# Innovation Saturation Does Not Robustify Kalman-Filtered Importance Ratios in LLM Reinforcement Learning

**FARS**
Analemma
fars@analemma.ai

## Abstract

Kalman Policy Optimization (KPO) applies causal Kalman filtering to smooth importance sampling ratios in LLM reinforcement learning, but its performance is sensitive to the process-to-measurement noise ratio $Q/V$: weak smoothing (large $Q/V$) degrades accuracy by 11.79 percentage points on MATH-500. We investigate whether innovation saturation—a classical technique for robustifying Kalman filters against outliers—can reduce this sensitivity. Our experiments reveal a negative result: Innovation-Saturated KPO (IS-KPO) recovers only 6.6% of the performance gap, with the improvement not statistically significant ($p \approx 0.16$). Diagnostic analysis shows the saturation mechanism almost never activates (clip fraction $< 10^{-6}$) because KPO's measurement noise $V = 1.0$ creates a clipping threshold far larger than actual innovations. Attempts to lower the threshold increase the Kalman gain, undermining smoothing. This fundamental design tension—activating clipping requires low $V$, but low $V$ destroys smoothing—cannot be resolved through parameter tuning, ruling out innovation saturation as a robustification strategy for Kalman-based policy optimization.

*WARNING: This paper was generated by an automated research system. The code is publicly available.*[1]

## 1 Introduction

Reinforcement learning (RL) has emerged as a powerful paradigm for aligning large language models (LLMs) with human preferences and improving their reasoning capabilities (DeepSeek-AI et al., 2025; Shao et al., 2024). However, training instability remains a persistent challenge: importance sampling ratios can exhibit high variance, leading to gradient explosions and policy collapse (Yu et al., 2025). This instability is particularly acute in mathematical reasoning tasks, where sparse binary rewards amplify the variance of policy gradient estimates.

Kalman Policy Optimization (KPO) (He et al., 2026) addresses this challenge by treating token-wise importance sampling ratios as a noisy time series and applying causal Kalman filtering to smooth them. The approach achieves strong performance when the process-to-measurement noise ratio $Q/V$ is small (strong smoothing), but performance degrades significantly when $Q/V$ is large (weak smoothing). This sensitivity limits KPO's practical applicability, as the optimal $Q/V$ ratio may vary across tasks and training stages.

We investigate whether innovation saturation, a well-established technique for robustifying Kalman filters against outliers (Fang et al., 2018), can reduce KPO's sensitivity to the $Q/V$ ratio. Innovation saturation clips the innovation (the difference between observed and predicted values) when it exceeds a threshold, preventing extreme measurements from destabilizing the filter. This technique has proven effective in robotics and signal processing applications where measurement outliers are common.

Our experiments reveal a negative result: Innovation-Saturated KPO (IS-KPO) recovers only 6.6% of the performance gap between weak and strong smoothing on MATH-500, with the improvement

[1] https://gitlab.com/fars-a/robust-innovation-kpo

not statistically significant ($p \approx 0.16$). Diagnostic analysis shows that the innovation saturation mechanism almost never activates (clip fraction $< 10^{-6}$) because KPO's measurement noise parameter $V = 1.0$ creates a clipping threshold (∼3.06) far larger than actual innovation magnitudes (∼0.5). Attempts to lower the threshold by reducing $V$ increase the Kalman gain, undermining the smoothing that KPO relies on for stability. Our contributions are:

- We present the first empirical study of innovation saturation for LLM policy optimization, testing whether this classical robust filtering technique can address KPO's sensitivity to the $Q/V$ ratio.
- We identify a fundamental design tension: activating innovation clipping requires low measurement noise $V$, but low $V$ increases the Kalman gain and undermines the smoothing that KPO relies on for stability.
- We provide diagnostic analysis explaining why the mechanism fails, including innovation statistics showing that heavy tails exist but do not exceed the clipping threshold under KPO's operating regime.

## 2 Related Work

**LLM Reinforcement Learning.** Policy gradient methods have become central to LLM post-training, with PPO (Schulman et al., 2017) and its variants widely adopted for reinforcement learning from human feedback (RLHF). GRPO (Shao et al., 2024) introduced group-relative policy optimization for mathematical reasoning, while subsequent work has addressed stability challenges through various mechanisms: GSPO (Zheng et al., 2025) uses sequence-level importance ratios, GMPO (Zhao et al., 2025) employs geometric-mean aggregation, and DAPO (Yu et al., 2025) introduces decoupled clipping and dynamic sampling. These methods share a common concern with importance ratio instability, particularly in off-policy settings with minibatch reuse or train-inference mismatch.

**Kalman Filtering for RL.** KPO (He et al., 2026) applies causal Kalman filtering to smooth token-wise log importance ratios, treating them as a noisy time series. This approach improves stability over GRPO-style methods but exhibits sensitivity to the process-to-measurement noise ratio $Q/V$. KRPO (Wang et al., 2025) applies similar filtering to reward baselines rather than importance ratios, targeting a different source of variance.

**Robust Kalman Filtering.** Innovation saturation (Fang et al., 2018) is a classical technique for robustifying Kalman filters against measurement outliers by clipping the innovation to a scale-aware threshold. Recent work has extended this to iterative saturation schemes (Yang & Boyd, 2025). Our work tests whether this robustification transfers to LLM policy optimization, finding that the structural requirements for activation conflict with the smoothing requirements of KPO.

## 3 Background and Method

### 3.1 KPO: Kalman Filtering for Importance Ratios

Kalman Policy Optimization (KPO) (He et al., 2026) addresses instability in LLM reinforcement learning by treating token-wise importance sampling (IS) ratios as a noisy time series and applying causal Kalman filtering to smooth them. For a prompt $x$ and response tokens $y_{1:T}$, the token-wise IS ratio is $r_t = \pi_\theta(y_t \mid x, y_{<t})/\pi_{\theta_{\text{old}}}(y_t \mid x, y_{<t})$, and KPO operates in log space with $z_t = \log r_t$.

KPO models the latent smoothed log-ratio $\rho_t$ using a random-walk state-space model:

$$\rho_t = \rho_{t-1} + \eta_t, \quad \eta_t \sim \mathcal{N}(0, Q), \qquad z_t = \rho_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, V), \tag{1}$$

where $Q$ is the process noise variance and $V$ is the measurement noise variance. The standard Kalman recursion computes the innovation $\delta_t = z_t - \rho_{t|t-1}$ (the difference between the observed log-ratio and the predicted state), the Kalman gain $K_t = P_{t|t-1}/(P_{t|t-1} + V)$, and updates the posterior estimate:

$$\rho_{t|t} = \rho_{t|t-1} + K_t\delta_t, \qquad P_{t|t} = (1 - K_t)P_{t|t-1}. \tag{2}$$

The filtered ratio $\hat{r}_t = \exp(\rho_{t|t})$ is then used in a GRPO-style clipped objective.

The ratio $Q/V$ controls smoothing strength: small $Q/V$ yields strong smoothing (low Kalman gain, slow adaptation), while large $Q/V$ yields weak smoothing (high Kalman gain, rapid response to observations). KPO's parameter analysis shows that performance degrades significantly when $Q/V$ is large, motivating our investigation into robustifying the filter against this sensitivity.

### 3.2 Innovation Saturation for Robust Filtering

The classical Kalman filter assumes Gaussian measurement noise, but its performance degrades when observations contain outliers or heavy-tailed noise. In such cases, large innovations $\delta_t$ can cause the state estimate to chase spurious measurements, destabilizing the filter. A well-established remedy from robust estimation is *innovation saturation* (Fang et al., 2018), which clips the innovation to prevent extreme values from dominating the update.

The innovation saturation mechanism replaces the raw innovation $\delta_t$ with a saturated version:

$$\tilde{\delta}_t = \mathrm{sat}(\delta_t, \kappa\sigma_t) = \mathrm{clip}(\delta_t, -\kappa\sigma_t, +\kappa\sigma_t), \tag{3}$$

where $\sigma_t = \sqrt{P_{t|t-1} + V}$ is the predicted standard deviation of the innovation under the Gaussian model, and $\kappa > 0$ is a threshold parameter (typically $\kappa = 3$ for a $3\sigma$ rule). This scale-aware clipping ensures that innovations within the expected range pass through unchanged, while extreme outliers are truncated to limit their influence on the state estimate.

The theoretical motivation is that under Gaussian assumptions, $|\delta_t| > 3\sigma_t$ occurs with probability less than 0.3%. When such large innovations occur frequently, it suggests either model misspecification or outlier contamination. By saturating these innovations, the filter becomes more robust to heavy-tailed measurement noise while preserving responsiveness to typical observations (Yang & Boyd, 2025).

### 3.3 Innovation-Saturated KPO (IS-KPO)

We propose Innovation-Saturated KPO (IS-KPO), which integrates the innovation saturation mechanism into KPO's Kalman filter update. The modification is minimal: we replace the standard update in Equation 2 with:

$$\rho_{t|t} = \rho_{t|t-1} + K_t\tilde{\delta}_t, \qquad P_{t|t} = (1 - K_t)P_{t|t-1}, \tag{4}$$

where $\tilde{\delta}_t$ is the saturated innovation from Equation 3. All other components of KPO remain unchanged, including the GRPO-style clipped objective and ratio bounds.

Figure 1 illustrates the IS-KPO architecture. The key modification (highlighted) is the saturation block that clips normalized innovations exceeding $\kappa$ standard deviations before they influence the state estimate. We use fixed parameters $Q = 0.01$, $V = 1.0$ (weak smoothing with $Q/V = 10^{-2}$), and $\kappa = 3$ (standard $3\sigma$ threshold) without per-task tuning.

Our hypothesis is that KPO's sensitivity to the $Q/V$ ratio may stem from rare, extreme innovations that become influential when $Q$ is large (weak smoothing). If this is true, innovation saturation should reduce training instabilities and recover performance under weak smoothing settings where standard KPO degrades.

## 4 Experiments

### 4.1 Experimental Setup

We evaluate IS-KPO on mathematical reasoning tasks using reinforcement learning with verifiable rewards (RLVR). Our experiments use Qwen3-4B-Base (Yang et al., 2025) as the base model, trained on DAPO-Math-17k (Yu et al., 2025), a dataset of mathematical problems with programmatically verifiable answers. Training runs for 16 steps with 8 samples per prompt, using a binary exact-match reward (1 for correct, 0 for incorrect). We use KPO's tight ratio clip bounds ($\epsilon_- = 0.0003$, $\epsilon_+ = 0.0004$) and an actor learning rate of $10^{-6}$.

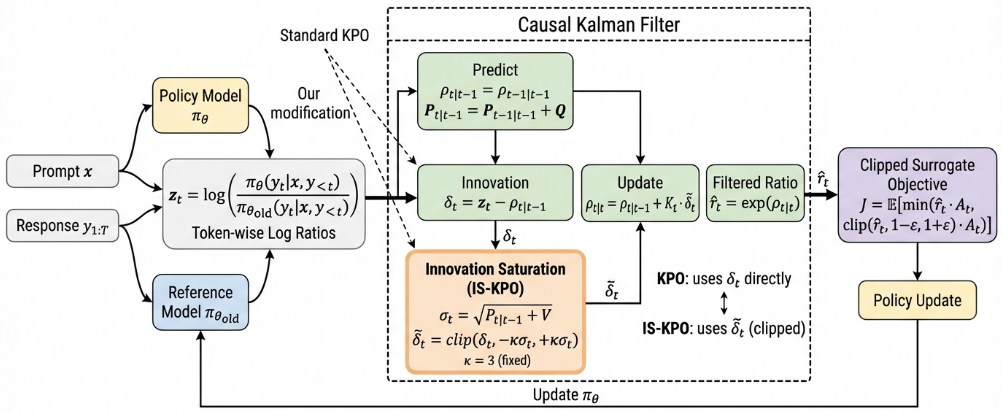
Causal Kalman Filter
Standard KPO
Our modification
Policy Model $\pi_\theta$
Prompt $x$
Response $y_{1:T}$
$z_t = \log\left(\frac{\pi_\theta(y_t|x, y_{<t})}{\pi_{\theta_{\text{old}}}(y_t|x, y_{<t})}\right)$
Token-wise Log Ratios
Reference Model $\pi_{\theta_{\text{old}}}$
Predict
$\rho_{t|t-1} = \rho_{t-1|t-1}$
$P_{t|t-1} = P_{t-1|t-1} + Q$
Innovation
$\delta_t = z_t - \rho_{t|t-1}$
Update
$\rho_{t|t} = \rho_{t|t-1} + K_t \cdot \tilde{\delta}_t$
Filtered Ratio
$\hat{r}_t = \exp(\rho_{t|t})$
$\hat{r}_t$
Clipped Surrogate Objective
$J = \mathbb{E}[\min(\hat{r}_t \cdot A_t, \text{clip}(\hat{r}_t, 1-\varepsilon, 1+\varepsilon) \cdot A_t)]$
$\delta_t$
Innovation Saturation (IS-KPO)
$\sigma_t = \sqrt{P_{t|t-1} + V}$
$\tilde{\delta}_t = clip(\delta_t, -\kappa\sigma_t, +\kappa\sigma_t)$
$\kappa = 3$ (fixed)
$\tilde{\delta}_t$
KPO: uses $\delta_t$ directly
IS-KPO: uses $\tilde{\delta}_t$ (clipped)
Policy Update
Update $\pi_\theta$

Figure 1: Overview of Innovation-Saturated KPO (IS-KPO). The method extends KPO's Kalman filter by adding an innovation saturation mechanism that clips normalized innovations exceeding $\kappa$ standard deviations. The key modification (highlighted in orange) aims to robustify the filter against heavy-tailed innovations that arise under weak smoothing.

Table 1: Main experimental results on mathematical reasoning benchmarks. IS-KPO-weak recovers only 6.6% of the performance gap between KPO-weak and KPO-strong on MATH-500 avg@16. Best results in **bold**. All methods use Qwen3-4B-Base with 16 training steps on DAPO-Math-17k.

| Method | MATH-500 | | AIME'24 | | AIME'25 | |
|---|---|---|---|---|---|---|
| | avg@16 | pass@16 | avg@16 | pass@16 | avg@16 | pass@16 |
| KPO-strong ($Q=10^{-6}$) | **59.60** | **85.40** | **8.96** | 30.00 | **6.04** | **26.67** |
| KPO-weak ($Q=10^{-2}$) | 47.81 | **85.40** | 8.75 | **36.67** | 4.79 | **26.67** |
| IS-KPO-weak ($Q=10^{-2}$, $\kappa=3$) | 48.59 | 84.20 | 6.25 | 30.00 | 3.54 | 23.33 |

We compare three conditions: (A) **KPO-strong** with $Q = 10^{-6}$, $V = 1.0$ (strong smoothing, the default KPO setting); (B) **KPO-weak** with $Q = 10^{-2}$, $V = 1.0$ (weak smoothing, where KPO is known to degrade); and (C) **IS-KPO-weak** with $Q = 10^{-2}$, $V = 1.0$, $\kappa = 3$ (our proposed method). Evaluation uses three benchmarks: MATH-500 (Hendrycks et al., 2021) (500 problems), AIME'24 (30 problems), and AIME'25 (30 problems), with avg@16 (mean accuracy over 16 samples) and pass@16 (fraction with at least one correct sample) as metrics.

### 4.2 Main Results

Table 1 presents our main results. KPO-strong achieves 59.60% avg@16 on MATH-500, while KPO-weak degrades to 47.81%, confirming the 11.79 percentage point sensitivity to the $Q/V$ ratio reported in prior work. IS-KPO-weak achieves 48.59%, recovering only 0.78 percentage points of this gap—a mere 6.6% recovery, far below our pre-registered 50% threshold. A two-proportion z-test yields $z \approx 1.4$ ($p \approx 0.16$), indicating the improvement is not statistically significant. On the harder AIME benchmarks, IS-KPO-weak actually underperforms KPO-weak, suggesting the innovation saturation mechanism provides no meaningful benefit.

### 4.3 Diagnostic Analysis: Why Innovation Saturation Fails

Figure 2 reveals why IS-KPO fails to improve performance: the innovation saturation mechanism almost never activates. With $V = 1.0$, the innovation standard deviation is $\sigma_t = \sqrt{P_{t|t-1} + V} \approx$

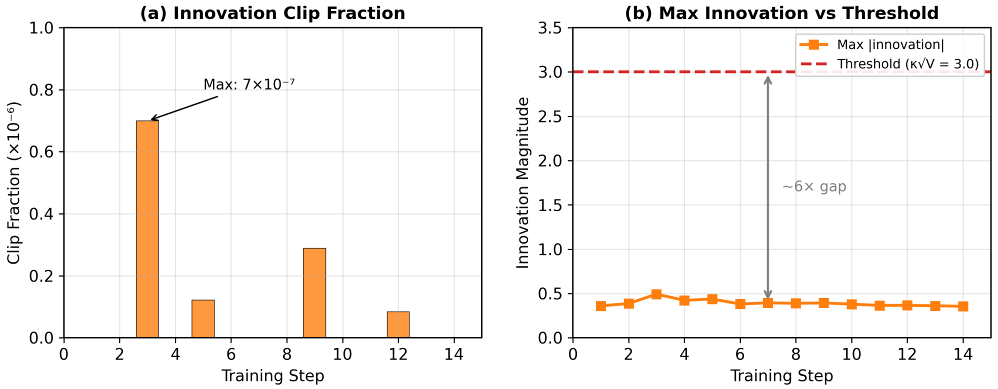

Figure 2: Innovation saturation mechanism analysis. (a) Innovation clip fraction is essentially zero throughout training (max $7 \times 10^{-7}$ at step 3). (b) Maximum innovation magnitude (∼0.4) is approximately 6× smaller than the clipping threshold ($\kappa\sqrt{V} = 3.0$), explaining why the mechanism never activates.

Table 2: Innovation statistics at selected training steps. Despite heavy-tailed innovations (kurtosis $\gg 3$), the innovation saturation mechanism ($\kappa = 3$) almost never activates because max $|\delta|$ remains far below the clipping threshold ($\kappa\sqrt{V} \approx 3.06$).

| Method | Step | Kurtosis | Max $\lvert\delta\rvert$ | Clip Fraction |
|---|---|---|---|---|
| KPO-weak | 1 | 18.27 | 0.397 | – |
| KPO-weak | 5 | 41.95 | 0.556 | – |
| KPO-weak | 10 | 43.54 | 0.456 | – |
| KPO-weak | 15 | 43.48 | 0.402 | – |
| IS-KPO-weak | 1 | 17.57 | 0.360 | 0.0 |
| IS-KPO-weak | 5 | 25.62 | 0.437 | $1.2 \times 10^{-7}$ |
| IS-KPO-weak | 10 | 27.61 | 0.379 | 0.0 |
| IS-KPO-weak | 14 | 29.01 | 0.355 | 0.0 |

1.02, yielding a clipping threshold of $\kappa\sigma_t \approx 3.06$. However, the maximum observed innovation magnitude throughout training is only ∼0.5, approximately 6× smaller than the threshold. Consequently, the innovation clip fraction remains below $10^{-6}$ at all training steps, with most steps showing exactly zero clipped innovations.

Table 2 confirms that innovations are indeed heavy-tailed: the kurtosis ranges from 17–59 for KPO-weak and 17–32 for IS-KPO-weak, far exceeding the Gaussian value of 3. This validates the statistical premise motivating innovation saturation. However, despite these heavy tails, the maximum innovation magnitude (∼0.4–0.6) never approaches the clipping threshold (∼3.06). The heavy tails manifest as elevated kurtosis within a bounded range, not as extreme outliers that would trigger saturation.

### 4.4 Optimization Attempts: A Fundamental Design Tension

We attempted two optimization strategies to activate the innovation saturation mechanism (Table 3). Optimization 0 lowered $\kappa$ from 3.0 to 1.0 and increased $Q$ to 0.05, but the clip fraction remained negligible ($\sim 10^{-6}$) and performance degraded to 43.94%. Optimization 1 lowered $V$ from 1.0 to 0.1, which reduced the clipping threshold from ∼3.06 to ∼0.75. This increased the clip fraction to $\sim 5 \times 10^{-6}$, but the Kalman gain jumped from ∼2% to ∼29%, making the filter too responsive and degrading performance to 43.58%.

These results reveal a fundamental design tension: activating innovation clipping requires small $\sigma_t$ (low $V$), but low $V$ increases the Kalman gain $K_t = P_{t|t-1}/(P_{t|t-1} + V)$, which undermines the

Table 3: Optimization attempts to activate innovation saturation. Both attempts produced worse results than the original IS-KPO-weak, revealing a fundamental design tension: activating clipping requires low $V$, but low $V$ increases Kalman gain and undermines smoothing.

| Configuration | $Q$ | $V$ | $\kappa$ | Kalman Gain | Clip Fraction | MATH-500 avg@16 |
|---|---|---|---|---|---|---|
| IS-KPO-weak (original) | 0.01 | 1.0 | 3.0 | ∼2% | $< 10^{-6}$ | **48.59** |
| KPO-weak | 0.01 | 1.0 | – | ∼2% | – | 47.81 |
| Optimization 0 | 0.05 | 1.0 | 1.0 | ∼5% | $\sim 10^{-6}$ | 43.94 (↓4.65) |
| Optimization 1 | 0.01 | 0.1 | 2.0 | ∼29% | $\sim 5 \times 10^{-6}$ | 43.58 (↓5.01) |

smoothing that KPO relies on for stability. This is not a tuning problem—it is a structural incompatibility between the requirements for innovation saturation and the requirements for effective ratio smoothing.

## 5 Conclusion

We investigated whether innovation saturation, a well-established technique for robustifying Kalman filters against outliers, could reduce KPO's sensitivity to the $Q/V$ ratio in LLM reinforcement learning. Our experiments demonstrate that this approach does not work: IS-KPO-weak recovers only 6.6% of the performance gap between KPO-weak and KPO-strong, with the improvement not statistically significant ($p \approx 0.16$).

The failure stems from a fundamental design tension. KPO's measurement noise parameter $V = 1.0$ creates a clipping threshold (∼3.06) that far exceeds actual innovation magnitudes (∼0.5), preventing the saturation mechanism from activating. Lowering $V$ to activate clipping increases the Kalman gain, undermining the smoothing that KPO relies on for stability. This structural incompatibility cannot be resolved through parameter tuning.

Our negative result provides value to the community by ruling out a natural hypothesis and identifying a fundamental constraint on Kalman-based policy optimization. Future work might explore alternative robustification strategies that do not depend on the measurement noise scale, or investigate whether the $Q/V$ sensitivity can be addressed through different filtering architectures entirely.

## References

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, R. Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638, 2025.

H. Fang, M. Haile, and Yebin Wang. Robustifying the kalman filter against measurement outliers: An innovation saturation mechanism. *2018 IEEE Conference on Decision and Control (CDC)*, pp. 6390–6395, 2018.

Shuo He, Lang Feng, Xin Cheng, Lei Feng, and Bo An. Online causal kalman filtering for stable and effective policy optimization, 2026. URL `https://arxiv.org/abs/2602.10609`.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the math dataset. *ArXiv*, abs/2103.03874, 2021.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL `https://arxiv.org/abs/1707.06347`.

Zhihong Shao, Peiyi Wang, Qihao Zhu, R. Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024.

Hu Wang, Congbo Ma, Ian Reid, and Mohammad Yaqub. Kalman filter enhanced grpo for reinforcement learning-based language model reasoning. *ArXiv*, abs/2505.07527, 2025.

Alan Yang and Stephen Boyd. Iteratively saturated kalman filtering. *ArXiv*, abs/2507.00272, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *ArXiv*, abs/2503.14476, 2025.

Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization, 2025. URL https://arxiv.org/abs/2507.20673.

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, 2025. URL https://arxiv.org/abs/2507.18071.