SEMIFACTUAL CREDIT-AUGMENTED POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has improved the reasoning capabilities of large language models (LLMs), yet their predictions remain sensitive to task-irrelevant prompt features. We investigate this sensitivity through semifactual prompt interventions that preserve the underlying problem and its answer. Our analysis reveals substantial variation in token-level sensitivity and shows that suppressing high-drift token candidates during decoding improves reasoning accuracy without updating model weights. These findings highlight a limitation of Group Relative Policy Optimization (GRPO), which assigns the same outcome-derived advantage to every response token and may reinforce potential spurious dependence alongside useful reasoning. Motivated by this observation, we introduce Semifactual Credit-Augmented Policy Optimization (SCAPO), a causally inspired variant of GRPO that incorporates semifactual stability into token-level credit assignment. SCAPO measures token probability drift for fixed responses under semifactual interventions and uses normalized stability scores to reduce advantages for relatively unstable tokens during early training, while granting no additional credit for stability alone. On Qwen3-4B-Base and Qwen3-1.7B-Base, SCAPO improves AIME 2024–2026 accuracy over GRPO by 5.63 and 4.17 percentage points, respectively. At both model scales, SCAPO achieves the best results on most evaluated mathematics benchmarks and all evaluated out-of-distribution benchmarks among the compared methods. These results suggest that semifactual stability provides an effective training signal for improving reasoning and generalization through finer-grained credit assignment in RLVR. Our code and data can be found in the Supplementary Material.

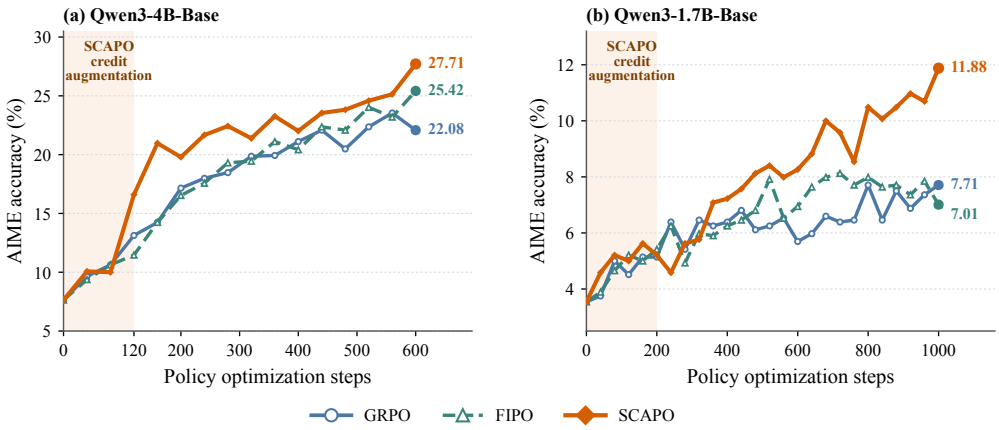

Figure 1: **SCAPO achieves the highest AIME accuracy on the two Qwen3 base models.** Aggregate AIME 2024–2026 accuracy for SCAPO, GRPO, and FIPO. Shaded regions mark SCAPO's semifactual credit augmentation phase. SCAPO reaches GRPO's final accuracy in fewer than half as many policy optimization steps and achieves higher final accuracy than both baselines at both model scales.

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.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has advanced the development of large language models (LLMs) with strong reasoning capabilities, such as OpenAI o1 (OpenAI, 2024) and DeepSeek-R1 (Guo et al., 2025). Group Relative Policy Optimization (GRPO) (Shao et al., 2024), a representative RLVR method, optimizes automatically checkable outcomes without process-level supervision. Recent evidence suggests that RLVR can diminish dependence on spurious correlations and improve generalization under distribution shifts (Fu et al., 2026). However, robustness evaluations still reveal sensitivity to irrelevant information and input perturbations (Mirzadeh et al., 2025; Huang et al., 2025). Understanding how training factors affect generalization in RLVR and how to further improve it remains underexplored.

We study spurious dependence in RLVR at the token level. Building on prior studies of spurious feature reliance (Wang et al., 2022; Fu et al., 2026), we probe causal invariance (Peters et al., 2016; Arjovsky et al., 2019) in LLM inference through semifactual prompt interventions (Goodman, 1947; Lu et al., 2022). Specifically, we alter task-irrelevant prompt features while preserving the mathematical problem and its answer, using drift in the token probabilities of a fixed response as a proxy for potential spurious dependence. We find that suppressing high-drift token candidates during decoding improves the accuracy of Qwen3-4B-Base (Yang et al., 2025) from 15.8% to 30.0% on a 1,000-question mathematical diagnostic panel without updating model weights (see Section 2). The above results suggest that LLMs are highly sensitive to token-level spurious features. However, GRPO assigns the same outcome-derived advantage to every valid token in a response, without directly accounting for this token-level sensitivity. As a result, positive outcome credit may reinforce potential spurious dependence alongside useful reasoning.

To address the above problem, one intuitive way is to change the RL training process, incorporating semifactual sensitivity into token-level credit assignment. In this paper, we introduce **Semifactual Credit-Augmented Policy Optimization (SCAPO)**, a causally inspired variant of GRPO that aims to turn semifactual stability into a token-level credit signal during training. This signal reveals differences in token sensitivity that final-answer correctness alone cannot distinguish. Verifiable rewards provide response-level supervision, while semifactual stability refines token-level credit assignment.

In particular, SCAPO uses the rollout policy to teacher-force each sampled response under the original prompt and its semifactual perturbations, measuring the probability drifts of each response token. After aggregating and normalizing these drifts within each prompt group, it adds only the negative part of the resulting stability score to the GRPO advantage as a detached token-level credit augmentation. Consequently, relatively unstable tokens receive lower advantages, while stability alone earns no additional credit, since stability does not imply correctness. We apply this credit augmentation during the early phase of training to shape trajectory selection, then continue optimizing the resulting policy with standard GRPO.

SCAPO improves AIME 2024–2026 accuracy (Mathematical Association of America, 2026) over GRPO by **+5.63** and **+4.17** points on Qwen3-4B-Base and Qwen3-1.7B-Base (Yang et al., 2025), respectively. As shown in Figure 1, SCAPO reaches GRPO's final AIME accuracy in fewer than half as many policy optimization steps. Moreover, across both model scales, SCAPO outperforms GRPO on all evaluated benchmarks and achieves the best performance on most of the competition-level mathematics benchmarks among all the compared RLVR methods. SCAPO also achieves the highest accuracy on out-of-distribution benchmarks at both model scales. In addition, our ablations further support the value of aligning credit corrections with semifactual sensitivity and selectively reducing credit for relatively unstable tokens. These results suggest that semifactual stability complements outcome rewards with an effective signal for finer-grained credit assignment in RLVR.

Our contributions can be summarized as:

- We study token-level spurious dependence in LLM inference through semifactual prompt interventions, revealing heterogeneous sensitivity and showing that suppressing high-drift tokens during decoding can improve reasoning accuracy without weight updates. (Sec. 2)
- We introduce **SCAPO**, an approach to token-level credit augmentation in GRPO using detached negative-only corrections derived from fixed-response semifactual stability, without requiring process supervision or an external reward model. (Sec. 3)

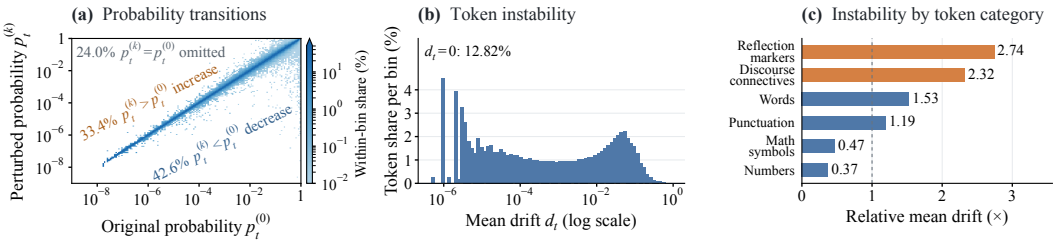

Figure 2: **Semifactual instability varies across tokens.** (a) Probability transitions under semifactual prompt interventions. (b) Distribution of mean token drift. (c) Relative category mean drift normalized by the overall mean, with reflection markers and discourse connectives showing higher sensitivity. Results use frozen Qwen3-4B-Base before RL.

- We empirically demonstrate SCAPO's effectiveness on Qwen3-4B-Base and Qwen3-1.7B-Base, achieving the best results on most benchmarks among the compared RLVR methods. These results show that semifactual stability can serve as an effective training-time signal for token-level credit assignment in RLVR. (Sec. 4)

2 Token-Level Semifactual Sensitivity

We probe potential token-level spurious dependence with semifactual sensitivity. Using a frozen base model, we characterize how this sensitivity varies across tokens and then examine whether suppressing unstable token candidates improves decoding.

Constructing semifactual perturbations. Semifactual perturbations alter incidental features of a prompt while preserving its underlying mathematical problem and final answer (Kenny & Keane, 2021). Following prior work on reasoning robustness under input perturbations (Mirzadeh et al., 2025; Huang et al., 2025), we construct four types of perturbations: paraphrase, minor typo noise, irrelevant scenario wrapping, and appended irrelevant context. We use GPT-5.5 (OpenAI, 2026) to generate these perturbations, instructing it to preserve quantities, mathematical expressions, constraints, and the requested quantity.

Example: Paraphrase

Original: What is the smallest odd number with four different prime factors?
Paraphrase: What is the smallest odd number **that has four distinct** prime factors?
Answer: 1155

All four perturbed prompts for this example and details of perturbation construction are provided in Appendix B.1.

Diagnostic setup and drift measurement. To characterize semifactual sensitivity before RL training, we sample one response per original prompt from frozen Qwen3-4B-Base on 1,000 questions selected from DAPO-Math-17K. Sampling details are provided in Appendix B.2. Let $p_t^{(0)}$ and $p_t^{(k)}$ denote the probabilities assigned to the identical sampled token at position t under the original and the k-th perturbed prompts, respectively. The probability drift measured by the bounded-symmetric distance is defined as follows:

$$d_t^{(k)} = 2\tanh\left(\left|\log p_t^{(0)} - \log p_t^{(k)}\right|/2\right) = \frac{|p_t^{(0)} - p_t^{(k)}|}{\left(p_t^{(0)} + p_t^{(k)}\right)/2}. \tag{1}$$

Normalizing by the local mean probability avoids the scale bias of absolute differences and preserves relative probability changes for low-probability tokens. The distance remains within $[0, 2]$, bounding the effect of extreme probability ratios for numerical robustness. We average across the four perturbations to obtain $d_t = \frac{1}{4}\sum_{k=1}^{4} d_t^{(k)}$.

Heterogeneity in token sensitivity. Figure 2(a) shows both increases and decreases in token probability across a wide range of original confidence levels. Figure 2(b) reveals heterogeneity in semifactual sensitivity. Nonzero probability drift magnitudes span several orders of magnitude, while 12.82% of positions show no recorded change. Semifactual sensitivity also varies across token categories, as shown in Figure 2(c). Reflection markers adapted from (Wang et al., 2025a) and discourse connectives drawn from (Das et al., 2018) exhibit mean drifts of 2.74× and 2.32× the overall mean, respectively, whereas mathematical symbols and numbers exhibit 0.47× and 0.37×, respectively. This contrast suggests that tokens lexically associated with organizing and reflection are more sensitive to these interventions than mathematical symbols or numbers. See Appendix B.3 for more details on token category definitions and statistics, and Appendix B.4 for illustrative case studies of token-level semifactual sensitivity.

Suppressing unstable token candidates improves rollout accuracy. To test whether this semifactual sensitivity signal can help improve reasoning accuracy, we evaluate Qwen3-4B-Base under standard sampling and an online d-filtered decoding strategy on the same 1,000 selected questions. Specifically, at each decoding step, we compute drift for every candidate token in the vocabulary, rank non-EOS candidates in descending drift order, and mask the longest prefix whose cumulative probability mass does not exceed 0.8. We always retain the EOS token, and then renormalize the remaining probabilities. With model weights unchanged, accuracy rises from 15.8% to 30.0% (+14.2 points), as shown in Figure 3. This result demonstrates that semifactual sensitivity provides an actionable token-level signal and motivates its use for finer-grained credit assignment during RLVR training. More details and additional results are provided in Appendix C.

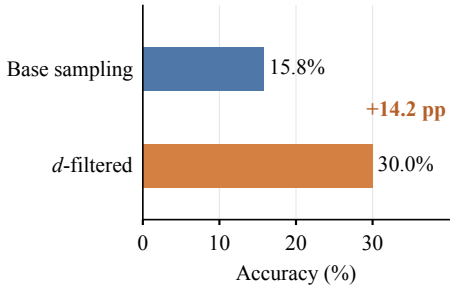

Figure 3: **Semifactual filtering improves rollout accuracy.** Suppressing high-drift token candidates during rollout generation improves Qwen3-4B-Base's accuracy by 14.2 percentage points on the same 1,000-question diagnostic panel.

3 Semifactual Credit-Augmented Policy Optimization

Building on the decoding benefit of suppressing unstable token candidates, we use semifactual stability to probe potential spurious dependence and refine token-level credit assignment during RLVR training. Our guiding principle is to reduce the advantages assigned to relatively unstable tokens without granting additional credit for stability alone, since stability does not imply correctness. In this section, we introduce Semifactual Credit-Augmented Policy Optimization (SCAPO), a GRPO-based algorithm that combines trajectory-level outcome advantages with token-level semifactual stability for finer-grained credit assignment. Figure 4 presents an overview of SCAPO, which consists of semifactual prompt intervention, token probability drift probing, relative stability signal construction, and policy optimization with GRPO advantages augmented by the semifactual credit signal.

3.1 Preliminaries: Group Relative Policy Optimization

Given a problem–answer pair $(x, a) \sim \mathcal{D}$, the rollout policy π_{old} samples a group of G responses $\{y_i\}_{i=1}^{G}$ under the original prompt x, where $y_i = (y_{i,1}, \ldots, y_{i,T_i})$ denotes the i-th response. Each response receives a verifiable outcome reward $R_i = \mathcal{V}(y_i, a) \in \{0, 1\}$. GRPO constructs a trajectory-level advantage from the relative rewards within the group (Shao et al., 2024):

$$A_i = \frac{R_i - \mu_R}{\sigma_R + \epsilon_A}, \qquad i = 1, \ldots, G, \tag{2}$$

where μ_R and σ_R are the mean and standard deviation of the group rewards, and $\epsilon_A > 0$ is a numerical stabilizer. The same A_i is applied to every valid token in y_i, so it cannot distinguish token-level dependence on task-irrelevant prompt features.

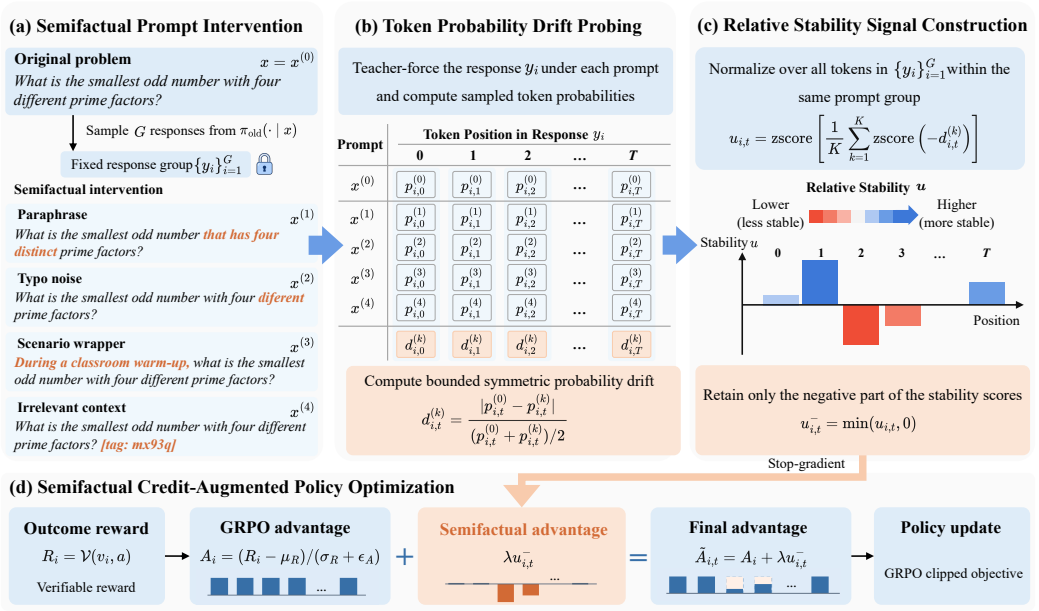

Figure 4: Overview of SCAPO. (a) Semifactual prompt interventions are paired with a fixed group of sampled responses. (b) Teacher forcing measures sampled-token probability drift under each intervention. (c) Group-wise normalization and aggregation yield relative stability scores, retaining only their negative part. (d) The scaled stability correction, with gradients stopped, is added to the GRPO advantage to produce token-level advantages for policy optimization.

For sampled token $y_{i,t}$, the importance ratio between the current policy π_θ and the rollout policy, together with its clipped counterpart, is

$$r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t} \mid x, y_{i,<t})}{\pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t})}, \qquad \bar{r}_{i,t}(\theta) = \text{clip}(r_{i,t}(\theta), 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}}). \tag{3}$$

Following the token-mean formulation (Yu et al., 2025), we update the policy by maximizing the clipped token-level policy gradient objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G} T_i} \sum_{i=1}^{G} \sum_{t=1}^{T_i} \min(r_{i,t}A_i, \bar{r}_{i,t}A_i)\right]. \tag{4}$$

3.2 Fixed-response semifactual probing

To obtain token-level credit signals, we examine how sampled-token probabilities change under semifactual prompt interventions while holding the response fixed. For each original prompt $x = x^{(0)}$, we construct K perturbed prompts $x^{(1)}, \ldots, x^{(K)}$ using the semifactual perturbations described in Section 2. As shown in Figure 4(a), the G responses sampled under the original prompt are held fixed across interventions.

We teacher-force the same response under each prompt, keeping both the sampled token $y_{i,t}$ and its response prefix $y_{i,<t}$ unchanged:

$$p_{i,t}^{(k)} = \pi_{\text{old}}\left(y_{i,t} \mid x^{(k)}, y_{i,<t}\right), \qquad k = 0, \ldots, K. \tag{5}$$

The rollout policy remains fixed during these probes and is updated as training proceeds.

Using the bounded-symmetric distance from Section 2, we convert these aligned probabilities into token-level drift, as illustrated in Figure 4(b):

$$d_{i,t}^{(k)} = \frac{\left| p_{i,t}^{(0)} - p_{i,t}^{(k)} \right|}{\left(p_{i,t}^{(0)} + p_{i,t}^{(k)} \right)/2}, \qquad k = 1, \ldots, K. \tag{6}$$

A larger $d_{i,t}^{(k)}$ indicates greater sensitivity of the sampled token's probability to the semifactual prompt intervention.

3.3 Group-relative stability estimation

We convert token-level drift into a relative stability signal within each prompt group. Let $\boldsymbol{d}^{(k)}$ collect the drifts at all valid token positions across the group's G responses for perturbation type k. For any vector $\boldsymbol{v}$ over the valid token positions, define the group-wise standardization

$$\mathrm{zscore}_x(\boldsymbol{v}) = \frac{\boldsymbol{v} - \mu_x(\boldsymbol{v})}{\sigma_x(\boldsymbol{v}) + \epsilon},$$

where the mean and standard deviation are computed over all valid response tokens in the group.

As shown in Figure 4(c), we standardize negative drift separately for each perturbation type, average the resulting scores, and normalize the aggregate:

$$\boldsymbol{u} = \mathrm{zscore}_x\left(\frac{1}{K} \sum_{k=1}^{K} \mathrm{zscore}_x\left(-\boldsymbol{d}^{(k)} \right) \right). \tag{7}$$

The resulting $u_{i,t}$ measures within-group relative stability, with negative values identifying relatively unstable tokens. Since stability alone does not imply correctness, we retain only the negative part of the signal:

$$u_{i,t}^{-} = \min(u_{i,t}, 0). \tag{8}$$

3.4 Policy optimization with semifactual credit

We add the semifactual credit signal to the original GRPO advantage, enabling tokens with the same outcome reward to receive different learning signals. As shown in Figure 4(d), the resulting token-level advantage is

$$\widetilde{A}_{i,t} = A_i + \lambda \, \mathrm{sg}\left(u_{i,t}^{-} \right), \tag{9}$$

where $\lambda \geq 0$ controls the augmentation strength and sg denotes stop-gradient. The semifactual correction is held fixed during each policy update.

This construction ensures $\widetilde{A}_{i,t} \leq A_i$, with equality when $u_{i,t} \geq 0$. When $A_i > 0$, the stability correction can weaken the positive reinforcement signal, and when $A_i < 0$, it strengthens the negative learning signal. The method thus refines outcome credit without treating semifactual stability as a token-level correctness label.

Substituting $\widetilde{A}_{i,t}$ into Eq. 4 gives SCAPO's policy objective:

$$\mathcal{J}_{\text{SCAPO}}(\theta) = \mathbb{E}\left[\frac{1}{\sum_{i=1}^{G} T_i} \sum_{i=1}^{G} \sum_{t=1}^{T_i} \min\left(r_{i,t} \widetilde{A}_{i,t}, \bar{r}_{i,t} \widetilde{A}_{i,t} \right) \right]. \tag{10}$$

SCAPO modifies only the advantage, leaving GRPO's importance ratio, clipping, and loss reduction unchanged. To guide early trajectory selection, we apply semifactual credit augmentation during an initial phase and then continue with standard GRPO. Specifically, at policy optimization step n, the coefficient is

$$\lambda = \begin{cases} \lambda_0, & 1 \leq n \leq N_0, \\ 0, & n > N_0, \end{cases} \tag{11}$$

where λ_0 and N_0 control the strength and duration of credit augmentation. Appendix D provides an analysis under a stochastic-gradient model.

4 EXPERIMENTS

4.1 SETUP

Datasets. We train on DAPO-Math-17K (Yu et al., 2025), a curated dataset of approximately 17,000 competition-level mathematics problems with integer answers. For SCAPO, each training prompt is paired with $K = 4$ precomputed perturbed prompts, corresponding to the four perturbation types introduced in Section 2.

Training. We initialize from Qwen3-4B-Base and Qwen3-1.7B-Base (Yang et al., 2025) and train directly with RL for 600 and 1,000 policy optimization steps, respectively. We use a rollout batch size of 128, an update batch size of 64, and 8 rollouts per prompt. The maximum response length is 16,384 tokens. We use the R1-style prompt template, binary rule-based rewards, and a constant learning rate of 10^{-6}. For SCAPO, we set $\lambda_0 = 0.01$ and apply credit augmentation for the first $N_0 = 120$ and 200 policy optimization steps on the 4B and 1.7B models, respectively. More detailed training hyperparameters are provided in Appendix E.1.

Evaluation. We evaluate on in-distribution competition-level mathematical reasoning benchmarks, including AIME 2024–2026 (Mathematical Association of America, 2026), AMC 2023–2025 (Mathematical Association of America, 2025), HMMT 2025–2026 (Harvard–MIT Mathematics Tournament, 2026), BRUMO 2025 (Brown University Math Olympiad, 2025), SMT 2025 (Stanford Math Tournament, 2025), Omni-Math (Gao et al., 2025a), Minerva (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). To assess out-of-distribution generalization, we evaluate NoOp-style distractor variants of AIME (Mirzadeh et al., 2025) and ThinkBench perturbations of AIME (Huang et al., 2025) for robustness to input perturbations. We further evaluate GPQA-Diamond (Rein et al., 2024), a graduate-level science question-answering benchmark, to examine transfer beyond mathematics. All main results use the final checkpoints with identical generation and scoring settings across methods within each benchmark. We sample at temperature 0.7 and top-$p = 0.9$ with a maximum response length of 16,384 tokens, score mathematical responses with Math-Verify (Kydlíček, 2025), and report average accuracy across sampled responses. Further evaluation details are provided in Appendix E.2.

Baselines. We consider the following representative RLVR methods: (1) *GRPO* (Shao et al., 2024): computes group-relative advantages from outcome rewards and applies token-level clipped policy updates. (2) *GSPO* (Zheng et al., 2025): uses sequence-level importance ratios and clipping for policy optimization. (3) *SAPO* (Gao et al., 2025b): replaces hard clipping with smooth, advantage-dependent gating. (4) *CF-GRPO* (Khandoga et al., 2026): estimates token-level credit through counterfactual masking of reasoning spans. (5) *FIPO* (Ma et al., 2026): reweights token advantages using discounted future KL. At each model scale, comparisons match the training data, reward function, prompt template, optimizer, rollout settings, training-step budget, and evaluation protocol. Method-specific hyperparameters are provided in Appendix E.1.

4.2 MAIN RESULTS

Mathematical reasoning. Table 1 shows that SCAPO improves over GRPO on all eight mathematical reasoning metrics at both model scales. AIME accuracy increases by 5.63 and 4.17 percentage points on the 4B and 1.7B models, respectively, while HMMT improves by 4.46 and 2.98 points. SCAPO achieves the highest accuracy on six of eight metrics at 4B and all eight at 1.7B, with the highest average accuracy at both scales. Additional results show that SCAPO achieves the highest Pass@128 on AIME and AMC at both model scales (see Appendix F.2).

Out-of-distribution generalization. SCAPO achieves the highest accuracy on all three out-of-distribution benchmarks at both model scales (Table 1). On GPQA-Diamond, accuracy improves over GRPO from 36.26% to 42.45% at 4B and from 27.42% to 31.25% at 1.7B, gains of 6.19 and 3.83 percentage points, respectively. Together with the improvements on NoOp-AIME and ThinkBench-AIME, these results show that the gains extend to both perturbed mathematical problems and scientific reasoning beyond the training domain.

Table 1: Overall accuracy on competition-level mathematical reasoning and out-of-distribution benchmarks (%). Base denotes the model before RL. AIME variants aggregate 2024–2026, while HMMT and AMC aggregate 2025–2026 and 2023–2025, respectively. Year-specific results are provided in Appendix F.1. Bold and underline indicate the best and second-best results, respectively.

	Benchmark	Base	GRPO	GSPO	SAPO	CF-GRPO	FIPO	SCAPO
	Qwen3-4B-Base							
In-Distribution	AIME 24–26	7.64	22.08	23.47	24.03	22.64	25.42	**27.71**
	AMC 23–25	37.40	59.74	64.71	65.45	62.06	64.57	**65.65**
	HMMT 25–26	2.08	11.71	13.69	15.18	14.09	15.28	**16.17**
	BRUMO 25	18.54	28.96	33.75	33.54	32.08	35.00	**37.71**
	SMT 25	12.74	27.12	26.06	27.12	26.77	28.18	**30.31**
	Omni-Math	26.48	35.45	38.71	38.18	36.41	38.60	**39.42**
	Minerva	31.62	40.07	39.71	41.18	41.54	**44.49**	43.38
	Olympiad	38.58	54.30	55.64	54.75	54.75	**57.57**	56.82
	Avg.	21.88	34.93	36.97	37.43	36.29	38.64	**39.65**
Out-of-Distribution	NoOp-AIME 24–26	6.74	20.56	20.14	20.42	21.74	21.74	**23.61**
	ThinkBench-AIME 24–26	6.53	22.64	22.08	23.47	23.54	23.89	**26.11**
	GPQA-Diamond	33.78	36.26	41.92	41.16	36.51	40.15	**42.45**
	Avg.	15.68	26.48	28.05	28.35	27.26	28.59	**30.72**
	Qwen3-1.7B-Base							
In-Distribution	AIME 24–26	3.61	7.71	7.36	10.00	6.74	7.01	**11.88**
	AMC 23–25	23.33	32.97	32.82	38.58	31.99	32.48	**41.39**
	HMMT 25–26	0.69	1.98	2.08	4.66	2.98	1.49	**4.96**
	BRUMO 25	6.88	13.54	11.88	17.71	14.37	10.00	**19.17**
	SMT 25	4.83	6.96	6.84	10.14	8.02	8.14	**10.85**
	Omni-Math	16.73	20.45	21.48	22.62	20.77	21.77	**24.74**
	Minerva	21.69	27.94	27.21	30.15	28.31	28.31	**30.88**
	Olympiad	21.81	31.31	33.68	37.83	33.98	32.64	**40.80**
	Avg.	12.45	17.86	17.92	21.46	18.39	17.73	**23.08**
Out-of-Distribution	NoOp-AIME 24–26	2.36	6.11	5.97	8.06	6.60	6.60	**8.40**
	ThinkBench-AIME 24–26	3.19	7.57	7.22	9.58	6.94	7.01	**12.01**
	GPQA-Diamond	22.37	27.42	29.21	30.96	27.61	28.24	**31.25**
	Avg.	9.31	13.70	14.13	16.20	13.72	13.95	**17.22**

4.3 Ablation Studies

We ablate the source and sign of token-level credit to examine their roles in SCAPO's performance gains, and the results are shown in Table 2.

Table 2: Ablations on Qwen3-1.7B-Base. SCAPO uses negative-only corrections.

Method	AIME 24–26	AMC 23–25	HMMT 25–26	BRUMO 25	SMT 25	Avg.
GRPO	7.71	32.97	1.98	13.54	6.96	12.63
SCAPO	**11.88**	**41.39**	**4.96**	**19.17**	**10.85**	**17.65**
Credit signal source						
Random shuffle	8.33	32.78	2.28	15.42	9.43	13.65
Counterfactual	6.32	30.51	1.69	12.92	6.96	11.68
Correction sign						
All-sign	10.76	38.83	3.77	15.21	9.67	15.65
Positive-only	9.79	34.10	2.58	15.42	8.37	14.05

Source of the credit signal. The random shuffle ablation permutes semifactual corrections within each response, preserving their values but breaking the original token alignment. The counterfactual ablation replaces answer-preserving semifactual prompts with answer-changing perturbations, while keeping the credit construction unchanged. SCAPO performs best on the five competition-level benchmarks, followed consistently by random shuffle and the counterfactual control. On AIME 24–26, SCAPO leads these controls by 3.55 and 5.56 percentage points, respectively. These results support aligning credit with semifactual sensitivity, since drift under answer-changing perturbations may reflect legitimate changes in token predictions rather than spurious dependence.

Sign of the credit correction. We compare negative-only, all-sign, and positive-only corrections by retaining negative, all, or positive stability scores. SCAPO's negative-only correction achieves the highest average accuracy across the five competition-level benchmarks. This supports reducing credit for relatively unstable tokens without rewarding stability alone.

4.4 Computational Efficiency

As shown in Figure 5, semifactual probing and credit construction account for only 1.8% of the measured training runtime in the 4B experiment, far less than autoregressive rollout. The probes reuse sampled responses through teacher forcing, avoiding fresh generation under each perturbed prompt. They are also confined to the initial credit augmentation phase and incur no further cost during subsequent optimization. Together, these properties keep probing a small fraction of total runtime despite evaluating multiple prompts. Training efficiency is therefore more strongly affected by major stages such as autoregressive generation, whose cost depends on response length.

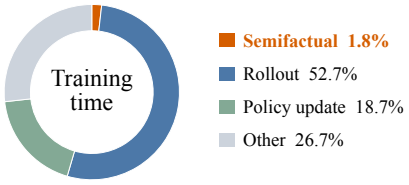

Figure 5: Training time breakdown of SCAPO on Qwen3-4B-Base. Other includes validation, reward computation, old policy log-probability computation, and miscellaneous overhead.

5 Related Work

Reinforcement learning with verifiable rewards. RLVR has advanced mathematical reasoning by optimizing automatically verifiable outcomes, with critic-free group-relative methods enabling direct RL training of base models (Shao et al., 2024; Guo et al., 2025; Zeng et al., 2025). Subsequent work improves normalization and sampling (Liu et al., 2025; Yu et al., 2025), adapts importance weighting and clipping (Zheng et al., 2025; Gao et al., 2025b), and promotes exploration through entropy control or off-policy guidance (Cui et al., 2025; Yan et al., 2025). However, response-level advantages remain too coarse to distinguish individual reasoning tokens. Studies of reasoning-path coverage further distinguish improved sampling efficiency from expansion of a model's reasoning repertoire (Yue et al., 2025). Like prior RLVR methods, our approach optimizes verifiable outcome rewards, but leverages semifactual stability to refine credit assignment at the token level.

Credit assignment for reasoning. Fine-grained feedback can be learned from process annotations or outcome labels (Lightman et al., 2024; Wang et al., 2024; Cui et al., 2026), or estimated through Monte Carlo rollouts (Kazemnejad et al., 2025). Token entropy, confidence, eligibility traces, and future policy divergence also guide token selection and weighting (Wang et al., 2025b; Xie et al., 2026; Mou et al., 2026; Ma et al., 2026). Perturbation- and gradient-based attribution methods estimate how reasoning tokens or spans contribute to final-answer predictions, providing signals for selective supervised fine-tuning and importance-weighted policy updates (Ruan et al., 2025; Khandoga et al., 2026; Li et al., 2026). Like prior work on fine-grained credit assignment, we use token-level signals to guide policy updates. Our contribution is to derive negative-only advantage corrections from the sensitivity of a fixed response to semifactual prompt interventions, using teacher-forced token probabilities rather than output-side attribution through span masking or continuation regeneration.

6 Conclusion

We presented SCAPO, a causally inspired approach for finer-grained token credit assignment in RLVR. By probing potential token-level spurious dependence through fixed-response sensitivity to semifactual prompt interventions, SCAPO applies negative-only corrections to GRPO advantages, selectively reducing credit for relatively unstable tokens. Across Qwen3-4B-Base and Qwen3-1.7B-Base, SCAPO outperforms GRPO on all reported benchmark metrics and achieves the best results on most benchmarks among the compared RLVR methods. These results highlight that semifactual interventions can serve not only as diagnostic probes of brittle reasoning but also as training signals that complement outcome rewards. Future work will explore this credit augmentation in larger models, more diverse architectures, and domains beyond mathematical reasoning.

AI USE STATEMENT

We used generative AI to generate semifactual perturbation data, polish writing, correct grammar, and assist with code debugging and script development. The authors take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Our work is easy to reproduce. The original datasets and pretrained models used in our experiments are publicly available, and detailed experimental hyperparameters are provided in Appendix E. We provide our code, developed on top of the open-source EasyR1 framework, together with all training and evaluation data in the Supplementary Material.

REFERENCES

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Brown University Math Olympiad. BrUMO 2025, 2025. URL https://www.brumo.org/archive. Accessed: 2026-07-15.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*, 2026.

Debopam Das, Tatjana Scheffler, Peter Bourgonje, and Manfred Stede. Constructing a lexicon of english discourse connectives. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 360–365, 2018.

Zhizhang Fu, Guangsheng Bao, Hongbo Zhang, Chenkai Hu, and Yue Zhang. Correlation or causation: Analyzing the causal structures of llm and lrm reasoning process. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2986–2999, 2026.

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, Zhengyang Tang, Benyou Wang, Daoguang Zan, Shanghaoran Quan, Ge Zhang, Lei Sha, Yichang Zhang, Xuancheng Ren, Tianyu Liu, and Baobao Chang. Omni-math: A universal olympiad level mathematic benchmark for large language models. In *International Conference on Learning Representations*, volume 2025, pp. 100540–100569, 2025a.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025b.

Nelson Goodman. The problem of counterfactual conditionals. *The journal of philosophy*, 44(5): 113–128, 1947.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean

Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Harvard–MIT Mathematics Tournament. HMMT: Past tournaments, 2026. URL https://beta.hmmt.org/www/archive/problems. Accessed: 2026-07-15.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 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.

Shulin Huang, Linyi Yang, Yan Song, Shawn Chen, Leyang Cui, Ziyu Wan, Qingcheng Zeng, Ying Wen, Kun Shao, Weinan Zhang, Jun Wang, and Yue Zhang. Thinkbench: Dynamic out-of-distribution evaluation for robust llm reasoning. *Advances in Neural Information Processing Systems*, 38, 2025.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. VinePPO: Refining credit assignment in RL training of LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29557–29590. PMLR, 2025.

Eoin M Kenny and Mark T Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11575–11585, 2021.

Mykola Khandoga, Rui Yuan, and Vinay Kumar Sankarapu. Beyond uniform credit: Causal credit assignment for policy optimization. *arXiv preprint arXiv:2602.09331*, 2026.

Hynek Kydlíček. Math-Verify: Math verification library, 2025. URL https://github.com/huggingface/math-verify. Version 0.9.0.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 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.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025.

Jinghui Lu, Linyi Yang, Brian Mac Namee, and Yue Zhang. A rationale-centric framework for human-in-the-loop machine learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6986–6996, 2022.

Chiyu Ma, Shuo Yang, Kexin Huang, Jinda Lu, Haoming Meng, Shangshang Wang, Bolin Ding, Soroush Vosoughi, Guoyin Wang, and Jingren Zhou. FIPO: Eliciting deep reasoning with future-KL influenced policy optimization. In *Conference on Language Modeling (COLM)*, 2026.

Mathematical Association of America. American Mathematics Competitions (AMC), 2025. URL https://maa.org/student-programs/amc/. Accessed: 2026-07-15.

Mathematical Association of America. American Invitational Mathematics Examination (AIME), 2026. URL https://maa.org/maa-invitational-competitions/. Accessed: 2026-07-15.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025.

Chaoli Mou, Zhan Zhuang, Xinning Chen, and Yu Zhang. Beyond uniform credit assignment: Selective eligibility traces for rlvr. *arXiv preprint arXiv:2605.05965*, 2026.

OpenAI. Learning to reason with LLMs, September 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

OpenAI. GPT-5.5 system card, April 2026. URL https://openai.com/index/gpt-5-5-system-card/.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

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

Zhiwen Ruan, Yixia Li, He Zhu, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. Enhancing large language model reasoning via selective critical token fine-tuning. *arXiv preprint arXiv:2510.10974*, 2025.

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

Stanford Math Tournament. Stanford Math Tournament 2025: Tests and solutions, 2025. URL https://www.stanfordmathtournament.org/past-tests/SMT/2025. Accessed: 2026-07-15.

Chenlong Wang, Yuanning Feng, Dongping Chen, Zhaoyang Chu, Ranjay Krishna, and Tianyi Zhou. Wait, we don't need to" wait"! removing thinking tokens improves reasoning efficiency. In *EMNLP (Findings)*, pp. 7459–7482, 2025a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38:115452–115486, 2025b.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *Findings of the association for computational linguistics: NAACL 2022*, pp. 1719–1729, 2022.

Xiaoxuan Wang, Han Zhang, Haixin Wang, Yidan Shi, Ruoyan Li, Kaiqiao Han, Chenyi Tong, Haoran Deng, Alexander K. Taylor, Renliang Sun, Yanqiao Zhu, Jason Cong, Yizhou Sun, and Wei Wang. ARLArena: A unified framework for stable agentic reinforcement learning. In *International Conference on Machine Learning*, 2026.

Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19057–19076, 2026.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. *Advances in Neural Information Processing Systems*, 38:117157–117186, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2025.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *Advances in Neural Information Processing Systems*, 38:57654–57689, 2025.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. SimpleRL-Zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Conference on Language Modeling (COLM)*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Limitations

Our training experiments are limited to mathematical reasoning with dense Qwen3 models of 1.7B and 4B parameters trained on DAPO-Math-17K. Although GPQA-Diamond provides evidence of transfer to scientific reasoning, the effectiveness of SCAPO at larger model and data scales and with broader training domains remains to be evaluated. Future work should also examine other architectures, including mixture-of-experts (MoE) and hybrid models.

B Semifactual Perturbations and Token-Level Analysis

B.1 Perturbation construction

We request GPT-5.5 (OpenAI, 2026) to generate four perturbed prompts per DAPO-Math-17K problem, using the full system and user prompts in Table 3.

Table 3: Full prompts used to generate semifactual perturbations. The user template is filled with the original problem and reference answer.

System prompt

```
Given a math problem and its original answer, produce exactly 4 prompt perturbations.

Core requirements:
- Preserve the same mathematical problem.
- Preserve all numeric values exactly as written.
- Preserve all constraints, operations, requested quantity, and the original answer.
- Produce exactly one perturbation for each perturbation_type: paraphrase, typo_noise,
scenario_wrap, irrelevant_context.
- Do not add any condition that helps solve the problem.
- Do not solve the problem.
- Do not reveal reasoning.
- Return JSON only.

Math and LaTeX preservation:
- Treat all math expressions as frozen text. This includes text inside $...$, $$...$$, \(...\),
\[...\], and standalone symbolic expressions such as variables, equations, inequalities,
fractions, powers, percentages, and units.
- Do not rewrite, reorder, escape, unescape, or modify any LaTeX/math content.
- Preserve every LaTeX command exactly as it appears in the original parsed string: examples
include \sin, \angle, \frac, \sqrt, \triangle, \overline.
- In JSON, escape backslashes correctly so that the parsed string contains exactly one backslash
before each LaTeX command. Never output doubled LaTeX commands like \\sin, and never output raw
control characters such as tab or form feed.
- Do not alter variables, point labels, entity labels, numbers, units, comparison signs,
operations, or answer format.

Type-specific requirements:

- paraphrase:
  Rewrite only the natural-language prose that sits outside math expressions. Keep the same
  scenario, the same entities, the same event and causal order, and every mathematical fact
  unchanged. Do not introduce a new setting or new objects.

- typo_noise:
  Introduce exactly one character-level edit (insert, delete, substitute, or swap two adjacent
  characters) in one ordinary English word. The chosen word must NOT be math, LaTeX, a variable,
  a number, a unit, a proper name, a label, or a mathematical keyword such as "mean", "median",
  "integer", "prime", "even", "odd". The perturbed text must differ from the original by at least
  that one non-whitespace character; whitespace-only differences are invalid. The word must still
  be recognizable to a human reader.

- scenario_wrap:
  Add a light scenario frame around the original problem. This can be a short prefix, a short
  suffix, or one brief background phrase inserted into the natural-language part of the problem.
  This should be a small irrelevant context shift, not a full rewrite into a new domain.

  Keep the original people, objects, quantities, units, formulas, variables, constraints,
  operations, event order, and requested quantity unchanged. Do not replace core entities or
  items. Do not change "clips" into "packages", "wallet" into "sensor kit", "flowers" into
  "crates", or a person into a machine.

  Good examples:
    * "At a school craft booth, [original problem]"
    * "During a budgeting exercise, [original problem]"
    * "While tracking a reading plan, [original problem]"
    * "[Original first sentence] This is part of a classroom activity. [remaining original
    problem]"
    * "[Original problem] This was recorded in a simple daily log."
    * "[Original problem] The setting is an ordinary after-school situation."

  Bad examples:
    * Rewriting the full story into a different domain.
    * Replacing the actors or objects with different actors or objects.
    * Adding hints, extra facts, new conditions, or solution-relevant context.
    * Adding numbers, formulas, variables, quantities, or mathematical facts.

- irrelevant_context:
```

System prompt (continued)

```
  Keep the original problem text exactly unchanged. Append exactly one short meaningless
  distractor at the very end (after the final question mark or full stop).

  Acceptable distractor shapes (pick one; vary the shape and content across samples):
    * A vacuous logical tautology, e.g. "and true is true", "note that 1 = 1", "recall that each
    thing equals itself".
    * A short random alphanumeric string of length 6 to 12 characters, e.g. "5XeflW1ZJc",
    "qP7mz39a".
    * A bracketed decorative tag, e.g. "[tag: mx93q]", "[marker-a7f]".

  The distractor MUST NOT contain:
    * Any mathematical fact, quantity, variable, unit, equation, or operator that could interact
    with the problem.
    * Any hint about the problem, the subject area, or the solution approach.
    * Any statement about the characters, objects, or scenario that appear in the problem.
    * Any framing or meta-language such as "question", "problem", "exercise", "consider",
    "observe", "for practice", "simple", "word problem".

Return this JSON shape:
{
  "perturbations": [
    {
      "perturbed_question": "...",
      "perturbation_type": "paraphrase | typo_noise | scenario_wrap | irrelevant_context"
    }
  ]
}
```

User prompt

```
Original problem:
{problem}

Original answer:
{answer}
```

Example: Four semifactual perturbations

Original: What is the smallest odd number with four different prime factors?

Paraphrase: What is the smallest odd number **that has four distinct** prime factors?

Typo noise: What is the smallest odd number with four **diferent** prime factors?

Scenario wrapper: **During a classroom warm-up,** what is the smallest odd number with four different prime factors?

Irrelevant context: What is the smallest odd number with four different prime factors? **[tag: mx93q]**

Answer: 1155

B.2 Diagnostic setup

We randomly select 1,000 DAPO-Math-17K questions and their four perturbed prompts to form a fixed diagnostic subset dataset.

Frozen Qwen3-4B-Base generates one response per original prompt, with temperature 1.0, top-$p = 1.0$, and a limit of 8,192 response tokens. We then teacher-force the same response tokens under the original prompt and all four perturbed prompts. Each response token contributes one mean drift d_t across the four perturbations, yielding 870,586 original response token drifts.

B.3 Token taxonomy and statistics

We assign each decoded response token to one of the eight mutually exclusive categories in Table 4. Reflection markers are identified using a 16-entry lexicon adapted from prior work (Wang et al., 2025a) and our generation traces. Discourse connectives are identified using 80 single-part DiMLex-Eng forms that contain no whitespace (Das et al., 2018). Lexical matching ignores case, surrounding whitespace, and edge punctuation, with reflection markers taking precedence over discourse connectives.

Table 4: Token categories and drift statistics underlying Figure 2(c). *Share* is the percentage of the 870,586 response token positions assigned to each category. *Mean* is the category's average drift d_t. *Relative mean* is this average divided by the overall mean (0.01640). ▁ denotes a space and `\n` a newline.

Category	Classification rule	Token examples	Share (%)	Mean d_t	Relative mean
Reflection marker[a]	Matches the reflection lexicon after lexical normalization.	▁any, ▁check, Now, ▁However, ▁verify, ▁again	0.29	0.04496	2.74
Discourse connective	Matches the 80-form DiMLex-Eng lexicon after normalization. Excludes reflection markers.	▁and, ▁for, ▁if, ▁or, Thus, ▁Therefore	2.68	0.03807	2.32
Word	Contains a Unicode letter and matches neither lexical category nor the math-symbol rule.	▁the, ▁of, ▁is, ▁to, ▁we, ▁in	40.92	0.02506	1.53
Punctuation	All non-whitespace characters have Unicode category P*. Excludes math symbols.	, . : #), :\n	6.75	0.01956	1.19
Math symbol	Begins with \ or contains only characters from the fixed math/LaTeX symbol set. Excludes numbers.	▁\, ▁=, ▁+, ▁-, {, }	26.62	0.00768	0.47
Number	Matches a signed integer, decimal, or percentage expression.	0, 1, 2, 3, 4, 5	15.31	0.00612	0.37
Whitespace-only	Decodes entirely to Unicode whitespace.	▁, ▁▁, \n, \n\n	6.71	0.00718	0.44
Other	Control tokens, undecodable UTF-8 fragments, or remaining forms not covered above.	<|endoftext|>, ```, ``, $,	0.72	0.02925	1.78

[a] The reflection lexicon contains the following 16 words: again, ah, alternative, alternatively, another, any, but, check, hmm, however, maybe, now, oh, other, verify, wait.

B.4 Complete-response case studies

Figures 6 and 7 illustrate two cases with their original and perturbed prompts. Response tokens are shaded by their mean probability drift across the four prompt perturbations, with darker shading indicating greater sensitivity. These examples show that even when the final answer is correct, the probabilities of individual response tokens can change substantially under answer-preserving prompt perturbations.

Case 1 | Compound interest

Original prompt

Kimberly borrows 1000 dollars from Lucy, who charged interest of 5% per month (which compounds monthly). What is the least integer number of months after which Kimberly will owe more than twice as much as she borrowed?

Paraphrase

Kimberly takes a loan of 1000 dollars from Lucy, who charged interest of 5% per month (which compounds monthly). What is the smallest integer number of months after which Kimberly will owe more than twice as much as she borrowed?

Typo noise

Kimberly borows 1000 dollars from Lucy, who charged interest of 5% per month (which compounds monthly). What is the least integer number of months after which Kimberly will owe more than twice as much as she borrowed?

Scenario wrapper

During a budgeting exercise, Kimberly borrows 1000 dollars from Lucy, who charged interest of 5% per month (which compounds monthly). What is the least integer number of months after which Kimberly will owe more than twice as much as she borrowed?

Irrelevant context

Kimberly borrows 1000 dollars from Lucy, who charged interest of 5% per month (which compounds monthly). What is the least integer number of months after which Kimberly will owe more than twice as much as she borrowed? qPzmXaRt

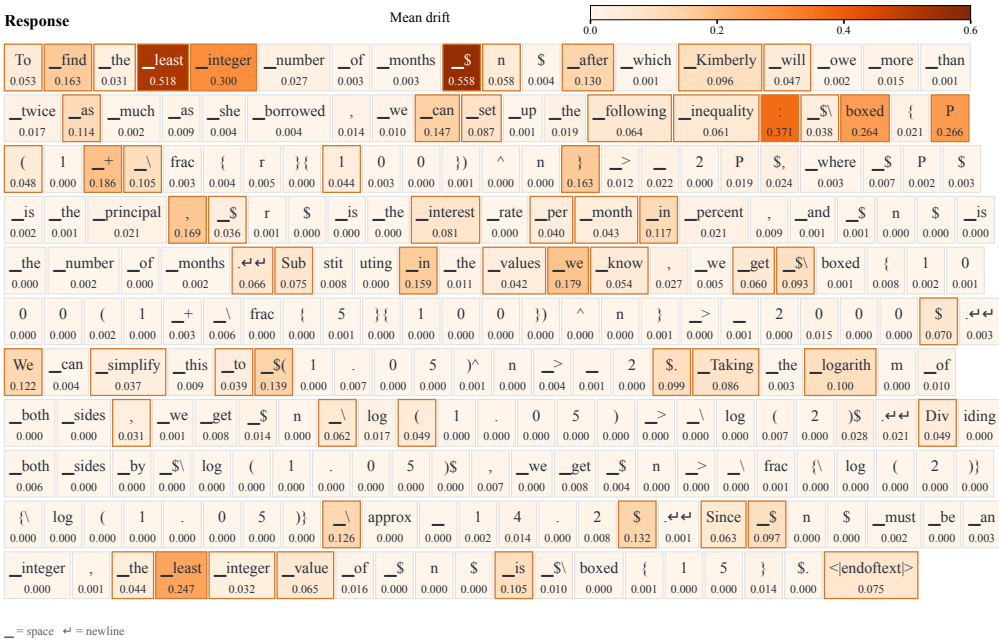

Figure 6: Token-level semifactual sensitivity in a compound-interest problem.

Case 2 | Five-digit numeral base

Original prompt

100_{10} in base b has exactly 5 digits. What is the value of b?

Paraphrase

In base b, 100_{10} has exactly 5 digits. What is the value of b?

Typo noise

100_{10} in base b has exatly 5 digits. What is the value of b?

Scenario wrapper

During a classroom warm-up, 100_{10} in base b has exactly 5 digits. What is the value of b?

Irrelevant context

100_{10} in base b has exactly 5 digits. What is the value of b? [marker-a7f]

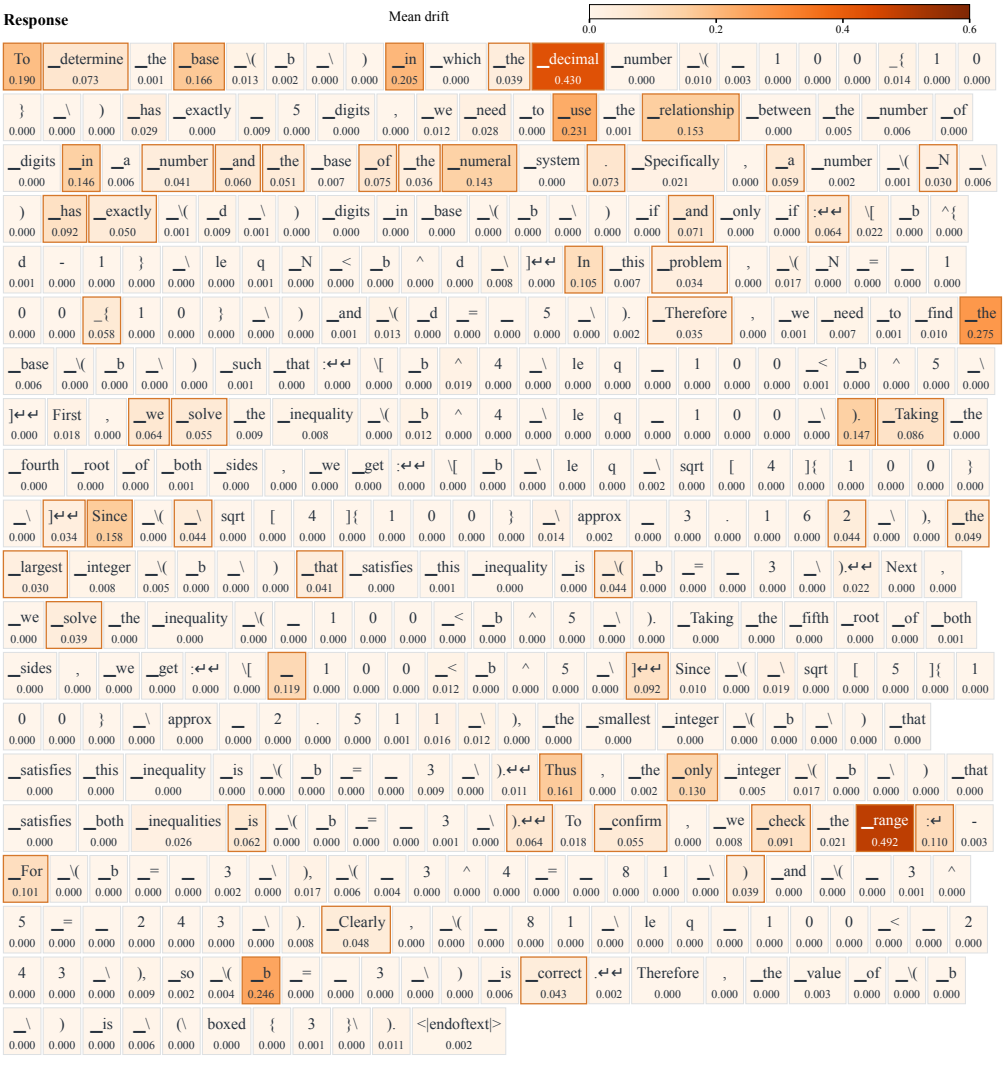

Figure 7: Token-level semifactual sensitivity in a numeral-base problem.

C d-FILTERED DECODING

C.1 DECODING SETUP

We use frozen Qwen3-4B-Base on the 1,000-question panel in Appendix B.2. Both base sampling and d-filtered decoding use the same prompt template, temperature of 1.0, top-$p = 1.0$, and a maximum response length of 8,192 tokens. We cap the cumulative clean-policy probability mass of masked tokens at 0.8, which is the sole filtering hyperparameter. For each question, we independently generate one response under each decoding strategy.

C.2 FILTERING PROCEDURE

At each decoding step, we compute the mean bounded-symmetric drift of every vocabulary token across the four perturbed prompts. We sort all token candidates by decreasing drift. We mask token candidates in this order until adding the next candidate would exceed the mass cap. The EOS token is always retained. The cap limits the total probability mass removed, not the fraction of tokens masked.

Writing $p(v)$ for the clean-policy probability at the current prefix and $\mathcal{M}$ for the masked set, the filtered sampling distribution is

$$p_{\text{filter}}(v) = \frac{p(v)\,\mathbf{1}\{v \notin \mathcal{M}\}}{1 - \sum_{w \in \mathcal{M}} p(w)}. \tag{12}$$

We sample a proposal from the clean policy and accept it if it is unmasked. Otherwise, we reject it and resample from the distribution renormalized over unmasked tokens. Drift and the masked set are recomputed at each step using the current filtered-response prefix.

C.3 RESULTS

As shown in Table 5, filtering yields 202 incorrect-to-correct and 60 correct-to-incorrect transitions, a gain of 14.20 percentage points. On average, 50.58 token proposals are rejected and resampled per filtered response (4.58% of decoding steps), with at least one such rejection in each of the 1,000 filtered responses.

Table 5: Decoding results on 1,000 questions with frozen Qwen3-4B-Base. The lower block reports question counts for the four combinations of baseline and filtered answer correctness.

Metric	Base sampling	d-filtered
Correct answers	158	300
Accuracy (%)	15.80	**30.00**
Paired correctness outcomes (question counts)		
	Filtered correct	Filtered incorrect
Baseline correct	98	60
Baseline incorrect	202	640

D OPTIMIZATION ANALYSIS

We analyze credit augmentation under a stochastic-gradient model with a fixed objective F.

Lemma D.1 (Bounded auxiliary gradient)**.** *Let $\widehat{g}_\lambda$ and $\widehat{g}_0$ be the batch gradients of Eqs. 10 and 4 on the same complete prompt groups, with advantages and stability scores held fixed. If importance ratios are at most R and token log-probability gradients have norm at most B, then*

$$\|\widehat{g}_\lambda - \widehat{g}_0\| \le \lambda H, \qquad H := RB/2. \tag{13}$$

Proof. Within each group, write $\langle\cdot\rangle$ for the valid-token mean and $c = (-u)_+$. Eq. 7 gives $\langle u\rangle = 0$ and $\langle u^2\rangle \le 1$, hence

$$\langle c\rangle = \tfrac{1}{2}\langle |u|\rangle \le \tfrac{1}{2}\sqrt{\langle u^2\rangle} \le \tfrac{1}{2}.$$

At differentiable points, the coefficient of a token's log-probability gradient in the clipped surrogate is rA, $r\max(A,0)$, or $r\min(A,0)$, each r-Lipschitz in A. Replacing A by $A-\lambda c$ therefore changes the group gradient by at most $\lambda RB\langle c\rangle \le \lambda H$. Token-mean aggregation across groups preserves this bound. □

Stochastic-gradient model. Consider

$$\theta_{n+1} = \theta_n + \eta\widehat{v}_n, \qquad \widehat{v}_n = \widehat{g}_n + \lambda_n\widehat{h}_n, \tag{14}$$

where $\widehat{g}_n$ is the baseline gradient and $\lambda_n\widehat{h}_n$ is the gradient difference induced by credit augmentation, with $\|\widehat{h}_n\| \le H$ by Lemma D.1. Let $\mathbb{E}_n$ denote the expectation conditional on the history before update n. Assume F is L-smooth, bounded above by $F_{\sup}$, and

$$\|\mathbb{E}_n\widehat{g}_n - \nabla F(\theta_n)\| \le \beta, \qquad \mathbb{E}_n\|\widehat{v}_n - \mathbb{E}_n\widehat{v}_n\|^2 \le \sigma^2. \tag{15}$$

The two gradient components may be correlated. The parameter β allows bias in the baseline direction.

Theorem D.2 (Finite-duration augmentation)**.** *Under the above assumptions, let* $0 < \eta \le 1/L$ *and use the schedule in Eq. 11. For any* $T \ge 1$*, set* $\Delta_F = F_{\sup} - \mathbb{E}F(\theta_1)$ *and* $m_T = \min(T, N_0)$*. Then*

$$\frac{1}{T}\sum_{n=1}^{T}\mathbb{E}\|\nabla F(\theta_n)\|^2 \le \frac{2\Delta_F}{\eta T} + L\eta\sigma^2 + \beta^2 + \left(2\beta H\lambda_0 + H^2\lambda_0^2\right)\frac{m_T}{T}. \tag{16}$$

Proof. Write $v_n = \nabla F(\theta_n)$ and $b_n = \mathbb{E}_n\widehat{v}_n - v_n$, so $\|b_n\| \le \beta + H\lambda_n$. Smoothness and Eq. 15 give

$$\begin{aligned}\mathbb{E}_n F(\theta_{n+1}) &\ge F(\theta_n) + \eta\langle v_n, v_n + b_n\rangle - \tfrac{L\eta^2}{2}\left(\|v_n + b_n\|^2 + \sigma^2\right)\\ &\ge F(\theta_n) + \tfrac{\eta}{2}\|v_n\|^2 - \tfrac{\eta}{2}(\beta + H\lambda_n)^2 - \tfrac{L\eta^2}{2}\sigma^2,\end{aligned}$$

where the second line uses $\eta L \le 1$ and $2\langle v, v+b\rangle = \|v\|^2 + \|v+b\|^2 - \|b\|^2$. Taking expectations, summing over n, and using $\mathbb{E}F(\theta_{T+1}) \le F_{\sup}$ yields the claim. □

Implication for the schedule. For an unbiased baseline ($\beta = 0$), the augmentation term is $H^2\lambda_0^2\min(T, N_0)/T$. With fixed λ_0 and N_0, this contribution to the average stationarity bound decays as N_0/T after augmentation ends. Here T is an arbitrary observation horizon, not a prescribed training budget.

E Additional Experimental Details

E.1 Training hyperparameters

Table 6 lists shared training hyperparameters and method-specific settings. A rule-based verifier assigns reward 1 to a correct boxed answer and 0 otherwise, and no separate format reward is used. All models share the following prompt template for mathematical reasoning during training and inference:

Prompt template

```
{question}
```

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

All methods, including GSPO and SAPO, use token-mean loss reduction for better performance on mathematical reasoning tasks (Yu et al., 2025; Wang et al., 2026).

Table 6: Shared and method-specific training hyperparameters. Batch sizes count prompts unless stated otherwise. Each rollout step contains two policy optimization steps. Clipping offsets are relative to 1.

Hyperparameter	**Qwen3-4B-Base**	**Qwen3-1.7B-Base**
Data and rollout settings		
Training dataset	DAPO-Math-17K (Yu et al., 2025)	
Maximum prompt length	1,024	
Maximum response length	16,384	
Rollout batch size (prompts)	128	
Responses per prompt (G)	8	
Sampling temperature / top-p	1.0 / 1.0	
Optimization settings		
Optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$	
Learning rate	10^{-6}	
Learning-rate schedule	Constant	
Weight decay	0.01	
Gradient clipping norm	1.0	
Update mini-batch size (prompts)	64	
Loss reduction	Token-mean	
KL regularization	Disabled	
Training settings		
Rollout steps	300	500
Policy optimization steps	600	1,000
Mixed precision	BF16	
Method-specific settings		
SCAPO: Correction strength (λ_0)	0.01	
SCAPO: Augmentation duration (N_0)	120	200
GRPO / CF-GRPO / FIPO / SCAPO:		
Policy clipping (lower / upper)	0.20 / 0.28	
Dual-clip coefficient	10.0	
GSPO: Policy clipping (lower / upper)	3×10^{-4} / 4×10^{-4}	
SAPO: Gate temperatures ($\tau_{\text{pos}}/\tau_{\text{neg}}$)	1.0 / 1.05	
CF-GRPO: Maximum probed spans per response	10	
CF-GRPO: Span length	5–50	
CF-GRPO: Token-weight bounds	$[0.5, 4.0]$	
FIPO: Future-KL decay	32.0	
FIPO: Influence-weight clipping	$[1.0, 1.2]$	
FIPO: Safety threshold	10.0	

Each rollout batch contains 128 prompt groups and is optimized once in mini-batches of 64 groups (512 responses), giving two policy optimization steps per rollout step. Thus, $N_0 = 120/200$ policy optimization steps correspond to 60/100 rollout steps for 4B/1.7B. Diagnostic checkpoint indices use rollout steps unless stated otherwise.

E.2 Evaluation protocol and training curves

Table 7 summarizes the evaluation datasets and sample counts.

Table 7: Evaluation benchmarks and sample counts. Slash-separated problem counts follow the listed years. *For GPQA-Diamond, we enumerate all $4! = 24$ permutations of the multiple-choice options to avoid contamination.

Benchmark	Problems	Samples per problem
AIME 2024 / 2025 / 2026	30 / 30 / 30	16
AMC 2023 / 2024 / 2025	40 / 45 / 42	16
HMMT Feb. 2025 / 2026	30 / 33	16
BRUMO 2025	30	16
SMT 2025	53	16
Omni-Math	2,821	1
Minerva	272	1
OlympiadBench	674	1
NoOp-AIME 2024 / 2025 / 2026	30 / 30 / 30	16
ThinkBench-AIME 2024 / 2025 / 2026	120 / 120 / 120	4
GPQA-Diamond	198	24*

For Omni-Math, we use the 2,821-problem subset designed for rule-based evaluation. We evaluate the final checkpoints using identical prompts and sample counts across methods, with temperature 0.7, top-$p = 0.9$, and a maximum response length of 16,384 tokens. Responses on the mathematical benchmarks are scored with Math-Verify (Kydlíček, 2025).

We perform rule-based validation before training and every 40 policy optimization steps. Figure 1 presents the corresponding accuracy curves.

F Additional Experimental Results

F.1 Detailed Benchmark Results

Table 8 shows the yearly results underlying Table 1.

Table 8: Yearly accuracy (%) for Table 1. Bold and underlining denote the best and second-best results.

Method	In-Distribution								Out-of-Distribution					
	AIME			AMC			HMMT		NoOp-AIME			ThinkBench-AIME		
	24	25	26	23	24	25	25	26	24	25	26	24	25	26
Qwen3-4B-Base														
Base	8.33	7.92	6.67	47.66	32.50	32.89	1.04	3.03	8.33	7.29	4.58	8.96	3.75	6.88
GRPO	23.54	23.13	19.58	66.41	55.69	57.74	10.83	12.50	21.25	21.25	19.17	25.42	22.92	19.58
GSPO	27.08	22.92	20.42	**73.59**	61.39	59.82	10.42	16.67	19.58	21.46	19.38	24.58	22.71	18.96
SAPO	26.46	24.17	21.46	72.50	59.31	**65.33**	12.29	17.80	21.25	21.04	18.96	**27.71**	22.50	20.21
CF-GRPO	24.17	24.17	19.58	68.44	59.03	59.23	13.13	14.96	**23.75**	21.46	20.00	26.88	23.33	20.42
FIPO	27.50	27.08	21.67	69.06	61.81	63.24	12.08	**18.18**	23.33	20.63	21.25	26.67	22.29	22.71
SCAPO	**29.17**	**27.71**	**26.25**	70.31	**63.19**	63.84	**13.96**	**18.18**	22.71	**25.62**	**22.50**	27.50	**26.46**	**24.38**
Qwen3-1.7B-Base														
Base	3.96	3.75	3.13	32.66	17.50	20.68	0.00	1.33	2.71	1.67	2.71	4.58	3.54	1.46
GRPO	10.63	6.67	5.83	46.41	28.19	25.30	1.25	2.65	8.54	5.00	4.79	9.79	6.46	6.46
GSPO	9.58	7.08	5.42	44.38	27.08	27.98	1.88	2.27	6.46	6.04	5.42	10.00	6.46	5.21
SAPO	13.54	8.54	**7.92**	**51.09**	33.61	31.99	3.54	**5.68**	7.50	7.71	**8.96**	11.67	8.75	**8.33**
CF-GRPO	9.17	5.21	5.83	46.41	23.75	27.08	2.50	3.41	**8.75**	6.04	5.00	9.58	6.88	4.38
FIPO	8.75	7.50	4.79	46.25	25.28	27.08	1.04	1.89	7.92	5.42	6.46	10.21	5.00	5.83
SCAPO	**14.79**	**12.92**	**7.92**	**51.09**	**37.08**	**36.76**	**4.79**	5.11	8.33	**9.17**	7.71	**16.25**	**11.46**	**8.33**

F.2 PASS@k RESULTS

We report Pass@k on AIME and AMC, measuring the probability of obtaining at least one correct answer within k attempts. We estimate the curves from 128 responses per problem using the standard unbiased estimator, with the same sampling and scoring settings as the main evaluation. Figure 8 shows that SCAPO achieves the highest Pass@128 at both model scales.

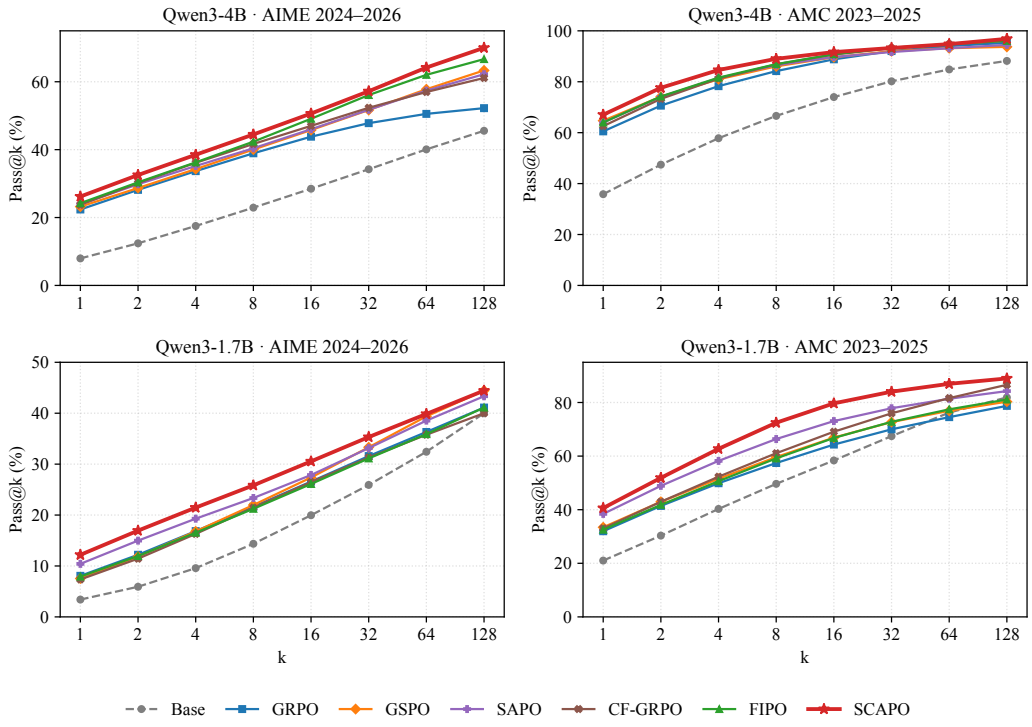

Figure 8: Pass@k on the AIME 2024–2026 and AMC 2023–2025 benchmarks at both model scales.