Calibration-Aware Policy Optimization for Reasoning LLMs

Ziqi Wang[1,2], Xingzhou Lou[1,2], Meiqi Wu[1], Zhengqi Wen[3], Junge Zhang[1,2]
[1]National Key Laboratory of Cognition and Decision Intelligence for Complex Systems, Institution of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Beijing National Research Center for Information Science and Technology, Tsinghua University
Correspondence to Junge Zhang: jgzhang@nlpr.ia.ac.cn

Abstract

Group Relative Policy Optimization (GRPO) enhances LLM reasoning but often induces overconfidence, where incorrect responses yield lower perplexity than correct ones, degrading relative calibration as described by the Area Under the Curve (AUC). Existing approaches either yield limited improvements in calibration or sacrifice gains in reasoning accuracy. We first prove that this degradation in GRPO-style algorithms stems from their uncertainty-agnostic advantage estimation, which inevitably misaligns optimization gradients with calibration. This leads to improved accuracy at the expense of degraded calibration. We then propose **C**alibration-**A**ware **P**olicy **O**ptimization **(CAPO)**. It adopts a logistic AUC surrogate loss that is theoretically consistent and admits regret bound, enabling uncertainty-aware advantage estimation. By further incorporating a noise masking mechanism, CAPO achieves stable learning dynamics that jointly optimize calibration and accuracy. Experiments on multiple mathematical reasoning benchmarks show that CAPO-1.5B significantly improves calibration by up to 15% while achieving accuracy comparable to or better than GRPO, and further boosts accuracy on downstream inference-time scaling tasks by up to 5%. Moreover, when allowed to abstain under low-confidence conditions, CAPO achieves a Pareto-optimal precision–coverage trade-off, highlighting its practical value for hallucination mitigation.

1 Introduction

Model calibration is defined as the correlation between a model's confidence in its answers and the ground truth correctness of those answers (Geng et al., 2024). From the era of conventional neural networks to the advent of Large Language Models (LLMs), calibration has consistently been a key focus due to its paramount importance in two aspects (Xiao et al., 2025; Tao et al., 2024; Kadavath et al., 2022).

First, calibration is fundamental to model trustworthiness, enabling reliable uncertainty estimation and abstention to mitigate hallucination—the generation of plausible but factually incorrect assertions. This is particularly vital in high-stakes domains such as finance and healthcare (Savage et al., 2025). Second, confidence estimates are widely used to guide algorithmic decisions. Multi-agent or cascading systems strengthen model collaboration when a single model is uncertain (Luo et al., 2025b; Warren and Dras, 2025; Chuang et al., 2025). Self-paced training prioritizes uncertain samples (Feng et al., 2025; Wang et al., 2025). And inference-time scaling strategies select candidate responses with high confidence (Stoisser et al., 2025; Vashurin et al., 2025; Zhou et al., 2025b). All these approaches critically depend on well-calibrated confidence estimates to reflect the true correctness of model outputs.

However, several studies have indicated that Reinforcement Learning from Verifiable Rewards (RLVR) (Shao et al., 2024;Yu et al., 2025;Zheng et al., 2025) can make models overconfident (Liu et al., 2025a;Kalai et al., 2025;Bereket and Leskovec, 2025). This often manifests as degradation in relative calibration, where the model outputs fluent but incorrect answers with higher perplexity than correct ones.

Existing efforts to address the overconfidence induced by algorithms such as GRPO (Shao et al., 2024) and GSPO (Zheng et al., 2025) primarily rely on heuristic designs. These works often lack quantitative analyses of calibration and, critically, provide no theoretical guarantees for calibration improvement. As a result, they either yield only limited calibration gains, as in CoDaPO (Zhou et al., 2025a) and CDE (Dai et al., 2025), or compromise the model's overall accuracy, as in SimKO (Peng et al., 2025).

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 18375–18390

To address these issues, we first experimentally show that during the GRPO optimization process, the typical relative calibration metric AUC (Ling et al., 2003; Vashurin et al., 2025; Zhou et al., 2025b) progressively worsens as accuracy improves. GRPO-style methods construct advantages from group-wise reward differences. We prove that reward-only sample evaluation induces optimization gradients corresponding to an inconsistent surrogate for calibration optimization (Gao and Zhou, 2012). Consequently, the learning process is not aligned with calibration improvement, which explains the degradation in experiments.

Building upon this discovery, we propose a novel approach **C**alibration-**A**ware **P**olicy **O**ptimization **(CAPO)**, which (1) adopts uncertainty-aware advantage estimation based on consistent logistic AUC surrogate, enabling joint optimization of calibration and accuracy, and (2) introduces a reference-model-based noise masking mechanism to maintain training stability. In addition to this theoretical justification, a complementary gradient analysis further illustrates how the learning dynamics naturally prioritizes the correction of responses with misaligned confidence, offering additional intuition.

Experimental results on Qwen2.5-Math-1.5B and 7B models demonstrate that CAPO enables stable and joint optimization of both calibration and accuracy. It successfully prevents calibration degradation and attains accuracy that matches or surpasses that of GRPO, as shown in Figure 1. Our contributions are summarized as follows:

- We provide **experiment evidence and theoretical explanation** of calibration degradation in GRPO-style algorithms.
- To address this issue, we introduce a **theoretically grounded and consistent optimization objective** that enables joint optimization of accuracy and calibration.
- **Extensive experiments across six benchmarks** demonstrate significant calibration improvements over GRPO, GSPO, and three additional baselines with preserved accuracy, yielding Pareto-optimal precision–coverage trade-offs for hallucination mitigation and improved inference-time scaling accuracy.

2 Preliminary and Related Work

As our approach builds upon AUC optimization techniques to analyze and mitigate relative calibration degradation in GRPO-style algorithms, this section reviews relevant background on GRPO and calibration, along with related prior work.

2.1 Group Relative Policy Optimization

The reasoning ability of LLMs can be optimized using outcome reward-based Reinforcement Learning, by modeling the generation process as an Markov Decision Process. GRPO achieves this efficiently by computing the advantage value using the reward differences among samples within a group, thereby obviating the need for a value model. For a specific question-answer pair (q, a), the behavior policy $\pi_{\theta_{\text{old}}}$ samples a group of G individual responses $\{o_i\}_{i=1}^{G}$. Then, the advantage $\hat{A}_{i,t}$ of the i-th response is calculated by normalizing the group-level rewards$\{R_i\}_{i=1}^{G}$:

$$\hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^{G})}{\text{std}(\{R_i\}_{i=1}^{G})}, \tag{1}$$

where $R = 1$ indicates a correct response and $R = 0$ indicates an incorrect one.

GRPO adopts a ppo-style clipped objective, together with a directly imposed KL penalty term:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}\big(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon\big)\hat{A}_{i,t}\Big) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Bigg], \tag{2}$$

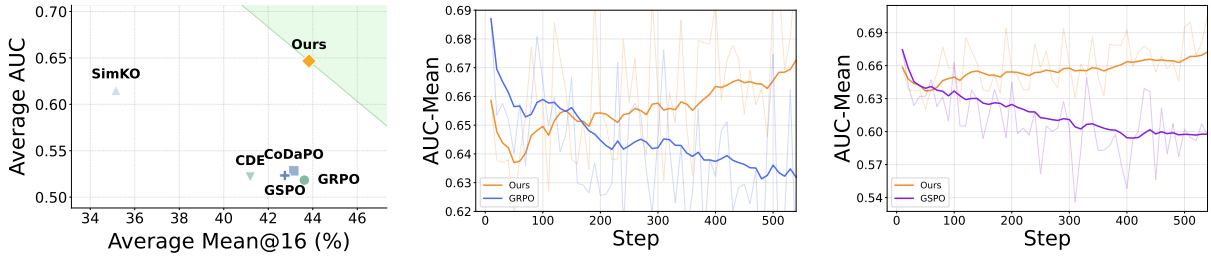
where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}. \tag{3}$$

The clipping operator $\text{clip}(\cdot,\, 1-\epsilon,\, 1+\epsilon)$ is introduced to constrain each policy update, preventing the new policy from drifting too far away from the previous one by restricting the policy ratio within the interval $[1-\epsilon,\, 1+\epsilon]$.

$$\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - \log\Big(\frac{\pi_{\text{ref}}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})}\Big) - 1. \tag{4}$$

In addition, the KL divergence between the current policy and the reference policy (i.e., the base model) is approximated as Equation (4).



(a) (b) (c)

Figure 1: (a) Comparison of average calibration (measured by AUC-mean) and accuracy (measured by mean@16) across six test benchmarks for our method and all baselines on Qwen2.5-Math-7B; (b)(c) Comparison of calibration dynamics on the validation set during training between our method and GRPO (left) / GSPO (right).

While base models are shown to be well-calibrated (Kalai et al., 2025), algorithms in the GRPO-like family are observed to cause model calibration collapse, where models generate confident, yet incorrect, responses (Liu et al., 2025a;Dai et al., 2025). Prior RL-based efforts to mitigate this focus on heuristics: reward/advantage shaping (Dai et al., 2025;Zhou et al., 2025a), regularization (Liu et al., 2025a), or label smoothing (Peng et al., 2025).

However, these works rely primarily on qualitative descriptions of calibration degradation. As a result, their heuristic approaches lack theoretical guarantees and generally fail to balance accuracy with calibration preservation.

Our work addresses these deficiencies from a principled quantitative perspective and offers a theoretical justification for calibration optimization, and a proper balance between performance and trustworthiness.

2.2 Calibration Metric

Perplexity (PPL) is a widely used uncertainty indicator in free-form generation (Stoisser et al., 2025; Vashurin et al., 2025; Zhou et al., 2025b):

$$\mathrm{PPL}(o) = \exp\left(-\frac{1}{|o|}\sum_{t=1}^{|o|}\log \pi(o_t \mid o_{<t})\right). \quad (5)$$

PPL reflects the model's intrinsic uncertainty without incurring additional computational overhead. In contrast, approaches that elicit explicit confidence via prompting (Zeng et al., 2025; Wen et al., 2024; Damani et al., 2025) are often sensitive to prompt design (Yang et al., 2024b) and self-consistency methods (Xiong et al., 2024; Manakul et al., 2023; Tanneru et al., 2024) increase inference costs, making them less general. We therefore consider them outside the scope of this work.

Model calibration can generally be quantified using two families of metrics (Geng et al., 2024): **relative calibration** (e.g., AUC) and **absolute calibration** (e.g., Expected Calibration Error).

Relative calibration emphasizes the model's ability to *rank* samples—assigning higher confidence to correct responses than to incorrect ones—formalized as (f denotes the confidence scoring function):

$$f(q, o_i) \le f(q, o_j) \iff R_i \le R_j, \quad (6)$$

A widely used metric for relative calibration, the Area Under the Curve (AUC) (Yang and Ying, 2022; Yuan et al., 2021; Zhu et al., 2022), directly quantifies the probability that a model assigns higher confidence to a correct response than to an incorrect one, thereby capturing the core notion of relative calibration:

$$\mathrm{AUC}(\pi, q, f) = \mathbb{E}_{o_i, o_j \sim \mathcal{D}}[\mathbb{I}((R_i - R_j)(f(o_i) - f(o_j)) > 0)], \quad (7)$$

where o_i and o_j denote a randomly sampled pair of responses generated by the model for the same question q. We further define AUC-mean as the average AUC score across all questions in a dataset, serving as an overall measure of the model's calibration performance:

$$\text{AUC-mean}(\pi, \mathcal{Q}, f) = \frac{1}{|\mathcal{Q}|}\sum_{q \in \mathcal{Q}} \mathrm{AUC}(\pi, q, f), \quad (8)$$

where $\mathcal{Q}$ denotes the set of questions.

Absolute calibration indicates that a model's returned confidence matches its true correctness likelihood:

$$P(R = 1 \mid f(q, o) = k) = k. \quad (9)$$

In this work, we focus on relative calibration rather than absolute calibration for two reasons:

(1) PPL is not directly comparable to correctness probability, making absolute calibration ill-defined, and (2) absolute calibration is an aggregate statistical property that does not ensure instance-level discrimination: for example, assigning identical confidence (e.g., 0.5) to both correct and incorrect responses perfectly satisfy absolute calibration criteria yet fails to distinguish them, limiting its usefulness for downstream tasks such as inference-time scaling. In contrast, relative calibration explicitly enforces pairwise separability.

2.3 Consistency of AUC Optimization

Maximizing the AUC is equivalent to minimizing the risk: $L(f) = -\mathrm{AUC}(\pi, q, f)$. However, this objective is non-differentiable and non-convex, thus a direct optimization typically results in NP-hard problems (Gao and Zhou, 2012).

To address this difficulty, prior work commonly replaces the indicator in Equation (7) with a pairwise differentiable surrogate loss:

$$\Psi(f, o_i, o_j) = \phi[(R_i - R_j)\,(f(o_i) - f(o_j))], \quad (10)$$

where ϕ is a convex function such as the exponential loss $\phi(t) = e^{-t}$, hinge loss $\phi(t) = \max(0, 1-t)$, or squared loss $\phi(t) = (1-t)^2$(Gao et al., 2013;Zhao et al., 2011;Kotlowski et al., 2011;Calders and Jaroszewicz, 2007;Charoenphakdee et al., 2019).

AUC consistency is defined as follows:

Definition 1. *The surrogate loss ϕ is said to be* consistent with AUC *if, for every sequence $\{f^{\langle n\rangle}(x)\}_{n\geq 1}$,*

$$L_\phi\left(f^{\langle n\rangle}\right) \to L_\phi^* \quad \Rightarrow \quad L\left(f^{\langle n\rangle}\right) \to L^*,$$

where L_ϕ is the surrogate risk and L is the true AUC risk. L_ϕ^* and L^* is the corresponding optimal value.

Intuitively, consistency emphasizes the *asymptotic correctness* of a surrogate loss: convergence to the optimal surrogate risk guarantees convergence to the optimal AUC. For this reason, consistency is widely regarded as one of the most important theoretical properties of surrogate losses.

A sufficient condition for achieving AUC consistency is provided by the following theorem (Gao and Zhou, 2012):

Theorem 1. *The surrogate loss $\Psi(f, o_i, o_j)$ is consistent with AUC if $\phi : \mathbb{R} \to \mathbb{R}$ is a convex, differentiable, and non-increasing function with $\phi'(0) < 0$.*

It can be readily shown that surrogate losses such as logistic loss and exponential loss are consistent with AUC according to the above theorem.

Furthermore, under the realizable setting where each response is assigned a correct or incorrect label with probability one, Gao and Zhou (2012) derives the following regret bound, which theoretically guarantees that minimizing the specified surrogate risk is aligned with optimizing AUC:

Theorem 2. *For exponential loss, hinge loss, general hinge loss, q-norm hinge loss, and least square loss, we have*

$$L(f) - L^* \leq L_\phi(f) - L_\phi^*, \quad (11)$$

and for logistic loss, we have

$$L(f) - L^* \leq \frac{1}{\ln 2}(L_\phi(f) - L_\phi^*). \quad (12)$$

Prior work on AUC optimization has primarily focused on training better classifiers on static datasets, whereas we study the relationship between reinforcement learning dynamics and relative calibration, and leverage AUC optimization techniques to achieve simultaneous improvements in both calibration and accuracy.

3 Why GRPO-style Objectives Degrade Calibration

Empirical Observation Prior works (Yu et al., 2025) have shown that the KL regularization term in GRPO is detrimental to optimization, while other studies (Liu et al., 2025b) indicate that removing the standard deviation term in advantage estimation yields an unbiased estimator. Accordingly, we adopt these optimized configurations when implementing GRPO. Empirically, as shown in Figure 1(b), during GRPO training, the test-set accuracy consistently improves, whereas the AUC-mean steadily deteriorates. Below, we provide a theoretical explanation for this phenomenon.

Theoretical Explanation At its core, the GRPO gradient is a REINFORCE-style gradient with group relative advantage estimation, as expressed in Equation (13).

$$\mathbb{E}_{o_{1:G}\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\nabla_\theta lpm_\theta(o_i)\right], \hat{A}_i = R_i - \bar{R}, \quad (13)$$

where $lpm_\theta(o_i) = \frac{1}{|o|}\sum_{t=1}^{|o|}\log\pi_\theta(o_t \mid o_{<t})$ (see Appendix A.1 for a detailed derivation). Consequently, this gradient can be equivalently rewritten

as Equation (14). Constant scaling factors are omitted for clarity in the subsequent analysis.

$$\mathbb{E}_{o_{1:G}\sim\mathcal{D}}\left[\sum_{1\le i<j\le G}\left(\nabla_\theta lpm_\theta(o_i)-\nabla_\theta lpm_\theta(o_j)\right)\left(R_i-R_j\right)\right]. \tag{14}$$

By invoking the unbiasedness property of U-statistics (see Appendix A.2 for a formal definition), namely that the expectation over a group of samples equals the expectation over a randomly sampled pair, it follows:

$$\begin{aligned}&\mathbb{E}_{o_1,o_2\sim\mathcal{D}}\Big[(\nabla_\theta lpm_\theta(o_1)-\nabla_\theta lpm_\theta(o_2))(R_1-R_2)\Big]\\&=\nabla_\theta\mathbb{E}_{o_1,o_2\sim\mathcal{D}}\Big[(lpm_\theta(o_1)-lpm_\theta(o_2))(R_1-R_2)\Big].\end{aligned} \tag{15}$$

Therefore, we have the following result:

Theorem 3. *The gradient of the GRPO objective coincides with the gradient of the AUC optimization objective with surrogate loss $\phi(t)=-t$ and scoring function $f=lpm$ (i.e., perplexity). This surrogate loss is inconsistent for AUC optimization (see Appendix A.3 for the proof).*

That is, while optimizing the GRPO objective (which effectively optimizes accuracy), the true AUC is not theoretically guaranteed to improve simultaneously. In practice, GRPO tends to overfit easy samples, leading to a sharpened output distribution, in which the perplexities of both positive and negative samples decrease concurrently, thereby degrading AUC. This theoretical prediction aligns well with our empirical observations shown in Figure 1(b).

The above analysis extends beyond GRPO to any algorithm relying on reward-only advantage estimator: for instance, Figure 1(c) demonstrates similar calibration degradation in GSPO, where although optimization is sequence-level, advantage estimation remains reward-based. Evaluating samples **purely based on reward while ignoring uncertainty** inevitably yields an AUC-inconsistent surrogate (proof in Appendix A.3), thereby training the model to pursue only high reward values rather than both honesty and accuracy.

4 Method

To address the issues in Section 3, we propose an uncertainty-aware advantage estimation derived from a consistent AUC surrogate.

4.1 Calibration-Aware Policy Optimization

Consistent Surrogate Objective Building upon the theoretical analysis in Section 3, we propose to replace the inconsistent surrogate loss implicitly induced by GRPO with a *logistic surrogate loss*: $\phi_\tau(t)=\log\left(1+\exp(-t/\tau)\right)$. The temperature parameter $\tau>0$ controls the smoothness of the surrogate. This surrogate is AUC-consistent and admits regret bound (Theorem 2). It theoretically guarantees that optimizing this surrogate objective leads to an improvement in AUC.

The corresponding policy optimization objective:

$$J_{\text{logistic}}(\theta)=-\mathbb{E}_{o_1,o_2\sim\mathcal{D}}\Big[\log\left(1+\exp(-t/\tau)\right)\Big], \tag{16}$$

where

$$t=\left(lpm_\theta(o_1)-lpm_\theta(o_2)\right)\left(R_1-R_2\right). \tag{17}$$

We rewrite the above objective as an expectation over a set of responses and take its gradient:

$$\nabla_\theta J_{\text{logistic}}(\theta)=\mathbb{E}_{o_{1:G}\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\tilde{A}_i\nabla_\theta lpm_\theta(o_i)\right], \tag{18}$$

$$\tilde{A}_i=\begin{cases}-\displaystyle\sum_{j:R_j=0}\phi'(lpm_\theta(o_i)-lpm_\theta(o_j)), & R_i=1,\\ \displaystyle\sum_{j:R_j=1}\phi'(lpm_\theta(o_j)-lpm_\theta(o_i)), & R_i=0,\end{cases} \tag{19}$$

where $\phi'(t)=-\sigma(-t)$, and $\sigma(\cdot)$ denotes the sigmoid function.

Denoising via Reference-Model-Based Masking. Outcome-based binary rewards fail to assess intermediate reasoning. Consequently, they risk penalizing near-correct logic due to minor errors, while simultaneously reinforcing spurious reasoning that yields correct answers by chance. Such gradient noise can destabilize training or even induce model collapse.

We leverage the inherent calibration of the reference model (i.e., base model) to assess the quality of the reasoning process: high reference-model perplexity (PPL) typically indicates syntactic or logical flaws, whereas low PPL suggests coherence. This approach is that it requires no additional training overhead while significantly improving training stability.

Based on this, we propose a masking strategy to filter noisy signals. Specifically, we exclude correct responses with reference-model PPL exceeding a threshold `ref-high` (likely lucky guesses) and incorrect responses with reference-model PPL below `ref-low` (likely near-correct reasoning penalized as failure). These thresholds are robust and simple to set, empirically corresponding to the upper

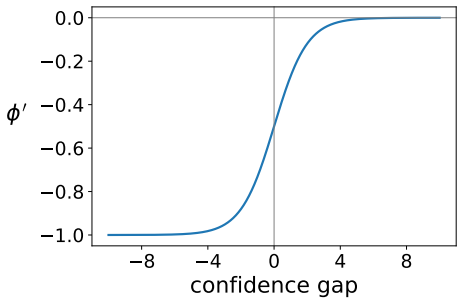


Figure 2: The relationship between advantage and the confidence gap of a correct–wrong response pair.

and lower quartiles of the reference model's PPL distribution.

Final Objective. By replacing the advantage estimation in the GRPO objective (2) with that in Equation (19), and applying noise mask, we obtain the objective under the behavior policy π_{old}:

$$J_{CAPO}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min \left(r_{i,t}(\theta)\hat{A}_i^{CAPO}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i^{CAPO} \right) \right], \tag{20}$$

where

$$\hat{A}_i^{CAPO} = m(o_i)\tilde{A}_i\,, \tag{21}$$

$$m(o) = \begin{cases} \mathbb{I}[\text{PPL}_{\text{ref}}(o) \le \texttt{ref-high}]\,, & \text{if } R(o)=1, \\ \mathbb{I}[\text{PPL}_{\text{ref}}(o) \ge \texttt{ref-low}]\,, & \text{if } R(o)=0. \end{cases} \tag{22}$$

This yields the optimization objective of CAPO. While the PPO-style formulation involves first-order approximation, the core logistic surrogate provides gradients aligned with relative calibration, offering strong theoretical justification. Subsection 4.2 and experiment results further corroborate this alignment.

4.2 Gradient Analysis

This subsection conducts a comparative analysis on the gradients between GRPO and our proposed objective function. The gradient induced by our objective is given by Equation (18) (mask term obviated for brevity). $\phi'(t)$ is illustrated in Figure 2. As can be seen, its magnitude decreases as the confidence gap between correct and incorrect responses increases. This indicates that the gradient places greater emphasis on correct samples with relatively high PPL and incorrect samples with low PPL, while suppressing the influence of samples that are already confidently and correctly ranked.

Samples near the misranking boundary in terms of reference-model PPL represent instances where the model's confidence estimates are inaccurate; they also represent highly informative samples in the model's decision space. Consequently, these are the key samples for improving both accuracy and AUC. Our method appropriately amplifies their gradients, while the masking mechanism removes update instability caused by extreme noisy samples. In contrast, GRPO assigns the same reward-based advantage to all positive and negative samples within a group, without accounting for their uncertainty or noise.

5 Experiment

5.1 Setting

Models and Datasets. We use Qwen2.5-Math-1.5B and Qwen2.5-Math-7B (Yang et al., 2024a) as the base models. The training and validation sets are constructed by randomly sampling 20k and 240 instances, respectively, from the DeepScaler dataset (Luo et al., 2025a). We evaluate all models on six benchmark datasets: AIME 2024, AIME 2025, MATH 500 (Lightman et al., 2023;Hendrycks et al., 2021), AMC 2023, Minerva (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024).

Baselines and Methods. The baselines we compare against fall into two categories: 1) methods designed to improve reasoning capability, including GRPO (Shao et al., 2024) and GSPO (Zheng et al., 2025). 2) methods proposed to address the calibration issues of GRPO, namely CDE (Dai et al., 2025), CoDaPO (Zhou et al., 2025a), and SimKO (Peng et al., 2025). For our method, τ is set to 0.6 for the 1.5B model, and 0.5 for the 7B. Hyperparameter details are provided in Appendix B.1.

Evaluation Metrics. Following Yue et al. (2025), we report mean@8 accuracy on the validation set and mean@16 accuracy on the test set. Model calibration is evaluated by AUC (Vashurin et al., 2025) and Hallucination mitigation by the Precision-Coverage curve.

5.2 Main Results

Improved Calibration with Preserved Accuracy Extensive experiments demonstrate that our method consistently improves calibration (AUC)

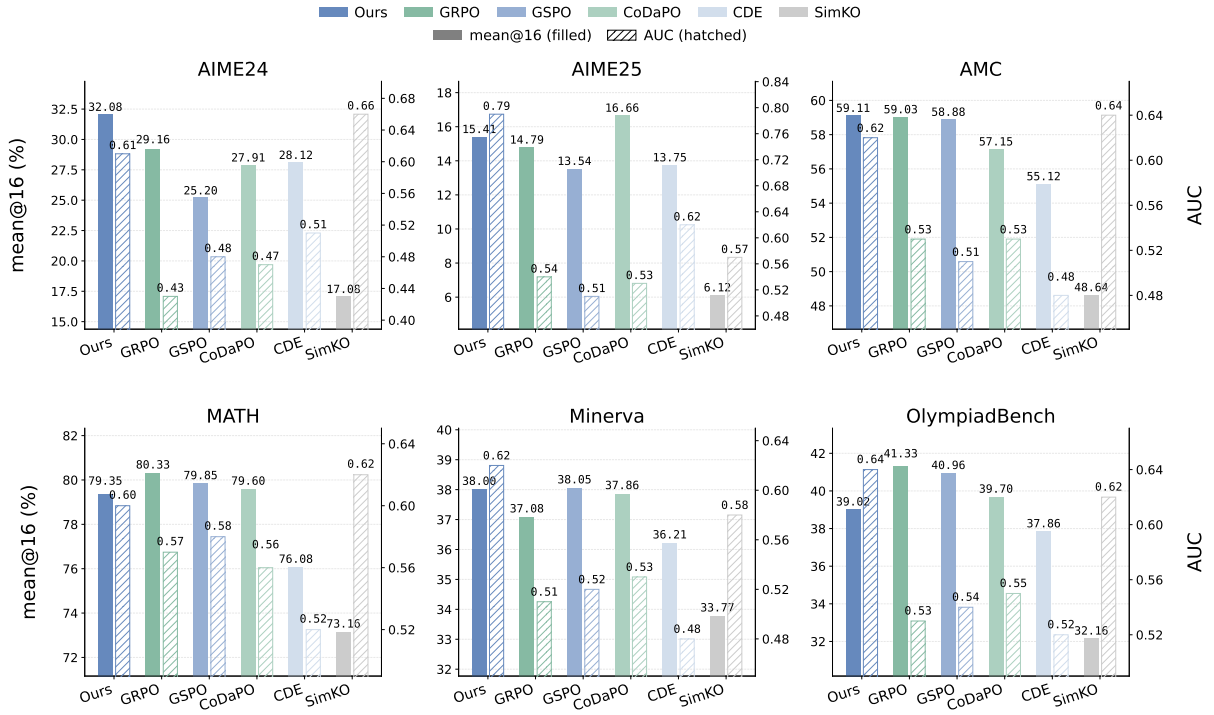


Figure 3: Results of calibration (measured by AUC-mean) and accuracy (measured by mean@16) for our method and all baselines on Qwen2.5-Math-7B across six test benchmarks.

Model	Ours	GRPO	GSPO	CoDaPO	CDE	SimKO
Qwen2.5-Math-1.5B	**25.33**	20.33	20.00	21.67	16.67	11.67
Qwen2.5-Math-7B	**38.33**	33.33	32.21	31.66	31.66	23.33

Table 1: Inference-time scaling accuracy (%) of different methods on the AIME 2024 and AIME 2025 datasets.

across all benchmarks without sacrificing accuracy. As shown in Figure 3 and Appendix B.2 Figure 5, compared to GRPO and GSPO, our method significantly improves calibration on all test datasets. In particular, on AIME 2025, CAPO increases AUC by approximately 15% for the 1.5B model (from 0.63 to 0.78) and by about 25% for the 7B model (from 0.54 to 0.79). Meanwhile, the mean@8 accuracy achieved by our method is comparable to or even better than that of GRPO, and it attains the highest accuracy on AIME 2024, AIME 2025, and Minerva.

In contrast, other methods aimed at addressing the calibration issue of GRPO provide only limited improvements in AUC and, in some cases, even slightly degrade it on certain datasets. Moreover, none of these methods are able to preserve consistent high accuracy. In particular, SIMKO severely damages accuracy, exhibiting an approximately 12% drop on AMC and a 7.7% drop on AIME 2024 compared to GRPO.

Stable and Consistent Optimization Dynamics We further present training dynamics to demonstrate the stability and consistency of our method. Figure 1(b)(c) shows AUC-mean on AIME 2024 and AIME 2025 throughout training. While GRPO and GSPO exhibit a gradual degradation in calibration as optimization proceeds, our method steadily improves AUC over training steps, demonstrating sustained calibration optimization rather than checkpoint-specific effects. Figure 6 in Appendix B.2 reports validation mean@8 accuracy, where only our method maintains stable performance comparable to GRPO for both 1.5B and 7B models. Together, these results indicate that our approach achieves more stable and reliable optimization behavior throughout training.

Hallucination Mitigation. We evaluate hallucination mitigation through the precision-coverage trade-off, where a model abstains from answering if its confidence falls below a threshold. Varying

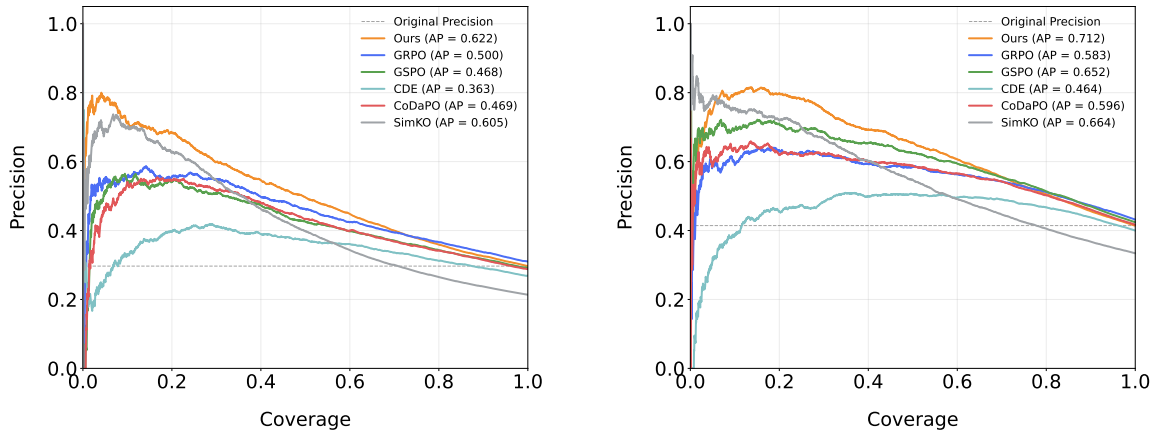


Figure 4: Precision-Coverage curves of our method and all baselines on six test benchmarks for the Qwen2.5-Math-1.5B (left) and Qwen2.5-Math-7B models (right).

this threshold traces a precision-coverage curve that characterizes the fraction of answered questions (coverage) and the accuracy among answered ones (precision). An ideal model should maintain high precision and coverage level. As shown in Figure 4, our method achieves a Pareto-optimal trade-off, consistently dominating GRPO, GSPO, and other baselines. Notably, while SimKO shows high precision at low coverage due to its ranking capability, its overall lower accuracy causes a sharp precision drop as coverage increases. Conversely, although GRPO matches our accuracy, its inferior calibration leads to consistently lower precision. By simultaneously enhancing calibration and accuracy, our method achieves a more favorable precision–coverage trade-off, enabling reliable hallucination mitigation with minimal abstention.

Impact of Calibration on Inference-Time Scaling The effectiveness of confidence-based inference-time scaling depends jointly on a model's accuracy and its calibration quality. We adopt the inference-time scaling algorithm proposed in Zhou et al. (2025b) (see Appendix A.4 for implementation details). As shown in Table 1, our method achieves the highest accuracy under inference-time scaling (5% over GRPO for 1.5B and 7B model), clearly demonstrating the critical role of calibration in inference-time scaling and the effectiveness of our approach.

5.3 Ablation

Sensitivity of Hyperparameters As detailed in Appendix B.2 Figure 8, we conduct sensitivity tests on $\tau \in \{0.4, 0.6, 1.0\}$ and different masking intervals. In both cases, the performance impact is marginal, confirming that our method is robust to the specific selection of τ and the `ref-high`/`ref-low` thresholds.

Ablation of the Denoising Strategy We further compare the performance of our method with and without the masking mechanism. As shown in Appendix B.2 Figure 9, removing the masking mechanism causes the model entropy to gradually increase, leading to early stagnation or even degradation in accuracy. In contrast, with the masking mechanism enabled, the model entropy remains stable and accuracy improves steadily. These observations demonstrate that noisy samples can severely disrupt optimization stability, and highlight the effectiveness of the proposed masking mechanism. Furthermore, applying only masking to GRPO fails to improve AUC, demonstrating that a consistent surrogate objective is essential.

6 Conclusion

In this paper, we have quantitatively analyzed and mitigated the calibration degradation induced by GRPO-style algorithms. We show that this degradation stems from reward-only advantage estimation. This estimation causes the gradient unaligned with relative calibration. To address this issue, we propose CAPO, which incorporates uncertainty-aware advantage estimation grounded in a consistent logistic AUC surrogate, together with a denoising mechanism to ensure stable training. Extensive empirical results demonstrate that CAPO trains models that achieve both high accuracy and calibration, leading to superior performance in hallucination mitigation and inference-time scaling.

Limitations

This study mainly evaluates the proposed method on mathematical reasoning benchmarks, which are also the primary focus of GRPO-style Reinforcement Learning with Verifiable Rewards (RLVR) methods. However our method is designed to be broadly applicable to training reasoning models. Its effectiveness on other reasoning tasks such as logical puzzles, commonsense reasoning, and open-domain question answering, remains to be further validated.

Acknowledgment

This work is supported by National Key R&D Program of China (2025ZD0122003), and in part by the Beijing Major Science and Technology Project under Contract no. Z251100008425009.

References

Michael Bereket and Jure Leskovec. 2025. Uncalibrated reasoning: Grpo induces overconfidence for stochastic outcomes. *arXiv preprint arXiv:2508.11800*.

Toon Calders and Szymon Jaroszewicz. 2007. Efficient auc optimization for classification. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 42–53. Springer.

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. 2019. On symmetric losses for learning from corrupted labels. In *International Conference on Machine Learning*, pages 961–970.

Yu-Neng Chuang, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. 2025. Learning to route llms with confidence tokens. *Preprint*, arXiv:2410.13284.

Runpeng Dai, Linfeng Song, Haolin Liu, Zhenwen Liang, Dian Yu, Haitao Mi, Zhaopeng Tu, Rui Liu, Tong Zheng, Hongtu Zhu, and 1 others. 2025. Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*.

Qi Feng, Yihong Liu, and Hinrich Schütze. 2025. Your pretrained model tells the difficulty itself: A self-adaptive curriculum learning paradigm for natural language understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 222–239.

Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. 2013. One-pass auc optimization. In *International Conference on Machine Learning*, pages 906–914. PMLR.

Wei Gao and Zhi-Hua Zhou. 2012. On the consistency of auc pairwise optimization. *arXiv preprint arXiv:1208.0645*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595.

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

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Wojciech Kotlowski, Krzysztof J Dembczynski, and Eyke Huellermeier. 2011. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1113–1120.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and 1 others. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Charles X Ling, Jin Huang, Harry Zhang, and 1 others. 2003. Auc: a statistically consistent and more discriminating measure than accuracy. In *International Joint Conference on Artificial Intelligence*, volume 3, pages 519–524.

Haotian Liu, Shuo Wang, and Hongteng Xu. 2025a. C^2 GSPG: Confidence-calibrated group sequence policy gradient towards self-aware reasoning. *arXiv preprint arXiv:2509.23129*.

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

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, and 1 others. 2025a. Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl. *Notion Blog*.

Yichen Luo, Yebo Feng, Jiahua Xu, Paolo Tasca, and Yang Liu. 2025b. Llm-powered multi-agent system for automated crypto portfolio management. *Preprint*, arXiv:2501.00826.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. 2025. Simko: Simple pass@ k policy optimization. *arXiv preprint arXiv:2510.14807*.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2025. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149.

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

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Josefa Lia Stoisser, Marc Boubnovski Martell, Lawrence Phillips, Gianluca Mazzoni, Lea Mørch Harder, Philip Torr, Jesper Ferkinghoff-Borg, Kaspar Martens, and Julien Fauqueur. 2025. Towards agents that know when they don't know: Uncertainty as a control signal for structured reasoning. *arXiv preprint arXiv:2509.02401*.

Sree Harsha Tanneru, Chirag Agarwal, and Himabindu Lakkaraju. 2024. Quantifying uncertainty in natural language explanations of large language models. In *International Conference on Artificial Intelligence and Statistics*, pages 1072–1080. PMLR.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. *arXiv preprint arXiv:2404.17287*.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. 2025. Uncertainty quantification for llms through minimum bayes risk: Bridging confidence and consistency. *arXiv preprint arXiv:2502.04964*.

Xiaoxuan Wang, Yihe Deng, Mingyu Derek Ma, and Wei Wang. 2025. Entropy-based adaptive weighting for self-training. *Preprint*, arXiv:2503.23913.

David Warren and Mark Dras. 2025. Bi-directional model cascading with proxy confidence. *Preprint*, arXiv:2504.19391.

Bingbing Wen, Chenjun Xu, Robert Wolfe, Lucy Lu Wang, Bill Howe, and 1 others. 2024. Mitigating overconfidence in large language models: A behavioral lens on confidence estimation and calibration. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J Su, and Li Shen. 2025. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. *arXiv preprint arXiv:2505.01997*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *Preprint*, arXiv:2306.13063.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024a. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *Preprint*, arXiv:2409.12122.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024b. On verbalized confidence scores for llms. *Preprint*, arXiv:2412.14737.

Tianbao Yang and Yiming Ying. 2022. Auc maximization in the era of big data and ai: A survey. *ACM computing surveys*, 55(8):1–37.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. 2021. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Qingcheng Zeng, Weihao Xuan, Leyang Cui, and Rob Voigt. 2025. Thinking out loud: Do reasoning models know when they're right? *arXiv preprint arXiv:2504.06564*.

Peilin Zhao, Steven CH Hoi, Rong Jin, and Tianbo Yang. 2011. Online auc maximization.

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. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Zhanke Zhou, Xiangyu Lu, Chentao Cao, Brando Miranda, Tongliang Liu, Bo Han, and Sanmi Koyejo. 2025a. Codapo: Confidence and difficulty-adaptive policy optimization for post-training language models. In *2nd AI for Math Workshop@ ICML 2025*.

Zhi Zhou, Tan Yuhao, Zenan Li, Yuan Yao, Lan-Zhe Guo, Xiaoxing Ma, and Yu-Feng Li. 2025b. Bridging internal probability and self-consistency for effective and efficient llm reasoning. *arXiv preprint arXiv:2502.00511*.

Dixian Zhu, Xiaodong Wu, and Tianbao Yang. 2022. Benchmarking deep auroc optimization: Loss functions and algorithmic choices. *arXiv preprint arXiv:2203.14177*.

A Additional Proof

A.1 Derivation of the Gradient of the GRPO Objective

Proof idea. We start from the full PPO clipped surrogate with importance sampling (IS) ratio. As proved by previous works (Schulman et al., 2017, 2015), IS ratio ensures the surrogate uses tokens drawn from the old policy while producing a unbiased gradient estimate for the current policy (given a fixed state). The clipping further enforces a trust-region-style constraint, and under a first-order approximation (small policy update), the clipped surrogate has the same leading-order gradient direction as the original optimization objective. Hence, PPO's gradient matches the REINFORCE-style gradient. GRPO differs from PPO only by replacing the advantage estimator with a group-wise baseline $\hat{A}_i = R_i - \bar{R}$, so substituting this advantage into the REINFORCE form yields the corresponding GRPO gradient.

Proof Let $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ denote the importance sampling ratio. The PPO clipped objective is

$$L^{\text{PPO}}(\theta) = \mathbb{E}_{(s_t,a_t)\sim\pi_{\theta_{\text{old}}}}\Big[\min\Big(r_t(\theta)A_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\, A_t\Big)\Big]. \tag{23}$$

To first order, the gradient of the above off-policy objective reduces to the on-policy REINFORCE-form gradient:

$$\nabla_\theta L^{\text{PPO}}(\theta) = \nabla_\theta \mathbb{E}_{(s_t,a_t)\sim\pi_\theta}\left[A_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\right], \tag{24}$$

The GRPO objective is obtained by estimating the advantage in PPO by group-wise reward differences. Given a group of G responses $\{o_i\}_{i=1}^G$ with rewards $\{R_i\}$, GRPO uses

$$\tilde{A}_i = R_i - \bar{R}, \qquad \bar{R} = \frac{1}{G}\sum_{i=1}^{G} R_i. \tag{25}$$

Substituting this $\tilde{A}_i$ into the REINFORCE-form gradient 24 yields:

$$\mathbb{E}_{o_{1:G}\sim\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}(R_i - \bar{R})\,\nabla_\theta \pi_\theta(o_{i,t} \mid o_{i,<t})\right]. \tag{26}$$

Finally, using the definition

$$lpm_\theta(o_i) = \frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\log \pi_\theta(o_{i,t} \mid o_{i,<t}), \tag{27}$$

we can obtain Equation 13.

A.2 Definition of U-Statistics

Let $\{Z_1, Z_2, \ldots, Z_n\}$ be i.i.d. random variables drawn from an underlying distribution $\mathcal{D}$. Consider a symmetric measurable function (kernel)

$$h : \mathcal{Z}^k \to \mathbb{R},$$

where $k \geq 1$ denotes the order of the kernel and symmetry means that

$$h(z_1, \ldots, z_k) = h(z_{\pi(1)}, \ldots, z_{\pi(k)})$$

for any permutation π of $\{1, \ldots, k\}$.

Definition (U-statistic). The *U-statistic* associated with kernel h is defined as

$$U_n \triangleq \binom{n}{k}^{-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} h(Z_{i_1}, \ldots, Z_{i_k}). \tag{28}$$

The U-statistic U_n is an estimator of the population quantity

$$\theta \triangleq \mathbb{E}_{(Z_1, \ldots, Z_k) \sim \mathcal{D}^k}[h(Z_1, \ldots, Z_k)]. \tag{29}$$

Unbiasedness. A key property of U-statistics is that they provide unbiased estimators of the target quantity θ.

Proposition 1 (Unbiasedness of U-statistics)**.** *For any symmetric kernel h with finite expectation,*

$$\mathbb{E}[U_n] = \theta. \tag{30}$$

Proof. By linearity of expectation,

$$\mathbb{E}[U_n] = \binom{n}{k}^{-1} \sum_{1 \leq i_1 < \cdots < i_k \leq n} \mathbb{E}[h(Z_{i_1}, \ldots, Z_{i_k})]. \tag{31}$$

Since $\{Z_i\}$ are i.i.d., each summand has expectation θ. There are exactly $\binom{n}{k}$ such terms, which yields

$$\mathbb{E}[U_n] = \binom{n}{k}^{-1} \cdot \binom{n}{k} \cdot \theta = \theta. \tag{32}$$

□

Relation to the main text. In the main text, Equation (14) can be viewed as a U-statistic with kernel order $k = 2$. Consequently, it is an unbiased estimator of the expectation in Equation (15).

A.3 Completion of the Proof of Theorem 3

Inconsistency of $\phi(t) = -t$ for AUC optimization. AUC depends only on the *ordering* induced by the scoring function. In particular, for any $\alpha > 0$, scaling the scores does not change AUC:

$$\mathrm{AUC}(f) = \mathrm{AUC}(\alpha f). \tag{33}$$

Configuration	Value
train_batch_size	128
ppo_mini_batch_size	64
max_prompt_length	1050
max_response_length	3046
`ref-high`	2.5
`ref-low`	1.05
τ (1.5B model)	0.6
τ (7B model)	0.5
ϵ	0.2
learning rate	1×10^{-6}
entropy_coeff	0
kl_loss_coef	0
rollout.n	8
validation.rollout.n	16
rollout.temperature	1.0
validation.temperature	1.0
total_steps(for 1.5B)	600
total_steps(for 7B)	400

Table 2: Key hyperparameter settings used in our experiments.

However, the surrogate risk with $\phi(t) = -t$ is *scale-sensitive* and *unbounded*:

$$\mathcal{L}_{-t}(\alpha f) = -\mathbb{E}[\alpha f(Z^+) - \alpha f(Z^-)] = \alpha\, \mathcal{L}_{-t}(f). \tag{34}$$

Take any nontrivial scoring function f with $\mathcal{L}_{-t}(f) < 0$ and $AUC(f) < 1$ (not optimal). Consider the sequence $f_m = \alpha_m f$ with $\alpha_m \uparrow +\infty$. Then $\mathrm{AUC}(f_m) = \mathrm{AUC}(f)$ for all m by (33), but $\mathcal{L}_{-t}(f_m) \to -\infty$ by (34). Hence, one can drive the surrogate objective arbitrarily close to its infimum without improving (or even changing) AUC. Therefore, minimizing the $\phi(t) = -t$ surrogate does not guarantee approaching an AUC-optimal scorer, i.e., this surrogate is not consistent for AUC optimization.

Generalizing to Reward-Only Advantage Estimators. It is worth noting that this result is not limited to GRPO, but applies to all algorithms that employ reward-only advantage estimators. Since the objectives induced by this class of advantage estimators are always linear with respect to the confidence function, they necessarily satisfy the property in (34). As a result, similar scaling counterexamples can be constructed to show that these objectives are not consistent AUC surrogate losses.

A.4 Pseudocode of the Inference-Time Scaling Algorithm

The inference-time scaling algorithm in Algorithm 1 is a direct instantiation of the *Perplexity Consistency* principle proposed in prior work Zhou et al., 2025b. In our implementation, N is 16.

The core assumption of Perplexity Consistency is that if the model is well-calibrated, responses corresponding to the correct answer should, on average, exhibit lower perplexity (higher likelihood) than incorrect ones.

Instead of selecting the single response with minimum perplexity, the proposed method aggregates confidence at the *answer level*. All responses that yield the same extracted final answer are grouped, and their probabilities are summed to form an aggregated confidence score. This aggregation achieves two desirable properties:

- **Self-consistency**: answers that are repeatedly produced by the model with high likelihood are reinforced.
- **Noise robustness**: low-probability or spurious generations contribute negligibly to the final decision.

Importantly, the entire procedure relies solely on the model's *endogenous confidence*, quantified by perplexity. As such, it constitutes a lightweight yet effective form of inference-time scaling that leverages both the model's calibration quality and reasoning ability.

Algorithm 1 Perplexity-Consistency Based Inference-Time Scaling

Input: question q, number of samples N, model π_θ
Output: final predicted answer $\hat{a}$
Sample N independent responses $\{y_i\}_{i=1}^{N} \sim \pi_\theta(\cdot \mid q)$
for $i \leftarrow 1$ **to** N **do**
    Extract final answer a_i from response y_i
    Compute token-level average log-probability:
    $\ell_i \leftarrow \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log p_\theta(y_{i,t} \mid y_{i,<t}, q)$
end for
Group responses by identical extracted answers:
    $\mathcal{G}(a) \leftarrow \{\, i \mid a_i = a \,\}$
for all candidate answers a **do**
    Compute aggregated confidence:
    $C(a) \leftarrow \sum_{i \in \mathcal{G}(a)} \exp(\ell_i)$
end for
$\hat{a} \leftarrow \arg\max_a C(a)$
return $\hat{a}$

B Additional Experiment details

B.1 Experimental Setup

Experimental configurations for the 1.5B and 7B models are detailed in Table 2. All methods share these identical settings, with method-specific hyperparameters following their original papers. For our method, we set the temperature parameter τ to 0.6 for 1.5B model and 0.5 for 7B model, `ref-high` and `ref-low` to the lower and upper quartiles of the reference model's PPL distribution over correct and incorrect responses, respectively. Experiments were conducted using the verl framework on 8 $\times$ A100 GPUs. Training a single CAPO/GRPO run to convergence takes approximately 24 hours for the 1.5B model and 48 hours for the 7B model.

B.2 Additional Experimental Results

Additional Experimental Results of Qwen2.5-Math-1.5B model and Accuracy Curves Figure 5 demonstrates the complete experimental results of Qwen2.5-Math-1.5B for all methods over six benchmarks. Our method achieves a significant improvement in calibration over all baselines on the 1.5B model, while maintaining accuracy gains that are comparable to or even surpass those of GRPO. In contrast, other methods either yield only limited improvements in calibration or incur a degradation in accuracy. Figure 6 shows the accuracy of all methods on the validation set as a function of training steps. Our method exhibits accuracy improvements comparable to GRPO on both model scales.

Ablation of Hyperparameters As illustrated in Figure 8, our method exhibits low sensitivity to the variation of the hyperparameter τ across different values ($\tau = 0.4, 0.6, 1.0$). Similarly, when adjusting the values of *ref-high* and *ref-low* to tighten the masking range from $[1.05, 2.5]$ to $[1.25, 2.1]$, Figure 9(b) demonstrates that the performance remains similarly insensitive to these hyperparameter settings.Figure 7 shows that applying the masking mechanism to GRPO alone does not improve calibration, highlighting the importance of the

calibration-aware advantage estimator.

Ablation of Noise Masking mechanism Comparing the performance of the algorithm with and without the masking mechanism, it can be observed from Figure 9(a) that removing the mask leads to a gradual increase in model entropy, causing the accuracy (acc) to stagnate prematurely or even decline. In contrast, with the inclusion of the masking mechanism, the model's entropy remains stable, and the accuracy exhibits a steady improvement. These observations underscore the disruptive impact of noisy data on experimental stability and optimization signals, as well as the effectiveness of the masking mechanism.

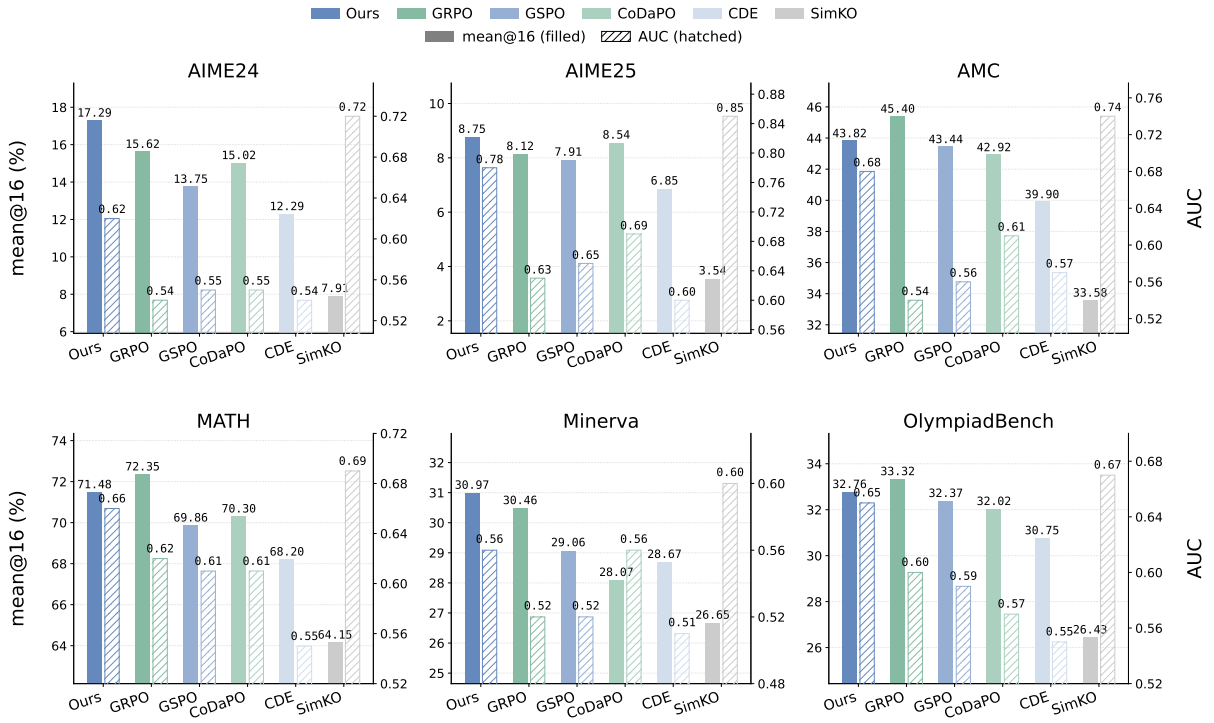
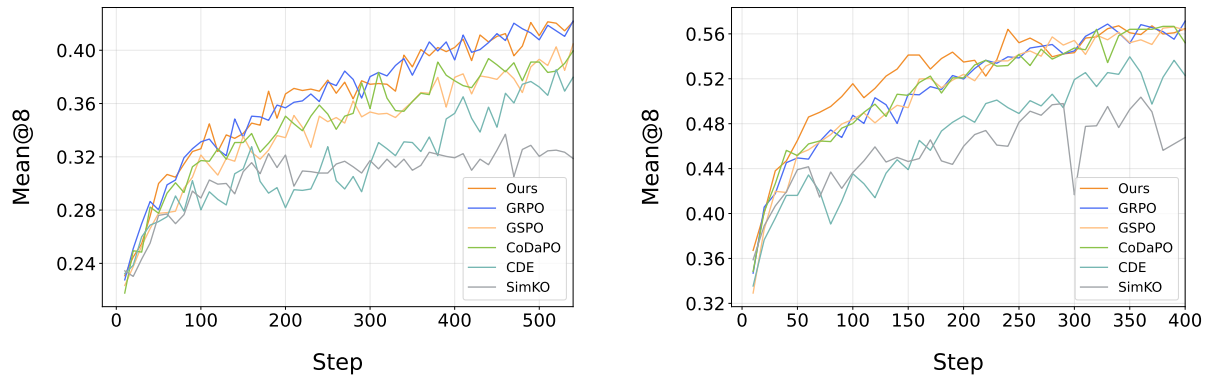
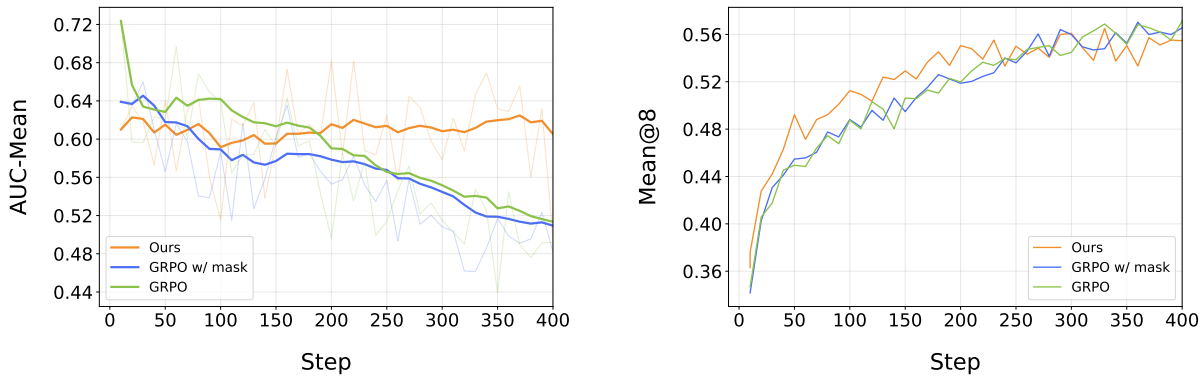


Figure 5: Results of calibration (measured by AUC-mean) and accuracy (measured by mean@16) for our method and all baselines on Qwen2.5-Math-1.5B across six test benchmarks.



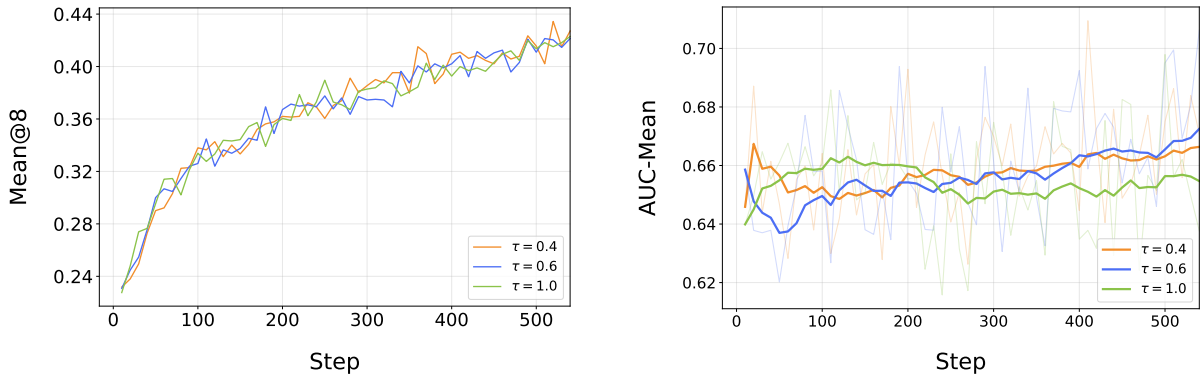
(a) (b)

Figure 6: Accuracy trajectories on the validation set over training steps for our method and all baselines on the Qwen2.5-Math-1.5B (left) and Qwen2.5-Math-7B models (right).



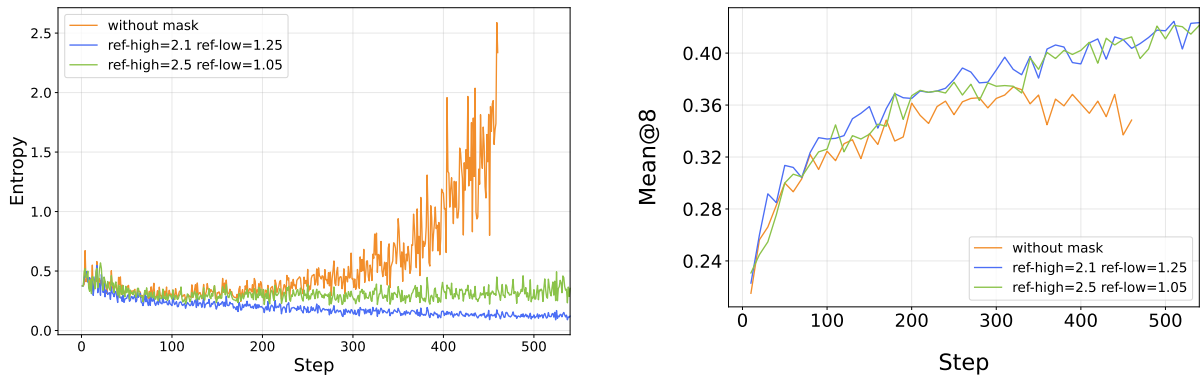
(a) (b)

Figure 7: Ablation studies on the effectiveness of applying the masking mechanism alone to GRPO. The results show that the mask by itself does not improve calibration, highlighting the necessity of the calibration-aware advantage estimator.



(a) (b)

Figure 8: Ablation studies on the sensitivity of accuracy improvement curves (a) and calibration metrics (b) to the hyperparameter τ.



(a) (b)

Figure 9: Ablation studies on the impact of the noise-masking mechanism on training stability (a) and the sensitivity to the `ref-high` and `ref-low` hyperparameters (b).