Stepwise Guided Policy Optimization: Coloring Your Incorrect Reasoning in GRPO

Anonymous authors
Paper under double-blind review

Abstract

Reinforcement learning (RL) has proven effective in strengthening the reasoning capabilities of large language models (LLMs). A widely adopted method, Group Relative Policy Optimization (GRPO) (Shao et al., 2024), has shown strong empirical results in training DeepSeek-R1 (Guo et al., 2025a). However, GRPO fails to update the policy when all responses within a group are incorrect (i.e., *all-negative-sample* groups). This limitation underscores a key gap between artificial and human intelligence: unlike humans, who can learn from mistakes, GRPO discards these signals. Our first contribution is to introduce a simple framework that mitigates the all-negative-sample issue by incorporating response diversity within groups using a *step-wise* judge model, which can be either directly trained or adapted from existing LLMs. We prove that this diversification can accelerate GRPO's learning dynamics in a simplified setting. We also empirically validate the proposed stepwise guided policy optimization (SGPO) method, demonstrating consistent gains across model sizes (7B, 14B, 32B) in offline and online training on 9 benchmarks, including base and distilled variants. Our results highlight two advantages: (i) SGPO surpasses GRPO, especially in the early and mid-training stages where all-negative-sample groups are prevalent; and (ii) SGPO does not require judge models to generate correct answers, differentiating it from knowledge distillation methods.

1 Introduction

The rise of OpenAI-o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025a), and Kimi-1.5 (Team et al., 2025) has highlighted the emergence of *large AI reasoning models*. Unlike instruction-tuned models (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; Achiam et al., 2023), which produce quick responses by statistically inferring the next token, these new reasoning models deliberately decompose complex prompts (e.g., mathematical problems) into intermediate steps and work through chain-of-thought reasoning (Wei et al., 2022; Yao et al., 2023; Besta et al., 2024; Xiang et al., 2025). This slower yet more rigorous process yields greater accuracy and makes them more human-like, enabling success on more complex and challenging tasks (Yang et al., 2018; Shi et al., 2024; Jain et al., 2025). As the generative AI applications move beyond simple conversational interfaces, these reasoning models are poised to become increasingly powerful and widely adopted, positioning them as a key frontier in practice.

At the heart of this revolution lies post-training with outcome-based and verifiable rewards (Cobbe et al., 2021; Uesato et al., 2022; Zelikman et al., 2022; Singh et al., 2023; Hosseini et al., 2024; Lightman et al., 2024; Wang et al., 2024; Setlur et al., 2025; Zhang et al., 2025b), together with reinforcement learning (RL) methods (Schulman et al., 2015; 2017; Li et al., 2024b; Ahmadian et al., 2024; Shao et al., 2024; Xiong et al., 2025a), appreciated for their simplicity, intuitiveness, and practicality. A leading approach is proximal policy optimization (PPO) (Schulman et al., 2017), which relies on a critic (or value) model to estimate advantages. While essential in general RL tasks, this critic is often unnecessary in large language models (LLMs) due to their deterministic transition dynamics (Li et al., 2024b). This observation has inspired alternatives such as group relative policy optimization (GRPO) (Shao et al., 2024) and its extensions (Yu et al., 2025b; Liu et al., 2025b; Chu et al., 2025; Zhang et al., 2025a), which estimate advantages directly in a group-relative fashion.

A major limitation of these methods arises when all sampled responses in a group are incorrect (i.e., *all-negative-sample* groups), which eliminates the learning signal and halts policy updates. In GRPO, given a prompt $\mathbf{x}$, responses $\{\mathbf{y}_i\}_{i=1}^G$ are drawn from the old policy π_{old} and assigned rewards $\{r_i\}_{i=1}^G$, where $r_i = 1$ if $\mathbf{y}_i$ is correct and 0 otherwise. Advantages are obtained by normalizing r_i within the group. If $r_i = 0$ for all i, the advantage vanishes, yielding no update. Such groups are frequent in early and mid-stages of training, when reasoning ability is weak[1]. This shortcoming highlights a gap between artificial and human intelligence: humans effectively learn from mistakes, which act as essential signals during cognitive development (Chialvo & Bak, 1999). In mathematical reasoning, all-negative-sample groups prompt a child to revise rules and strengthen reasoning ability.

Recent studies suggest that negative samples in RL-based large reasoning model training carry more nuanced value than previously assumed (Xiong et al., 2025a). Instead of treating negative samples uniformly, they advocate for principled mechanisms to distinguish negative samples. One prominent direction is process reward models (PRMs) (Lightman et al., 2024; Wang et al., 2024; Luo et al., 2024; Setlur et al., 2025; Zhang et al., 2025b), which estimate either the probability of final success or its change after each reasoning step. However, their reliance on speculative value functions makes them prone to reward hacking (Skalse et al., 2022).

The key insight is that many reasoning tasks possess a structure where step-level correctness can be explicitly defined. This motivates the use of a step-wise judge model that evaluates trajectories by labeling each step as correct (1) or incorrect (0). Such a model can be trained directly (Xiong et al., 2025b) or adapted from existing LLMs (Zha et al., 2025; He et al., 2025)[2]. By grounding rewards in step-level correctness rather than speculative value estimates, our method mitigates reward hacking and yields clearer signals. Intuitively, this allows negative samples to be differentiated through their trajectories: while early-stage reasoning trajectories are of low-quality, these remain informative – much like partial credit in education, where intermediate steps still guide learning.

Our approach enables a holistic evaluation of multi-step reasoning by transforming negative samples from binary outcome rewards into graded, step-level rewards. Consider a negative sample with five reasoning steps $(a_1, a_2, a_3, a_4, a_5)$. If the first error occurs at a_3, then a_1 and a_2 are correct, yielding a correctness proportion of $\frac{2}{5}$. To improve reliability, we adopt a Grok4-Heavy -inspired strategy where multiple independent judgments are obtained from the judge model, and the error position is determined by the majority vote. We further introduce two scaling parameters β and γ to downweight noisy or unreliable signals (see Eq. (2)). Unlike PRMs, our approach avoids memory overhead and does not require costly step-level human annotations, thereby accelerating training.

Contribution. We propose and analyze a simple and efficient framework that introduces response diversity within all-negative-sample groups. It is both theoretically grounded in the simplified setting and empirically effective on various models, distinguishing our approach from existing heuristics. Our contributions can be summarized as follows:

1. We propose a *Stepwise Guided Policy Optimization* (SGPO) framework that leverages a step-wise judge model that identifies the first incorrect step that causes the trajectory to deviate from correctness. This makes evaluation computationally tractable and reliable. *It is important to emphasize that our contribution lies not in designing effective judge models, but in introducing a framework that leverages step-wise judges to effectively distinguish negative samples.* We also prove that SGPO outperforms GRPO in a simplified setting.

2. We conduct the experiments demonstrating the effectiveness of our approach in improving LLM reasoning. Evaluations are undertaken across various model sizes (7B, 14B, 32B) in both offline and online settings with nine benchmarks, including base and distilled variants. Our results reveal two key benefits: (i) SGPO delivers improvements beyond the reach of GRPO, especially in the early and mid-stages of training where all-negative-sample groups are common; (ii) SGPO does not

[1] To reduce computational cost, training often uses small group sizes and short rollouts, further increasing the likelihood of all-negative-sample groups.

[2] We do not have access to their judge models as it's not publicly released, so we adapt our own from existing LLMs.

rely on more powerful judge models generating correct answers, allowing it to be distinguish from knowledge distillation methods.

The additional overhead from all-negative-sample groups remains modest, since the correctness can be efficiently verified against reference solutions, enabling rapid assessment of reasoning steps. As the computational and financial costs of closed-source judge models (`o4-mini`, `Claude3.7`) rise, SGPO accelerates learning dynamics, making the trade-off worthwhile. SGPO also outperforms GRPO with less powerful and more affordable open-source judge models (`DeepSeek-V3-0324`, `Qwen3-235B-A22B`, `QwQ-32B`), confirming that SGPO remains effective even without cutting-edge LLMs and underscoring its practicality in lower-resource settings.

2 Preliminaries and Technical Background

Modern LLMs are built based on the Transformer architecture (Vaswani et al., 2017) and generate responses $\mathbf{y} = (a_1, \ldots, a_H)$ to user prompts $\mathbf{x}$, where each token $a_h \in \mathcal{V}^\star$, with $\mathcal{V}$ denoting the vocabulary and $\mathcal{V}^\star$ the set of all possible token sequences. We view the LLM as a policy $\pi_\theta(\mathbf{y}|\mathbf{x})$ parameterized by θ, assigning probabilities to responses $\mathbf{y}$ given $\mathbf{x}$. The policy operates in an auto-regressive way as follows (Agarwal et al., 2020; Mei et al., 2021; Li et al., 2024b):

$$\pi_\theta(\mathbf{y}|\mathbf{x}) = \prod_{h=1}^{H} \pi_\theta(a_h \mid \mathbf{x}, a_1, \ldots, a_{h-1}).$$

For a prompt $\mathbf{x}$ with ground-truth response $\mathbf{y}^\star_{\mathbf{x}}$, performance is evaluated using a regular-expression match on the final answer: $r(\mathbf{x}, \mathbf{y}) = 1$ if $\mathbf{y}$ matches $\mathbf{y}^\star_{\mathbf{x}}$ and $r(\mathbf{x}, \mathbf{y}) = 0$ otherwise (Hendrycks et al., 2021). We consider the reasoning tasks defined over a dataset $\mathcal{D} = (\mathbf{x}, \mathbf{y}^\star_{\mathbf{x}})$, where each $\mathbf{x}$ is a problem and $\mathbf{y}^\star_{\mathbf{x}}$ its ground-truth solution.

The policy gradient methods (Williams, 1992; Sutton & Barto, 1998) aim to maximize the objective $J(\theta) = \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_\theta(\cdot|\mathbf{x})}[r(\mathbf{x}, \mathbf{y})]$ where ρ is the prompt distribution and π_θ is an LLM policy. Parameters are updated via $\theta \leftarrow \theta + \eta \nabla_\theta J(\theta)$. In practice, trajectories are sampled from an old policy $\pi_{\theta_{\text{old}}}$, which is different from π_θ, motivating the use of importance sampling as follows:

$$J(\theta) = \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{x})} \left[\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}|\mathbf{x})} r(\mathbf{x}, \mathbf{y}) \right].$$

However, this estimator suffers from high variance when π_θ deviates from $\pi_{\theta_{\text{old}}}$. To stabilize training, clipped surrogate objectives are used as follows:

$$J(\theta) = \mathbb{E}_{\mathbf{x}\sim\rho, \mathbf{y}\sim\pi_{\theta_{\text{old}}}(\cdot|\mathbf{x})} \left[\min \left\{ \frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}|\mathbf{x})} r(\mathbf{x}, \mathbf{y}), \texttt{clip}\left(\frac{\pi_\theta(\mathbf{y}|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}|\mathbf{x})}, 1-\epsilon, 1+\epsilon \right) r(\mathbf{x}, \mathbf{y}) \right\} \right].$$

The group relative policy optimization (GRPO) and its variants (Yu et al., 2025b; Liu et al., 2025b; Chu et al., 2025; Zhang et al., 2025a) adopt this framework but estimate gradients using groups of samples. For each prompt $\mathbf{x}$, GRPO samples responses $\mathbf{y}_1, \ldots, \mathbf{y}_G$ from $\pi_{\theta_{\text{old}}}$. We aim at maximizing the objective function in the form of

$$J(\theta) = \mathbb{E}_{\mathbf{x}\sim\rho, \{\mathbf{y}_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot|\mathbf{x})} \left[\frac{1}{G} \sum_{i=1}^{G} \min \left\{ \frac{\pi_\theta(\mathbf{y}_i|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}_i|\mathbf{x})} A_i, \texttt{clip}\left(\frac{\pi_\theta(\mathbf{y}_i|\mathbf{x})}{\pi_{\theta_{\text{old}}}(\mathbf{y}_i|\mathbf{x})}, 1-\epsilon, 1+\epsilon \right) A_i \right\} \right],$$

where $\epsilon \in (0, 1)$ and the advantage A_i is computed as

$$A_i = \frac{r(\mathbf{x}, \mathbf{y}_i) - \texttt{mean}(\{r(\mathbf{x}, \mathbf{y}_1), \ldots, r(\mathbf{x}, \mathbf{y}_G)\})}{\texttt{std}(\{r(\mathbf{x}, \mathbf{y}_1), \ldots, r(\mathbf{x}, \mathbf{y}_G)\})}, \tag{1}$$

where $r(\mathbf{x}, \mathbf{y}_i) = 1$ if $\mathbf{y}_i$ matches the ground-truth answer and 0 otherwise.

Remark 2.1. *When rewards are identical across all samples within a group, $A_i = 0$ and no update occurs. This is appropriate for all-positive groups but constitutes a critical limitation for all-negative groups, where GRPO fails to exploit mistakes as learning signals.*

3 Main Results

We propose the Stepwise Guided Policy Optimization (SGPO) framework, which employs the step-wise judge model to detect the first incorrect step that leads a trajectory away from correctness. In a simplified setting, we prove that SGPO consistently accelerates GRPO's learning dynamics.

3.1 A Step-wise Judge Model

We propose a principled reward mechanism for negative samples, wherein the step-wise judge model differentiates between structurally sound but partially incorrect reasoning and entirely erroneous responses. This design is motivated by the intuition that an incorrect final answer does not invalidate the entire reasoning process. For instance, a model may follow a logically coherent sequence of steps yet make a minor error – such as an arithmetic slip – that leads to an incorrect conclusion. Treating such cases the same as fundamentally flawed or incoherent reasoning does not make sense. This refinement remains effective under constraints such as reduced output length, where a model may be unable to complete the full solution but still demonstrates a valid reasoning trajectory.

Our step-wise judge model can be either trained directly or adapted from existing LLMs. It evaluates responses sequentially, identifying the first substantive error – such as a computational slip or a logical fallacy – that causes the trajectory to deviate from correctness. To formalize this, we define the *Reasoning Trajectory Score* (RTS) for an incorrect response $\mathbf{y}$, denoted as $\texttt{RTS}(\mathbf{y}) \in [0, 1]$. The judge model checks each step in order, pinpoints the first error, and treats all preceding steps as the valid reasoning segment. $\texttt{RTS}(\mathbf{y})$ is then computed as the ratio of the valid segment length to the total trajectory length. For example, if $\mathbf{y}$ consists of five steps $(a_1, a_2, a_3, a_4, a_5)$ and the first error occurs at a_4, then $\texttt{RTS}(\mathbf{y}) = \frac{3}{5}$, indicating that three steps of reasoning are correct before erroneous.

In our experiment, we adapt the judge model from existing LLMs, either closed-source (`o4-mini`, `Claude3.7`) or open-source (`DeepSeek-V3-0324`, `Qwen3-235B-A22B`, `QwQ-32B`). To enhance reliability and further reduce variance in the reward signal, we employ the following protocol: (i) alongside the candidate response, we provide a reference solution drawn from a supervised fine-tuning dataset with correct answers and reasoning trajectories, anchoring the intended solution path and enabling error localization; and (ii) we elicit step-wise evaluation rather than holistic evaluation. The judge model justifies correctness or flags an error sentence by sentence, identifies the first clear mistake, and then traces how this error propagates to the final incorrect conclusion.

Based on the reasoning trajectory score, we introduce a new outcome reward function:

$$r_{\texttt{SGPO}}(\mathbf{y}) = \begin{cases} 1, & \text{if the final answer of } \mathbf{y} \text{ is correct,} \\ \frac{1}{1+\exp(-\beta(\texttt{RTS}(\mathbf{y})-\gamma))}, & \text{otherwise.} \end{cases} \tag{2}$$

where $\gamma > 0$ and $\beta > 0$ are two parameters to decide scale threshold and scale intensity, respectively. This design ensures that the model receives a more informative gradient signal during training, thereby encouraging refinement of partially correct reasoning rather than indiscriminate penalization of all incorrect outputs. This specification of $r_{\texttt{SGPO}}$ can be directly incorporated into the advantage calculation in Eq. (1). As a consequence, we refer to SGPO as GRPO using Eq. (2).

Remark 3.1. *Our approach differs from process reward models (PRMs) (e.g. Lightman et al., 2024). For a prompt* $\mathbf{x}$ *and a prefix of reasoning steps* $(a_1, \ldots, a_t)$*, a PRM typically predicts either (i) a prefix-level value* $V(\mathbf{x}, a_{1:t}) = \mathbb{P}(\text{final answer correct} \mid \mathbf{x}, a_{1:t})$*, or (ii) a step-level progress signal such as* $\Delta_t = V(\mathbf{x}, a_{1:t}) - V(\mathbf{x}, a_{1:t-1})$*. In practice, PRMs are trained by supervised ranking of intermediate steps and are used to re-rank trajectories or shape training at the* prefix *level, acting as approximate value (or Q-value) functions over prefixes. In contrast, SGPO introduces a different way of producing and using feedback signals: (i) Policy-guided rollouts without search. All trajectories are sampled from the current policy, without PRM-guided exploration or trajectory alteration; (ii) Post-hoc first-error identification. A step-wise judge inspects the entire trajectory, pinpoints the earliest error relative to a reference solution, and converts this into a calibrated scalar reward* $r_{\texttt{SGPO}}(\mathbf{y})$ *via the reasoning trajectory score; (iii) Stable credit assignment in all-negative-sample groups. By locating the first definitive mistake only after observing the full trace, SGPO*

eliminates the look-ahead ambiguity and feedback loops inherent to PRM-guided search (Zhang et al., 2024a), while avoiding the need for the judge to solve the problem or approximate a value function.

Remark 3.2. *Our approach differs from knowledge distillation (e.g. Kang et al., 2023; Gu et al., 2024). The student model trained via knowledge distillation inherits the judge model's failure, since it only imitates the judge model's outputs. For instance, consider the AIME problem: "The twelve letters $\{A, B, C, D, E, F, G, H, I, J, K, L\}$ are randomly grouped into six pairs. Each pair is ordered alphabetically to form a two-letter word, and the six words are listed alphabetically, such as AB, CD, EF, GH, IJ, KL. The probability that the last word listed contains G is $\frac{m}{n}$ with m, n coprime. Find $m + n$". Neither the student model (`DeepSeek-R1-Distill-Qwen-7B`) nor the judge models (`DeepSeek-V3-0324`) can solve this problem within 16 rollouts. In contrast, SGPO leverages the judge model to identify mistakes in the student's reasoning, providing learning signals that go beyond imitation and enabling improvements unattainable by knowledge distillation.*

3.2 Accelerating Learning Dynamics

We present a theoretical analysis to explain why SGPO outperforms GRPO. To this end, we consider a simplified setting with a reasoning horizon of $H = 2$, where each step admits two possible actions $a_h \in 1, 2$ for $h = 1, 2$. This configuration follows prior works (Dayan, 1991; Li et al., 2024b), where analogous examples were employed to validate theoretical insights. Without loss of generality, we assume a unique ground-truth response $\mathbf{y}^\star_{\mathbf{x}} = (2, 2)$ for the prompt $\mathbf{x}$. For clarity, we restrict the sample space to $(1, 1), (2, 1), (2, 2)$, excluding $(1, 2)$ since a correct reasoning step is unlikely to, and should not, follow an incorrect precursor.

To illustrate the effect of SGPO, we analyze the learning dynamics of SGPO and GRPO in this simplified setting. Under GRPO, the rewards are assigned as $r((2, 2)) = 1$ and $r((2, 1)) = r((1, 1)) = 0$, meaning that only selecting the "good" action 2 at both steps yields a positive reward. In contrast, SGPO assigns $r_{\texttt{SGPO}}((2, 2)) = 1$, $r_{\texttt{SGPO}}((2, 1)) = \frac{1}{2}$ and $r_{\texttt{SGPO}}((1, 1)) = 0$. The difference is that partial progress – choosing the "good" action 2 in the first step but failing at the second – receives no credit in GRPO yet proportional credit in SGPO. Here, $\frac{1}{2}$ is chosen for illustrative purposes to convey the qualitative behavior of the reward mechanism, while the exact values used in experiments are determined by Eq. (2).

The algorithm iteratively updates the policy parameter θ using samples drawn from the current policy π_θ. We rewrite the generic GRPO update with a step size $\eta > 0$ as follows,

$$\theta^{(k+1)} = \theta^{(k)} + \eta \cdot g(\theta), \quad \text{where } g(\theta) = \frac{1}{NGH} \left(\sum_{i=1}^{N} \sum_{k=1}^{G} \sum_{h=1}^{H} s_\theta(\mathbf{x}^i, a^{i,k}_{1:h-1}) A_{i,k} \right),$$

where N is the number of prompts, G is the number of groups, H is the number of reasoning steps in each response, $s_\theta(\mathbf{x}^i, a^{i,k} 1 : h - 1) := \nabla\theta \log \pi_\theta(a_t|\mathbf{x}, a_{1:h-1})$ is the score function, and the advantage $A_{i,k}$ is defined by Eq. (1) for each question $\mathbf{x}^i$. To distinguish, we denote $g_{\texttt{GRPO}}(\cdot)$ as the gradient estimator using classical outcome reward model r, and $g_{\texttt{SGPO}}(\cdot)$ as the gradient estimator using the reward $r_{\texttt{SGPO}}$ as proposed in Section 3.1.

In our analysis, we examine the population-level learning dynamics with $G = 2$, omitting clipping and importance sampling. For simplicity, we perform our analysis in the likelihood space rather than in the parameter space directly. Indeed, we define the key quantities as follows,

$$p \doteq \pi_{\theta_1}(a_1 = 2, |, \mathbf{x}) = \frac{e^{\theta_1^{\mathbf{x},2}}}{e^{\theta_1^{\mathbf{x},1}} + e^{\theta_1^{\mathbf{x},2}}}, \qquad q \doteq \pi_{\theta_2}(a_2 = 2, |, \mathbf{x}, a_1 = 2) = \frac{e^{\theta_2^{\mathbf{x},2,2}}}{e^{\theta_2^{\mathbf{x},2,1}} + e^{\theta_2^{\mathbf{x},2,2}}}.$$

Note that the original 4-dimensional parameter space defined by $\theta_1^{\mathbf{x},1}$, $\theta_1^{\mathbf{x},2}$, $\theta_2^{\mathbf{x},2,1}$ and $\theta_2^{\mathbf{x},2,2}$ in $\mathbb{R}$ is reduced to a 2-dimensional likelihood space defined by $p, q \in [0, 1]$.

For our simple stylized model, we compute the score functions in terms of likelihood parameters p, q as follows,

$$s(a_1 = 1 \,|\, \mathbf{x}) = \begin{bmatrix} p \\ -p \\ 0 \\ 0 \end{bmatrix}, \quad s(a_1 = 2 \,|\, \mathbf{x}) = \begin{bmatrix} p - 1 \\ 1 - p \\ 0 \\ 0 \end{bmatrix},$$

and

$$s(a_2 = 1 \,|\, \mathbf{x}, a_1 = 2) = \begin{bmatrix} 0 \\ 0 \\ q \\ -q \end{bmatrix}, \quad s(a_2 = 2 \,|\, \mathbf{x}, a_1 = 2) = \begin{bmatrix} 0 \\ 0 \\ q-1 \\ 1-q \end{bmatrix}.$$

The responses can be drawn i.i.d. from the distribution as follows,

$$(a_1, a_2) = \begin{cases} (1,1), & \text{w.p. } 1-p, \\ (2,1), & \text{w.p. } p(1-q), \\ (2,2), & \text{w.p. } pq. \end{cases}.$$

The SGPO and GRPO training dynamics with population-level policy gradient can be computed exactly for the stylized model as follows,

$$\bar{g}_{\texttt{SGPO}}(\theta) = \mathbb{E}[g_{\texttt{SGPO}}(\theta)] = \tfrac{1}{2} \begin{bmatrix} p(p-1) \\ p(1-p) \\ p^2 q(q-1) \\ p^2 q(1-q) \end{bmatrix}, \quad \bar{g}_{\texttt{GRPO}}(\theta) = \mathbb{E}[g_{\texttt{GRPO}}(\theta)] = \tfrac{1}{2} \begin{bmatrix} p(p-1)q \\ p(1-p)q \\ pq(q-1) \\ pq(1-q) \end{bmatrix}.$$

Since $g_{\texttt{GRPO}}(\theta)$ and $g_{\texttt{SGPO}}(\theta)$ concentrate around $\bar{g}_{\texttt{GRPO}}(\theta)$ and $\bar{g}_{\texttt{SGPO}}(\theta)$ when the number of samples in each group is sufficiently large, it is reasonable to analyze the population-level dynamics for illustration. Note that the high-probability guarantees for the sample-level dynamics can be derived using concentration inequalities under certain conditions.

Denote $p^{(k)}_{\texttt{GRPO}}$ and $q^{(k)}_{\texttt{GRPO}}$ as the value of quantity p and q at iteration k under GRPO. Analogously, $p^{(k)}_{\texttt{SGPO}}$ and $q^{(k)}_{\texttt{SGPO}}$ are the corresponding probabilities p and q at iteration k under SGPO. We can explicitly write down the SGPO and GRPO update rules with $\eta = 1$ as follows,

$$\begin{cases} p^{(k+1)}_{\texttt{SGPO}} = \exp(f_{11}(p^{(k)}_{\texttt{SGPO}})), \\ q^{(k+1)}_{\texttt{SGPO}} = \exp(f_{12}(p^{(k)}_{\texttt{SGPO}}, q^{(k)}_{\texttt{SGPO}})), \end{cases} \quad \text{and} \quad \begin{cases} p^{(k+1)}_{\texttt{GRPO}} = \exp(f_{21}(p^{(k)}_{\texttt{GRPO}}, q^{(k)}_{\texttt{GRPO}})), \\ q^{(k+1)}_{\texttt{GRPO}} = \exp(f_{22}(p^{(k)}_{\texttt{GRPO}}, q^{(k)}_{\texttt{GRPO}})), \end{cases}, \tag{3}$$

where the functions f_{ij} are defined by

$$\begin{aligned} f_{11}(p) &= \log(p) + p(1-p) - \log(1 - p + pe^{p(1-p)}), \\ f_{21}(p,q) &= \log(p) + p(1-p)q - \log(1 - p + pe^{p(1-p)q}), \\ f_{12}(p,q) &= \log(q) + p^2 q(1-q) - \log(1 - q + qe^{p^2 q(1-q)}), \\ f_{22}(p,q) &= \log(q) + pq(1-q) - \log(1 - q + qe^{pq(1-q)}). \end{aligned} \tag{4}$$

Our theoretical findings are summarized in the following theorem.

Theorem 3.3. *Suppose that $p^{(0)}_{GRPO} = q^{(0)}_{GRPO} = p^{(0)}_{SGPO} = q^{(0)}_{SGPO} = \frac{1}{2}$ and $\eta = 1$[3] for GRPO and SGPO. Then, we have that (i) GRPO and SGPO achieve successful learning: $p^{(k)}_{GRPO}, q^{(k)}_{GRPO}, p^{(k)}_{SGPO}, q^{(k)}_{SGPO} \to 1$ as $k \to +\infty$; (ii) SGPO outperforms GRPO in learning the "good" action in the first step: $p^{(k)}_{SGPO} > p^{(k)}_{GRPO}$ for all $k \geq 1$; (iii) SGPO outperforms GRPO in learning the optimal policy: $p^{(k)}_{SGPO} q^{(k)}_{SGPO} > p^{(k)}_{GRPO} q^{(k)}_{GRPO}$ for all $k \geq 1$.*

Proof. To show (i), recall that the sequence $(p^{(k)}_{\texttt{SGPO}})_{k \in \mathbb{N}}$ is strictly increasing and bounded in $(0,1)$ from Lemmas B.4(i) and B.4(ii), so it converge to some value $c \in (0,1]$. Take limit as $k \to \infty$:

$$1 = \lim_{k \to \infty} \frac{p^{(k+1)}_{\texttt{SGPO}}}{p^{(k)}_{\texttt{SGPO}}} = \lim_{k \to \infty} \frac{1}{(1 - p^{(k)}_{\texttt{SGPO}}) e^{-\Delta^{(k)}_{\texttt{SGPO},p}} + p^{(k)}_{\texttt{SGPO}}} = \frac{1}{(1-c)e^{-c(1-c)} + c}.$$

Using the simple Taylor lower bound $e^{-x} \geq 1 - x$, we have

$$1 = \frac{1}{(1-c)e^{-c(1-c)} + c} \geq \frac{1}{(1-c)(1-c(1-c)) + c} \implies (c-1)^2 \leq 0 \implies c = 1.$$

[3] We use the unit stepsize for simplicity. Our results are valid for any sufficiently small step size.

This shows $p_{\texttt{SGPO}}^{(k)} \to 1$ as $k \to \infty$. Similarly, we can show $q_{\texttt{GRPO}}^{(k)}, p_{\texttt{SGPO}}^{(k)}, q_{\texttt{SGPO}}^{(k)} \to 1$ as $k \to \infty$.

To show (ii), consider the base case:

$$\begin{aligned} p_{\texttt{SGPO}}^{(1)} &= \exp(f_{11}(p_{\texttt{SGPO}}^{(0)})) = \exp(\log p_{\texttt{SGPO}}^{(0)} + h_{p_{\texttt{SGPO}}^{(0)}}(p_{\texttt{SGPO}}^{(0)}(1-p_{\texttt{SGPO}}^{(0)}))) \\ &= \exp(\log p_{\texttt{GRPO}}^{(0)} + h_{p_{\texttt{GRPO}}^{(0)}}(p_{\texttt{GRPO}}^{(0)}(1-p_{\texttt{GRPO}}^{(0)}))) \\ &> \exp(\log p_{\texttt{GRPO}}^{(0)} + h_{p_{\texttt{GRPO}}^{(0)}}(p_{\texttt{GRPO}}^{(0)}(1-p_{\texttt{GRPO}}^{(0)})q_{\texttt{GRPO}}^{(0)})) = \exp(f_{21}(p_{\texttt{GRPO}}^{(0)}, q_{\texttt{GRPO}}^{(0)})) = p_{\texttt{GRPO}}^{(1)}, \end{aligned}$$

where the inequality follows from Lemma B.1(ii). Thus, we use induction and assume $p_{\texttt{SGPO}}^{(k)} > p_{\texttt{GRPO}}^{(k)}$ for some $k \geq 1$. Then we have,

$$\begin{aligned} p_{\texttt{SGPO}}^{(k+1)} &= \exp(f_{11}(p_{\texttt{SGPO}}^{(k)})) > \exp(f_{11}(p_{\texttt{GRPO}}^{(k)})) = \exp(\log p_{\texttt{GRPO}}^{(k)} + h_{p_{\texttt{GRPO}}^{(k)}}(p_{\texttt{GRPO}}^{(k)}(1-p_{\texttt{GRPO}}^{(k)}))) \\ &> \exp(\log p_{\texttt{GRPO}}^{(k)} + h_{p_{\texttt{GRPO}}^{(k)}}(p_{\texttt{GRPO}}^{(k)}(1-p_{\texttt{GRPO}}^{(k)})q_{\texttt{GRPO}}^{(k)})) = \exp(f_{21}(p_{\texttt{GRPO}}^{(k)}, q_{\texttt{GRPO}}^{(k)})) = p_{\texttt{GRPO}}^{(k+1)}, \end{aligned}$$

where the first inequality uses Lemma B.1(i) and the second one uses Lemma B.1(ii). Thus, $p_{\texttt{SGPO}}^{(k+1)} > p_{\texttt{GRPO}}^{(k+1)}$ and induction completes. We have proved that $p_{\texttt{SGPO}}^{(k)} > p_{\texttt{GRPO}}^{(k)}$ for all $k \geq 1$.

To show (iii), first notice that we can show $p_{\texttt{GRPO}}^{(k)} = q_{\texttt{GRPO}}^{(k)}$ for all $k \geq 0$ by induction. The base case is trivial by initialization. Suppose $p_{\texttt{GRPO}}^{(k)} = q_{\texttt{GRPO}}^{(k)}$ for some $k \geq 0$, then by noticing that $f_{21}(p,p) = f_{22}(p,p)$, we have

$$\begin{aligned} p_{\texttt{GRPO}}^{(k+1)} &= \exp(f_{21}(p_{\texttt{GRPO}}^{(k)}, q_{\texttt{GRPO}}^{(k)})) = \exp(f_{21}(p_{\texttt{GRPO}}^{(k)}, p_{\texttt{GRPO}}^{(k)})) \\ &= \exp(f_{22}(p_{\texttt{GRPO}}^{(k)}, p_{\texttt{GRPO}}^{(k)})) = \exp(f_{22}(p_{\texttt{GRPO}}^{(k)}, q_{\texttt{GRPO}}^{(k)})) = q_{\texttt{GRPO}}^{(k+1)}. \end{aligned}$$

Thus, by induction, $p_{\texttt{GRPO}}^{(k)} = q_{\texttt{GRPO}}^{(k)}$ for all $k \geq 0$. Now, we can reduce the update rule of $p_{\texttt{GRPO}}^{(k)}$ as

$$p_{\texttt{GRPO}}^{(k+1)} = \frac{1}{(1/p_{\texttt{GRPO}}^{(k)} - 1)\exp(-(p_{\texttt{GRPO}}^{(k)})^2(1-p_{\texttt{GRPO}}^{(k)}))+1}.$$

Also recall the update rule of $p_{\texttt{SGPO}}^{(k)}$ and $q_{\texttt{SGPO}}^{(k)}$:

$$\begin{aligned} p_{\texttt{SGPO}}^{(k+1)} &= \frac{1}{(1/p_{\texttt{SGPO}}^{(k)} - 1)\exp(-p_{\texttt{SGPO}}^{(k)}(1-p_{\texttt{SGPO}}^{(k)}))+1} \\ q_{\texttt{SGPO}}^{(k+1)} &= \frac{1}{(1/q_{\texttt{SGPO}}^{(k)} - 1)\exp(-(p_{\texttt{SGPO}}^{(k)})^2 q_{\texttt{SGPO}}^{(k)}(1-q_{\texttt{SGPO}}^{(k)}))+1} \end{aligned},$$

and it suffices to show $p_{\texttt{SGPO}}^{(k)} q_{\texttt{SGPO}}^{(k)} > (p_{\texttt{GRPO}}^{(k)})^2$ for all $k \geq 1$. We prove by induction. For the base case,

$$\sqrt{p_{\texttt{SGPO}}^{(1)} q_{\texttt{SGPO}}^{(1)}} = \sqrt{\frac{1}{1+(1/2)e^{-1/4}} \cdot \frac{1}{1+(1/2)e^{-1/16}}} > \frac{1}{1+(1/2)e^{-1/8}} = p_{\texttt{GRPO}}^{(1)}.$$

The above inequality holds true since Lemma B.1(iv) implies

$$2\log(1+(1/2)e^{-1/8}) > \log(1+(1/2)e^{-1/4}) + \log(1+(1/2)e^{-1/16}),$$

It remains to show that $p_{\texttt{SGPO}}^{(k)} q_{\texttt{SGPO}}^{(k)} > (p_{\texttt{GRPO}}^{(k)})^2$ implies $p_{\texttt{SGPO}}^{(k+1)} q_{\texttt{SGPO}}^{(k+1)} > (p_{\texttt{GRPO}}^{(k+1)})^2$ for $k \geq 1$. By Lemma B.4(iii), we know $p_{\texttt{SGPO}}^{(k)} > q_{\texttt{SGPO}}^{(k)}$ for all $k \geq 1$. Thus, Lemma B.2 implies that

$$p_{\texttt{SGPO}}^{(k+1)} q_{\texttt{SGPO}}^{(k+1)} = \frac{1}{A(p_{\texttt{SGPO}}^{(k)})B(p_{\texttt{SGPO}}^{(k)}, q_{\texttt{SGPO}}^{(k)})} > \frac{1}{C\left(\sqrt{p_{\texttt{SGPO}}^{(k)} q_{\texttt{SGPO}}^{(k)}}\right)^2}.$$

Using Lemma B.1(iii), we complete the induction by applying our induction hypothesis:

$$\frac{1}{C\left(\sqrt{p_{\texttt{SGPO}}^{(k)} q_{\texttt{SGPO}}^{(k)}}\right)^2} = \exp\left(f_{21}\left(\sqrt{p_{\texttt{SGPO}}^{(k)} q_{\texttt{SGPO}}^{(k)}}, \sqrt{p_{\texttt{SGPO}}^{(k)} q_{\texttt{SGPO}}^{(k)}}\right)\right) > \exp(f_{21}(p_{\texttt{GRPO}}^{(k)}, p_{\texttt{GRPO}}^{(k)})) = (p_{\texttt{GRPO}}^{(k+1)})^2.$$

This completes the proof. □

Table 1: Evaluation results on offline RL training. For each model, we report the baseline performance before RL training. We then report RL training results that uses only negative samples and positive samples, respectively. Performance across validation and training dataset (`LIMO`) is shown.

	`AMC23` avg@16	`AIME24` avg@16	`MATH500` pass@1	`Olympiads` pass@1	`LIMO` pass@1
Qwen2.5-14B-Instruct					
Baseline	58.59	14.58	**80.40**	41.78	31.70
Negative Samples only	**61.88**	**15.21**	**80.40**	**42.37**	30.11
Positive Samples only	61.72	14.58	79.80	42.07	**38.68**
Qwen2.5-32B-Instruct					
Baseline	64.22	17.08	**83.60**	45.93	34.64
Negative Samples only	**69.53**	**20.42**	83.00	46.37	36.47
Positive Samples only	66.87	18.75	**83.60**	**47.41**	**41.86**

Remark 3.4. *Theorem 3.3 presents one of the first theoretical analyses of GRPO with multiple samples and multi-step reasoning in the context of LLM reasoning. The first part establishes that SGPO converges to the optimal policy. The second and third parts demonstrate that SGPO both accelerates the acquisition of partially correct reasoning steps and preserves partial reasoning ability even when the final answer is incorrect. Importantly, the theorem provides a* ***per-iteration*** *comparison of learning under different reward mechanisms – an aspect rarely examined in previous works. The provable improvement in learning the optimal policy is also consistent with our numerical findings. We plot the resulting learning curves of our numerical simulation in Figure 1. The left panel shows the probability of selecting the "good" action in the first step at iteration k (i.e., $p^{(k)}_{\texttt{SGPO}}$ vs. $p^{(k)}_{\texttt{GRPO}}$), while the right panel shows the probability of learning the optimal policy (i.e., $p^{(k)}_{\texttt{SGPO}} q^{(k)}_{\texttt{SGPO}}$ vs. $p^{(k)}_{\texttt{GRPO}} q^{(k)}_{\texttt{GRPO}}$). The results align with the predictions of Theorem 3.3, demonstrating that the likelihood of learning the optimal policy under SGPO consistently exceeds that of GRPO across training.*

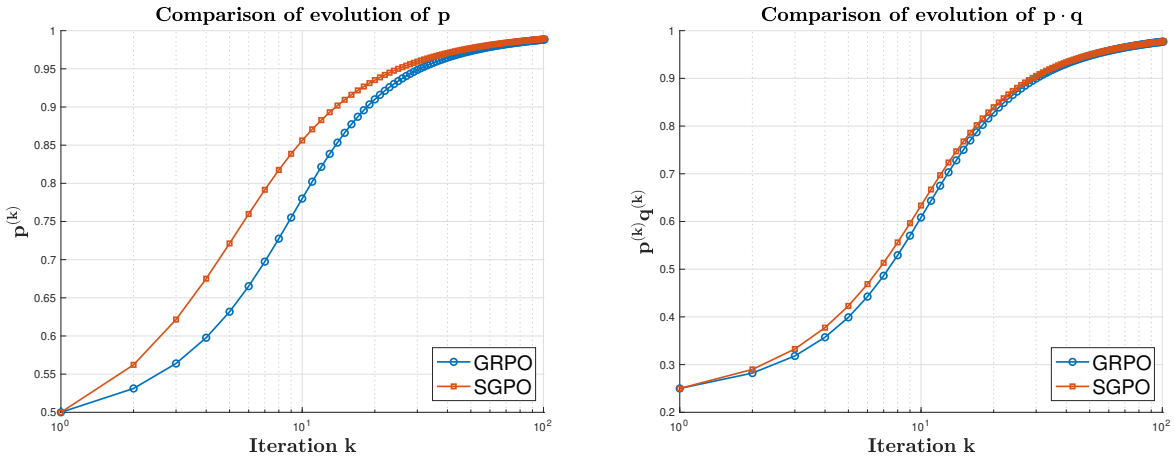

Figure 1: Learning dynamics of GRPO and SGPO in the simplified setting.

4 Experiments

We demonstrate the benefits of differentiating negative samples through experiments in both offline and online settings. Offline RL is more computationally efficient, offering faster training, reduced memory consumption, and improved stability. In contrast, online RL provides greater flexibility and learning capacity,

and has become the standard approach in large-scale reasoning models such as DeepSeek-R1 (Guo et al., 2025a).

4.1 Offline Training

For baselines, we consider strong models without further fine-tuned on math-specific SFT datasets, namely `Qwen2.5-14B-Instruct` and `Qwen2.5-32B-Instruct`. Prior work showed that a small set of carefully curated prompts significantly enhance the reasoning capability. Accordingly, we adopt the `GAIR/LIMO` dataset (Ye et al., 2025) as the training set, which has demonstrated strong potential for improving the reasoning performance of large-scale (32B) models in offline SFT. Evaluation is conducted on four standard math reasoning benchmarks: `AIME24`, `AMC23`, `MATH500` (Hendrycks et al., 2021), and `OlympiadBench` (He et al., 2024). Our aim is to highlight the rich learning signal contained in all-negative-sample groups, showing that training exclusively on them can still yield performance gains. For benchmarks with fewer than 100 questions (`AMC23`, `AIME24`), we report `avg@16` results with a decoding temperature of 0.6 and `Top_P` $= 0.95$. For benchmarks with more than 100 questions, we report `pass@1` results using greedy decoding. Across all experiments, the maximum decoding length is set to 32768 tokens.

We conduct all response generation and model updates using offline RL (Peters & Schaal, 2007) with the standard GRPO mechanism. Specifically, the model is updated with advantages estimated from the offline dataset (see, e.g., Peng et al., 2019; Li et al., 2024b). For each prompt, we sample six responses per group and identify all-negative-sample groups in which all responses yield incorrect answers. Within these groups, we apply the step-wise judge model to assign differentiated rewards to negative samples, which are then used for offline RL updates. The model is trained for three epochs with a learning rate of 2×10^{-6}. As a contrastive baseline, we also perform offline RL using only positive rollouts with correct answers. This parallel setup enables a direct comparison between learning from exclusively negative reasoning trajectories and from exclusively positive ones.

We conduct offline RL training to demonstrate that utilizing all-negative-sample groups can enhance the reasoning abilities of LLMs. For comparison, we also include positive-only offline RL training. As shown in Table 1, SGPO with negative samples consistently improve performance across most of benchmarks, in some cases even surpassing models trained solely on positive samples. Notably, in the 14B model experiment, training on negative samples yields improvements on four benchmarks relative to the positive-sample baseline. These findings underscore the utility of negative samples, which should not be discarded in online GRPO training; see more discussions in Section 4.4.

4.2 Online Training

For baselines, we consider applying `Qwen2.5-14B-Instruct`, `Qwen2.5-32B-Instruct`, `QwQ-32B`, `DeepSeek-R1-Distill-Qwen-7B` and `DeepSeek-R1-Distill-Llama-8B`. Online GRPO training is implemented using the `verl` framework (Sheng et al., 2025). For the step-wise judge model, we adopt a diverse set of LLMs, ranging from closed-source models with strong reasoning capabilities (`o4-mini`, `Claude3.7`) to open-source models that are more accessible to the community, including `DeepSeek-V3-0324`, `Qwen3-235B-A22B`, and `QwQ-32B`.

Compared to offline RL, online RL yields larger improvements in a model's reasoning capabilities. Since our baselines already include strong distillation models, some benchmarks used in offline evaluation are nearing saturation. To provide a better assessment, we expand our evaluation suite beyond `AMC23`, `AIME24`, `MATH500`, and `OlympiadBench` by including `AIME25`, `GradeSchool` (Ye et al., 2025), `CHMath24`, `Kaoyan`, and `Gaokao`. Specifically, `CHMath24` is the benchmark from the 2024 Chinese High School Mathematics League Competition, `Gaokao` from China's 2024 National College Entrance Examination, `Kaoyan` from the Chinese Graduate School Entrance Examinations, and `GradeSchool` targets elementary-level mathematical reasoning. Among these, `CHMath24` and `Gaokao` each contain fewer than 100 questions, for which we apply the temperature-based decoding for evaluation.

For GRPO training, we use the `AIME` collections from 1997 to 2023 provided in DeepScaler (Luo et al., 2025b), training for 12 epochs. All training questions are in English, while evaluation benchmarks include

Table 2: Evaluation results on online RL training. We refer to Baseline as the performance of the original model without RL finetuning. **Overall** is average performance across all the benchmarks. Note that the training dataset is `AIME1997-2023`. For `DeepSeek-R1-Distill-Qwen-7B`, we report additional results, including (i) compatibility with more judge models and (ii) ablation on the stability parameters β and γ.

	Kaoyan pass@1	GradeMath pass@1	MATH500 pass@1	Olympiads pass@1	CHMath24 avg@16	AIME25 avg@16	AIME24 avg@16	GaoKao avg@16	AMC23 avg@16	Overall avg
DeepSeek-R1-Distill-Qwen-7B										
Baseline	50.25	41.43	87.00	49.93	73.75	**40.62**	52.92	80.22	89.53	62.85
GRPO	55.78	43.33	89.40	**56.00**	71.04	36.68	52.08	80.30	88.91	63.72
SGPO+o4-mini-0416	57.79	46.19	90.80	54.67	75.00	38.33	54.58	81.33	90.00	65.41
SGPO+DeepSeek-V3-0324	54.77	**47.17**	91.00	55.11	**77.29**	40.42	**56.87**	**82.28**	90.83	**66.19**
SGPO+Qwen3-235B-A22B	56.78	46.67	**92.00**	54.67	73.33	37.92	55.63	81.17	90.63	65.42
SGPO+QwQ-32B	52.26	45.24	**92.00**	53.78	75.00	35.21	56.46	**82.28**	**91.88**	64.91
SGPO+QwQ-32B w/o $\{\beta,\gamma\}$	**58.29**	42.38	90.20	55.11	74.58	38.69	53.63	81.24	88.75	65.08
DeepSeek-R1-Distill-Llama-8B										
Baseline	29.15	23.81	77.40	41.48	**61.46**	27.92	**42.29**	**72.78**	87.97	51.58
GRPO	35.68	28.33	**84.00**	46.32	57.08	**28.33**	42.08	68.99	86.72	53.06
SGPO+Claude-3.7	**39.70**	**29.05**	83.60	**48.44**	58.96	24.58	39.37	71.52	**89.06**	**53.81**
Qwen2.5-14B-Instruct										
Baseline	37.69	49.52	80.40	41.78	21.88	13.13	**14.58**	**41.14**	58.59	39.85
GRPO	**43.22**	47.14	80.20	43.11	21.88	13.13	13.33	39.16	**59.84**	40.11
SGPO+o4-mini-0416	38.69	**53.33**	**81.00**	**44.00**	**22.92**	**16.67**	14.17	39.00	59.22	**41.00**
Qwen2.5-32B-Instruct										
Baseline	45.73	**53.81**	**83.60**	45.93	26.87	12.29	17.08	44.15	64.22	43.74
GRPO	**48.24**	52.86	83.20	45.93	22.50	12.08	**21.67**	**45.73**	67.34	44.39
SGPO+o4-mini-0416	**48.24**	**53.81**	83.00	**46.81**	**29.79**	**14.58**	19.58	45.09	**69.53**	**45.06**
QwQ-32B										
Baseline	64.32	62.38	94.60	68.74	**89.39**	**68.54**	77.71	86.88	97.03	78.84
GRPO	71.36	63.81	94.60	69.48	88.75	64.38	75.83	**87.11**	97.03	79.15
SGPO+DeepSeek-V3-0324	**73.37**	**64.76**	**95.00**	**70.22**	88.33	66.46	**78.33**	**87.11**	**97.97**	**80.17**

multilingual questions. Notably, negative samples learned during training generalize well to out-of-domain mathematical reasoning tasks. SGPO training follows the same setup. With batch-simultaneous processing, judge model calls take 90 seconds per batch of negatives, adding 10% wall-clock time relative to rollout and update. Step-wise supervision is applied only to all-negative-sample groups during the first three epochs, as we expect this duration to suffice for the model to internalize corrective signals; beyond this point, unresolved examples are more indicative of model capacity limits than learnability. For all models, rollout length is fixed at 8192 tokens and group size at 8. Models less than 8B are trained on 8 H100, 14B models on 16 H100, and 32B models on 32 H200. We adopt the default KL coefficient and learning rate from the `verl` training script (Sheng et al., 2025), and use the LIMO evaluation script (Ye et al., 2025), both of which are standard practices in the community.

A key insight from Table 2 is that stronger models generate higher-quality negative samples, which substantially aid learning. As model capability improves, so does the informativeness of its mistakes. Negative samples broadly fall into two categories: (i) correct reasoning trajectories truncated by output length limits, and (ii) trajectories containing logical errors. The first type remains highly valuable – yet discarded in GRPO – since it preserves meaningful reasoning steps, motivating our step-wise judge model. The second type, though incorrect, still provides informative signals, especially when all samples fail on genuinely challenging problems. Notably, stronger distilled models average 6K tokens per response, compared to only 1K tokens for weaker base models, making truncated but informative negative samples more common in the stronger case. Likewise, their erroneous responses also tend to be richer and more useful for step-level judgment.

4.3 Other Ablation Studies

To assess reliability of judge models, we evaluate our approach not only with strong closed-source reasoning models but also with publicly available models of weaker capacity: `DeepSeek-V3`, `Qwen3-235B` and `QwQ-32B`. As shown in Table 2 (best-tuned results) and Table 3 (multiple runs with weaker judges), performance remains

Table 3: Evaluation results are reported for `DeepSeek-R1-Distill-Qwen-7B` across four independent runs. First column indicates judge models and its corresponding reward stability setup.

	Kaoyan pass@1	GradeMath pass@1	MATH500 pass@1	Olympiads pass@1	CHMath24 avg@16	AIME25 avg@16	AIME24 avg@16	GaoKao avg@16	AMC23 avg@16	Overall avg
DeepSeek-R1-Distill-Qwen-7B-SGPO										
+Qwen3-235B-A22B	53.90 ± 2.10	46.55 ± 0.24	91.30 ± 0.87	53.45 ± 1.35	74.48 ± 1.10	37.40 ± 1.05	55.73 ± 1.39	81.33 ± 0.18	90.19 ± 0.48	64.92 ± 0.37
+QwQ-32B	53.89 ± 1.66	44.88 ± 1.36	91.15 ± 0.84	53.71 ± 0.74	74.33 ± 1.29	37.03 ± 1.23	54.76 ± 1.87	81.91 ± 0.53	90.08 ± 1.23	64.64 ± 0.41
+QwQ-32B w/o $\{\beta,\gamma\}$	56.14 ± 2.76	44.53 ± 2.41	90.10 ± 0.66	53.64 ± 1.08	73.89 ± 1.13	38.70 ± 1.80	53.63 ± 2.15	81.24 ± 0.50	88.83 ± 0.81	64.52 ± 0.57

Table 4: Evaluation results are reported for `DeepSeek-R1-Distill-Qwen-7B` as base model and `QwQ-32B` as judge model with and without majority voting.

	Kaoyan pass@1	GradeMath pass@1	MATH500 pass@1	Olympiads pass@1	CHMath24 avg@16	AIME25 avg@16	AIME24 avg@16	GaoKao avg@16	AMC23 avg@16	Overall avg
DeepSeek-R1-Distill-Qwen-7B-SGPO										
+QwQ-32B	53.89 ± 1.66	44.88 ± 1.36	91.15 ± 0.84	53.71 ± 0.74	74.33 ± 1.29	37.03 ± 1.23	54.76 ± 1.87	81.91 ± 0.53	90.08 ± 1.23	64.64 ± 0.41
+QwQ-32B with voting	56.66 ± 1.66	45.24 ± 2.16	91.35 ± 0.50	53.82 ± 0.94	74.19 ± 0.92	37.35 ± 2.25	55.81 ± 0.99	82.12 ± 0.43	90.53 ± 1.07	65.23 ± 0.18

Table 5: Evaluation results are reported in terms of `pass@16` across benchmarks. The first two columns show the total number of questions and the number solved within 16 attempts, while the last two columns report the number of unique questions solved by one method but not the other.

	SGPO - pass@16	GRPO - pass@16	SGPO \ GRPO	GRPO \ SGPO
AIME24	23/30	19/30	4	0
AIME25	21/30	21/30	1	1
Gaokao	70/79	68/79	2	0
AMC23	39/40	38/40	1	0
CHMath24	27/30	25/30	2	0

stable, indicating that weaker judges do not significantly degrade outcomes. We attribute this reliability to two design choices: (i) first-step error identification with a reference answer. SGPO requires the judge only to verify each step against the reference, not to solve the problem, thereby reducing task difficulty and avoiding the pitfalls of generic PRMs; (ii) reward stability parameters β and γ, which filter and smooth noisy signals. As confirmed by ablations, removing β and γ increases variance and weakens performance. To improve verification, we incorporate a Grok4−Heavy-inspired strategy: multiple independent evaluations by the judge model, with the error position selected by majority voting. Using `QwQ-32B` as the judge model, `DeepSeek-R1-Distill-Qwen-7B` as the base model, and four rollouts per judgment, we have observed noticeable gains in consistency and stability (see Table 4).

While `avg@16` measures average performance across rollouts, `pass@16` reflects the ability to solve new questions with multiple attempts. As shown in Table 5, SGPO's gains in `pass@16` stem directly from leveraging negative samples. Learning only from solvable problems reinforces existing ability, whereas all-negative-sample groups correspond to genuinely difficult questions where GRPO consistently fails. These are precisely the cases where additional feedback can be most valuable. By providing step-level signals, SGPO rewards near-misses by reinforcing correct reasoning up to the first error, penalizes early failures by discouraging persistent error modes, and exposes blind spots by turning hard cases into informative training signals. In this regard, SGPO provides benefits that GRPO cannot match, covering more hard problems and providing sharper credit assignment, which translates to faster and more reliable learning under realistic compute budgets.

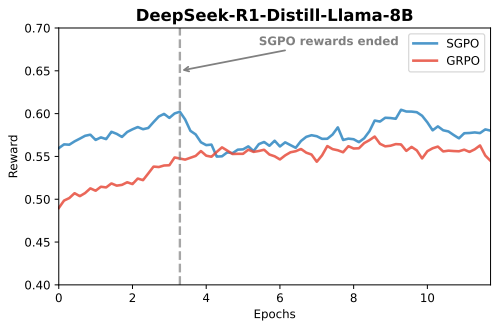

Figure 2: Evaluation results on GRPO and SGPO. SGPO rewards end at epoch 3.

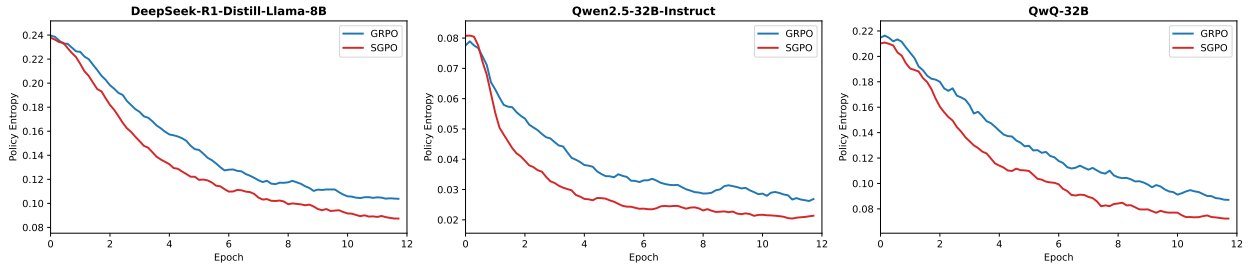

Figure 3: Policy entropy levels during training for GRPO and SGPO across different base models.

By leveraging richer early-stage signals from negative samples, SGPO can achieve faster and stronger performance than GRPO. As shown in Figure 2, SGPO continues improving beyond epoch 5 by solving several additional hard training problems, whereas GRPO plateaus. This improvement stems from informative negative samples that help resolve previously unsolved problems as also shown in Table 5. In line with our theoretical finding that SGPO converges faster than GRPO, empirical metrics offer supporting evidence. Prior work on RLVR entropy highlights its link to performance: Cui et al. (2025) showed that lower policy entropy under correct signals correlates with stronger policies, while Agarwal et al. (2025) demonstrated that directly minimizing entropy can improve performance. As shown in Figure 3, SGPO reduces policy entropy more rapidly than GRPO, indicating faster convergence toward deterministic RLVR behavior with higher rollout confidence. This matches our theoretical results, confirming that step-wise signals accelerate convergence.

4.4 Discussions

We highlight the motivation for evaluating both offline and online RL. In the offline setup, training uses **only** negative samples, allowing us to directly test whether incorrect or incomplete reasoning trajectories can improve performance. In the online setup, we simulate realistic GRPO training, where batches contain a random mix of positive and negative samples. This demonstrates that negative samples are not only effective in isolation but also remain valuable in practical settings with noisier, mixed data. While mixing positives and negatives introduces noise, simply discarding negative samples does not stabilize training; in several cases, the performance of GRPO even drops below baseline, as the model overfits to problems it can already solve.

This instability arises from limited out-of-domain generalization and catastrophic forgetting. Without exposure to challenging or partially correct reasoning, the model risks overfitting to easy cases, reinforcing shallow heuristics instead of developing robust problem-solving skills. The absence of diverse failure cases can also cause catastrophic forgetting, degrading performance on previously solvable tasks. Incorporating negative samples mitigates these issues, as SGPO consistently outperforms GRPO on Chinese OOD math benchmarks. Nonetheless, further work is needed to design more stable training frameworks, including richer reward diversification mechanisms for handling negative samples and corresponding efficient RL methods beyond GRPO.

5 Conclusion

We propose a simple and efficient framework that introduces response diversity within all-negative-sample groups and prove, in a simplified setting, that such diversification can accelerate the learning dynamic of GRPO. Empirically, our approach can yield consistent improvements across model sizes in both offline and online training over nine benchmarks, including base and distilled variants. Future works include extending theoretical results to broader multi-step reasoning tasks, applying response diversity to accelerate other RL methods, and designing lightweight, task-specific reward models that evaluate reasoning steps correctly even if they cannot solve the full problem.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *ArXiv Preprint: 2303.08774*, 2023. (Not cited.)

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. Optimality and approximation with policy gradient methods in markov decision processes. In *COLT*, pp. 64–66, 2020. (Not cited.)

S. Agarwal, Z. Zhang, L. Yuan, J. Han, and H. Peng. The unreasonable effectiveness of entropy minimization in LLM reasoning. *ArXiv Preprint: 2505.15134*, 2025. (Not cited.)

A. Ahmadian, C. Cremer, M. Gallé, M. Fadaee, J. Kreutzer, O. Pietquin, A. Üstün, and S. Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *ACL*, pp. 12248–12267, 2024. (Not cited.)

D. Arora and A. Zanette. Training language models to reason efficiently. *ArXiv Preprint: 2502.04463*, 2025. (Not cited.)

M. G. Azar, Z. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. A general theoretical paradigm to understand learning from human preferences. In *AISTATS*, pp. 4447–4455, 2024. (Not cited.)

Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *ArXiv Preprint: 2212.08073*, 2022. (Not cited.)

M. Besta, N. Blach, A. Kubicek, R. Gerstenberger, M. Podstawski, L. Gianinazzi, J. Gajda, T. Lehmann, H. Niewiadomski, P. Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI*, pp. 17682–17690, 2024. (Not cited.)

X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu, et al. Deepseek LLM: Scaling open-source language models with longtermism. *ArXiv Preprint: 2401.02954*, 2024. (Not cited.)

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *NeurIPS*, pp. 1877–1901, 2020. (Not cited.)

H. Chen, G. He, L. Yuan, G. Cui, H. Su, and J. Zhu. Noise contrastive alignment of language models with explicit rewards. In *NeurIPS*, pp. 117784–117812, 2024a. (Not cited.)

H. Chen, H. Zhao, H. Lam, D. Yao, and W. Tang. MallowsPO: Fine-tune your LLM with preference dispersions. In *ICLR*, 2025a. URL https://openreview.net/forum?id=d8cnezVcaW. (Not cited.)

P. Chen, X. Chen, W. Yin, and T. Lin. ComPO: Preference alignment via comparison oracles. *ArXiv Preprint: 2505.05465*, 2025b. (Not cited.)

Q. Chen, L. Qin, J. Liu, D. Peng, J. Guan, P. Wang, M. Hu, Y. Zhou, T. Gao, and W. Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *ArXiv Preprint: 2503.09567*, 2025c. (Not cited.)

X. Chen, J. Xu, T. Liang, Z. He, J. Pang, D. Yu, L. Song, Q. Liu, M. Zhou, Z. Zhang, et al. Do not think that much for 2+3 = ? on the overthinking of o1-like LLMs. *ArXiv Preprint: 2412.21187*, 2024b. (Not cited.)

J. Cheng and B. Van Durme. Compressed chain of thought: Efficient reasoning through dense representations. *ArXiv Preprint: 2412.13171*, 2024. (Not cited.)

D. R. Chialvo and P. Bak. Learning from mistakes. *Neuroscience*, 90(4):1137–1148, 1999. (Not cited.)

A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. (Not cited.)

P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. In *NeurIPS*, pp. 4302–4310, 2017. (Not cited.)

X. Chu, H. Huang, X. Zhang, F. Wei, and Y. Wang. GPG: A simple and strong reinforcement learning baseline for model reasoning. *ArXiv Preprint: 2504.02546*, 2025. (Not cited.)

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *ArXiv Preprint: 2110.14168*, 2021. (Not cited.)

G. Cui, Y. Zhang, J. Chen, L. Yuan, Z. Wang, Y. Zuo, H. Li, Y. Fan, H. Chen, W. Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *ArXiv Preprint: 2505.22617*, 2025. (Not cited.)

P. Dayan. Reinforcement comparison. In *Connectionist Models*, pp. 45–51. Elsevier, 1991. (Not cited.)

B. Ding, C. Qin, L. Liu, Y. K. Chia, B. Li, S. Joty, and L. Bing. Is GPT-3 a good data annotator? In *ACL*, pp. 11173–11195, 2023. (Not cited.)

H. Dong, W. Xiong, D. Goyal, Y. Zhang, W. Chow, R. Pan, S. Diao, J. Zhang, K. Shum, and T. Zhang. RAFT: Reward ranked fine-tuning for generative foundation model alignment. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=m7p5O7zblY. (Not cited.)

H. Dong, W. Xiong, B. Pang, H. Wang, H. Zhao, Y. Zhou, N. Jiang, D. Sahoo, C. Xiong, and T. Zhang. RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a13aYUU9eU. (Not cited.)

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Model alignment as prospect theoretic optimization. In *ICML*, pp. 12634–12651, 2024. (Not cited.)

G. Feng, B. Zhang, Y. Gu, H. Ye, D. He, and L. Wang. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *NeurIPS*, pp. 70757–70798, 2023. (Not cited.)

K. Gandhi, D. Lee, G. Grand, M. Liu, W. Cheng, A. Sharma, and N. D. Goodman. Stream of search (SOS): Learning to search in language. *ArXiv Preprint: 2404.03683*, 2024. (Not cited.)

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *ICML*, pp. 10835–10866, 2023. (Not cited.)

F. Gilardi, M. Alizadeh, and M. Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. (Not cited.)

Z. Gou, Z. Shao, Y. Gong, Y. Shen, Y. Yang, M. Huang, N. Duan, and W. Chen. ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *ICLR*, 2024. URL https://openreview.net/forum?id=Ep0TtjVoap. (Not cited.)

Y. Gu, L. Dong, F. Wei, and M. Huang. MiniLLM: Knowledge distillation of large language models. In *ICLR*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ. (Not cited.)

D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *ArXiv Preprint: 2501.12948*, 2025a. (Not cited.)

J. Guo, Y. Wu, J. Qiu, K. Huang, X. Juan, L. Yang, and M. Wang. Temporal consistency for LLM reasoning process error identification. *ArXiv Preprint: 2503.14495*, 2025b. (Not cited.)

S. Hao, Y. Gu, H. Ma, J. Hong, Z. Wang, D. Wang, and Z. Hu. Reasoning with language model is planning with world model. In *EMNLP*, pp. 8154–8173, 2023. (Not cited.)

S. Hao, Y. Gu, H. Luo, T. Liu, X. Shao, X. Wang, S. Xie, H. Ma, A. Samavedhi, Q. Gao, Z. Wang, and Z. Hu. LLM reasoners: New evaluation, library, and analysis of step-by-step reasoning with large language models. In *COLM*, 2024a. URL https://openreview.net/forum?id=b0y6fbSUG0. (Not cited.)

S. Hao, S. Sukhbaatar, D. Su, X. Li, Z. Hu, J. Weston, and Y. Tian. Training large language models to reason in a continuous latent space. *ArXiv Preprint: 2412.06769*, 2024b. (Not cited.)

A. Havrilla, Y. Du, S. C. Raparthy, C. Nalmpantis, J. Dwivedi-Yu, M. Zhuravinskyi, E. Hambro, S. Sukhbaatar, and R. Raileanu. Teaching large language models to reason with reinforcement learning. *ArXiv Preprint: 2403.04642*, 2024. (Not cited.)

C. He, R. Luo, Y. Bai, S. Hu, Z. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, et al. OlympiadBench: A challenging benchmark for promoting AGI with Olympiad-level bilingual multimodal scientific problems. In *ACL*, pp. 3828–3850, 2024. (Not cited.)

T. He, R. Mu, L. Liao, Y. Cao, M. Liu, and B. Qin. Good learners think their thinking: Generative PRM makes large reasoning model more efficient math learner. *ArXiv Preprint: 2507.23317*, 2025. (Not cited.)

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe. (Not cited.)

J. Hong, N. Lee, and J. Thorne. ORPO: Monolithic preference optimization without reference model. In *EMNLP*, pp. 11170–11189, 2024. (Not cited.)

A. Hosseini, X. Yuan, N. Malkin, A. Courville, A. Sordoni, and R. Agarwal. V-STar: Training verifiers for self-taught reasoners. In *COLM*, 2024. URL https://openreview.net/forum?id=stmqBSW2dV. (Not cited.)

J. Huang and K. C-C. Chang. Towards reasoning in large language models: A survey. In *ACL*, pp. 1049–1065, 2023. (Not cited.)

K. Huang, J. Guo, Z. Li, X. Ji, J. Ge, W. Li, Y. Guo, T. Cai, H. Yuan, R. Wang, Y. Wu, M. Yin, S. Tang, Y. Huang, C. Jin, X. Chen, C. Zhang, and M. Wang. MATH-Perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations. *ArXiv Preprint: 2502.06453*, 2025. (Not cited.)

S. Huang, Z. Ma, J. Du, C. Meng, W. Wang, and Z. Lin. Mirror-consistency: Harnessing inconsistency in majority voting. In *EMNLP*, pp. 2408–2420, 2024. (Not cited.)

H. Hwang, D. Kim, S. Kim, S. Ye, and M. Seo. Self-explore: Enhancing mathematical reasoning in language models with fine-grained rewards. In *EMNLP*, pp. 1444–1466, 2024. (Not cited.)

A. Jaech, A. Kalai, A. Lerer, A. Richardson, A. El-Kishky, A. Low, A. Helyar, A. Madry, A. Beutel, A. Carney, et al. OpenAI o1 system card. *ArXiv Preprint: 2412.16720*, 2024. (Not cited.)

N. Jain, K. Han, A. Gu, W-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. In *ICLR*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL. (Not cited.)

M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. In *NeurIPS*, pp. 48573–48602, 2023. (Not cited.)

T. Khot, H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, and A. Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *ICLR*, 2023. URL https://openreview.net/forum?id=_nGgzQjzaRy. (Not cited.)

M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh. Reward design with language models. In *ICLR*, 2023. URL https://openreview.net/forum?id=10uNUgI5Kl. (Not cited.)

A. K. Lampinen, I. Dasgupta, S. C. Y. Chan, H. R. Sheahan, A. Creswell, D. Kumaran, J. L. McClelland, and F. Hill. Language models, like humans, show content effects on reasoning tasks. *PNAS Nexus*, 3(7): pgae233, 2024. (Not cited.)

Y. LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1): 1–62, 2022. (Not cited.)

H. Lee, S. Phatale, H. Mansoor, T. Mesnard, J. Ferret, K. Lu, C. Bishop, E. Hall, V. Carbune, A. Rastogi, et al. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *ICML*, pp. 26874–26901, 2024. (Not cited.)

L. Lehnert, S. Sukhbaatar, D. Su, Q. Zheng, P. McVay, M. Rabbat, and Y. Tian. Beyond $A^\star$: Better planning with transformers via search dynamics bootstrapping. In *COLM*, 2024. URL https://openreview.net/forum?id=SGoVIC0u0f. (Not cited.)

Z. Li, H. Liu, D. Zhou, and T. Ma. Chain of thought empowers transformers to solve inherently serial problems. In *ICLR*, 2024a. URL https://openreview.net/forum?id=3EWTEy9MTM. (Not cited.)

Z. Li, T. Xu, Y. Zhang, Z. Lin, Y. Yu, R. Sun, and Z-Q. Luo. ReMax: A simple, effective, and efficient reinforcement learning method for aligning large language models. In *ICML*, pp. 29128–29163, 2024b. (Not cited.)

Z. Li, C. Chen, T. Xu, Z. Qin, J. Xiao, Z-Q. Luo, and R. Sun. Preserving diversity in supervised fine-tuning of large language models. In *ICLR*, 2025. URL https://openreview.net/forum?id=NQEe7B7bSw. (Not cited.)

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *ICLR*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi. (Not cited.)

T. Liu, Y. Zhao, R. Joshi, M. Khalman, M. Saleh, P. J. Liu, and J. Liu. Statistical rejection sampling improves preference optimization. In *ICLR*, 2024a. URL https://openreview.net/forum?id=xbjSwwrQOe. (Not cited.)

T. Liu, Z. Qin, J. Wu, J. Shen, M. Khalman, R. Joshi, Y. Zhao, M. Saleh, S. Baumgartner, J. Liu, et al. LiPO: Listwise preference optimization through learning-to-rank. In *NAACL*, pp. To appear, 2025a. (Not cited.)

Z. Liu, M. Lu, S. Zhang, B. Liu, H. Guo, Y. Yang, J. Blanchet, and Z. Wang. Provably mitigating overoptimization in RLHF: Your SFT loss is implicitly an adversarial regularizer. In *NeurIPS*, pp. 138663–138697, 2024b. (Not cited.)

Z. Liu, C. Chen, W. Li, P. Qi, T. Pang, C. Du, W. S. Lee, and M. Lin. Understanding R1-zero-like training: A critical perspective. *ArXiv Preprint: 2503.20783*, 2025b. (Not cited.)

H. Luo, L. Shen, H. He, Y. Wang, S. Liu, W. Li, N. Tan, X. Cao, and D. Tao. o1-pruner: Length-harmonizing fine-tuning for o1-like reasoning pruning. *ArXiv Preprint: 2501.12570*, 2025a. (Not cited.)

L. Luo, Y. Liu, R. Liu, S. Phatale, M. Guo, H. Lara, Y. Li, L. Shu, Y. Zhu, L. Meng, et al. Improve mathematical reasoning in language models by automated process supervision. *ArXiv Preprint: 2406.06592*, 2024. (Not cited.)

M. Luo, S. Tan, J. Wong, X. Shi, W. Y. Tang, M. Roongta, C. Cai, J. Luo, L. E. Li, R. A. Popa, and I. 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, 2025b. Notion Blog. (Not cited.)

A. Madaan, K. Hermann, and A. Yazdanbakhsh. What makes chain-of-thought prompting effective? a counterfactual study. In *EMNLP*, pp. 1448–1535, 2023. (Not cited.)

J. Mei, Y. Gao, B. Dai, C. Szepesvari, and D. Schuurmans. Leveraging non-uniformity in first-order non-convex optimization. In *ICML*, pp. 7555–7564, 2021. (Not cited.)

Y. Meng, M. Xia, and D. Chen. SimPO: Simple preference optimization with a reference-free reward. In *NeurIPS*, pp. 124198–124235, 2024. (Not cited.)

W. Merrill and A. Sabharwal. The expressive power of transformers with chain of thought. In *ICLR*, 2024. URL https://openreview.net/forum?id=NjNGlPh8Wh. (Not cited.)

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. In *NeurIPS*, pp. 27730–27744, 2022. (Not cited.)

A. Pal, D. Karkhanis, S. Dooley, M. Roberts, S. Naidu, and C. White. Smaug: Fixing failure modes of preference optimisation with DPO-positive. *ArXiv Preprint: 2402.13228*, 2024. (Not cited.)

R. Y. Pang, W. Yuan, H. He, K. Cho, S. Sukhbaatar, and J. Weston. Iterative reasoning preference optimization. In *NeurIPS*, pp. 116617–116637, 2024. (Not cited.)

R. Park, R. Rafailov, S. Ermon, and C. Finn. Disentangling length from quality in direct preference optimization. In *ACL*, pp. 4998–5017, 2024. (Not cited.)

X. Peng, A. Kumar, G. Zhang, and S. Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *ArXiv Preprint: 1910.00177*, 2019. (Not cited.)

J. Peters and S. Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *ICML*, pp. 745–750, 2007. (Not cited.)

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, pp. 53728–53741, 2023. (Not cited.)

R. Rafailov, J. Hejna, R. Park, and C. Finn. From r to q^*: Your language model is secretly a Q-function. In *COLM*, 2024. URL https://openreview.net/forum?id=kEVcNxtqXk. (Not cited.)

N. Razin, S. Malladi, A. Bhaskar, D. Chen, S. Arora, and B. Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *ICLR*, 2025. URL https://openreview.net/forum?id=uaMSBJDnRv. (Not cited.)

P. Roit, J. Ferret, L. Shani, R. Aharoni, G. Cideron, R. Dadashi, M. Geist, S. Girgin, L. Hussenot, O. Keller, et al. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *ACL*, pp. 6252–6272, 2023. (Not cited.)

J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel. Trust region policy optimization. In *ICML*, pp. 1889–1897, 2015. (Not cited.)

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *ArXiv Preprint: 1707.06347*, 2017. (Not cited.)

A. Setlur, C. Nagpal, A. Fisch, X. Geng, J. Eisenstein, R. Agarwal, A. Agarwal, J. Berant, and A. Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *ICLR*, 2025. URL https://openreview.net/forum?id=A6Y7AqlzLW. (Not cited.)

Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. RL on incorrect synthetic data scales the efficiency of LLM math reasoning by eight-fold. In *NeurIPS*, pp. 43000–43031, 2024. (Not cited.)

Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *ArXiv Preprint: 2402.03300*, 2024. (Not cited.)

G. Sheng, C. Zhang, Z. Ye, X. Wu, W. Zhang, R. Zhang, Y. Peng, H. Lin, and C. Wu. HybridFlow: A flexible and efficient RLHF framework. In *EuroSys*, pp. 1279–1297. ACM, 2025. (Not cited.)

B. Shi, M. Tang, K. R. Narasimhan, and S. Yao. Can language models solve Olympiad programming? In *COLM*, 2024. URL https://openreview.net/forum?id=kGa4fMtP9l. (Not cited.)

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. (Not cited.)

A. Singh, J. D. Co-Reyes, R. Agarwal, A. Anand, P. Patil, X. Garcia, P. J. Liu, J. Harrison, J. Lee, K. Xu, A. T. Parisi, A. Kumar, A. A. Alemi, A. Rizkowsky, A. Nova, B. Adlam, B. Bohnet, G. F. Elsayed, H. Sedghi, I. Mordatch, I. Simpson, I. Gur, J. Snoek, J. Pennington, J. Hron, K. Kenealy, K. Swersky, K. Mahajan, L. A. Culp, L. Xiao, M. Bileschi, N. Constant, R. Novak, R. Liu, T. Warkentin, Y. Bansal, E. Dyer, B. Neyshabur, J. Sohl-Dickstein, and N. Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=lNAyUngGFK. Expert Certification. (Not cited.)

I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg. ProgPrompt: Generating situated robot task plans using large language models. In *ICRA*, pp. 11523–11530. IEEE, 2023. (Not cited.)

J. Skalse, N. H. R. Howe, D. Krasheninnikov, and D. Krueger. Defining and characterizing reward hacking. In *NeurIPS*, pp. 9460–9471, 2022. (Not cited.)

C. V. Snell, J. Lee, K. Xu, and A. Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *ICLR*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n. (Not cited.)

F. Song, B. Yu, M. Li, H. Yu, F. Huang, Y. Li, and H. Wang. Preference ranking optimization for human alignment. In *AAAI*, pp. 18990–18998, 2024. (Not cited.)

S. Srivastava, A. PV, S. Menon, A. Sukumar, A. Philipose, S. Prince, S. Thomas, et al. Functional benchmarks for robust evaluation of reasoning performance, and the reasoning gap. *ArXiv Preprint: 2402.19450*, 2024. (Not cited.)

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. In *NeurIPS*, pp. 3008–3021, 2020. (Not cited.)

D. Su, S. Sukhbaatar, M. Rabbat, Y. Tian, and Q. Zheng. Dualformer: Controllable fast and slow thinking by learning with randomized reasoning traces. In *ICLR*, 2025. URL https://openreview.net/forum?id=bmbRCRiNDu. (Not cited.)

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*, volume 1. MIT Press, 1998. (Not cited.)

F. Tajwar, A. Singh, A. Sharma, R. Rafailov, J. Schneider, T. Xie, S. Ermon, C. Finn, and A. Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. In *ICML*, pp. 47441–47474, 2024. (Not cited.)

Y. Tang, Z. Guo, Z. Zheng, D. Calandriello, R. Munos, M. Rowland, P. H. Richemond, M. Valko, B. Pires, and B. Piot. Generalized preference optimization: A unified approach to offline alignment. In *ICML*, pp. 47725–47742, 2024. (Not cited.)

K. Team, A. Du, B. Gao, B. Xing, C. Jiang, C. Chen, C. Li, C. Xiao, C. Du, C. Liao, et al. Kimi k1.5: Scaling reinforcement learning with LLMs. *ArXiv Preprint: 2501.12599*, 2025. (Not cited.)

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *ArXiv Preprint: 2302.13971*, 2023. (Not cited.)

J. Uesato, N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins. Solving math word problems with process-and outcome-based feedback. *ArXiv Preprint: 2211.14275*, 2022. (Not cited.)

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, pp. 6000–6010, 2017. (Not cited.)

H. Wang, C. Qian, W. Zhong, X. Chen, J. Qiu, S. Huang, B. Jin, M. Wang, K-F. Wong, and H. Ji. OTC: Optimal tool calls via reinforcement learning. *ArXiv Preprint: 2504.14870*, 2025. (Not cited.)

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *ACL*, pp. 9426–9439, 2024. (Not cited.)

X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw. (Not cited.)

J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, pp. 24824–24837, 2022. (Not cited.)

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992. (Not cited.)

Y. Wu, Z. Sun, S. Li, S. Welleck, and Y. Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving. In *ICLR*, 2025. URL https://openreview.net/forum?id=VNckp7JEHn. (Not cited.)

V. Xiang, C. Snell, K. Gandhi, A. Albalak, A. Singh, C. Blagden, D. Phung, R. Rafailov, N. Lile, D. Mahan, et al. Towards system 2 reasoning in LLMs: Learning how to think with meta chain-of-though. *ArXiv Preprint: 2501.04682*, 2025. (Not cited.)

J. Xiao, Z. Li, X. Xie, E. Getzen, C. Fang, Q. Long, and W. J. Su. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *ArXiv Preprint: 2405.16455*, 2024. (Not cited.)

Y. Xie, K. Kawaguchi, Y. Zhao, J. X. Zhao, M-Y. Kan, J. He, and M. Q. Xie. Self-evaluation guided beam search for reasoning. In *NeurIPS*, pp. 41618–41650, 2023. (Not cited.)

W. Xiong, H. Dong, C. Ye, Z. Wang, H. Zhong, H. Ji, N. Jiang, and T. Zhang. Iterative preference learning from human feedback: Bridging theory and practice for RLHF under KL-constraint. In *ICML*, pp. 54715–54754, 2024. (Not cited.)

W. Xiong, J. Yao, Y. Xu, B. Pang, L. Wang, D. Sahoo, J. Li, N. Jiang, T. Zhang, C. Xiong, and H. Dong. A minimalist approach to LLM reasoning: From rejection sampling to reinforce. *ArXiv Preprint: 2504.11343*, 2025a. (Not cited.)

W. Xiong, W. Zhao, W. Yuan, O. Golovneva, T. Zhang, J. Weston, and S. Sukhbaatar. StepWiser: Stepwise generative judges for wiser reasoning. *ArXiv Preprint: 2508.19229*, 2025b. (Not cited.)

H. Xu, A. Sharaf, Y. Chen, W. Tan, L. Shen, B. Van Durme, K. Murray, and Y. J. Kim. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. In *ICML*, pp. 55204–55224, 2024. (Not cited.)

Y. Yang, D. Campbell, K. Huang, M. Wang, J. Cohen, and T. Webb. Emergent symbolic mechanisms support abstract reasoning in large language models. *ArXiv Preprint: 2502.20332*, 2025. (Not cited.)

Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. (Not cited.)

S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, pp. 11809–11822, 2023. (Not cited.)

Y. Ye, Z. Huang, Y. Xiao, E. Chern, S. Xia, and P. Liu. LIMO: Less is more for reasoning. *ArXiv Preprint: 2502.03387*, 2025. (Not cited.)

F. Yu, A. Gao, and B. Wang. OVM, Outcome-supervised value models for planning in mathematical reasoning. In *NAACL*, pp. 858–875, 2024a. (Not cited.)

F. Yu, L. Jiang, H. Kang, S. Hao, and L. Qin. Flow of reasoning: Efficient training of LLM policy with divergent thinking. In *ICML*, pp. To appear, 2025a. (Not cited.)

L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. Kwok, Z. Li, A. Weller, and W. Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *ICLR*, 2024b. URL https://openreview.net/forum?id=N8N0hgNDRt. (Not cited.)

Q. Yu, Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, T. Fan, G. Liu, L. Liu, X. Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *ArXiv Preprint: 2503.14476*, 2025b. (Not cited.)

H. Yuan, Z. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang. RRHF: Rank responses to align language models with human feedback. In *NeurIPS*, pp. 10935–10950, 2023a. (Not cited.)

L. Yuan, G. Cui, H. Wang, N. Ding, X. Wang, B. Shan, Z. Liu, J. Deng, H. Chen, R. Xie, Y. Lin, Z. Liu, B. Zhou, H. Peng, Z. Liu, and M. Sun. Advancing LLM reasoning generalists with preference trees. In *ICLR*, 2025. URL https://openreview.net/forum?id=2ea5TNVR0c. (Not cited.)

Z. Yuan, H. Yuan, C. Li, G. Dong, K. Lu, C. Tan, C. Zhou, and J. Zhou. Scaling relationship on learning mathematical reasoning with large language models. *ArXiv Preprint: 2308.01825*, 2023b. (Not cited.)

X. Yue, X. Qu, G. Zhang, Y. Fu, W. Huang, H. Sun, Y. Su, and W. Chen. MAmmoTH: Building math generalist models through hybrid instruction tuning. In *ICLR*, 2024. URL https://openreview.net/forum?id=yLClGs770I. (Not cited.)

E. Zelikman, Y. Wu, J. Mu, and N. D. Goodman. STaR: self-taught reasoner bootstrapping reasoning with reasoning. In *NeurIPS*, pp. 15476–15488, 2022. (Not cited.)

K. Zha, Z. Gao, M. Shen, Z-W. Hong, D. S. Boning, and D. Katabi. RL Tango: Reinforcing generator and verifier together for language reasoning. *ArXiv Preprint: 2505.15034*, 2025. (Not cited.)

D. Zhang, S. Zhoubian, Z. Hu, Y. Yue, Y. Dong, and J. Tang. ReST-MCTS*: LLM self-training via process reward guided tree search. In *NeurIPS*, pp. 64735–64772, 2024a. (Not cited.)

K. Zhang, Y. Hong, J. Bao, H. Jiang, Y. Song, D. Hong, and H. Xiong. GVPO: Group variance policy optimization for large language model post-training. *ArXiv Preprint: 2504.19599*, 2025a. (Not cited.)

L. Zhang, A. Hosseini, H. Bansal, M. Kazemi, A. Kumar, and R. Agarwal. Generative verifiers: Reward modeling as next-token prediction. In *The NeurIPS Workshop on System-2 Reasoning at Scale*, 2024b. URL https://openreview.net/forum?id=aLgXy8A7k7. (Not cited.)

Z. Zhang, C. Zheng, Y. Wu, B. Zhang, R. Lin, B. Yu, D. Liu, J. Zhou, and J. Lin. The lessons of developing process reward models in mathematical reasoning. *ArXiv Preprint: 2501.07301*, 2025b. (Not cited.)

H. Zhao, G. I. Winata, A. Das, S-X. Zhang, D. Yao, W. Tang, and S. Sahu. RainbowPO: A unified framework for combining improvements in preference optimization. In *ICLR*, 2025. URL https://openreview.net/forum?id=trKee5pIFv. (Not cited.)

Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *ArXiv Preprint: 2305.10425*, 2023. (Not cited.)

D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le, and E. H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*, 2023. URL https://openreview.net/forum?id=WZH7099tgfM. (Not cited.)

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *ArXiv Preprint: 1909.08593*, 2019. (Not cited.)

A Related Works

We comment on all related topics, including reasoning through test-time compute, chain-of-thought and its variants, direct preference alignment methods, reward models and reinforcement learning from AI feedback. For an overview of more reasoning models and methods, we refer to two recent surveys (Huang & Chang, 2023; Chen et al., 2025c).

Reasoning through test-time compute. OpenAI-o1 (Jaech et al., 2024) is among the first large-scale applications of RL to reasoning, and achieved state-of-the-art performance upon release. Following this trend, DeepSeek-R1 (Guo et al., 2025a) is the first open-weight model to match or exceed OpenAI-o1. Their real-world success stories have involved several simple yet novel techniques that enhance LLM reasoning through more test-time compute, including chain-of-thought (Wei et al., 2022), self-consistency (Wang et al., 2023), best-of-N sampling (Snell et al., 2025), process reward models (Lightman et al., 2024), Monte Carlo tree search (Silver et al., 2016; Hao et al., 2023), tree-of-thought (Yao et al., 2023), and recent works on preventing overthinking (Chen et al., 2024b; Team et al., 2025; Luo et al., 2025a; Arora & Zanette, 2025) and compressing chain-of-thought (Hao et al., 2024b; Cheng & Van Durme, 2024). More specifically, *chain-of-thought* is a reasoning approach where intermediate steps are explicitly written to make complex problem-solving processes more transparent and logical. *Self-consistency* suggests generating multiple final answers and returning the mode of an empirical distribution, enhancing test-time performance when test-time verifiers are unavailable. Unfortunately, it is computationally expensive and effective only when answers can be clustered. *Best-of-N sampling* resolves this issue by sampling answers from the model and selecting the best at test time according to the scoring function; however, it is sensitive to the accuracy of test-time scoring functions (Gao et al., 2023). *Process reward models* offer fine-grained supervision of chain-of-thought reasoning, but they might be vulnerable to reward hacking and introduce computation overhead. *Monte Carlo tree search* is a generic technique that allocates computational resources toward the most promising regions of the search space, and *tree-of-thought* and its extension (Besta et al., 2024; Gandhi et al., 2024) simplified this idea by exploring multiple reasoning paths in a specific structure, allowing language models to select the most promising line of thought for complex problem-solving. Both *length regularization* and *compressed chain-of-thought* are developed to reduce inference costs for reasoning, which is crucial for the economic feasibility, user experience and environmental sustainability of LLMs. In addition, several works have focused on specific reasoning tasks (Lampinen et al., 2024; Yang et al., 2025; Srivastava et al., 2024; Huang et al., 2025; 2024; Guo et al., 2025b; Gou et al., 2024; Wang et al., 2025), demonstrating promising performance. The recent findings Xiong et al. (2025a) have shown that the REINFORCE-type methods (including GRPO (Shao et al., 2024)) can not effectively learn from all-negative-sample groups. Our work alleviates this issue by leveraging AI feedback to differentiate negative samples. We also provide a theoretical analysis through a stylized model, explaining why such diversification improves GRPO's learning dynamics.

Chain-of-Thought and its variants. Chain-of-thought (CoT) refers to as a broad class of methods that generate an intermediate reasoning process before arriving at a final answer. These approaches either prompt LLMs (Wei et al., 2022; Khot et al., 2023; Zhou et al., 2023) or train LLMs to generate reasoning chains through supervised fine-tuning (SFT) (Yue et al., 2024; Yu et al., 2024b; Li et al., 2025) and/or RL (Wang et al., 2024; Shao et al., 2024; Havrilla et al., 2024; Yu et al., 2025a). While CoT has proven effective for certain tasks, its auto-regressive generation nature makes it challenging to mimic human reasoning on more complex problems (LeCun, 2022; Hao et al., 2023), which require planning and search. Recent efforts were devoted to equipping LLMs with tree search methods (Xie et al., 2023; Yao et al., 2023; Hao et al., 2024a) or training LLMs on search trajectories (Lehnert et al., 2024; Gandhi et al., 2024; Su et al., 2025). Several other works have investigated why CoT is effective. For example, (Madaan et al., 2023) used a counterfactual prompting approach to examine the relative contributions of prompt elements, including symbols (digits, entities) and patterns (equations). (Feng et al., 2023; Merrill & Sabharwal, 2024; Li et al., 2024a) analyzed CoT from the perspective of model expressivity, and (Feng et al., 2023) showed that employing CoT increases the effective depth of a transformer since the generated outputs are looped back to the input. This insight motivated the chain-of-continuous-thought paradigm (Hao et al., 2024b), and a related approach has been proposed in (Cheng & Van Durme, 2024).

Direct preference alignment methods. These methods (e.g., DPO (Rafailov et al., 2023)) are simple and stable offline alternatives to online RLHF. Various DPO variants with other objectives have been proposed, including ranking ones beyond pairwise preference data (Dong et al., 2023; Yuan et al., 2023a; Song et al., 2024; Chen et al., 2024a; Liu et al., 2025a) and simple ones that do not rely on a reference model (Hong et al., 2024; Meng et al., 2024). Since DPO does not train a reward model, the limited size of human labels becomes a bottleneck. To alleviate this limitation, subsequent works proposed to augment preference data using a trained SFT policy (Zhao et al., 2023) or a refined SFT policy with rejection sampling (Liu et al., 2024a). The DPO loss was recently extended to token-level MDP (Rafailov et al., 2024) given that the transition is deterministic – which has covered the fine-tuning of LLMs – and more general RL problems (Azar et al., 2024). There are other DPO variants (Ethayarajh et al., 2024; Park et al., 2024; Xu et al., 2024; Tang et al., 2024; Meng et al., 2024; Chen et al., 2025a; Zhao et al., 2025). For example, (Ethayarajh et al., 2024) designed the specific loss using a prospect theory, (Tang et al., 2024) optimized a general preference loss instead of the log-likelihood loss, and (Meng et al., 2024) aligned the reward function in the preference optimization objective with the generation metric. Dong et al. (2024) and (Xiong et al., 2024) proposed to generate human feedback in an online fashion to mitigate the distribution-shift and over-parameterization phenomenon. This improves DPO for complex reasoning tasks (Pang et al., 2024). Several other works focus on *unintentional alignment* of DPO and developing new methods (Pal et al., 2024; Tajwar et al., 2024; Liu et al., 2024b; Xiao et al., 2024; Yuan et al., 2025; Razin et al., 2025; Chen et al., 2025b). Among these works, (Razin et al., 2025) proposed to measure the similarity between preferred and dispreferred responses using the centered hidden embedding similarity (CHES) score and showed that filtering out preference pairs with small CHES score improves DPO, while (Chen et al., 2025b) proposed to use comparison oracles, and showed that combining it with DPO effectively alleviated the issue of unintentional alignment.

Reward models. For the prompt $\mathbf{x}$ with a ground-truth response $\mathbf{y}^\star_{\mathbf{x}}$, we evaluate by implementing a regular expression match on the final answer (Hendrycks et al., 2021): $r(\mathbf{x}, \mathbf{y}) = 1$ if $\mathbf{y}$ matches $\mathbf{y}^\star_{\mathbf{x}}$ on the *final answer* and $r(\mathbf{x}, \mathbf{y}) = 0$ otherwise. An *outcome reward* model (ORM) (Cobbe et al., 2021; Uesato et al., 2022) is trained for estimating $r(\mathbf{x}, \mathbf{y})$. In particular, we first choose $\mathbf{x} \in \mathcal{D}$ and collect training samples $(\mathbf{x}, \mathbf{y} \sim \pi_\theta(\cdot|\mathbf{x}), r(\mathbf{x}, \mathbf{y}))$. Then, we take $(\mathbf{x}, \mathbf{y})$ as input and train an ORM to predict $r(\mathbf{x}, \mathbf{y})$. This can be done using binary classification (Cobbe et al., 2021; Yu et al., 2024a), direct preference optimization (Hosseini et al., 2024) or next-token prediction (Zhang et al., 2024b). Previous works also train LLMs on self-generated data using the ground-truth outcome reward model with either supervised fine-tuning (Singh et al., 2024; Yuan et al., 2023b; Zelikman et al., 2022) or online RL (Bi et al., 2024; Guo et al., 2025a). A *process reward* model (PRM) is trained to score a_h at $\mathbf{s}_h = (\mathbf{x}, a_1, \ldots, a_{h-1})$ either using human annotations (Lightman et al., 2024) or the value functions based on LLM-generated data (Wang et al., 2024; Luo et al., 2024; Setlur et al., 2025); indeed, PRMs estimate either the likelihood of future success or the change in the likelihood of future success before and after taking a_h. In addition, PRMs were also developed to improve search methods (Snell et al., 2025; Wu et al., 2025), and to identify the "first pit" in an incorrect reasoning trajectory to construct preference pairs for direct preference alignment (Hwang et al., 2024; Setlur et al., 2024).

Reinforcement learning from AI feedback. Reinforcement learning from human feedback (RLHF) uses human-preference-aligned reward models to evaluate response quality (Christiano et al., 2017; Ziegler et al., 2019; Stiennon et al., 2020; Ouyang et al., 2022). A key barrier to scale RLHF is the need for high-quality human labels. Previous studies (Gilardi et al., 2023; Ding et al., 2023) have shown that modern LLMs exhibit strong alignment with human judgments, suggesting that AI-generated labels can serve as a viable alternative. In this context, (Bai et al., 2022) was the first to explore RLAIF, jointly optimizing helpfulness and harmlessness using both human and AI-generated labels, and (Roit et al., 2023; Kwon et al., 2023; Lee et al., 2024) showed that LLMs can produce informative reward signals for RL post-training. Our approach can leverage AI feedback to introduce response diversity within all-negative-sample groups by assigning intermediate binary rewards to reasoning steps. Indeed, one identifies the proportion of correct steps in the reasoning trajectory and use it to compute a reward $r_i \in [0, 1)$.

B Missing Proofs

We provide several technical lemmas that are important to the proof of Theorem 3.3. Indeed, the first lemma summarizes the properties of particular functions related to the aforementioned functions f_{11}, f_{21}, f_{12} and f_{22} from Eq. (4).

Lemma B.1. *The following statements hold true,*

(i) The function f_{11} is strictly increasing on $(0,1)$.
(ii) The function $h_p(x) := x - \log(1-p+pe^x)$ is strictly increasing for any fixed $p \in (0,1)$.
(iii) The function f_{21} is strictly increasing in either p for any fixed q or q for any fixed p on $(0,1)$.
(iv) The function $\varphi(x) := \log(1+(1/2)e^{-e^x})$ is strictly concave on $(-\infty, 0)$.

Proof. First of all, we have

$$f_{11}'(p_1) = \frac{1+(1-2p_1)p_1(1-p_1)}{p_1(1+p_1+p_1e^{p_1(1-p_1)})} > \frac{3}{4p_1(1+p_1+p_1e^{p_1(1-p_1)})} > 0.$$

Thus, the function f_{11} is strictly increasing on $(0,1)$.

Furthermore, we have

$$h_p'(x) = 1 - \frac{pe^x}{1-p+pe^x} = \frac{1-p}{1-p+pe^x} \overset{0<p<1}{>} 0.$$

Thus, the function $h_p(x)$ is strictly increasing.

Moreover, we have

$$\begin{aligned} \frac{\partial f_{21}(p_1,p_2)}{p_1} &= \frac{1+p_2(1-2p_1)p_1(1-p_1)}{p_1(1+p_1+p_1e^{p_2p_1(1-p_1)})} > \frac{3}{4p_1(1+p_1+p_1e^{p_1(1-p_1)})} > 0, \\ \frac{\partial f_{21}(p_1,p_2)}{p_2} &= \frac{p_1(1-p_1)^2}{1+p_1+p_1e^{p_2p_1(1-p_1)}} > 0. \end{aligned}$$

Thus, the function f_{21} is strictly increasing in either p for any fixed q or q for any fixed p on $(0,1)$.

Finally, we have

$$\varphi''(x) = \frac{(e^{x+e^x}/2 - e^{e^x}/2 - 1/4)e^x}{e^{2e^x}+e^{e^x}+1/4}.$$

Since $u = e^x \in (0,1)$ for $x < 0$, we have

$$(ue^u/2 - e^u/2 - 1/4)u = (e^u(u-1)/2 - 1/4)u < -(1/4)u < 0.$$

Thus, $\varphi''(x) < 0$ for all $x < 0$ which shows that f is strictly concave on $(-\infty, 0)$. □

The second lemma presents an inequality which plays a key role in the proof of Theorem 3.3.

Lemma B.2. *We define the auxiliary functions as follows,*

$$A(x) = 1 + \left(\tfrac{1}{x} - 1\right) e^{-x(1-x)},\ B(x,y) = 1 + \left(\tfrac{1}{y} - 1\right) e^{-x^2y(1-y)},$$
$$C(z) = 1 + \left(\tfrac{1}{z} - 1\right) e^{-z^2(1-z)}.$$

Then, we have $C(\sqrt{xy})^2 > A(x)B(x,y)$ for all x and y satisfying $1/2 < y < x < 1$.

Proof. We consider the lower and upper bound of e^{-u} when $u > 0$:

$$1 - u + \frac{u^2}{2} - \frac{u^3}{6} < e^{-u} < 1 - u + \frac{u^2}{2}.$$

Since $1/x - 1$, $1/y - 1$ and $1/\sqrt{xy} - 1$ are all positive, we have

$$A(x) \le 1 + \frac{1-x}{x}\left(1 - x(1-x) + \frac{x^2(1-x)^2}{2}\right) = \frac{1}{x} - (1-x)^2 + \frac{x(1-x)^3}{2}.$$

$$B(x,y) \le 1 + \tfrac{1-y}{y}\left(1 - x^2y(1-y) + \tfrac{x^4y^2(1-y)^2}{2}\right) = \tfrac{1}{y} - x^2(1-y)^2 + \tfrac{x^4y(1-y)^3}{2}.$$

$$\begin{aligned} C(z) &\ge 1 + \tfrac{1-z}{z}\left(1 - z^2(1-z) + \tfrac{z^4(1-z)^2}{2} - \tfrac{z^6(1-z)^3}{6}\right) \\ &= \tfrac{1}{z} - z(1-z)^2 + \tfrac{z^3(1-z)^3}{2} - \tfrac{z^5(1-z)^4}{6}. \end{aligned}$$

Set $z^2 = xy$, the original statement is equivalent to $(zC(z))^2 > (xA(x))(yB(x,y))$. Using the above upper and lower bound, it suffices to show $C_1(\sqrt{xy})^2 > A_1(x)B_1(x,y)$ where

$$A_1(x) = 1 - x(1-x)^2 + x^2(1-x)^3/2,\ B_1(x,y) = 1 - x^2y(1-y)^2 + x^4y^2(1-y)^3/2,$$
$$C_1(z) = 1 - z^2(1-z)^2 + z^4(1-z)^3/2 - z^6(1-z)^4/6.$$

By Lemma B.3, this is indeed true. This completes the proof. □

Lemma B.3. *Define functions*

$$A_1(x) = 1 - x(1-x)^2 + x^2(1-x)^3/2,\ B_1(x,y) = 1 - x^2y(1-y)^2 + x^4y^2(1-y)^3/2,$$
$$C_1(z) = 1 - z^2(1-z)^2 + z^4(1-z)^3/2 - z^6(1-z)^4/6.$$

Then, $C_1(\sqrt{xy})^2 > A_1(x)B_1(x,y)$ *for all* $1/2 < y < x < 1$.

Proof. Let $x = u^2$ and $y = v^2$, then $1 > u > v > 1/\sqrt{2}$ and $z = uv$. We next show the desired inequality holds on a larger region, i.e., $1 > u > v > 2/3$. On this larger region, we have the reparameterization as follows:

$$u = \tfrac{2s+3}{3s+3}, \quad v = \tfrac{2r+2s+3}{3r+3s+3}, \quad s, r \in (0, +\infty),$$

or equivalently,

$$s = \tfrac{3(1-u)}{3u-2}, \quad r = \tfrac{3(u-v)}{(3u-2)(3v-2)}, \quad 1 > u > v > \tfrac{2}{3}.$$

It is easy to see this defines a one-to-one correspondence from (u,v)-space to (s,r)-space. Thus, we aim to prove the following function f is positive:

$$F(s,r) := C_1\left(\tfrac{2s+3}{3s+3}\cdot\tfrac{2r+2s+3}{3r+3s+3}\right)^2 - A_1\left(\left(\tfrac{2s+3}{3s+3}\right)^2\right)B_1\left(\left(\tfrac{2s+3}{3s+3}\right)^2, \left(\tfrac{2r+2s+3}{3r+3s+3}\right)^2\right).$$

By leveraging Sympy's symbolic engine, the function expands and simplifies to:

$$F(s,r) = \tfrac{f(s,r)}{c(s+1)^{20}(r+s+1)^{20}},\ \text{where } f(s,r) := \sum_{k=0}^{20} c_{20-k}(s)r^{20-k},$$

where $c > 0$ is a universal constant, and single-variable polynomials $c_{20}(s), \ldots, c_2(s), c_0(s) > 0$ and $\Delta_2 := c_1(s)^2 - 4c_2(s)c_0(s) < 0$, for all $s > 0$ (see Table 6 for details). Notice that from the table, we can see the only nontrivial parts are $c_3(s) > 0$ and $c_2(s) > 0$ because only these two contain negative coefficients. The positivity of $c_3(s)$ is simple because there is only one term (s^9) with negative coefficient and for all $s > 0$,

$$194714567104543630051526 64s^{10} + 968458837773164307192723 6s^8 > 144138231093502245414997 26s^9.$$

To see this, simple estimation and AM-GM inequality yield

$$\text{LHS} > 1.9 \times 10^{25}s^{10} + 9.6 \times 10^{24}s^8 > 2\sqrt{182.4} \times 10^{24}s^9 > 2.7 \times 10^{25}s^9 > \text{RHS}.$$

The positivity of $c_2(s)$ is more complicated because it has 4 negative terms s^{10}, s^9, s^8, s^7. However, we can use similar idea, i.e., choosing a pair of positive terms to bound a negative term:

$$957910627865555087240887423 20s^{15} + 571809550541807937530952s^5 > 702735952364323683297077 16s^{10};$$
$$437858623301624990522095297 68s^{14} + 184789343789534461150530s^4 > 998540727043228715373926 04s^9;$$
$$163267368535271229917151558 24s^{13} + 354883755696224721692 40s^3 > 357260313779697920881889 25s^8;$$
$$460821905008432679074893315 3s^{12} + 436295085881317044922 8s^2 > 609903789530767014228760 8s^7.$$

To see this, simple estimation and AM-GM inequality yield

$$\begin{aligned}
&\text{LHS} > 9.5 \times 10^{28} s^{15} + 5.7 \times 10^{23} s^5 > 2\sqrt{541.5} \times 10^{25} s^{10} > 4.6 \times 10^{26} s^{10} > \text{RHS};\\
&\text{LHS} > 4.3 \times 10^{28} s^{14} + 1.8 \times 10^{23} s^4 > 2\sqrt{77.4} \times 10^{25} s^9 > 1.7 \times 10^{26} s^{10} > \text{RHS};\\
&\text{LHS} > 1.6 \times 10^{28} s^{13} + 3.5 \times 10^{22} s^3 > 2\sqrt{5.6} \times 10^{25} s^9 > 4.7 \times 10^{25} s^8 > \text{RHS};\\
&\text{LHS} > 4.6 \times 10^{27} s^{12} + 4.3 \times 10^{21} s^2 > 2\sqrt{19.7} \times 10^{24} s^7 > 8.8 \times 10^{24} s^7 > \text{RHS}.
\end{aligned}$$

In conclusion, we have all coefficient $c_i(s)$ positive except $c_1(s)$, but it doesn't affect the positivity of f because $\Delta_2 < 0$. This completes the proof. $\square$

Table 6: Coefficient Lists of $F(s,r)$

Notation	Value
c	437675956526049436836
c_{20}	$2(2s+3)^2(1714774320744848750s^{18}+26610409260691576200s^{17}+191777874646880231718 1s^{16}+850194149855082319224s^{15}+2587082434290045806049s^{14}+5704415906039160731874s^{13}+9366197581963232054460s^{12}+11563054951307567026248s^{11}+10670965452123886149660s^{10}+7187176769582075261292s^{9}+3372972168996579430017s^{8}+1072082836158220703952s^{7}+370302094042890771285s^{6}+329901677376902425818s^{5}+282799986805616267862s^{4}+151168270170893365008s^{3}+49139849518345513368s^{2}+9090603727935062976s+742484948385838248)$
c_{19}	$8(2s+3)^2(8573871603724243750s^{19}+14118375184879884045 0s^{18}+10850654575356110840 97s^{17}+5159905424662678527663s^{16}+1696201782101404135528 5s^{15}+407615158929069302613 93s^{14}+73777358861762677983126s^{13}+10196816347629194227764 3s^{12}+10771955018327920294533 6s^{11}+8595187100533224794234 7s^{10}+5047714242076387274703 9s^{9}+2122846171048422727081 2s^{8}+717557625336028620219 3s^{7}+382164211944790881013 8s^{6}+334314060253961507098 2s^{5}+230805874138031094614 4s^{4}+104652869156562147134 4s^{3}+30076928586074414602 8s^{2}+5040301464344059293 6s+3785769852305984190)$
c_{18}	$2(2s+3)^2(32580712094152126250 0s^{20}+56739873809773123962 00s^{19}+46320143614200183575358s^{18}+23516462486845543431474 0s^{17}+83033078234649987863140 2s^{16}+21591058927666966255084 32s^{15}+42680663647777106288781 12s^{14}+65211777262765861916282 64s^{13}+77429334197200253596601 31s^{12}+71109113618735821090519 92s^{11}+49764748763830706631955 17s^{10}+26002307191355911482692 22s^{9}+10353079281161503861096 95s^{8}+4204952201581657836723 00s^{7}+2874363458621680262094 21s^{6}+2257499121175270471203 54s^{5}+1322055539247657570232 86s^{4}+525462620987478955328 64s^{3}+135964972588619305441 08s^{2}+20873057772457298889 36s+1452933291809671974 54)$
c_{17}	$12(2s+3)^2(32580712094152126250 0s^{21}+59829921917002688553 00s^{20}+51700930512613327403406s^{19}+27907978077725947794 4590s^{18}+10531873179450453647679 66s^{17}+29454589028098495868768 10s^{16}+63109003318653630127438 44s^{15}+10554182290179327214273 482s^{14}+13894164004300292404029 789s^{13}+14397511755502695216399 777s^{12}+11648187532971253859583 195s^{11}+7253672843249894875203 621s^{10}+3463934076062711984148 183s^{9}+1401311871124636922510 679s^{8}+7096350734538033843306 63s^{7}+5266013007817227189696 21s^{6}+3664114629209035570141 20s^{5}+1877909383669174506336 06s^{4}+6682826578897923841868 4s^{3}+1577499396431898546259 2s^{2}+2238177282159012945966s+1453112757439619700 78)$
c_{16}	$3(2s+3)^2(55387210560058614625 00s^{22}+10696394904119483034 4800s^{21}+97537241738707509555 7110s^{20}+55777018695676356245 16312s^{19}+22401070138854784562671 602s^{18}+67034725561809321462257 220s^{17}+15468959266006177578068 3034s^{16}+28088953980133276709409 1608s^{15}+40566432899300593609815 8220s^{14}+46743223459745032337765 4624s^{13}+42816291170183647045348 8816s^{12}+30885119317741985964812 0664s^{11}+17392371150762459541255 5792s^{10}+7861104390229527730501 4364s^{9}+3450272393210865996806 8191s^{8}+2078486967808784701135 0512s^{7}+1521107049127527089926 0479s^{6}+9439795169478817720557 390s^{5}+4323642777299210867466 840s^{4}+1398336255225225668686 176s^{3}+3041651033834191453666 80s^{2}+4017017098446888839688 0s+2445688799534592091926)$

c_{15}	$12(2s+3)^2(4430976844804689170000s^{23}+8977362465878807211960 0s^{22}+861449789967478967902968s^{21}+52020523405703639617992 72s^{20}+221509586747221476368878 80s^{19}+706107689774315971385024 16s^{18}+174547747719038741579249148s^{17}+341846798830220759744006802s^{16}+537017897494050473220887475s^{15}+680386317088643909072652357s^{14}+694900917445527709102606203s^{13}+568854294351452352559155357s^{12}+370162500862325208186949407s^{11}+192089934615274810143113637s^{10}+85560000374404871078044743s^{9}+42188979069325595690484894s^{8}+27986482983635448916149477s^{7}+19349473000527948423160098s^{6}+10826549522417650582347903s^{5}+4499119009419911850147903s^{4}+1337268081929646109116429s^3+2701887274473978701502 99s^2+3341219144872220287179 3s+191648133946778922704 7)$
c_{14}	$3(2s+3)^2(4430976844804689170000 0s^{24}+9397609008462027996336 00s^{23}+946588223144997927058161 6s^{22}+601894156442872654302402 24s^{21}+270831851136366961169859120s^{20}+916082243422713193340650464s^{19}+241460571833561888085397003 2s^{18}+507178480505041003582316757 6s^{17}+860600335178662801913981102 4s^{16}+1188207272155926800257859433 6s^{15}+1337313573606658891526722519 2s^{14}+1223378975301732738139865659 2s^{13}+903808724360213876819507870 4s^{12}+536928655269087344627611732 8s^{11}+262353585257354926709155530 0s^{10}+119607178098293965186514409 6s^{9}+66007824462349678026358807 5s^{8}+45139319599766620866745885 2s^{7}+29034367707326894186404630 5s^{6}+14804204283262980396732105 0s^{5}+5646419333104340411674178 4s^{4}+1555947843019285082981882 4s^{3}+2939179015494775847121192s^2+3420469295854970612531 76s+1855791464680027845905 4)$
c_{13}	$6(2s+3)^2(4430976844804689170000 0s^{25}+9817855551045248780712 00s^{24}+1035713233442216953911268 8s^{23}+691661317833842749970623 20s^{22}+327906609514495942625889552s^{21}+1172872936442019296825004672s^{20}+328306372162931054591158932 0s^{19}+736033323922078596671432221 2s^{18}+1341123809195951980117385531 4s^{17}+2003087433125863944761640543 0s^{16}+2461263206658443506304569389 6s^{15}+2486180794324324784856470272 8s^{14}+2055461630993867656408819363 2s^{13}+1382870177464444560719848929 6s^{12}+759330701801977694759131669 2s^{11}+357316443768253965114029891 4s^{10}+171120540089863474143258870 9s^{9}+102643620132548287322773118 1s^{8}+69334669815813021335144258 7s^{7}+41372946832745234109278382 3s^{6}+19407764383083192653596067 4s^{5}+6856190522023177561905108 0s^{4}+1764097050880333254639794 4s^{3}+3132633769453198732801032s^2+3445604242270009355658 60s+1774509692548916842362 0)$
c_{12}	$3(2s+3)^2(1440067474561523980250 00s^{26}+3327383180429252608653 600s^{25}+3668680667792192157502 4412s^{24}+2567126637857826878686070 40s^{23}+1278878488443918869718977532s^{22}+4822501965635852078399708760s^{21}+1428493716760540762512361303 2s^{20}+3403967573402053121871363996 0s^{19}+6626919408734780484264189093 6s^{18}+1064202511819990594837395914 64s^{17}+1416737183478865486115278969 44s^{16}+1565122537347448498315035920 64s^{15}+1431194586986727902840203591 56s^{14}+1077729194579570065355629032 36s^{13}+6658356331681893190511728762 5s^{12}+3420802797778710148000657507 2s^{11}+1582143051029433989141678174 1s^{10}+8030913522092664743057080482s^{9}+5050521785734143734257145460s^{8}+3292374301511579644939102872s^{7}+1827212365211212115171329068s^{6}+7951466255776474175898377 40s^{5}+2621421821039038791088123 89s^{4}+6335442721429618093777958 4s^{3}+1062561041972189809432027 2s^{2}+1108770079900593722922360s+5437162259941865052170 7)$

c_{11}	$2(2s+3)^2(28801349491230479605000 0s^{27}+6927926613537598727151600s^{26}+$ $7968499405072408869622956 0s^{25}+5830135753486700217327189 84s^{24}+$ $3044719667796555717818159544s^{23}+1207115288169085268602840692 0s^{22}+$ $3771975180228306456186022831 2s^{21}+9518657075468497895431568072 4s^{20}+$ $1971414432505091371211002980 18s^{19}+3386213982946173069363008884 86s^{18}+$ $4853203941721254940783622901 42s^{17}+5817879064516111538357630359 26s^{16}+$ $5828027860833244086757468021 92s^{15}+4859938328494178086778700964 80s^{14}+$ $3355944446899925164463735414 70s^{13}+1918186782349513132753181871 66s^{12}+$ $9321550814168794157984604359 1s^{11}+4299548855661907296542112784 5s^{10}+$ $2299333420544570449863855165 9s^{9}+1470200287638349161916712529 3s^{8}+$ $9151802478157350485470780638s^{7}+4746495171599998113140409294s^{6}+$ $1930238883493800552821467836s^{5}+5976617058265502770521785 82s^{4}+$ $1363694210951181524092878 75s^{3}+2169005140767851682417301 5s^{2}+$ $2154400447443907595487183s+1008731317586940460287 45)$
c_{10}	$(2s+3)^2(6336296888070705513100 00s^{28}+1584239110567672292139120 0s^{27}+$ $1897621049684466450141042 96s^{26}+1448899233419905865696230704s^{25}+$ $7915017749037631812296989272s^{24}+3291137409296388117112885 1808s^{23}+$ $1081841136804198369570625711 20s^{22}+2881814574059132039155601476 56s^{21}+$ $6325630775471257682897241758 52s^{20}+1156966248842167228077853758672s^{19}+$ $1775610230389152091840398326292s^{18}+2294661769917892997575576903488s^{17}+$ $2498192965867701162268688835924s^{16}+2285659112305489254434518711416s^{15}+$ $1748991971403770816761602941934s^{14}+1113797684163785846264967323376s^{13}+$ $5924913951597435743900967421 11s^{12}+2744766007155415656150112137 96s^{11}+$ $1270446195852713506501888454 52s^{10}+705385823261477238654077656 80s^{9}+$ $4488968110254001706707249846 5s^{8}+2660289888716075956783633452 0s^{7}+$ $1296776974656174947900170128 0s^{6}+4960789745148078555411299844s^{5}+$ $1450732098731590049595423084s^{4}+3139308944683290821452849 56s^{3}+$ $4752567062061261105861801 9s^{2}+4506795195444098216749962s+2019810664389200880741 21)$
c_{9}	$2(2s+3)^2(28801349491230479605000 00s^{29}+7474247118895785746840400s^{28}+$ $9308513797882609298968418 4s^{27}+7404022441838906136102386 16s^{26}+$ $4222492016436341447317422840s^{25}+1837374587737775714357659 7976s^{24}+$ $6337431126498352669577742938 4s^{23}+1776892961989847704086466054 92s^{22}+$ $4119921298476838644538559778 90s^{21}+7992705700875121584791677780 14s^{20}+$ $1307454563265970530661662310 488s^{19}+1811440906673507850196226957 292s^{18}+$ $2129008833631551803404401966 618s^{17}+2120318902997465121470552812 398s^{16}+$ $1782752578142774040882126045 846s^{15}+1258348636755929955587480574 282s^{14}+$ $7422229506057785006481351208 08s^{13}+3687595326777177603238716204 48s^{12}+$ $1632923401776166653452978780 34s^{11}+7562813554420816525428037104 0s^{10}+$ $4285873595461389310248944256 9s^{9}+2681560986049355543901710696 7s^{8}+$ $1515406630759595340156585468 0s^{7}+6986502898234763627448498006s^{6}+$ $2529774938694747662152377363s^{5}+7023305018106747704878627 23s^{4}+$ $1447326550278639488278622 61s^{3}+2092454720837126248556029 5s^{2}+$ $1899529777608715114521399s+8166786538840395351817 5)$

c_8	$3(2s+3)^2(1440067474561523980250 00s^{30}+3873703685787439628342400s^{29}+5008695060602392586847903 6s^{28}+4143472933431033114133572 00s^{27}+2462455716417895553466814404s^{26}+1119031183906634425953671402 4s^{25}+4040929597277317705010456655 6s^{24}+1189471348113066694902726162 00s^{23}+2904604480014499379866096419 84s^{22}+5956509828769617004442479161 36s^{21}+1034394503332837044832871973 660s^{20}+1529122522398341452237748264 328s^{19}+1929170112144408531295247240 352s^{18}+2077275729835779898627972452 564s^{17}+1904400092584849147398803257 857s^{16}+1479295019246765244421617960 408s^{15}+9672325875815231341287676907 37s^{14}+5298271494562945135798668577 62s^{13}+2458501493711512202476023508 38s^{12}+1035861819814159623650793501 36s^{11}+4744196983100521076585665149 0s^{10}+2701737032861373124298872611 6s^{9}+1654360407462134231653483389 6s^{8}+8959808104815503968120901160s^{7}+3934148480503417587877177653s^{6}+1356708501926603220744080694s^{5}+3593560100508123307874732 79s^{4}+7080039802536542890191925 2s^{3}+9805705639441860010993074s^{2}+8542944279890061972050 52s+3530618607561154024305 6)$
c_7	$12(2s+3)^2(2215488422402344585000 00s^{31}+6169667403272286743484 00s^{30}+8270907073696162683430488s^{29}+7105488837495844741933140 0s^{28}+4393191571105468038118969 68s^{27}+2081153035704566658130624 800s^{26}+7851699408458680802242634 484s^{25}+2420751676062542307220707 3414s^{24}+6209251742880587586679467 4881s^{23}+1341913491590212396884148 57431s^{22}+2465183787323117750294642 65197s^{21}+3872290488006535317381147 39999s^{20}+5218490775618946576363984 49655s^{19}+6040104996590269514608976 07741s^{18}+5997061939521369643211276 94249s^{17}+5089326000927003623080146 94866s^{16}+3669238458301067350871990 88303s^{15}+2229706788782718855351590 28828s^{14}+1135107514048166915897613 55815s^{13}+4895962158015973384528988 4633s^{12}+1934381973317095554431955 2163s^{11}+8570967648243419298660377 211s^{10}+4842281705213527108416513 279s^{9}+2909138374562317990360031 394s^{8}+1522921303596173996138575 905s^{7}+6422129655286405723104368 79s^{6}+2123474087926098366448153 59s^{5}+5394050486913816598335555 6s^{4}+1020098142154886627227202 1s^{3}+1357609423748226427157778s^{2}+1137825877924574767888 65s+4528476210896135134182)$
c_6	$3(2s+3)^2(4430976844804689170000 0s^{32}+1275958134912779427134400s^{31}+1771212464874352037424600 0s^{30}+1578007833031923424848077 76s^{29}+1013476002096822653934247 536s^{28}+4996346811471994438240970 400s^{27}+1965689104900375999773821 9664s^{26}+6334301858598682308261972 9288s^{25}+1702585370436327792294737 23584s^{24}+3867192559701889939172245 74336s^{23}+7491959730371222021272021 35768s^{22}+1245952789865163842159925 500256s^{21}+1785990969360715046000117 363568s^{20}+2210911897085152430737822 315488s^{19}+2363325656949833631869930 474232s^{18}+2176386534475868543250178 452024s^{17}+1718519269597730769046219 931925s^{16}+1154898997656814617038309 424924s^{15}+6539937996823707165850409 76045s^{14}+3090978123875898429940490 88882s^{13}+1226499865148059279758110 13678s^{12}+4413932159338648643754130 4232s^{11}+1819460051207313117965247 4218s^{10}+1008072945524717714946636 0108s^{9}+6012154495009832863143860 703s^{8}+3086713901904485302077423 912s^{7}+1264536513002644891553573 532s^{6}+4042690043135359977090641 88s^{5}+9908531855413739589346726 5s^{4}+1806686959296584557373176 8s^{3}+2318037559219684454533893s^{2}+1873358777849595772989 78s+7192069815364796066229)$

c_5	$6(2s+3)^2(8861953689609378340000s^{33}+26359655783422030111440 0s^{32}+3784460184258343211722032s^{31}+34920466929143934879808368s^{30}+23264213016687712042575297 6s^{29}+119169780094536708645805507 2s^{28}+48806947518804415592184766 32s^{27}+164069508244695476805607787 64s^{26}+46112865280418851489322074134s^{25}+109812274944960351155991786018s^{24}+22372585643351671399884980563 2s^{23}+39266014091709822788394101988 0s^{22}+59645033289927608500675400043 6s^{21}+78624155913942530616774743732 4s^{20}+90016247244570627082720573540 8s^{19}+89411845619881376237474487484 2s^{18}+76801976570035682154104806421 1s^{17}+56718891951502940479512463981 5s^{16}+35689284917369058878427986592 3s^{15}+18888772644005809091812356980 7s^{14}+82758519677573937797801658678s^{13}+29755841968295155180938377160s^{12}+92656551239853153882294277 62s^{11}+32698047051242709609461794 92s^{10}+17453787504376746722545253 43s^{9}+10671241715712761650793291 61s^{8}+5531923274479354261466695 44s^{7}+2244718815472323109755889 2s^{6}+7027128711382933309284923 4s^{5}+1675985651729512089493665 6s^{4}+2963427571961631174417201s^{3}+3679880939623824678753 21s^{2}+2875049397368658429499 8s+1066338997876680421860)$
c_4	$3(2s+3)^2(5538721056005861462500s^{34}+17000093042867794800120 0s^{33}+2521542870629614052494182s^{32}+24069033861073900393194288s^{31}+16611214023947173022412795 0s^{30}+88286198474324819522022706 8s^{29}+37581355594818610758601905 60s^{28}+131557876582587412358907558 00s^{27}+38587134973448786366330248032s^{26}+96129109345583765886638321232s^{25}+20544714774638033566476875495 4s^{24}+37944945337592677887102869236 8s^{23}+60877678098043538114517793164 6s^{22}+85125183973189749706760054107 2s^{21}+10391283871123404096276184222 11s^{20}+11073389076832412879108175746 16s^{19}+10282364259109118891102115423 27s^{18}+82865338871955099784133459981 4s^{17}+57567986009921446099133676712 2s^{16}+34111398025087438571519549875 2s^{15}+16960211216096687324467255456 2s^{14}+69019121880524178108395152092s^{13}+22192275965509681657763172138s^{12}+55130393244696074306221011 84s^{11}+13085527335820566166857177 99s^{10}+62282242696827293561075056 2s^{9}+4412082186155643660214786 80s^{8}+25108662674499791880622183 2s^{7}+1055427263431979624476683 30s^{6}+33182712041751038192684904s^{5}+7819078036046387444720541s^{4}+13533729394126374890003 88s^{3}+1635933722331722818732 02s^{2}+12396727299661082512092s+4448137902935067283 68)$
c_3	$6(2s+3)^2(6516142418830425250 00s^{35}+20618119083643318565400s^{34}+3156213349996927861511 16s^{33}+31130651785720476660078 60s^{32}+2222979754518395268052028 4s^{31}+12242256173994148159983548 4s^{30}+5408405145287421117617769 12s^{29}+19683840172480631833231049 76s^{28}+6014334916396782599305697730s^{27}+15642733509059107994947830402s^{26}+3499106342792033984966407009 8s^{25}+6783451224245674486841205351 0s^{24}+1146107356416358232326805582 64s^{23}+1694202892194124436822450064 36s^{22}+2196301294579059436764574758 06s^{21}+2499148884589175758632804858 78s^{20}+2494580325094377405252407777 31s^{19}+2179203808681064934815157564 21s^{18}+1658682998392915454101611337 71s^{17}+1091968041316217564562484267 89s^{16}+6145574397104443845110768323 2s^{15}+2901082085745016093381214013 6s^{14}+1111172476982043580911263970 0s^{13}+3230978841571830756780993930s^{12}+5977340713166300300960714 52s^{11}+1947145671045436300515266 4s^{10}-1441382310935022454149972 6s^{9}+9684588377731643071927236s^{8}+1256021958303918550403991 0s^{7}+6629813221106696409536742s^{6}+2267528918316638233964400s^{5}+5495108516124471979461 00s^{4}+9516959371683531754669 8s^{3}+1133337907636920896160 6s^{2}+8377615455010504271 46s+2911835924761782947 4)$

c_2	$(2s+3)^2(65161424188304252500 0s^{36}+2123612870508923148360 0s^{35}+3351768289135045172584 92s^{34}+3412480272349060314590760s^{33}+2518418762184767430421861 2s^{32}+1435324779126332041071110 40s^{31}+6571995334134877377460058 40s^{30}+2483057802762141981372604872s^{29}+7890444256414221204984097182s^{28}+2138670662626443899749925750 4s^{27}+4996862660986503205494353272 2s^{26}+1014439442075341249520382837 48s^{25}+1800227084981344254247742573 58s^{24}+2804733165486533307037649392 32s^{23}+3847701887175612902111068827 24s^{22}+4655680614090253461407950552 68s^{21}+4970705288178906475764968742 21s^{20}+4678534555507223380165602910 68s^{19}+3872550916392453613427525516 70s^{18}+2806584923777792155024177616 76s^{17}+1768405807430274222159766204 77s^{16}+9579106278655550872408874232 0s^{15}+4378586233016249905220952976 8s^{14}+1632673685352712299171515582 4s^{13}+4608219050084326790748933153s^{12}+7630908802856105790078631 08s^{11}-7027359523643236832970771 6s^{10}-9985407270432287153739260 4s^{9}-3572603137796979208818892 5s^{8}-6099037895307670142287608s^{7}+3859970007112238323561 68s^{6}+5718095505418079375309 52s^{5}+1847893437895344611505 30s^{4}+3548837556962247216924 0s^{3}+4362950858813170449228s^{2}+3206021409943901224 56s+10806813741383936712)$
c_1	$2s^2(2s+3)^2(5s+6)^2(1371819456595879000s^{33}+4271634557055966808 8s^{32}+6435444751783137019 08s^{31}+6247566555702580389060s^{30}+4391676556526924901666 0s^{29}+2381294370524934521705 40s^{28}+1036074226442125183419168s^{27}+3714930106381256786671824s^{26}+1118772868707257505376370 4s^{25}+2869670825417555723588648 4s^{24}+6335289140657084244295687 8s^{23}+1213301616269188484657558 398s^{22}+2027652551311758119054867 52s^{21}+2969521561596848429994761 88s^{20}+3821939060079245330600661 96s^{19}+4329771063731550080610336 36s^{18}+4318829425555373349451823 76s^{17}+3789227783217399510763191 60s^{16}+2916937858513690287806626 44s^{15}+1961472241580118928394333 94s^{14}+1144103496246446264232127 29s^{13}+5724888433243097639645162 7s^{12}+2413061795050375698405100 8s^{11}+8287189086050003227856022s^{10}+2152391916458370195915867s^{9}+3251846145974111403146 01s^{8}-3300797235187890347540 4s^{7}-4179251093868663889730 4s^{6}-1583118597299644973035 8s^{5}-3877006197061115690130s^{4}-6655060240921758556 80s^{3}-7841981439220991112 0s^{2}-5759186746951521828s-200126180395998828)$
c_0	$s^4(2s+3)^2(5s+6)^4(5487277826383516s^{30}+1629002070479364 48s^{29}+2337377142714373098s^{28}+2158816572989759829 6s^{27}+1442107043736372374 22s^{26}+7422068524214498072 76s^{25}+3061285798160471289822s^{24}+1039189563664458602011 2s^{23}+2958836372773561206903 6s^{22}+7165140410814613889709 6s^{21}+1491176200730274615474 36s^{20}+2688067051789587271872 48s^{19}+4221859922152866254547 36s^{18}+5801917523864989863237 12s^{17}+6996947243102316240642 72s^{16}+7417572618016561112259 84s^{15}+6916662471035836623516 12s^{14}+5670330877797167100233 52s^{13}+4080471316446565805592 96s^{12}+2570366447945656343830 08s^{11}+1411452641418268040735 76s^{10}+6717846053228888490951 6s^{9}+2749966305794295114104 1s^{8}+9582491278489242855672s^{7}+2803365139419823150782s^{6}+6756829533138365528 76s^{5}+1306594986712056470 52s^{4}+1948956305690965449 6s^{3}+2105305456045150908s^{2}+1465944860513900 88s+4941387170271576)$

Δ_2	$-36s^4(2s+3)^4(5s+6)^4(188188862149501274759871978264100000s^{66}$ $+1171982279367338935485204891772187040 0s^{65}$ $+3590340595158437649151724049466274089 92s^{64}$ $+7212088571266452516322419090023965301952s^{63}$ $+10683913202536878409344821445611408008 9896s^{62}$ $+12446539307453434624006123411233664305561 12s^{61}$ $+11874531272586759157048871425273776419646480s^{60}$ $+9539717463978084911740021792529086079754307 2s^{59}$ $+6586014211961623946750060924092706557833870 96s^{58}$ $+3968022355733696091765263130542489256493297296s^{57}$ $+2111718215787453698016965546515609174427692038 0s^{56}$ $+1002346955322263064366492155231201633504320407 84s^{55}$ $+427723545408455739277440670239201960498527445168s^{54}$ $+165169320489858311000198383167538370383880311128 0s^{53}$ $+58038813051285159280644727598895039061348407197 88s^{52}$ $+18645262399111732404927179918528703191829668776944s^{51}$ $+5498216092397086290715585398715769282805100212946 8s^{50}$ $+1493403271312119095060211916193032794760693070014 80s^{49}$ $+374738908504925756556865067810806785110272348870998s^{48}$ $+87095917136338876712545518393847454868443303273397 6s^{47}$ $+18791254231056775235326951135911186532510305432196 50s^{46}$ $+3770893425502604495219029177077658520798571516698604s^{45}$ $+7050015221784174490794108388251813073259643663091008s^{44}$ $+1229751697814540088737743012540597239130953523640481 6s^{43}$ $+2003816448416423665928915846433322332297986145636955 8s^{42}$ $+3053220493065830192217996030789798685096097311040106 0s^{41}$ $+4353974016564230774744937671615227473899797195238128 8s^{40}$ $+5814794272234047468845528446358546692682944712469086 4s^{39}$ $+7276560981967024598164222622942133354281616530078647 8s^{38}$ $+8535222627194411740352205447227499061598741207765802 0s^{37}$ $+9386118512358018098652883224139505412412521033278797 8s^{36}$ $+9677367622870745479418594983628077002220070263778913 6s^{35}$ $+9353550534056682983528096959246190247987486973707032 4s^{34}$ $+8472657307033122222401048138408764617479499951727572 0s^{33}$ $+7189292302539646724238556713520429645936760693117107 2s^{32}$ $+5710750538582167076586109442116935829639912622263872 0s^{31}$ $+4242947557936920029498039681532283983986502171148171 2s^{30}$ $+2945340224867131137091897558637195040010087585918442 4s^{29}$ $+1907688741881547559139122534482867333198543394985974 5s^{28}$ $+1150963090261374243005671011302597994636892936359156 0s^{27}$ $+6455314570589262935425947995355398219841083619460275s^{26}$ $+3357432561061020603580382785922585900536396393910862s^{25}$ $+1614483581569691108434974127018929362203153022365980s^{24}$ $+7151843921239467644322741543961403561291367742781 52s^{23}$ $+2905615937078824111549712703038991345504578369200 80s^{22}$ $+1076866147250386418961409060935085408978065543348 84s^{21}$ $+361751237678278765712871685717720960666418421897 15s^{20}$ $+109362954136214178401472446585845791272071743893 52s^{19}$ $+29562314469519458854524531776006502223600315078 53s^{18}$ $+7143283154950873347674525118298108303417045322 94s^{17}$ $+1578622563973412013157962625289520289452101601 78s^{16}$ $+3466918481540893224348429068849868462291600055 2s^{15}$ $+879220885968824907033894237443409495692996843 4s^{14}$ $+272895891349790467960562930434828306766214915 6s^{13}$ $+90851447731067951899123953408477649856954913 0s^{12}$ $+28199738465157182443141692544401588181 01696600s^{11}$ $+7659571256132994444113894245418887273849699 2s^{10}$

	$+1782323536880591650543354401619541599668158 4s^9$ $+3522507407940359387775769801382214828420516s^8$ $+5867820050646080008320434741079193955195 52s^7$ $+8148452315276322918847755584680364066167 2s^6$ $+9275689993243887053899463828498929963344s^5$ $+8438081673219230573799541298816126340 96s^4$ $+5898808758941212962363346647613729785 6s^3$ $+2972581287433071568118115850779633072s^2$ $+959233912036916287714016721581540 16s$ $+148335141036636539337219080656939 2)$

The last lemma presents some properties for the population-level SGPO and GRPO dynamics.

Lemma B.4. *Under the assumptions from Theorem 3.3, the following statements hold true,*

(i) $p_{SGPO}^{(k)}, q_{SGPO}^{(k)}, p_{GRPO}^{(k)}, q_{GRPO}^{(k)} \in (0,1)$ *for all* $k \geq 0$.
(ii) $p_{SGPO}^{(k)}, q_{SGPO}^{(k)}, p_{GRPO}^{(k)}, q_{GRPO}^{(k)}$ *are strictly increasing in* k *and lie in* $(\frac{1}{2}, 1)$ *for all* $k \geq 1$.
(iii) $p_{SGPO}^{(k)} > q_{SGPO}^{(k)}$ *for all* $k \geq 1$.

Proof. We first rewrite the update rule in Eq. (3) as follows,

$$
\begin{aligned}
p_{\texttt{SGPO}}^{(k+1)} &= p_{\texttt{SGPO}}^{(k)} \frac{e^{\Delta_{\texttt{SGPO},p}^{(k)}}}{1-p_{\texttt{SGPO}}^{(k)}+p_{\texttt{SGPO}}^{(k)} e^{\Delta_{\texttt{SGPO},p}^{(k)}}}, \quad \text{where} \quad \Delta_{\texttt{SGPO},p}^{(k)} = p_{\texttt{SGPO}}^{(k)}(1-p_{\texttt{SGPO}}^{(k)}), \\
q_{\texttt{SGPO}}^{(k+1)} &= q_{\texttt{SGPO}}^{(k)} \frac{e^{\Delta_{\texttt{SGPO},q}^{(k)}}}{1-q_{\texttt{SGPO}}^{(k)}+q_{\texttt{SGPO}}^{(k)} e^{\Delta_{\texttt{SGPO},q}^{(k)}}}, \quad \text{where} \quad \Delta_{\texttt{SGPO},q}^{(k)} = (p_{\texttt{SGPO}}^{(k)})^2 q_{\texttt{SGPO}}^{(k)}(1-q_{\texttt{SGPO}}^{(k)}), \\
p_{\texttt{GRPO}}^{(k+1)} &= p_{\texttt{GRPO}}^{(k)} \frac{e^{\Delta_{\texttt{GRPO},p}^{(k)}}}{1-p_{\texttt{GRPO}}^{(k)}+p_{\texttt{GRPO}}^{(k)} e^{\Delta_{\texttt{GRPO},p}^{(k)}}}, \quad \text{where} \quad \Delta_{\texttt{GRPO},p}^{(k)} = p_{\texttt{GRPO}}^{(k)}(1-p_{\texttt{GRPO}}^{(k)}) q_{\texttt{GRPO}}^{(k)}, \\
q_{\texttt{GRPO}}^{(k+1)} &= q_{\texttt{GRPO}}^{(k)} \frac{e^{\Delta_{\texttt{GRPO},q}^{(k)}}}{1-q_{\texttt{GRPO}}^{(k)}+q_{\texttt{GRPO}}^{(k)} e^{\Delta_{\texttt{GRPO},q}^{(k)}}}, \quad \text{where} \quad \Delta_{\texttt{GRPO},q}^{(k)} = p_{\texttt{GRPO}}^{(k)} q_{\texttt{GRPO}}^{(k)}(1-q_{\texttt{GRPO}}^{(k)}).
\end{aligned}
$$

First of all, the uniform initialization yields the desired result for $k = 0$. Suppose $p_{\texttt{SGPO}}^{(k)} \in (0,1)$ for some $k \geq 0$. Then, we have

$$
1 - p_{\texttt{SGPO}}^{(k)} + p_{\texttt{SGPO}}^{(k)} e^{\Delta_{\texttt{SGPO},p}^{(k)}} > p_{\texttt{SGPO}}^{(k)} e^{\Delta_{\texttt{SGPO},p}^{(k)}} > 0,
$$

which implies $p_{\texttt{SGPO}}^{(k+1)} \in (0,1)$. By induction, we have $p_{\texttt{SGPO}}^{(k)} \in (0,1)$ for all $k \geq 0$. Similarly, we can show that $q_{\texttt{SGPO}}^{(k)}, p_{\texttt{GRPO}}^{(k)}, q_{\texttt{GRPO}}^{(k)} \in (0,1)$ for all $k \geq 0$.

Furthermore, we have $\Delta_{\texttt{SGPO},p}^{(k)} > 0$ since $p_{\texttt{SGPO}}^{(k)} \in (0,1)$. This implies

$$
\frac{p_{\texttt{SGPO}}^{(k+1)}}{p_{\texttt{SGPO}}^{(k)}} = \frac{1}{(1-p_{\texttt{SGPO}}^{(k)}) e^{-\Delta_{\texttt{SGPO},p}^{(k)}} + p_{\texttt{SGPO}}^{(k)}} > \frac{1}{1-p_{\texttt{SGPO}}^{(k)}+p_{\texttt{SGPO}}^{(k)}} = 1.
$$

Since $p_{\texttt{SGPO}}^{(0)} = \frac{1}{2}$, we have $p_{\texttt{SGPO}}^{(k)} \in (\frac{1}{2}, 1)$ for all $k \geq 1$. Similarly, we can show that $q_{\texttt{SGPO}}^{(k)}, p_{\texttt{GRPO}}^{(k)}, q_{\texttt{GRPO}}^{(k)}$ are strictly increasing and lie in $(\frac{1}{2}, 1)$.

Finally, we have $p_{\texttt{SGPO}}^{(0)} \geq q_{\texttt{SGPO}}^{(0)}$. Thus, it suffices to show that $p_{\texttt{SGPO}}^{(k)} \geq q_{\texttt{SGPO}}^{(k)}$ implies $p_{\texttt{SGPO}}^{(k+1)} > q_{\texttt{SGPO}}^{(k+1)}$ for all $k \geq 0$. Indeed, Lemma B.1(i) and $p_{\texttt{SGPO}}^{(k)} \geq q_{\texttt{SGPO}}^{(k)}$ yield

$$
p_{\texttt{SGPO}}^{(k+1)} = \exp(f_{11}(p_{\texttt{SGPO}}^{(k)})) \geq \exp(f_{11}(q_{\texttt{SGPO}}^{(k)})) = \exp(\log(q_{\texttt{SGPO}}^{(k)}) + h_{q_{\texttt{SGPO}}^{(k)}}(q_{\texttt{SGPO}}^{(k)}(1-q_{\texttt{SGPO}}^{(k)}))).
$$

Then, Lemma B.1(ii) and $p_{\texttt{SGPO}}^{(k)} \in (0,1)$ yield

$$
\exp(\log(q_{\texttt{SGPO}}^{(k)}) + h_{q_{\texttt{SGPO}}^{(k)}}(q_{\texttt{SGPO}}^{(k)}(1-q_{\texttt{SGPO}}^{(k)}))) > \exp(\log q_{\texttt{SGPO}}^{(k)} + h_{q_{\texttt{SGPO}}^{(k)}}((p_{\texttt{SGPO}}^{(k)})^2 q_{\texttt{SGPO}}^{(k)}(1-q_{\texttt{SGPO}}^{(k)}))).
$$

In addition, we have

$$q_{\texttt{SGPO}}^{(k+1)} = \exp(f_{12}(p_{\texttt{SGPO}}^{(k)}, q_{\texttt{SGPO}}^{(k)})) = \exp(\log q_{\texttt{SGPO}}^{(k)} + h_{q_{\texttt{SGPO}}^{(k)}}((p_{\texttt{SGPO}}^{(k)})^2 q_{\texttt{SGPO}}^{(k)}(1 - q_{\texttt{SGPO}}^{(k)}))).$$

Putting these pieces together yields $p_{\texttt{SGPO}}^{(k+1)} > q_{\texttt{SGPO}}^{(k+1)}$. □