DisCO: Reinforcing Large Reasoning Models with Discriminative Constrained Optimization

Gang Li[1] **Ming Lin** **Tomer Galanti**[1] **Zhengzhong Tu**[1] **Tianbao Yang**[1]
[1] Texas A&M University
{gang-li, galanti, tzz, tianbao-yang}@tamu.edu
linming04@gmail.com

Abstract

The recent success and openness of DeepSeek-R1 have brought widespread attention to Group Relative Policy Optimization (GRPO) as a reinforcement learning method for large reasoning models (LRMs). In this work, we analyze the GRPO objective under a binary reward setting and reveal an inherent limitation of question-level difficulty bias arising from its group relative advantage function. We also identify a connection between GRPO and traditional discriminative methods in supervised learning. Motivated by these insights, we introduce a new **Discriminative Constrained Optimization (DisCO)** framework for reinforcing LRMs, grounded in the principle of discriminative learning: increasing the scores of positive answers while decreasing those of negative ones. The main differences between DisCO and GRPO and its recent variants are: (1) it replaces the group relative objective with a discriminative objective defined by a scoring function; (2) it abandons clipping-based surrogates in favor of non-clipping RL surrogate objectives used as scoring functions; (3) it employs a simple yet effective constrained optimization approach to enforce the KL divergence constraint. As a result, DisCO offers notable advantages over GRPO and its variants: (i) it completely eliminates difficulty bias by adopting discriminative objectives; (ii) it addresses the entropy instability in GRPO and its variants through the use of non-clipping scoring functions and a constrained optimization approach, yielding long and stable training dynamics; (iii) it allows the incorporation of advanced discriminative learning techniques to address data imbalance, where a significant number of questions have more negative than positive generated answers during training. Our experiments on enhancing the mathematical reasoning capabilities of SFT-finetuned models show that DisCO significantly outperforms GRPO and its improved variants such as DAPO, achieving average gains of 7% over GRPO and 6% over DAPO across six benchmark tasks for a 1.5B model.[1]

1 Introduction

The recent success and openness of DeepSeek-R1 have sparked a surge of interest in large reasoning models (LRMs), particularly in the context of fine-tuning via reinforcement learning (RL) [23]. The core approach involves iteratively generating synthetic data using the reasoning model and applying a rule-based reward mechanism to label the outputs. These rewards are then used to update the policy, *i.e.*, the reasoning model itself. Notably, this framework, featuring a novel policy optimization method called Group Relative Policy Optimization (GRPO), has enabled DeepSeek-R1 to achieve performance comparable to advanced proprietary LRMs at that time such as OpenAI-o1 on many reasoning benchmarks. As a result, GRPO has rapidly become a focal point for advancing LRM capabilities, particularly in domains like mathematics and scientific reasoning.

[1]The code is available at: https://github.com/Optimization-AI/DisCO

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-1921

Several efforts have sought to replicate the performance of DeepSeek-R1 or to further enhance reasoning models using GRPO [67, 31, 42, 27, 69, 74, 5], while few others have tried to identify its inherent limitations with potential remedies [40, 79, 39]. Useful tricks have been introduced to improve GRPO [40, 41, 79, 27, 13, 82]. For instance, DAPO [79] employs two distinct clipping hyperparameters to mitigate *entropy collapse*, encouraging exploration. Dr. GRPO [40] removes the variance normalization in advantage function, aiming to mitigate the issue of *difficulty bias*. However, these approaches remain heuristic and ad-hoc, lacking a principled foundation and falling short of fully addressing GRPO's inherent limitations. Our analysis identifies that Dr. GRPO continues to suffer from the difficulty bias issue, while our experiments show that DAPO may induce excessive entropy growth, producing highly random outputs. This motivates us to explore a central question:

How can we design more effective optimization methods for reinforcing large reasoning models in a principled manner without inheriting the limitations of GRPO?

This paper addresses the above question through a complete redesign of the objective function, grounded in the principles of discriminative learning. Specifically, we first analyze the objective function of GRPO and its variants under **a binary reward setting**, leading to two key insights: **(1)** the root cause of GRPO's difficulty bias lies in its group relative advantage function, which induces disproportionately small weights to questions that are either too easy or too hard; and **(2)** there exists a conceptual connection to traditional discriminative approaches in AUC maximization, which aim to increase the scores of positive outputs while decreasing that of negative outputs.

Building upon these insights, we propose a principled optimization framework for reinforcing large reasoning models based on discriminative learning. Specifically, we optimize a discriminative objective using a proper scoring function over input-output pairs, which increases the score of positive outputs and decreases that of negative ones. The flexibility of our framework allows us to leverage simple non-clipping RL surrogate objectives as scoring functions without suffering from entropy instability, and to incorporate advanced discriminative techniques to address data imbalance in generated rollouts. To ensure training stability, we adopt a simple yet effective constrained optimization method to enforce a trust region constraint bounding the KL divergence between the updated model and the old model. Our experiments for mathematical reasoning show that DisCO significantly outperforms all baselines for fine-tuning DeepSeek-R1-Distill-Qwen and -Llama models with a maximum 8k response length for both training and inference, and also achieves a better performance than GRPO that uses a maximum 24k length for training and 32k length for inference.

Our main contributions are summarized as follows:

- We present an **analysis of GRPO's objective function**, identifying the root cause of difficulty bias and revealing its conceptual connection to classic discriminative methods for AUC maximization.
- We introduce a **principled discriminative constrained optimization framework** for reinforcing large reasoning models, which avoids both difficulty bias and training instability. This framework gives rise to a family of methods we refer to as **DisCO**.
- We demonstrate **significant improvements** of our DisCO method over GRPO and four other baselines, including DAPO, through experiments for fine-tuning LRMs on mathematical reasoning tasks, with evaluations across six benchmarks.

2 Related Work

Large Reasoning Models (LRMs). Recent advances of LRMs, such as OpenAI o1 [51], DeepSeek-R1 [23] and Kimi K1.5 [63], have demonstrated strong reasoning capability in solving complex tasks. Departing from earlier approaches in LLMs, such as Chain-of-thought (CoT) prompting [66, 48, 81], Tree-of-Thought [78], Monte Carlo Tree Search [18, 64, 71], a major breakthrough was achieved by scaling RL training using verifiable rewards to incentivize LLMs to learn through self-exploration [23, 63]. Inspired by DeepSeek-R1's core algorithm GRPO [59], the research community has actively pursued improved techniques for large-scale RL training, focusing primarily on three directions: algorithm design [79, 40, 13, 39, 63, 61], reward curation [82, 67, 79], and sampling strategies [79, 27, 83, 30]. Our work falls under the category of algorithm design.

Among these, Dr. GRPO [40] identifies response-level length bias and question-level difficulty bias in GRPO algorithm, advocating the removal of length and advantage normalization to improve token efficiency. DAPO [79] highlights several limitations of GRPO, such as entropy collapse, training instability, and biased loss, and addresses them through techniques like decoupled clipping,

dynamic sampling, and a token-level policy loss. GPG [13] introduces a simplified REINFORCE-based objective that eliminates the need for both the critic and reference models, thereby enhancing scalability for RL training. TRPA [61] simply uses the Direct Preference Optimization (DPO) objective and a KL divergence regularization for fine-tuning LRMs. It can be recovered from our basic approach, which uses a logistic function as the surrogate loss, the log of likelihood ratio with respect to a frozen reference model as the scoring function, and the KL divergence as a regularization rather than a constraint. However, it does not address the imbalanced rollouts. The uniqueness and significance of our contributions lie in the analysis of GRPO objective and its variants that reveal key limitations, and the integration of advanced discriminative learning approaches for handling imbalanced rollouts and efficient constrained optimization technique for ensuring training stability.

Reinforcement Learning (RL). RL is a learning paradigm centered on control and decision-making, in which an agent optimizes a target objective through trial-and-error interactions with its environment [8]. RL approaches are typically categorized into model-based [60, 53, 49, 17] and model-free methods [68, 62, 46, 56, 57, 38, 20]. Among model-free methods, the evolution from Vanilla Policy Gradient [68, 62] to TRPO [56] and PPO [57] has influenced the development of GRPO. In the context of fine-tuning LLMs, another line of work is RL from human feedback (RLHF). An early example of connecting RL with LLMs dates back to OpenAI's work on integrating human preferences to improve text generation tasks, such as summarization using the PPO algorithm [85]. This approach was later extended to fine-tune LLMs for instruction following and/or alignment on helpfulness and harmlessness [52, 3, 22, 75]. Due to the high data requirements and training costs of standard RLHF, off-policy methods like DPO [54] and its variants [2, 16, 72, 44, 24], have been proposed to reduce reliance on explicit reward models. Another line of on-policy algorithms for RLHF, such as RLOO [1], ReMax [36], and REINFORCE++ [29], has been introduced to reduce the computational burden by removing the critic network in PPO. While some works [43, 32, 29, 10, 70] attempt to adapt RLHF techniques for reasoning tasks, they have not yielded significant improvements.

Discriminative Learning. Parallel to RL, discriminative learning is another classical learning paradigm, that has been studied extensively for many traditional tasks, including multi-class classification [14, 15, 6], AUC maximization [76, 80], and learning to rank [9, 19, 7]. These methods are grounded in the common principle of increasing prediction scores for positive (relevant) labels (data) while decreasing scores for negative (irrelevant) ones. Nevertheless, discriminative learning remains under-explored in the cotext of LLM training. Recently, Guo et al. [25] proposed discriminative probabilistic approaches for supervised fine-tuning of LLMs. However, unlike our approach, they did not employ an RL framework with verifiable rewards to fine-tune LRMs.

3 Preliminaries

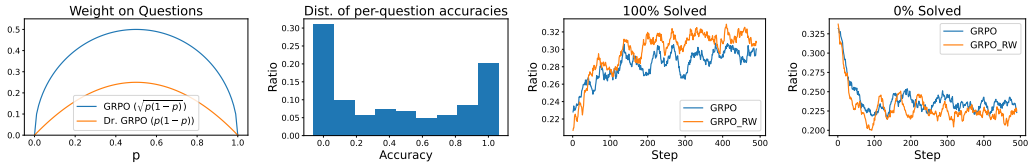
We consider fine-tuning a generative reasoning model π_θ parameterized by θ. The old model in one step of learning is denoted by π_{old}. It is used to generate answers for a set of input questions. Given a question q (with prompt included), the generated output o follows the distribution $\pi_{\text{old}}(\cdot|q)$, which includes reasoning traces and the final answer. Specifically, output o is generated token by token, i.e., $o_t \sim \pi_{\text{old}}(\cdot|q, o_{<t})$, for $t = 1, \cdots, |o|$. We consider a rule-based reward mechanism that returns a binary value for a given question q and its corresponding answer in the output o, which uses either exact match against extracted answer or a formal verification tool [23, 33, 55]. Let $r(o|q) \in \{1, 0\}$ denote the reward assigned to an output o with respect to the input q. Let $p(q) = \mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}[r(o|q)] \in [0, 1]$, which quantifies the difficulty of the question q under the model π_{old}. We denote by $\pi^+_{\text{old}}(\cdot|q)$ the conditional distribution of outputs when the reward is one (i.e., positive answers) and by $\pi^-_{\text{old}}(\cdot|q)$ the conditional distribution of outputs when the reward is zero (i.e., negative answers). By the law of total expectation, for any function $g(o, q)$ we have

$$\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}[g(o, q)] = p(q)\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)}[g(o, q)] + (1 - p(q))\mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)}[g(o, q)]. \tag{1}$$

Group Relative Policy Optimization (GRPO). The key idea of GRPO is to generate multiple outputs for an input q and define a group relative advantage function. For analysis, we consider the expectation formulation instead of empirical average of the GRPO objective for maximization:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_q\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}\left[\frac{1}{|o|}\sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\text{old}}(o_t|q, o_{<t})}, A(o|q)\right)\right] - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}), \tag{2}$$

 https://doi.org/10.52202/085713-1921



(a) (b) (c) (d)

Figure 1: (a) Weight on questions based on correctness probability p; (b) Histogram of per-question accuracy evaluated in the GRPO learning; (c) Comparison of the ratio of questions with 100% correctness probability; (d) Comparison of the ratio of questions with 0% correctness probability.

where $f(x,y) = \min(xy, \text{clip}(x, 1-\epsilon, 1+\epsilon)y)$, $A(o|q) = \frac{(r(o|q) - \mathbb{E}_{o' \sim \pi_{\text{old}}(\cdot|q)} r(o'|q))}{\sqrt{\text{Var}_{o' \sim \pi_{\text{old}}(\cdot|q)} r(o'|q)}}$ is the advantage function that quantifies how much better the reward of o is compared to average reward, π_{ref} is a frozen reference model.

Recently, several variants of GRPO have been introduced [79, 13, 39, 82, 40]. Many of them retain the advantage function $A(o|q)$ while modifying other components such as hyper-parameter ϵ, the normalization factor and the likelihood ratio. Several works employ an unnormalized advantage function $\hat{A}(o|q) = r(o|q) - \mathbb{E}_{o' \sim \pi_{\text{old}}(\cdot|q)} r(o'|q)$ [40, 13].

4 Analysis of GRPO and its variants

In the following analysis we assume $p(q) \in (0,1)$; otherwise we can remove them from consideration as done in practice [23, 42, 59].

Proposition 1 *Let us consider the objective of GRPO and its variants with the following form:*

$$\mathcal{J}_0(\theta) = \mathbb{E}_q \mathbb{E}_{o \sim \pi_{old}(\cdot|q)} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{old}(o_t|q, o_{<t})}, A(o|q) \right) \right]. \tag{3}$$

Assume that $f(x,y)$ is non-decreasing function of x such that $f(x,y) = \mathbb{I}(y > 0) y f^+(x,1) - \mathbb{I}(y \leq 0)|y| f^-(x,1)$, where both f^+, f^- are non-decreasing functions of x, then we have

$$\mathcal{J}_0(\theta) = \mathbb{E}_q \sqrt{p(q)(1-p(q))} \mathbb{E}_{o \sim \pi^+_{old}(\cdot|q), o' \sim \pi^-_{old}(\cdot|q)} [s^+_\theta(o,q) - s^-_\theta(o',q)], \tag{4}$$

where $s^+_\theta(o,q) = \frac{1}{|o|} \sum_{t=1}^{|o|} f^+\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, 1 \right)$ and $s^-_\theta(o,q) = \frac{1}{|o|} \sum_{t=1}^{|o|} f^-\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, 1 \right)$. In particular, for GRPO we have

$$f^+(x,1) = \min(x, 1+\epsilon), \quad f^-(x,1) = \max(x, 1-\epsilon). \tag{5}$$

Remark: The assumption of $f(x,y)$ indeed holds for GRPO and its variants. We will present the analysis for several variants of GRPO in Appendix B.3.

The proof of the above proposition is included in Appendix B.1 and is inspired by [47] with differences that lead to two **new insights** from Proposition 1 regarding the two components of $\mathcal{J}_0$. First, let us consider the component $\mathbb{E}_{o \sim \pi^+_{\text{old}}(\cdot|q), o' \sim \pi^-_{\text{old}}(\cdot|q)} [s^+_\theta(o,q) - s^-_\theta(o',q)]$. Since both f^+ and f^- are non-decreasing functions of the first argument, then both $s^+_\theta(o,q)$ and $s^-_\theta(o,q)$ are non-decreasing functions of $\pi_\theta(o_t|q,o_{<t})$. Hence, maximizing $\mathcal{J}_0$ would increase the likelihood of tokens in the positive answers and decrease the likelihood of tokens in the negative answers. This makes sense as we would like the new model to have a high likelihood of generating a positive (correct) answer and a low likelihood of generating a negative (incorrect) answer. This mechanism is closely related to traditional discriminative methods of supervised learning in the context of AUC maximization [77], which aims to maximize the scores of positive samples $o \sim \pi^+_{\text{old}}(\cdot|q)$ while minimizing scores of negative samples $o' \sim \pi^-_{\text{old}}(\cdot|q)$, where the q acts like the classification task in the AUC maximization. Hence, in the context of discriminative learning, we refer to $s^+(o,q)$ and $s^-(o,q)$ as scoring functions. Therefore, $\mathbb{E}_{o \sim \pi^+_{\text{old}}(\cdot|q), o' \sim \pi^-_{\text{old}}(\cdot|q),} [s^+(o,q) - s^-(o',q)]$ is a discriminative objective.

Second, let us consider the component $\omega(q) = \sqrt{p(q)(1-p(q))}$, which acts like a weight scaling the discriminative objective for each individual input question. It is this component that leads to difficulty bias. As shown in Figure 1(a), questions with very high $p(q)$ values (close to 1) or very

low $p(q)$ values (close to 0) receive small weights for their discriminative objectives, causing the optimization to focus primarily on questions of intermediate difficulty while paying little attention to hard questions ($p(q) \approx 0$) and easy questions ($p(q) \approx 1$). This mechanism may significantly hinder the learning efficiency. Intuitively, if the generated answers have only one correct solution out of 10 trials, i.e. $p(q) = 0.1$, we should grasp this chance to enhance the model instead of overlooking it. On the other hand, even when we encounter an easy question with a probability of $p(q) = 0.9$, we should keep improving the model rather than being satisfied because it still makes mistakes with respect to this question. Our hypothesis is that removing this weight could accelerate the training. To validate this hypothesis, we conducted a series of empirical experiments for fine-tuning a 1.5B model as described in Section 6. We start by examining whether a substantial number of questions have correctness probabilities ($p(q)$) near 0 or 1. As shown in Figure 1(b), during GRPO training, the correctness probabilities across individual questions appear broadly distributed, with many near 0 or 1. Then, we compare the original GRPO with a variant that removes weight $\sqrt{p(q)(1-p(q))}$:

$$\mathcal{J}_{\text{GRPO\_RW}} = \mathbb{E}_q \mathbb{E}_{o \sim \pi^+_{\text{old}}, o' \sim \pi^-_{\text{old}}} [s^+(o, q) - s^-(o', q)] - \beta \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{ref}). \tag{6}$$

The results are shown in Figure 1(c) and 1(d). We can observe that the variant without the weighting mechanism quickly achieves a higher ratio of 100% correctness and a lower ratio of 0% correctness, confirming the detrimental impact of the inappropriate weighting.

We note that the difficulty bias has been pointed out in a recent work Dr. GRPO [40]. To mitigate this issue, Dr. GRPO uses the un-normalized advantage function $\hat{A}(o|q)$. However, with a similar analysis as above (cf. Appendix B.3), we can derive that Dr. GRPO still has a question-level weight $\omega(q) = p(q)(1 - p(q))$ before the discriminative objective. As shown in Figure 1(a), this weight mitigates but does not eliminate the imbalanced weight across questions.

5 A Discriminative Constrained Optimization Framework

While the last section has suggested a tangible remedy to address the difficulty bias of GRPO and its variants by removing the weight before the discriminative objective, there are other issues of the scoring function of GRPO and its variants. Next, we propose a general discriminative learning framework for reinforcing LRMs and incorporate advanced techniques to facilitate the learning.

5.1 A basic approach

Motivated by the connection with AUC maximization, we redesign the objective directly from the principle of discriminative learning. For a given question q, let $s_\theta(o, q)$ denote a scoring function that measures how likely the model π_θ "predicts" the output o for a given input q [2]. Then the AUC score for the "task" q is equivalent to $\mathbb{E}_{o \sim \pi^+_{\text{old}}, o' \sim \pi^-_{\text{old}}} [\mathbb{I}(s_\theta(o, q) > s_\theta(o', q))]$. Using a continuous surrogate function ℓ, we form the following objective (in expectation form) for maximization:

$$\mathcal{J}_1(\theta) = \mathbb{E}_q \mathbb{E}_{o \sim \pi^+_{\text{old}}(\cdot|q), o' \sim \pi^-_{\text{old}}(\cdot|q)} \ell(s_\theta(o, q) - s_\theta(o', q)). \tag{7}$$

Different surrogate functions $\ell(\cdot)$ can be used. For comparison, with GRPO, we simply use the identity function $\ell(s) = s$. One difference from the discriminative objective (6) is that we use a single scoring function $s_\theta(o, q)$ for both positive outputs o and negative outputs o'. It is notable that the different scoring functions for positive and negative outputs in (6) actually arise from the clipping operations of GRPO objective. Recent works have found that the clipping could lead to entropy collapse [79]. In addition, the clipping could cause the vanishing gradient, which may also slow down the learning process. To avoid these issues, we consider non-clipping scoring functions.

Scoring functions. We consider two choices of scoring functions, i.e., log-likelihood and likelihood ratio. The log-likelihood (log-L) scoring function is defined by $s_\theta(o, q) = \frac{1}{|o|} \sum_{t=1}^{|o|} \log \pi_\theta(o_t|q, o_{<t})$. The likelihood ratio (L-ratio) scoring function is computed by $s_\theta(o, q) = \frac{1}{|o|} \sum_{t=1}^{|o|} \frac{\pi_\theta(o_t|q, o_{<t})}{\pi_{\text{old}}(o_t|q, o_{<t})}$. In Appendix B.2, we discuss the connection between the two scoring functions and the surrogate objectives of vanilla policy gradient methods [68] and TRPO [56], respectively.

Stabilize the training with Constrained Optimization. Training instability is a long-standing issue in RL [56, 57]. Different methods have been introduced to ensure stability. Recent RL-based methods for learning reasoning models either follow the clipping operation of PPO [57] or use the KL divergence regularization $\mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}})$ or $\mathbb{D}_{\text{KL}}(\pi_{\text{old}} || \pi_\theta)$ [59, 61, 52]. However, the clipping

[2] in the context of generative models, "predicts" is like "generates".

https://doi.org/10.52202/085713-1921

operation could lead to the entropy collapse [79], which we try to avoid by using the non-clipped scoring function. The KL divergence regularization $\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})$ while being used in traditional RL is not effective for preventing entropy collapse [27, 79]. Regarding the regularization with $\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta)$, earlier studies [56, 57] has found that it would be difficult to choose a single value of the regularization parameter that performs well across different problems or even within a single problem where the the characteristics change over the course of learning. To tackle this issue, we revisit the idea of trust region constraint of TRPO [56], i.e., restricting the updated model θ in the trust region $\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta) \leq \delta$. As a result, we solve the following **discriminative constrained optimization** problem:

$$\begin{aligned} &\max_\theta \mathcal{J}_1(\theta) := \mathbb{E}_q \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q), o'\sim\pi^-_{\text{old}}(\cdot|q)} \ell(s_\theta(o,q) - s_\theta(o',q)) \\ &s.t. \quad \mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta) \leq \delta. \end{aligned} \tag{8}$$

For sake of efficiency, we use a different optimization approach from TRPO to solve the above constrained optimization. Inspired by the recent advances of non-convex ineqaulity constrained optimization algorithm [35], we adopt a squared-hinge penalty function for the constraint and solve the following problem with an appropriate penalty parameter β:

$$\max_\theta \mathbb{E}_q \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q), o'\sim\pi^-_{\text{old}}(\cdot|q)} \ell(s_\theta(o,q) - s_\theta(o',q)) - \beta[\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta) - \delta]^2_+, \tag{9}$$

where $[\cdot]_+ = \max\{\cdot, 0\}$. It has been shown that under an appropriate assumption regarding the constraint function and β, solving the above squared-hinge penalized objective (9) can return a KKT solution of the original constrained problem (8). We refer the readers to [35] for more in-depth analysis of this approach.

Finally, we would like to emphasize the difference between using the squared-hinge penalty function and the regular KL divergence regularization $\beta\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta)$. The squared-hinge penalty function has a dynamic weighting impact for the gradient, $\nabla\beta[\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta) - \delta]^2_+ = 2\beta[\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta) - \delta]_+\nabla\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta)$, such that if the constraint is satisfied then the weight $2\beta[\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta) - \delta]_+$ before the gradient of the regularization term $\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta)$ becomes zero. This means the KL divergence regularization is only effective when the constraint is violated. In contrast, the regular KL divergence regularization $\beta\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta)$ always contributes a gradient $\beta\nabla\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta)$ no matter whether the constraint is satisfied or not, which could harm the learning.

5.2 An improved approach for tackling imbalanced rollouts

One advantage of designing the objective based on the principle of discriminative learning is the ability to leverage a wide range of advanced techniques from the literature to improve training. A key challenge in RL fine-tuning for reasoning models is the sparse rewards, which lead to imbalance in generated rollouts. Specifically, for some questions where $p(q) \ll 1$, the number of negative outputs can significantly exceed the number of positive ones. This reflects a classic data imbalance issue, which has been extensively studied in the discriminative learning community [84, 58, 50]. To address this issue, we consider distributionally robust optimization (DRO) [84, 50].

Let us first discuss why the basic approach could be ineffective for combating the imbalanced rollouts. The objective function $\mathcal{J}_1$ is motivated by maximizing AUC for each question q, i.e., $\mathbb{E}_{o\sim\pi^+_{\text{old}}, o'\sim\pi^-_{\text{old}}}[\mathbb{I}(s_\theta(o,q) > s_\theta(o',q))]$. However, when there is much more negative data than positive data, AUC is not a good measure. For example, let us consider a scenario where there are 1 positive o_+ and 100 negatives $\{o^1_-, \ldots, o^{100}_-\}$. If the scores of these data are $s(o^1_-, q) = 0.9, s(o_+, q) = 0.5, s(o^2_-, q) = s(o^3_-, q) \ldots = s(o^{100}_-, q) = 0.001$, then the AUC score is $\frac{99}{100} = 0.99$. The AUC score is high but is not informative as the model still generates the negative data o^1_- more likely than the positive data o_+. In the literature, this issue has been addressed by maximizing a partial AUC score, which considers the pairwise order between all positives and the top ranked negatives. We utilize a surrogate function of partial AUC score formulated from the perspective of DRO [84].

Consider a question q and a positive data o. We denote by $\mathcal{Q}$ the set of probability measures Q on negative data given q (absolutely continuous with respect to $\pi^-_{\text{old}}(\cdot|q)$). Denote by $\mathbb{D}_{\text{KL}}(Q, \pi^-_{\text{old}}(\cdot|q))$ the KL divergence between a distribution Q and the negative data distribution $\pi^-_{\text{old}}(\cdot|q)$. A DRO

Table 1: Comparison of different methods for reinforcing large reasoning models. "L-ratio" means likelihood ratio, "log-L" means log-likelihood, "proper" means any proper scoring function.

Method	Difficulty Bias	Clipping	KL Divergence	Score Function	Tackles Imbalanced Rollouts
GRPO [23]	Yes	Yes	regularization, π_{ref}	clipped L-ratio	No
Dr. GRPO [40]	Yes	Yes	No	clipped L-ratio	No
DAPO [79]	Yes	Yes	No	clipped L-ratio	No
GPG [13]	Yes	No	No	log-L	No
TRPA [61]	No	No	regularization, π_{old}	log L-ratio	No
DisCO	No	No	constraint, π_{old}	proper	Yes

formulation for partial AUC maximization is given by [84][Theorem 2]:

$$\inf_{Q\in\mathcal{Q}} \tau\mathbb{D}_{\text{KL}}(Q, \pi^-_{\text{old}}(\cdot|q)) + \mathbb{E}_{o'\sim Q}[s_\theta(o,q) - s_\theta(o',q)] : \\ = -\tau \log\left(\mathbb{E}_{o'\sim\pi^-_{\text{old}}(\cdot|q)} \exp\left(\frac{s_\theta(o',q) - s_\theta(o,q)}{\tau}\right)\right).$$

As a result, we construct the following DRO-based objective for maximization:

$$\mathcal{J}_2(\theta) = -\mathbb{E}_q\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)}\tau\log\left(\mathbb{E}_{o'\sim\pi^-_{\text{old}}(\cdot|q)} \exp\left(\frac{s_\theta(o',q) - s_\theta(o,q)}{\tau}\right)\right). \tag{10}$$

It is easy to show that $\mathcal{J}_2(\theta) \leq \mathcal{J}_1(\theta)$ by Jensen's inequality for the convex function $-\log$. Hence, maximizing $\mathcal{J}_2(\theta)$ will automatically increasing $\mathcal{J}_1(\theta)$. However, the reverse is not true. This also explains why maximizing $\mathcal{J}_2(\theta)$ could be more effective than maximizing $\mathcal{J}_1(\theta)$. The above risk function is also known as optimized certainty equivalents (OCE) in mathematical finance [4]. We would like to point out that although OCE or DRO has been considered for RL in existing works [65, 73], they differ from our work for addressing different issues. Wang et al. [65] apply the OCE to compute a robust reward, replacing the standard expected reward used in traditional RL settings. Xu et al. [73] adopt DRO for direct preference optimization of LLMs, aiming to mitigate the noise in human preference data by addressing the distributional shift between the empirical distribution of (q, o, o') and its true underlying distribution.

Finally, we solve the following discriminative constrained optimization problem by using the same squared-hinge penalty method:

$$\max_\theta \mathcal{J}_2(\theta) := -\mathbb{E}_q\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)}\tau\log\left(\mathbb{E}_{o'\sim\pi^-_{\text{old}}(\cdot|q)} \exp\left(\frac{s_\theta(o',q) - s_\theta(o,q)}{\tau}\right)\right), \\ s.t. \quad \mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta) \leq \delta. \tag{11}$$

To differentiate the approach for solving (8) and (11), we refer to the former as **DisCO-b** and the latter as **DisCO**. In practice, all expectations will be replaced by empirical averages and the KL divergence is also estimated at each iteration by using sampled data following [52]. We present a full algorithm in Algorithm 1 in Appendix. Finally, we give a comparison between DisCO and existing RL fine-tuning methods for reinforcing LRMs from different aspects in Table 1.

6 Experiments

In this section, we empirically evaluate the effectiveness of the proposed DisCO by comparing with GRPO and other variants for reinforcing SFT-finetuned models.

Task Setting. We validate our method on mathematical reasoning tasks. Specifically, we use the DeepScaleR-Preview-Dataset [42] for training, which includes AIME problems from 1984 to 2023, AMC problems before 2023, and questions from the Omni-MATH [21] and Still [45] datasets, totaling approximately 40.3k unique problem-answer pairs. We evaluate models on six benchmark datasets: AIME 2024, AIME 2025, MATH 500 [28, 37], AMC 2023, Minerva [34], and Olympiad Bench (O-Bench) [26]. Following [23, 42], we adopt the pass@1 metric [11] averaged over $k = 16$ responses for each question to ensure the reliability of model performances. The metric for each question is calculated as $\frac{1}{k}\sum_{i=1}^k \mathbb{I}(o_i \text{ is correct})$, where o_i denotes the i-th generated response. For both the training and evaluation of our method and the baselines (unless otherwise specified), the maximum response length is limited to 8k tokens. To verify the generalizability of our method to other datasets, we also conducted experiments on DAPO-Math-17k [79] dataset, which is included in the Appendix A.3.

 https://doi.org/10.52202/085713-1921

Table 2: Comparison with baseline models and baseline methods for fine-tuning 1.5B models. OpenAI-o1-preview is included as a reference. MRL denotes Max Response Length utilized in training/testing. The shaded models are trained by other works and the shaded numbers are reported in their original works or in [42]. All other results are either evaluated on existing models or on the models trained by us using different approaches. Methods in the bottom area are all for fine-tuning DeepSeek-R1-Distill-Qwen-1.5B model on the same DeepScaleR dataset. DS is short for DeepSeek-R1, DSR is short for DeepScaleR.

Model/Method	MRL(Train/Test)	AIME 2024	AIME 2025	MATH 500	AMC 2023	Minerva	O-Bench	Avg.
OpenAI-o1-Preview	-	0.4	-	0.814	-	-	-	-
DS-Distill-Qwen-1.5B	32k+ / 32k	0.288	0.263	0.828	0.629	0.265	0.433	0.451
DS-Distill-Qwen-1.5B	32k+ / 8k	0.181	0.215	0.758	0.515	0.237	0.353	0.376
STILL-3-1.5B-preview	29k / 32k	0.325	0.248	0.844	0.667	0.290	0.454	0.471
DSR-1.5B-Preview	24k / 32k	**0.431**	0.304	**0.878**	0.736	0.302	0.500	0.525
DSR-1.5B-Preview	24k / 8k	0.358	0.258	0.860	0.679	0.297	0.473	0.488
GRPO	8k / 8k	0.277	0.242	0.838	0.647	0.276	0.462	0.457
GRPO-ER	8k / 8k	0.298	0.242	0.839	0.649	0.279	0.452	0.460
Dr. GRPO	8k / 8k	0.252	0.238	0.831	0.631	0.268	0.440	0.443
DAPO	8k / 8k	0.310	0.252	0.848	0.675	0.296	0.456	0.473
TRPA	8k / 8k	0.354	0.235	0.835	0.653	0.283	0.458	0.470
DisCO (L-ratio)	8k / 8k	0.381	0.306	**0.878**	0.746	0.319	**0.512**	0.524
DisCO (log-L)	8k / 8k	0.404	**0.317**	0.876	**0.758**	**0.333**	0.509	**0.533**

Table 3: Comparison with baseline models and baseline methods for fine-tuning 7B models. Methods in the bottom area are all for fine-tuning DeepSeek-R1-Distill-Qwen-7B model on the same DeepScaleR dataset.

Model/Method	MRL(Train/Test)	AIME 2024	AIME 2025	MATH 500	AMC 2023	Minerva	O-Bench	Avg.
DS-Distill-Qwen-7B	32k+ / 32k	0.560	0.396	0.923	0.825	0.380	0.568	0.609
DS-Distill-Qwen-7B	32k+ / 8k	0.402	0.292	0.873	0.688	0.355	0.471	0.513
GRPO-LEAD-7B	8k / 8k	0.470	0.345	0.893	0.748	0.372	0.500	0.555
TRPA	8k / 8k	0 .570	-	0.870	0.780	0.360	0.550	-
GRPO	8k / 8k	0.498	0.394	0.916	0.807	0.381	0.555	0.592
GRPO-ER	8k / 8k	0.515	0.381	0.916	0.825	0.376	0.544	0.593
Dr. GRPO	8k / 8k	0.488	0.346	0.910	0.792	0.368	0.546	0.575
DAPO	8k / 8k	0.454	0.335	0.907	0.799	0.388	0.535	0.570
TRPA	8k / 8k	0.510	0.367	0.898	0.779	0.379	0.534	0.578
DisCO (L-ratio)	8k / 8k	**0.583**	**0.421**	0.923	0.852	0.399	0.585	**0.627**
DisCO (log-L)	8k / 8k	0.558	0.410	**0.927**	**0.854**	**0.410**	**0.592**	0.625

Table 4: Comparison with baseline models and baseline methods for fine-tuning 8B models. Methods in the bottom area are all for fine-tuning DeepSeek-R1-Distill-Llama-8B model on the same DeepScaleR dataset.

Model/Method	MRL(Train/Test)	AIME 2024	AIME 2025	MATH 500	AMC 2023	Minerva	O-Bench	Avg.
DS-Distill-Llama-8B	32k+ / 32k	0.506	0.346	0.896	0.815	0.295	0.541	0.566
DS-Distill-Llama-8B	32k+ / 8k	0.348	0.238	0.825	0.652	0.267	0.440	0.462
GRPO	8k / 8k	0.410	0.240	0.873	0.759	0.307	0.506	0.516
GRPO+ER	8k / 8k	0.408	0.277	0.882	0.785	0.311	0.511	0.529
Dr. GRPO	8k / 8k	0.423	0.285	0.867	0.786	0.300	0.497	0.526
DAPO	8k / 8k	0.333	0.308	0.879	0.794	0.325	0.522	0.527
TRPA	8k / 8k	0.454	0.279	0.864	0.756	0.289	0.518	0.527
DisCO (L-ratio)	8k / 8k	0.506	**0.356**	**0.900**	0.831	0.326	0.553	0.579
DisCO (log-L)	8k / 8k	**0.523**	0.354	0.896	**0.843**	**0.331**	**0.560**	**0.584**

Models. We conduct experiments with fine-tuning three models: DeepSeek-R1-Distill-Qwen-1.5B model (Q1.5B), DeepSeek-R1-Distill-Qwen-7B model (Q7B), and DeepSeek-R1-Distill-Llama-8B (L8B). All are distilled reasoning models.

Baselines. We primarily compare our methods with five most recent state-of-the-art reinforcement learning methods, including (1) GRPO [23]; (2) GRPO with an entropy regularization (GRPO-ER) that adds an entropy on probabilities of output tokens as a regularization to prevent entropy collapse, which is used by DeepScaleR [42]; (3) Dr. GRPO [40]; (4) DAPO's objective [79]; (5) TRPA [61]. For a comprehensive evaluation, we also include a set of reasoning models that are trained from the same base model by other studies with various techniques, such as (6) STILL-3-1.5B-preview [12], which adapt GRPO by periodically replacing the reference model after a fixed number of training steps; (7) DeepScaleR(DSR)-1.5B-Preview that uses maximum response length of 24k for training [42]; (8) GRPO-LEAD-7B [82], which extends GRPO by incorporating length-dependent

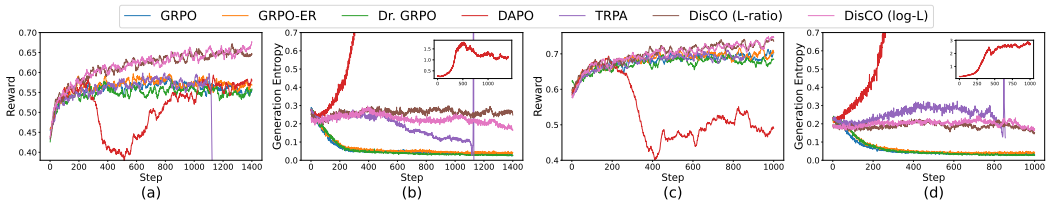


Figure 2: Training dynamics of different methods: left two are for fine-tuning DeepSeek-R1-Distill-Qwen-1.5B model and right two are for fine-tuning DeepSeek-R1-Distill-Qwen-7B model. (a), (c) plot the training reward (averaged over generated outputs for questions used in each step) vs the number of training steps (cf. Algorithm 1); (b), (d) plot the generation entropy vs training steps.

rewards, explicit penalty terms, and difficulty-based advantage reweighting to encourage concise and precise reasoning.

Training Details. For all the methods, we tune the constant learning rate in $[5\text{e}^{-7}, 1\text{e}^{-6}, 2\text{e}^{-6}]$ with AdamW optimizer with weight decay as 0.01. Generally, a learning rate of 2e^{-6} works better for the Q1.5B model, 1e^{-6} for the Q7B model, and 5e^{-7} for the L8B model. We employ a training batch size of 128, a mini-batch size of 32, and 8 responses for each question. The temperature is set to 0.6 for both training and evaluation, following the usage recommendation from [23]. For GRPO, β is set to 0.001 as commonly used [12, 42]. For GRPO-ER, we use a coefficient of 0.001 for the entropy regularization [42]. For DAPO, we set ϵ_{low} to 0.2 and ϵ_{high} to 0.28 by following their paper. For our method, δ is set to 10^{-4} based on the empirical observation that the average KL divergence is around $2 * 10^{-5}$ and β is set to 10^3 such that the effective weight of the KL regularization when the constraint is violated by δ is on the order of $\beta * \delta = 0.1$. Since L-ratio and log-L scoring functions have different orders, we choose $\tau = 1$ for L-ratio and $\tau = 10$ for log-L scoring function, from $\{0.5, 1, 5, 10\}$. For fair comparisons, we do not implement Dynamic Sampling [79] for DAPO and other methods, as it introduces approximately three times the sampling cost at each training step. All methods are run for 1,400 steps on Q1.5B models and 1,000 steps on Q7B/L8B models. Evaluations are conducted every 200 steps, and the best performance for each method is reported.

6.1 Comparison with Baselines

Performance. We evaluate all the models across six mathematics-focused benchmark datasets to demonstrate the effectiveness of DisCO. The results are summarized in Table 2, 3 and 4. From Table 2 for Q1.5B models, we can observe that our proposed DisCO methods consistently outperform other baselines by a large margin. Notably, DisCO (log-L) with 8k length for both training and inference achieves an 7% average improvement over GRPO and surpasses DeepScaleR-1.5B-Preview that was trained with maximum 24k length and evaluated with 32k length. A similar trend is observed for Q7B models and L8B models (Table 3 and Table 4), where DisCO significantly outperforms all competing approaches.

Training Dynamics. We compare the training dynamics of different methods in terms of training rewards and generation entropy. From Figure 2 for fine-tuning Qwen-1.5B and Qwen-7B models, we can see that all baselines suffer from premature saturation due to either entropy collapse for GRPO, GRPO-ER, Dr. GRPO or excessive entropy growth of DAPO, which leads to an early deterministic or highly random policy. The entropy collapse phenomenon is also observed by [79, 27, 42]. TRPA that uses a KL divergence regularization is also observed with instability in the generation entropy in later steps (around 1100 for Q1.5B model and around 800 for Q7B model). In contrast, our methods with the two scoring functions are most stable, with training rewards kept increasing and generation entropy maintained around 0.22. We also include training dynamics for the L8B model in Appendix A.2, which follows a similar trend.

6.2 Ablation Studies

DisCO vs DisCO-b. Figure 3 (left) compares DisCO with DisCO-b using the L-ratio scoring function for training Q7B models for 1000 steps. The comparison clearly demonstrates the significant improvements of DisCO over DisCO-b, especially on the difficult AIME datasets. We also compare DisCO with DisCO-b for other settings in Appendix A.1, and observe that DisCO is consistently better than DisCO-b on average in all settings. It is also notable that DisCO-b with different scoring functions are also better than other baselines trained or evaluated by us in Table 2 and Table 3.

https://doi.org/10.52202/085713-1921

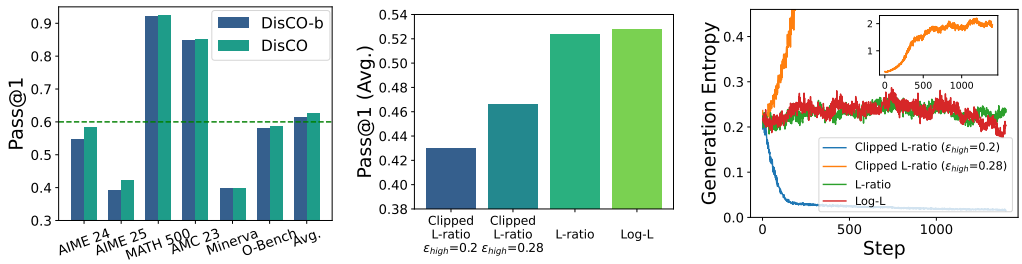


Figure 3: Ablation studies: left for comparing DisCO vs DisCO-b; middle and right for comparing clipping with non-clipping scoring functions.

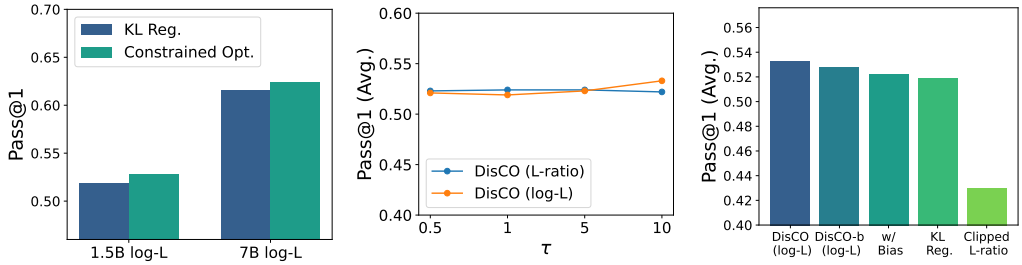


Figure 4: Ablation studies: left for comparing KL regularization vs constrained optimization; middle for sensitivity of DisCO w.r.t. the hyperparameter τ; right for contribution of each component.

Clipping vs Non-Clipping scoring functions. We compare non-clipping scoring functions L-ratio, log-L with clipped L-ratio (5) in our DisCO-b approach for training Q1.5B models in Figure 3 (middle and right). For the clipped L-ratio, we adopt two versions: one with $\epsilon_{high} = 0.2$ to align with GRPO objective, and another with $\epsilon_{high} = 0.28$ similar to DAPO objective. We can see that clipped L-ratio with $\epsilon_{high} = 0.2$ causes entropy collapse while clipped L-ratio with $\epsilon_{high} = 0.28$ leads to excessively high entropy level, both yielding worse performance than non-clipping scoring functions.

KL Regularization vs Constrained Optimization. We investigate the advantages of constrained optimization over KL regularization for DisCO. Specifically, the KL regularization weight is set to the commonly used 0.001 [12, 42]. As shown in Figure 4 (left), constrained optimization performs better than KL regularization on both Q1.5B and Q7B models. Moreover, during our experiments, we observed that KL regularization leads to instability in training on Q7B models, similar to TRPA, which indicates that KL regularization is not sufficient to stabilize training.

Sensitivity of hyperparameter τ. We study the sensitivity of DisCO to hyperparameter τ on training Q1.5B models. Similar to above experiments, we run DisCO for 1400 steps with different $\tau \in \{0.5, 1, 5, 10\}$. The result shown in Figure 3 (right) indicates that DisCO is not sensitive to τ in these ranges.

Effect of each design choice. We analyze the individual contribution of each component in DisCO by replacing its components separately with other designs. We experiment on Q1.5B models and compare with (1) DisCO-b that removes hard negative weighting; (2) adding question-level weight bias $\sqrt{p(q)(1-p(q)}$ to DisCO-b, (3) replacing the KL-divergence constraint with a KL-divergence regularization in DisCO-b, and (4) using a clipping scoring function with $\epsilon_{high} = 0.2$ in DisCO-b, respectively. From Figure 4 (right), we can see that each of our proposed components is important in DisCO's improvement, where the use of a non-clipping scoring function is of vital importance.

7 Conclusion

In this work, we have proposed a novel discriminative constrained optimization framework for reinforcing large reasoning models, motivated by the analysis of the GRPO objective. The proposed framework is grounded in the principle of discriminative learning, avoiding difficulty bias and enhancing training stability with constrained trust region optimization. The experiments on mathematical reasoning demonstrated the significant superiority of our approaches, compared with GRPO and its recent variants. While this work focuses on binary rewards, future extensions could incorporate discriminative ranking objectives, like [9], to handle non-binary rewards. It would be interesting to apply the proposed approaches for fine-tuning larger models or other reasoning tasks.

https://doi.org/10.52202/085713-1921

Acknowledgements

We are grateful for the reviewers' constructive comments. G. Li and T. Yang were partially supported by NSF grant #2306572.

References

[1] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

[2] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[3] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[4] Aharon Ben-Tal and Marc Teboulle. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17:449–476, 02 2007.

[5] Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, Gerald Shen, Ameya Sunil Mahabaleshwarkar, Bilal Kartal, Yoshi Suhara, Olivier Delalleau, Zijia Chen, Zhilin Wang, David Mosallanezhad, Adi Renduchintala, Haifeng Qian, Dima Rekesh, Fei Jia, Somshubra Majumdar, Vahid Noroozi, Wasi Uddin Ahmad, Sean Narenthiran, Aleksander Ficek, Mehrzad Samadi, Jocelyn Huang, Siddhartha Jain, Igor Gitman, Ivan Moshkov, Wei Du, Shubham Toshniwal, George Armstrong, Branislav Kisacanin, Matvei Novikov, Daria Gitman, Evelina Bakhturina, Jane Polak Scowcroft, John Kamalu, Dan Su, Kezhi Kong, Markus Kliegl, Rabeeh Karimi, Ying Lin, Sanjeev Satheesh, Jupinder Parmar, Pritam Gundecha, Brandon Norick, Joseph Jennings, Shrimai Prabhumoye, Syeda Nahida Akter, Mostofa Patwary, Abhinav Khattar, Deepak Narayanan, Roger Waleffe, Jimmy Zhang, Bor-Yiing Su, Guyue Huang, Terry Kong, Parth Chadha, Sahil Jain, Christine Harvey, Elad Segal, Jining Huang, Sergey Kashirsky, Robert McQueen, Izzy Putterman, George Lam, Arun Venkatesan, Sherry Wu, Vinh Nguyen, Manoj Kilaru, Andrew Wang, Anna Warno, Abhilash Somasamudramath, Sandip Bhaskar, Maka Dong, Nave Assaf, Shahar Mor, Omer Ullman Argov, Scot Junkin, Oleksandr Romanenko, Pedro Larroy, Monika Katariya, Marco Rovinelli, Viji Balas, Nicholas Edelman, Anahita Bhiwandiwalla, Muthu Subramaniam, Smita Ithape, Karthik Ramamoorthy, Yuting Wu, Suguna Varshini Velury, Omri Almog, Joyjit Daw, Denys Fridman, Erick Galinkin, Michael Evans, Shaona Ghosh, Katherine Luna, Leon Derczynski, Nikki Pope, Eileen Long, Seth Schneider, Guillermo Siman, Tomasz Grzegorzek, Pablo Ribalta, Monika Katariya, Chris Alexiuk, Joey Conway, Trisha Saar, Ann Guan, Krzysztof Pawelec, Shyamala Prayaga, Oleksii Kuchaiev, Boris Ginsburg, Oluwatobi Olabiyi, Kari Briski, Jonathan Cohen, Bryan Catanzaro, Jonah Alben, Yonatan Geifman, and Eric Chung. Llama-nemotron: Efficient reasoning models, 2025.

[6] C.M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

[7] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, 2005.

[8] Yuji Cao, Huan Zhao, Yuheng Cheng, Ting Shu, Yue Chen, Guolong Liu, Gaoqi Liang, Junhua Zhao, Jinyue Yan, and Yun Li. Survey on large language model-enhanced reinforcement learning: Concept, taxonomy, and methods. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[9] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: From pairwise approach to listwise approach. volume 227, pages 129–136, 06 2007.

 https://doi.org/10.52202/085713-1921

[10] Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Step-level value preference optimization for mathematical reasoning. *arXiv preprint arXiv:2406.10858*, 2024.

[11] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[12] Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, et al. An empirical study on eliciting and improving r1-like reasoning models. *arXiv preprint arXiv:2503.04548*, 2025.

[13] Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025.

[14] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[15] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, March 2002.

[16] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[17] Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I Jordan, Joseph E Gonzalez, and Sergey Levine. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*, 2018.

[18] Xidong Feng, Ziyu Wan, Muning Wen, Stephen Marcus McAleer, Ying Wen, Weinan Zhang, and Jun Wang. Alphazero-like tree-search can guide large language model decoding and training. *arXiv preprint arXiv:2309.17179*, 2023.

[19] Yoav Freund, Raj Iyer, Robert E Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *Journal of machine learning research*, 4(Nov):933–969, 2003.

[20] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[21] Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024.

[22] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[23] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[24] Siqi Guo, Ilgee Hong, Vicente Balmaseda, Tuo Zhao, and Tianbao Yang. Discriminative finetuning of generative large language models without reward models and preference data. *arXiv preprint arXiv:2502.18679*, 2025.

[25] Siqi Guo, Ilgee Hong, Vicente Balmaseda, Tuo Zhao, and Tianbao Yang. Discriminative finetuning of generative large language models without reward models and preference data, 2025.

[26] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

[27] Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner series. https://capricious-hydrogen-41c.notion.site/Skywork-Open-Reaonser-Series-1d0bc9ae823a80459b46c149e4f51680, 2025. Notion Blog.

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

[29] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

[30] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

[31] HuggingFace. Open r1: A fully open reproduction of deepseek-r1. https://huggingface.co/blog/open-r1, 2025. Blog.

[32] Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Unlocking rl potential for llm reasoning through refined credit assignment. *arXiv preprint arXiv:2410.01679*, 2024.

[33] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025.

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

[35] Gang Li, Wendi Yu, Yao Yao, Wei Tong, Yingbin Liang, Qihang Lin, and Tianbao Yang. Model developmental safety: A safety-centric method and applications in vision-language models. *arXiv preprint arXiv:2410.03955*, 2024.

[36] Ziniu Li, Tian Xu, Yushun Zhang, Zhihang Lin, Yang Yu, Ruoyu Sun, and Zhi-Quan Luo. Remax: A simple, effective, and efficient reinforcement learning method for aligning large language models. *arXiv preprint arXiv:2310.10505*, 2023.

[37] 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 *The Twelfth International Conference on Learning Representations*, 2023.

[38] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[39] Zhihang Lin, Mingbao Lin, Yuan Xie, and Rongrong Ji. Cppo: Accelerating the training of group relative policy optimization-based reasoning models. *arXiv preprint arXiv:2503.22342*, 2025.

[40] 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. *arXiv preprint arXiv:2503.20783*, 2025.

[41] Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-mini-Level-1cf81902c14680b3bee5eb349a512a51, 2025. Notion Blog.

 https://doi.org/10.52202/085713-1921

[42] Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c1468005bed8ca303013a4e2, 2025. Notion Blog.

[43] Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*, 3, 2024.

[44] Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

[45] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, et al. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*, 2024.

[46] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016.

[47] Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification, 2025.

[48] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[49] Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 7559–7566. IEEE, 2018.

[50] Hongseok Namkoong and John C. Duchi. Variance-based regularization with convex objectives. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 2975–2984, Red Hook, NY, USA, 2017. Curran Associates Inc.

[51] OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024. Blog.

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

[53] Sébastien Racanière, Théophane Weber, David Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adrià Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al. Imagination-augmented agents for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[54] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[55] Z. Z. Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, Z. F. Wu, Zhibin Gou, Shirong Ma, Hongxuan Tang, Yuxuan Liu, Wenjun Gao, Daya Guo, and Chong Ruan. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition, 2025.

[56] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

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

[58] Shai Shalev-Shwartz and Yonatan Wexler. Minimizing the maximal loss: How and why? *CoRR*, abs/1602.01690, 2016.

[59] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[60] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*, 2017.

[61] Xuerui Su, Shufang Xie, Guoqing Liu, Yingce Xia, Renqian Luo, Peiran Jin, Zhiming Ma, Yue Wang, Zun Wang, and Yuting Liu. Trust region preference approximation: A simple and stable reinforcement learning algorithm for llm reasoning. *arXiv preprint arXiv:2504.04524*, 2025.

[62] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

[63] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[64] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[65] Kaiwen Wang, Dawen Liang, Nathan Kallus, and Wen Sun. A reductions approach to risk-sensitive reinforcement learning with optimized certainty equivalents, 2025.

[66] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[67] Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, et al. Light-r1: Curriculum sft, dpo and rl for long cot from scratch and beyond. *arXiv preprint arXiv:2503.10460*, 2025.

[68] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[69] Xiaomi LLM-Core Team. Mimo: Unlocking the reasoning potential of language model – from pretraining to posttraining, 2025.

[70] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

[71] Huajian Xin, ZZ Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, et al. Deepseek-prover-v1. 5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. *arXiv preprint arXiv:2408.08152*, 2024.

[72] Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In *International Conference on Machine Learning*, pages 55204–55224. PMLR, 2024.

[73] Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. Distributionally robust direct preference optimization, 2025.

 https://doi.org/10.52202/085713-1921

[74] Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and Ping Luo. Dancegrpo: Unleashing grpo on visual generation, 2025.

[75] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[76] Tianbao Yang and Yiming Ying. Auc maximization in the era of big data and ai: A survey, 2022.

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

[78] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[79] 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 preprint arXiv:2503.14476*, 2025.

[80] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. 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, 2021.

[81] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[82] Jixiao Zhang and Chunsheng Zuo. Grpo-lead: A difficulty-aware reinforcement learning approach for concise mathematical reasoning in language models. *arXiv preprint arXiv:2504.09696*, 2025.

[83] Xiaojiang Zhang, Jinghui Wang, Zifei Cheng, Wenhao Zhuang, Zheng Lin, Minglei Zhang, Shaojie Wang, Yinghan Cui, Chao Wang, Junyi Peng, et al. Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. *arXiv preprint arXiv:2504.14286*, 2025.

[84] Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, and Tianbao Yang. When AUC meets DRO: Optimizing partial AUC for deep learning with non-convex convergence guarantee. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27548–27573. PMLR, 17–23 Jul 2022.

[85] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A More Experimental Results

For all the experiments on 1.5B models, each run consumes 4*2 40G A100 GPUs and each training step takes approximately 6 minutes. For all the experiments on 7B models, each run consumes 1*8 80G H100 GPUs and each training step takes approximately 6.5 minutes.

A.1 Detailed comparison between DisCO and DisCO-b

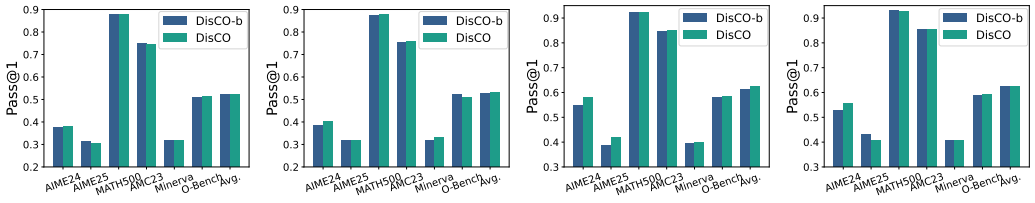
In this part, we compare DisCO and DisCO-b with different score functions on different models. As shown in Figure 5, DisCO consistently demonstrates better performance compared to DisCO-b across all settings, with higher average scores observed in each case. This consistent advantage highlights the effectiveness of the full DisCO framework. Additionally, it is worth emphasizing that even the DisCO-b variants, with L-ratio or log-L scoring functions, outperform all other baseline methods that are presented in Table 2 and Table 3. These results collectively underscore the robustness and general effectiveness of the DisCO approach.

A.2 Training dynamics for fine-tuning 8B model.

In this part, we present the training dynamics of different methods for fine-tuning the DeepSeek-R1-Distill-Llama-8B model in Figure 6. Similar to observation in Figure 2 for fine-tuning 1.5B and 7B models, we can see that GRPO, GRPO-ER, and Dr. GRPO still suffer from entropy collapse while DAPO leads to excessive entropy growth, all accompanied by premature saturation in training reward. TRPA with a KL divergence regularization is also observed with instability in the training, indicating the insufficiency of KL regularization to stabilize training. In contrast, our methods with the two scoring functions and the KL constraint demonstrate the greatest stability, with training rewards continuing to rise and generation entropy remaining around 0.2.

A.3 Experiments on DAPO-Math-17K dataset.

In order to demonstrate that the improvements achieved by DisCO are fundamental, rather than relying on specific properties of the dataset, we conducted additional experiments on the DAPO-Math-17K dataset [79] using 1.5B models, training them for 1400 steps. As shown in Table 5, DisCO



(a) L-ratio on 1.5B (b) log-L on 1.5B (c) L-ratio on 7B (d) log-L on 7B

Figure 5: Comparison between DisCO-b and DisCO on different models with different score functions.

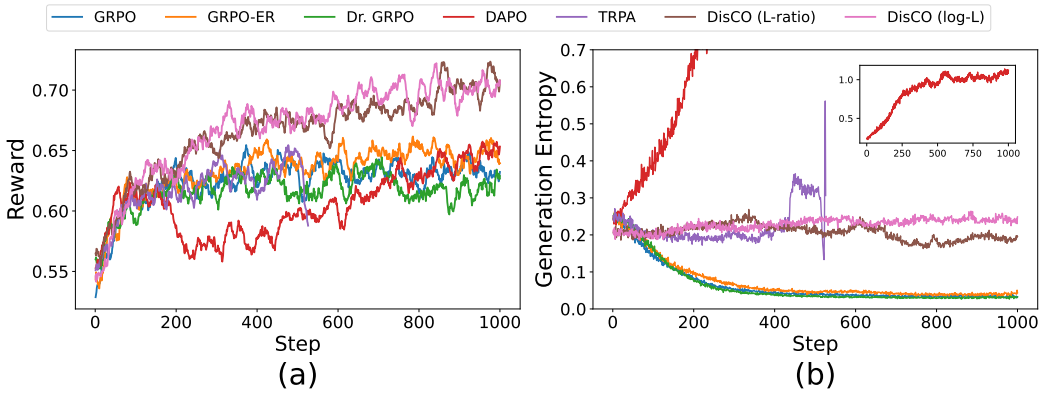


Figure 6: Training dynamics of different methods for fine-tuning DeepSeek-R1-Distill-Llama-8B model. (a) plots the training reward (averaged over generated outputs for questions used in each step) vs the number of training steps; (b) plots the generation entropy vs training steps.

https://doi.org/10.52202/085713-1921

Table 5: Comparison with baseline methods for fine-tuning DeepSeek-R1-Distill-Qwen-1.5B models on DAPO-Math-17K dataset.

Model/Method	MRL(Train/Test)	AIME 2024	AIME 2025	MATH 500	AMC 2023	Minerva	O-Bench	Avg.
DS-Distill-Qwen-1.5B	32k+ / 32k	0.288	0.263	0.828	0.629	0.265	0.433	0.451
DS-Distill-Qwen-1.5B	32k+ / 8k	0.181	0.215	0.758	0.515	0.237	0.353	0.376
GRPO	8k / 8k	0.342	0.256	0.842	0.672	0.267	0.458	0.473
GRPO-ER	8k / 8k	0.290	0.260	0.852	0.681	0.287	0.463	0.472
Dr. GRPO	8k / 8k	0.300	0.250	0.849	0.705	0.292	0.464	0.477
DAPO	8k / 8k	0.275	0.229	0.812	0.653	0.256	0.441	0.444
TRPA	8k / 8k	0.346	0.279	0.836	0.683	0.281	0.450	0.479
DisCO (L-ratio)	8k / 8k	0.413	0.310	**0.874**	**0.775**	0.307	0.495	0.529
DisCO (log-L)	8k / 8k	**0.460**	**0.317**	0.873	**0.775**	**0.320**	**0.502**	**0.541**

methods still outperform other baselines by a large margin, demonstrating the generalizability of the proposed method to other datasets.

B More Theoretical Results

B.1 Proof of Proposition 1

Proof *Since* $\mathbb{E}_{o\sim\pi_{old}(\cdot|q)} r(o|q) = p(q)$, $Var_{o\sim\pi_{old}(\cdot|q)} r(o|q) = p(q)(1-p(q))$, *we have*

$$A(o|q) = \begin{cases} \sqrt{\frac{1-p(q)}{p(q)}}, & \text{if } r(o|q) = 1, \\ -\sqrt{\frac{p(q)}{1-p(q)}}, & \text{if } r(o|q) = 0 \end{cases} \tag{12}$$

According to (1), we have

$$\begin{aligned}
&\mathbb{E}_q \mathbb{E}_{o\sim\pi_{old}(\cdot|q)} \left[\frac{1}{|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, A(o|q) \right) \right] \\
&= \mathbb{E}_q \Bigg[p(q) \mathbb{E}_{o\sim\pi^+_{old}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, A(o|q) \right) \\
&\qquad + (1-p(q)) \mathbb{E}_{o\sim\pi^-_{old}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, A(o|q) \right) \Bigg] \\
&= \mathbb{E}_q \Bigg[p(q) \mathbb{E}_{o\sim\pi^+_{old}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, \sqrt{\frac{1-p(q)}{p(q)}} \right) \\
&\qquad + (1-p(q)) \mathbb{E}_{o\sim\pi^-_{old}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, -\sqrt{\frac{p(q)}{1-p(q)}} \right) \Bigg] \\
&= \mathbb{E}_q \sqrt{p(q)(1-p(q))} \Bigg[\mathbb{E}_{o\sim\pi^+_{old}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} f^+\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, 1 \right) \\
&\qquad - \mathbb{E}_{o\sim\pi^-_{old}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} f^-\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{old}(o_t|q,o_{<t})}, 1 \right) \Bigg]
\end{aligned} \tag{13}$$

where the last equality is due to the assumption about $f(x,y)$. *For GPRO, we have* $f^+(x,1) = \min(x, \textit{clip}(x, 1-\epsilon, 1+\epsilon)) = \min(x, 1+\epsilon)$ *and* $f^-(x,1) = \max(x, \textit{clip}(x, 1-\epsilon, 1+\epsilon)) = \max(x, 1-\epsilon)$.

Algorithm 1 Discriminative Constrained Optimization

Input: Initial policy model π_0, reward function r, question set $\mathcal{D}$, hyperparameter δ, β, τ.
Policy model $\pi_\theta = \pi_0$
for Step $= 1, \cdots, T$ **do**
  Sample a batch of questions $\mathcal{B}$ from $\mathcal{D}$
  Update the old policy model $\pi_{\text{old}} = \pi_\theta$
  For each question $q \in \mathcal{B}$, sample n responses $\{o_i\}_{i=1}^n \sim \pi_{\text{old}}(\cdot|q)$ denoted by S_q and partition it into S_q^+ and S_q^- based on rewards $r(o_i|q) \in \{0, 1\}$
  for minibatch $\mathcal{B}_m \in \mathcal{B}$ **do**
    Compute KL divergence estimator by

$$\hat{\mathbb{D}}_{KL} = \frac{1}{\sum_{q\in\mathcal{B}_m}\sum_{o\in S_q}|o|} \sum_{q\in\mathcal{B}_m}\sum_{o\in S_q}\sum_{t=1}^{|o|} \log \frac{\pi_{\text{old}}(o_t|q,o_{<t})}{\pi_\theta(o_t|q,o_{<t})}$$

    Compute gradient estimator of $\mathcal{J}_2(\theta)$ by

$$G_1 = \frac{1}{|\mathcal{B}_m|}\sum_{q\in\mathcal{B}_m}\frac{1}{|S_q^+|}\sum_{o\in S_q^+}\left(\nabla s_\theta(o,q) - \nabla\Big(\tau\log\sum_{o'\in S_q^-}\exp(\frac{s_\theta(o',q)}{\tau})\Big)\right)$$

    Compute gradient estimator of constraint by $G_2 = 2\beta[\hat{\mathbb{D}}_{KL} - \delta]_+\nabla\hat{\mathbb{D}}_{KL}$
    Update π_θ with Adam-W using the gradient estimator $G = G_1 + G_2$
  end for
end for

B.2 Connection between discriminative objectives and surrogate objectives in RL

The score function L-ratio is inspired by the same principle as the surrogate objective in TRPO [56]. TRPO aims to maximize the following objective subject to a constraint:

$$\begin{aligned}\max_\theta \mathbb{E}_q\mathbb{E}_{o\sim\pi_{\theta_{old}}(\cdot|q)}&\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}A(o_t)\\ s.t.\quad &\mathbb{D}_{\text{KL}}(\pi_{\text{old}}||\pi_\theta)\le\delta.\end{aligned} \tag{14}$$

When we apply the advantage function (12) to the objective, we have

$$\begin{aligned}&\mathbb{E}_q\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}\left[\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}A(o|q)\right]\\
&=\mathbb{E}_q\Bigg[p(q)\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)}\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}A(o|q)\\
&\qquad+(1-p(q))\mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)}\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}A(o|q)\Bigg]\\
&=\mathbb{E}_q\Bigg[p(q)\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)}\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}*\sqrt{\frac{1-p(q)}{p(q)}}\\
&\qquad+(1-p(q))\mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)}\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}*(-\sqrt{\frac{p(q)}{1-p(q)}})\Bigg]\\
&=\mathbb{E}_q\sqrt{p(q)(1-p(q))}\left[\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)}\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}-\mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)}\frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}\right]\end{aligned} \tag{15}$$

This gives the exact scoring function L-ratio: $s_\theta(o,q) = \frac{1}{|o|}\sum_{t=1}^{|o|}\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}$. After removing the improper weight $\sqrt{p(q)(1-p(q))}$, Eqn. (15) is same as Eqn. (7) with $\ell(s) = s$.

In the reinforcement learning literature, vanilla policy gradient methods also gain significant attention due to their simplicity and remarkable performance. The vanilla policy gradient(VPG) methods work

 https://doi.org/10.52202/085713-1921

Table 6: Weighted discriminative objectives and their scoring functions $s^+(o,q)$ and $s^-(o,q)$ for different methods, where $\sigma(\cdot)$ is the sigmoid function.

Objective	$\mathbb{E}_q \omega(q) \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot\|q), o'\sim\pi^-_{\text{old}}(\cdot\|q)} \ell\big(s^+_\theta(o,q) - s^-_\theta(o',q)\big)$
GRPO	$\omega(q) = \sqrt{p(q)(1-p(q))}, \quad \ell(s) = s$ $s^+_\theta(o,q) = \frac{1}{\|o\|}\sum_{t=1}^{\|o\|} \min(\frac{\pi_\theta(o_t\|q,o_{<t})}{\pi_{\text{old}}(o_t\|q,o_{<t})}, 1+\epsilon), \quad s^-_\theta(o',q) = \frac{1}{\|o'\|}\sum_{t=1}^{\|o\|} \max(\frac{\pi_\theta(o'_t\|q,o'_{<t})}{\pi_{\text{old}}(o'_t\|q,o'_{<t})}, 1-\epsilon)$
Dr. GRPO	$\omega(q) = p(q)(1-p(q), \quad \ell(s) = s$ $s^+_\theta(o,q) = \sum_{t=1}^{\|o\|} \min(\frac{\pi_\theta(o_t\|q,o_{<t})}{\pi_{\text{old}}(o_t\|q,o_{<t})}, 1+\epsilon), \quad s^-_\theta(o',q) = \sum_{t=1}^{\|o'\|} \max(\frac{\pi_\theta(o'_t\|q,o'_{<t})}{\pi_{\text{old}}(o'_t\|q,o'_{<t})}, 1-\epsilon)$
DAPO	$\omega(q) = \sqrt{p(q)(1-p(q))}, \quad \ell(s) = s$ $s^+_\theta(o,q) = \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot\|q)}\|o\|}\sum_{t=1}^{\|o\|} \min(\frac{\pi_\theta(o_t\|q,o_{<t})}{\pi_{\text{old}}(o_t\|q,o_{<t})}, 1+\epsilon_{high}), \quad s^-_\theta(o',q) = \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot\|q)}\|o\|}\sum_{t=1}^{\|o'\|} \max(\frac{\pi_\theta(o'_t\|q,o'_{<t})}{\pi_{\text{old}}(o'_t\|q,o'_{<t})}, (1-\epsilon_{low}))$
GPG	$\omega(q) = \alpha p(q)(1-p(q), \quad \ell(s) = s$ $s^+_\theta(o,q) = \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot\|q)}\|o\|}\sum_{t=1}^{\|o\|} \log \pi_\theta(o_t\|q,o_{<t}), \quad s^-_\theta(o',q) = \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot\|q)}\|o\|}\sum_{t=1}^{\|o'\|} \log \pi_\theta(o'_t\|q,o'_{<t})$
TRPA	$\omega(q) = 1, \quad \ell(s) = \log(\sigma(\beta(o)s))$ $s^+_\theta(o,q) = \sum_{t=1}^{\|o\|} \log \frac{\pi_\theta(o_t\|q,o_{<t})}{\pi_{ref}(o_t\|q,o_{<t})}, \quad s^-_\theta(o',q) = \sum_{t=1}^{\|o'\|} \log \frac{\pi_\theta(o'_t\|q,o'_{<t})}{\pi_{ref}(o'_t\|q,o'_{<t})}$

by computing an estimator of the policy gradient and plugging it into a stochastic gradient algorithm. The most commonly used surrogate objective function for gradient estimator has the form:

$$\mathcal{J}_{\text{VPG}} = \mathbb{E}_q \mathbb{E}_{o\sim\pi_{\theta_{old}}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} \log \pi_\theta(o_t|q,o_{<t}) A(o_t) \tag{16}$$

Similar to the derivation above, by plugging in the advantage estimator Eqn. (12), we have:

$$\begin{aligned} \mathcal{J}_{\text{VPG}} = \mathbb{E}_q \sqrt{p(q)(1-p(q))} \Bigg[& \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} \log \pi_\theta(o_t|q,o_{<t}) \\ & - \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} \frac{1}{|o|} \sum_{t=1}^{|o|} \log \pi_\theta(o_t|q,o_{<t}) \Bigg] \end{aligned} \tag{17}$$

This directly motivate the score function log-L: $s_\theta(o,q) = \frac{1}{|o|}\sum_{t=1}^{|o|} \log \pi_\theta(o_t|q,o_{<t})$. After removing the inappropriate weight on questions, Eqn. (17) is same as Eqn. (7) with $\ell(s) = s$.

B.3 Analysis of other variants of GRPO

In this part, we show that the other variants of GRPO still have difficulty bias on questions.

Let's start with Dr. GRPO. In Dr. GRPO, the un-normalized advantage function is employed:

$$\hat{A}(o|q) = \begin{cases} 1-p(q), & \text{if } r(o|q) = 1, \\ -p(q), & \text{if } r(o|q) = 0 \end{cases} \tag{18}$$

With $f(x,y) = \min(xy, \text{clip}(x, 1-\epsilon, 1+\epsilon)y)$, we have

$$\begin{aligned} \mathcal{J}_{\text{Dr.GRPO}} =& \mathbb{E}_q \mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)} \left[\sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, \hat{A}(o|q) \right) \right] \\ =& \mathbb{E}_q \Bigg[p(q) \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, \hat{A}(o|q) \right) \\ & + (1-p(q)) \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, \hat{A}(o|q) \right) \Bigg] \\ =& \mathbb{E}_q \Bigg[p(q) \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, 1-p(q) \right) \\ & + (1-p(q)) \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, -p(q) \right) \Bigg] \\ =& \mathbb{E}_q p(q)(1-p(q)) \Big[\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} s^+_\theta(o,q) - \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} s^-_\theta(o,q) \Big] \end{aligned} \tag{19}$$

where $s_\theta^+(o,q) = \sum_{t=1}^{|o|} \min(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, 1+\epsilon)$, $s_\theta^-(o,q) = \sum_{t=1}^{|o|} \max(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, 1-\epsilon)$. We can see that the difficult bias $\omega(q) = p(q)(1-p(q))$ on questions persists in Dr. GRPO.

Secondly, let's reformulate the DAPO objective to show the question-level bias. With $f(x,y) = \min(xy, \text{clip}(x, 1-\epsilon_{low}, 1+\epsilon_{high})y)$ and advantage estimator (12), the expected version of DAPO is

$$
\begin{aligned}
\mathcal{J}_{\text{DAPO}} =& \mathbb{E}_q \mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)} \left[\frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, A(o|q) \right) \right] \\
=& \mathbb{E}_q \Bigg[p(q) \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, A(o|q) \right) \\
& + (1-p(q)) \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, A(o|q) \right) \Bigg] \\
=& \mathbb{E}_q \Bigg[p(q) \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, \sqrt{\frac{1-p(q)}{p(q)}} \right) \\
& + (1-p(q)) \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} f\left(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, -\sqrt{\frac{p(q)}{1-p(q)}} \right) \Bigg] \\
=& \mathbb{E}_q \sqrt{p(q)(1-p(q))} \left[\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} s_\theta^+(o,q) - \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} s_\theta^-(o,q) \right]
\end{aligned}
\tag{20}
$$

where $s_\theta^+(o,q) = \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \min(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, 1+\epsilon)$, $s_\theta^-(o,q) = \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \max(\frac{\pi_\theta(o_t|q,o_{<t})}{\pi_{\text{old}}(o_t|q,o_{<t})}, 1-\epsilon)$. We can see that the difficult bias $\omega(q) = \sqrt{p(q)(1-p(q))}$ is placed on questions persists in DAPO.

Thirdly, we show the difficult bias in GPG objective. In GPG, $\alpha\hat{A}(o|q)$ is employed as their advantage estimator. Thus, we have

$$
\begin{aligned}
\mathcal{J}_{\text{GPG}} =& \mathbb{E}_q \mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)} \left[\frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \alpha \log \pi_\theta(o_t|q,o_{<t}) \hat{A}(o|q) \right] \\
=& \mathbb{E}_q \Bigg[p(q) \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \alpha \log \pi_\theta(o_t|q,o_{<t}) \hat{A}(o|q) \\
& + (1-p(q)) \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \alpha \log \pi_\theta(o_t|q,o_{<t}) \hat{A}(o|q) \Bigg] \\
=& \mathbb{E}_q \Bigg[p(q) \mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \alpha \log \pi_\theta(o_t|q,o_{<t}) * (1-p(q)) \\
& + (1-p(q)) \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \alpha \log \pi_\theta(o_t|q,o_{<t}) * (-p(q)) \Bigg] \\
=& \mathbb{E}_q \alpha p(q)(1-p(q)) \left[\mathbb{E}_{o\sim\pi^+_{\text{old}}(\cdot|q)} s_\theta^+(o,q) - \mathbb{E}_{o\sim\pi^-_{\text{old}}(\cdot|q)} s_\theta^-(o,q) \right]
\end{aligned}
\tag{21}
$$

where $s_\theta^+(o,q) = \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \log \pi_\theta(o_t|q,o_{<t})$, and $s_\theta^-(o,q) = \frac{1}{\mathbb{E}_{o\sim\pi_{\text{old}}(\cdot|q)}|o|} \sum_{t=1}^{|o|} \log \pi_\theta(o_t|q,o_{<t})$. We can see that the difficult bias $\omega(q) = \alpha p(q)(1-p(q))$ is on questions in GPG.

Finally, we summarize the question-level weights and their score functions for other variants of GRPO in Table 6.

https://doi.org/10.52202/085713-1921

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's contributions are clearly stated in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in the last section of the paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All assumptions are clearly stated and proofs are provided in Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Detailed instructions to reproduce the results are provided in the experimental part of the main paper. Furthermore, we provide source code in the supplementary materials for reproducing the algorithm.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

 https://doi.org/10.52202/085713-1921

Answer: [Yes]

Justification: Source code is provided in the supplementary materials for reproducing the proposed algorithm.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed information about experimental settings and hyperparameters is provided in the experimental part of the paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer:[No]

   Justification: We follow the practice in this research area. Error bars are not reported because it would be too computationally expensive for training large language models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided the experimental platform and running time of one experiment for our algorithm in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Research conducted in the paper conform with the NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We haven't identified any social impacts of this work.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/085713-1921

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer:[NA]

    Justification: This paper proposes an algorithm without posing such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Datasets and methods utilized in this paper are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

 https://doi.org/10.52202/085713-1921

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The LLM is used only for writing and editing purposes.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.