned shift in the group-mean advantage. This distinguishes it from sign-symmetric sample-level reweighting in which all candidates of a sample are multiplied by the same scalar.

Corollary 1: group-mean trend consistency. Suppose that the auxiliary streak term is inactive, so that $\zeta_i = 0$, and that the history-conditioning gate is active, with $\rho_i d_i \lambda > 0$. For any group satisfying

$$\frac{1}{K} \sum_{j=1}^{K} |\widetilde{A}_{i,j}| > 0, \tag{29}$$

Equations 26 and 28 give

$$\frac{1}{K} \sum_{j=1}^{K} A_{i,j} = \frac{\rho_i d_i \lambda}{2} g_i \frac{1}{K} \sum_{j=1}^{K} |\widetilde{A}_{i,j}|. \tag{30}$$

It follows that

$$\operatorname{sign}\left(\frac{1}{K}\sum_{j=1}^{K} A_{i,j}\right) = \operatorname{sign}(g_i). \tag{31}$$

Hence, performance above the sample's historical level induces a positive shift in the group-mean shaped advantage, whereas performance below its historical level induces a negative shift. If the group contains no non-zero GRPO advantages or the history-conditioning gate is inactive, no such shift is introduced.

Initial update-direction decomposition. We next characterize the update direction conditional on a fixed rollout batch and at $\theta = \theta_{\text{old}}$. At this point, $r_{i,j}(\theta_{\text{old}}) = 1$, and clipping is inactive. Define

$$\psi_{i,j} = \nabla_\theta \log r_{i,j}(\theta)|_{\theta=\theta_{\text{old}}} \tag{32}$$

as the score associated with candidate j. The per-group update direction is then

$$G_i^{\text{Trend}} = \frac{1}{K}\sum_{j=1}^{K} A_{i,j}\psi_{i,j} = \bar{h}_i G_i^{\text{GRPO}} + \Delta h_i G_i^{\text{abs}}, \tag{33}$$

where

$$G_i^{\text{GRPO}} = \frac{1}{K}\sum_{j=1}^{K} \widetilde{A}_{i,j}\psi_{i,j}, \qquad G_i^{\text{abs}} = \frac{1}{K}\sum_{j=1}^{K} |\widetilde{A}_{i,j}|\psi_{i,j}. \tag{34}$$

If $h_i^+ = h_i^-$, the update reduces to a uniform rescaling of the sample's GRPO contribution. When $h_i^+ \neq h_i^-$, the second term introduces an additional sign-asymmetric correction determined by the sample's temporal state.

Equation 33 characterizes the initial update direction induced by the fixed rollout batch. Away from θ_{old}, clipping may become active and this score-function decomposition need not retain the same unclipped form. Nevertheless, the exact clipped-surrogate reweighting in Equation 19 continues to hold throughout the update.

Sign, ordering, and zero-signal preservation. Since h_i^+ and h_i^- are strictly positive,

$$\operatorname{sign}(A_{i,j}) = \operatorname{sign}(\widetilde{A}_{i,j}). \tag{35}$$

Candidates with positive advantages share the same multiplier h_i^+, while candidates with negative advantages share h_i^-. Therefore, their within-sign ordering is unchanged. Together with sign preservation, this preserves the weak pairwise ordering of candidates induced by GRPO:

$$\widetilde{A}_{i,j_1} \le \widetilde{A}_{i,j_2} \quad \Longrightarrow \quad A_{i,j_1} \le A_{i,j_2}. \tag{36}$$

Moreover,

$$\widetilde{A}_{i,j} = 0 \quad \Longrightarrow \quad A_{i,j} = 0. \tag{37}$$

In particular, if all candidates receive identical rewards and GRPO assigns zero advantage to the entire group, TrendGRPO also produces zero advantages. The method therefore reconstructs existing relative learning signals rather than creating a gradient for zero-information groups.

Bounded update amplification. Assuming $0 \le \rho_i, d_i \le 1$, $|g_i| \le 1$, $\lambda, \mu \ge 0$, and $0 \le s_i/(s_i + 1) < 1$, the shaping weights satisfy

$$1 \le h_i^+ \le 1 + \lambda, \qquad 1 \le h_i^- \le 1 + \lambda + \mu. \tag{38}$$

Let $M = 1 + \lambda + \mu$. Then

$$|A_{i,j}| \le M|\widetilde{A}_{i,j}|, \qquad A_{i,j}^2 \le M^2 \widetilde{A}_{i,j}^2. \tag{39}$$

Furthermore, because $m_{i,j}$ is held fixed during the update, Equation 18 gives, wherever the clipped surrogate is differentiable,

$$\nabla_\theta \ell_\delta\left(r_{i,j}(\theta), A_{i,j}\right) = m_{i,j}\nabla_\theta \ell_\delta\left(r_{i,j}(\theta), \widetilde{A}_{i,j}\right). \tag{40}$$

Consequently, each candidate-wise gradient contribution satisfies

$$\|\nabla_\theta \ell_\delta\left(r_{i,j}(\theta), A_{i,j}\right)\| \leq M \left\|\nabla_\theta \ell_\delta\left(r_{i,j}(\theta), \widetilde{A}_{i,j}\right)\right\|. \tag{41}$$

An analogous statement holds for a valid subgradient at the clipping boundaries. Thus, TrendGRPO can amplify candidate-wise update contributions, but the amplification is explicitly bounded by its hyperparameters. This candidate-wise result does not imply the same norm ratio for the aggregated gradient, whose direction and magnitude also depend on cancellation among candidates.

Estimator interpretation and limitations. The multiplier $m_{i,j}$ depends on the observed candidate group and the stored history of the corresponding sample. It is therefore generally correlated with both $\widetilde{A}_{i,j}$ and the candidate score function. As a result, TrendGRPO is not designed as an unbiased estimator of the original GRPO gradient, nor does the bound above guarantee lower gradient variance. Instead, it deliberately optimizes a history-conditioned, sign-specific reweighted surrogate.

Accordingly, the role of this analysis is to characterize how TrendGRPO changes the optimization signal and to establish its structural properties, rather than to claim equivalence to the original GRPO objective, universal variance reduction, or guaranteed performance improvement.