TAPO: Dynamic Teacher and Perturbed Answer Injection for Policy Optimization

Maowei Jiang[1*†], Zihang Wang[2*], Qi Wang[1*], Peter Búš[1*†], Moquan Cheng[4*], Yifan Wang[3*], Quangao Liu[2], Ruiqi Li[2], Pengyu Zeng[1], Ruikai Liu[2], Alan Liang[2], Yansong Xu[2], Yusong Hu[1], Chaoran Zhang[1], Zhiyong Dong[1]

[1]Tsinghua University, Shenzhen International Graduate School, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]The Chinese University of Hong Kong, Shenzhen
[4]China Mobile Communications Corporation, China
jmw24@mails.tsinghua.edu.cn, peter_bus@sz.tsinghua.edu.cn

Abstract

Reinforcement learning (RL) has emerged as a powerful framework to improve the reasoning performance of large language models (LLMs), with approaches such as Group Relative Policy Optimization (GRPO) showing promising results. However, GRPO and its variants struggle with collapsed groups (i.e., all-correct or all-incorrect completions), leading to zero-variance rewards and ineffective gradient signals. Moreover, focusing solely on final answer correctness while ignoring the reasoning process, along with rigid length penalties, can hinder training stability and output quality. To address these issues, we introduce TAPO, a reinforcement learning framework that enhances optimization signals by modifying sampled completions within training groups. TAPO incorporates three core techniques: (1) *Dynamic Teacher Injection (DTI)*, which selectively injects high-quality or adversarial examples to restore effective gradient signals in collapsed groups; (2) *Perturbed Answer Injection (PAI)*, which makes partially correct completions to provide contrastive supervision separating reasoning correctness but wrong answer from the trajectories; and (3) *InfoLen-Aware Reward Shaping*, a fine-grained reward strategy that penalizes outputs based on both length and semantic redundancy, encouraging concise yet informative responses. Extensive experimental results demonstrate that TAPO significantly improves the mathematical reasoning capabilities of LLMs across multiple challenging benchmarks, outperforming the GRPO baseline by a substantial margin. Componentwise ablations further validate the contribution of each proposed technique.

1 Introduction

In recent years, the rapid advancements in large-scale language models (LLMs) have enabled significant breakthroughs in complex reasoning tasks. Notable examples include systems such as OpenAI-o1 (OpenAI 2024), Deepseek-R1 (Shao et al. 2024), and Kimi-k1.5 (Team et al. 2025), which have achieved impressive results across challenging domains like mathematical reasoning (Liu et al. 2025; Seßler et al. 2024), code generation (Zheng et al. 2024; Tong and Zhang 2024; Li et al. 2023), and scientific problem solving (Wang et al. 2023; Ma et al. 2024; Lu et al. 2024).

Building on these advances, reinforcement learning (RL) techniques have played a central role in aligning model behavior with desired reasoning patterns (Ziegler et al. 2019; Bai et al. 2022; Wu et al. 2023; Chaudhari et al. 2024). Among them, PPO (Schulman et al. 2017) is a classic RL algorithm (Fig. 1(a)) widely used to fine-tune LLMs for reasoning and alignment, but it relies on an auxiliary value model (critic) to estimate returns, which adds complexity and potential instability in training. Group Relative Policy Optimization (GRPO) (Shao et al. 2024) has emerged as a strong alternative to traditional critic-based methods (Haarnoja et al. 2018; Fujimoto, van Hoof, and Meger 2018; Sutton et al. 2000) by directly computing relative rewards over sampled completion groups (Fig. 1(b)).

Despite promising progress in reinforcement learning (RL) for large language models (LLMs), several key challenges remain to be fully addressed. First, existing RL frameworks like GRPO often encounter "collapsed" groups (Yu et al. 2025) where all completions are either correct or incorrect. These groups yield zero-variance rewards and thus provide no effective gradient signal, stalling learning progress. DAPO (Yu et al. 2025) alleviates this issue through dynamic sampling, selectively preserving only groups with mixed correctness to ensure non-zero reward variance. However, this approach may fail in early-stage training or on particularly challenging tasks, where correct completions are extremely sparse and valid groups are nearly impossible to collect. In such cases, the model enters a regime of sample starvation, where no eligible group can be formed to trigger updates. This severely limits training efficiency and scalability in low-signal regimes. Second, existing reward functions in RL pipelines (Uesato et al. 2022; Pan et al. 2023; Guo et al. 2025) typically assign scores based only on the final answer correctness, neglecting the reasoning process and output semantics. This coarse-grained reward signal may fail to distinguish between completions that demonstrate reasonable intermediate reasoning but yield incorrect answers. Moreover, prior GRPO-based methods (Zhang and Zuo 2025; Dai, Yang, and Si 2025; Yu et al. 2025) penalize over-length completions by applying

[1]Equal contribution.

[2]Corresponding author.

hard truncation strategies, wherein outputs exceeding a predefined maximum length are forcibly clipped and assigned punitive rewards. To improve reward continuity, DAPO introduces length-based soft penalties that scale linearly with the excess length. However, both strategies focus solely on output length and ignore semantic redundancy. As a result, verbose yet informative completions may be over-penalized, while long but repetitive outputs can still receive moderate rewards.

Motivated by the above limitations, we propose TAPO, a reinforcement learning framework that improves training stability and reasoning supervision through structured interventions on sampled completions. TAPO integrates three components. Dynamic Teacher Injection (DTI) repairs collapsed groups by injecting contrastive teacher samples and directly assigning fixed advantages, enabling gradient updates even when all completions are correct or incorrect. Perturbed Answer Injection (PAI) introduces minimally altered completions to distinguish valid reasoning from incorrect final answers, and similarly uses fixed advantage assignment to provide stable feedback. InfoLen further improves reward quality by jointly considering output length and semantic redundancy, reducing noisy supervision and encouraging concise yet informative generation. Together, these techniques improve optimization dynamics without modifying model size.

We conduct extensive experiments across multiple challenging mathematical reasoning benchmarks to evaluate the effectiveness of TAPO. Using DeepSeek-R1-Distill-Qwen-1.5B (Guo et al. 2025) as the backbone, TAPO achieves an average accuracy of 62.87%, significantly outperforming the Naive GRPO baseline (34.3%) and other 1.5B and 7Bscale models such as Still-1.5B (Chen et al. 2025) and Qwen2.5-Math-7B-Instruct (Yang et al. 2024a). In particular, TAPO improves the accuracy of AIME-2024 (of Problem Solving 2024) by 15.4 points over the base model and 10 points over Naive GRPO baseline.

2 Related Work

Large Scale Reasoning Models. Recent advances have demonstrated that scaling up language models not only improves fluency and knowledge retention, but also unlocks emergent capabilities in multi-step reasoning (Wei et al. 2022b; Wang et al. 2022b; Zhou et al. 2022). Instruction-tuned models such as FLAN (Longpre et al. 2023) and T0 (Sanh et al. 2022) improve generalization to unseen reasoning tasks by fine-tuning on broad collections of instructional data. Chain-of-thought (CoT) prompting (Wei et al. 2022a; Zhang et al. 2022; Wang et al. 2022a; Sanwal 2025) further enables large models to produce intermediate reasoning steps, which significantly boosts performance on arithmetic, symbolic, and commonsense benchmarks. Complementary to prompting, several works enhance reasoning through external tool use and structured intermediate computation. Program-Aided Language Models (PAL) (Gao et al. 2022) generate Python programs as structured reasoning traces and execute them to solve problems more reliably. Toolformer (Schick et al. 2023) learns to invoke external APIs, such as calculators and search engines, in a self-supervised way to augment the model's responses. Minerva (Lewkowycz et al. 2022) scales PaLM-540B using math-focused corpora and achieves state-of-the-art quantitative reasoning performance without tool usage. Code LLaMA (Roziere et al. 2023) leverages code pretraining to enhance symbolic reasoning across programming and mathematical domains. Recent proprietary systems such as OpenAI-o1 reportedly achieve strong performance on complex reasoning tasks via large-scale reinforcement learning (OpenAI 2024). However, the lack of public training details hinders analysis and reproducibility.

RL for LLMs. RL has emerged as a powerful framework for aligning language models with human preferences and reasoning quality (Ouyang et al. 2022; Xu et al. 2022; Bai et al. 2022; Wu et al. 2023; Christiano et al. 2017; Stiennon et al. 2020; Wang et al. 2024; Cao et al. 2024; Yan et al. 2025). Proximal Policy Optimization(PPO) (Schulman et al. 2017) was first applied to large-scale instruction tuning in InstructGPT (Ouyang et al. 2022), where it optimized outputs using reward models trained from human feedback. To address the inefficiency and instability of reward-based methods, Direct Preference Optimization (DPO) (Rafailov et al. 2023; Liu, Sun, and Zheng 2024; Pal et al. 2024) reformulated the problem into direct preference optimization without an explicit reward model (Aksitov et al. 2023; Yang et al. 2024b). Group Relative Policy Optimization (GRPO) (Shao et al. 2024) extended this idea to group-level comparisons, enabling multi-candidate optimization with improved variance reduction.

Sample-Level Interventions for Optimizing Reasoning Recent efforts in RL for LLMs have explored various strategies to improve the training efficiency and reasoning performance by operating on sampled completions. DAPO (Yu et al. 2025) introduces a dynamic sampling mechanism to discard degenerate reward groups (e.g., all-correct or all-wrong), aiming to maintain effective gradient signals. CPPO (Lin et al. 2025) further filters out low-advantage completions to concentrate optimization on stronger samples. In the supervised setting, Rejection Sampling Fine-Tuning (RFT) (Yuan et al. 2023) selects correct and diverse reasoning paths to improve generalization. However, these methods either discard potentially useful samples or rely on external selection heuristics, which may lead to sample inefficiency or biased supervision. In contrast, our TAPO framework directly modifies sampled completions via structured injection of contrastive examples, improving reward variance and training dynamics without increasing model size or data scale.

3 Preliminaries

Reinforcement Learning for Large Language Models. RL has recently become vital for aligning and enhancing the reasoning abilities of LLMs. In principle, the objective is to maximize an expected cumulative reward R over a dataset D via optimizing a policy model π_θ parameterized by θ:

$$J(\theta) = E_{q\sim D,\, a\sim\pi_\theta(\cdot|q)}\left[R_\theta(a \mid q)\right], \tag{1}$$

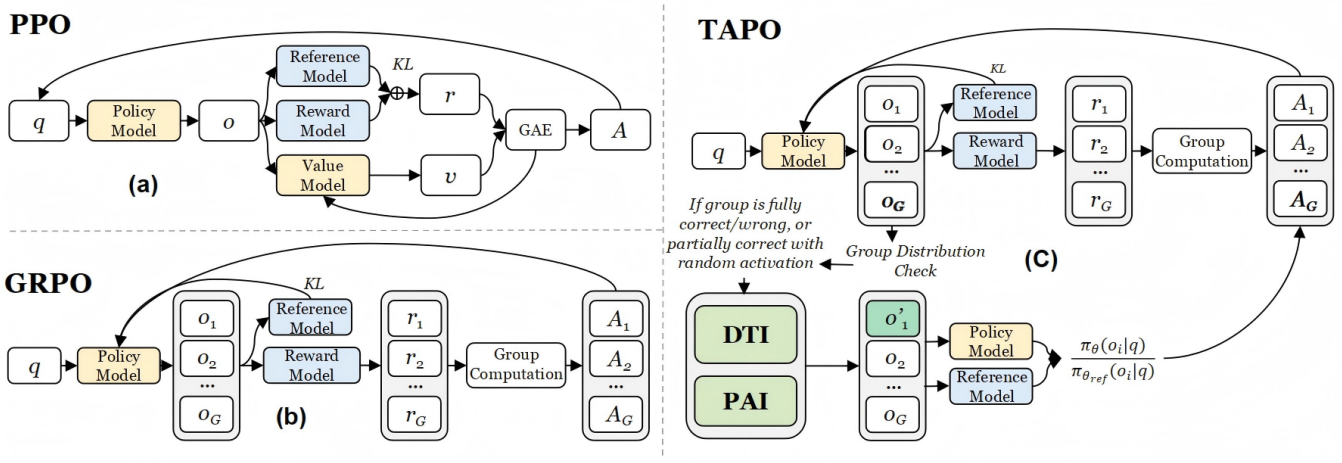

Figure 1: Comparison of reinforcement learning frameworks for LLMs. (a) PPO relies on a value model (critic) to estimate token-level advantages. (b) GRPO simplifies this by computing normalized group-level rewards across sampled completions. (c) TAPO builds on GRPO by introducing three techniques: Dynamic Teacher Injection (DTI) and Perturbed Answer Injection (PAI), which modify group composition and advantage assignment, and InfoLen, which applies semantics-aware reward shaping.

where q denotes an input instruction or question, and a is the generated completion. The reward function $R_\theta(a|q)$ measures the quality of the response a with respect to the task goal, and can be instantiated via rule-based criteria (e.g., format correctness, factual accuracy), learned reward models, or preference comparisons.

Group Relative Policy Optimization (GRPO). GRPO (Shao et al. 2024) extends PPO by sampling multiple completions $\{o_i\}_{i=1}^{G}$ per query q and optimizing the policy based on their relative advantages. Formally:

$$J_{\text{GRPO}}(\theta) = E_{q\sim D,\{o_i\}\sim\pi_{\theta_{\text{old}}}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}\Big)\Bigg] - \lambda\cdot E_q\left[\text{KL}\left[\pi_\theta(\cdot|q)\,\|\,\pi_{\text{ref}}(\cdot|q)\right]\right] \tag{2}$$

where $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$ is the token-level importance ratio, and $\hat{A}_{i,t} = \frac{r_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}$ denotes the completion-level relative advantage applied across tokens. The KL penalty term is computed with respect to a reference policy π_{ref}, which typically corresponds to the initial pretrained model or an earlier policy checkpoint. This term constrains the updated policy π_θ from deviating too far from π_{ref}, helping maintain training stability.

4 Our Method: TAPO

Unlike the RL methods mentioned above that treat sampled completions uniformly, TAPO explicitly intervenes on sampled groups and completions to improve reasoning optimization (Fig. 1(c)). The key point herein is that both degenerate groups (where all outputs are either correct or incorrect) and ambiguous completions (e.g., valid reasoning but incorrect final answer) impede the gradient guidance quality and learning efficiency. Specifically, our approach consists of three components: (1) *Dynamic Teacher Injection (DTI)*, which repairs collapsed groups by inserting a contrastive teacher sample with fixed advantage; (2) *Perturbed Answer Injection (PAI)*, which perturbs partially correct completions to provide process-sensitive supervision; and (3) *InfoLen-Aware Reward Shaping*, which reduces reward noise by penalizing redundant overlength outputs. These techniques operate under a unified objective that enhances optimization without modifying the reward model or increasing model size. In the following subsections, we first introduce the optimizaiton objective of TAPO, and then detail each component.

TAPO Objective. TAPO extends GRPO by introducing targeted interventions during advantage computation. Such interventions include injecting teacher/perturbed samples and adjusting advantage values correspondingly. The according objective is defined as:

$$J_{\text{TAPO}}(\theta) = E_{q\sim D,\{o_i\}\sim\pi_{\theta_{\text{old}}}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(r_{i,t}(\theta)\hat{A}_{i,t}, \text{clip}(r_{i,t}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}\Big)\Bigg] - \lambda\cdot E_q\left[\text{KL}\left(\pi_\theta(\cdot|q)\,\|\,\pi_{\text{ref}}(\cdot|q)\right)\right] \tag{3}$$

where the advantage $\hat{A}_{i,t}$ is defined as:

$$\hat{A}_{i,t} = \begin{cases} a_{\text{teacher}}, & \text{if } C=0 \text{ or } C=G \quad \text{(DTI)} \\ a_{\text{perturbed}}, & \text{if } 0<C<G \text{ and } s<\lambda \quad \text{(PAI)} \\ \frac{r_i-\mu_G}{\sigma_G}, & \text{otherwise} \end{cases} \tag{4}$$

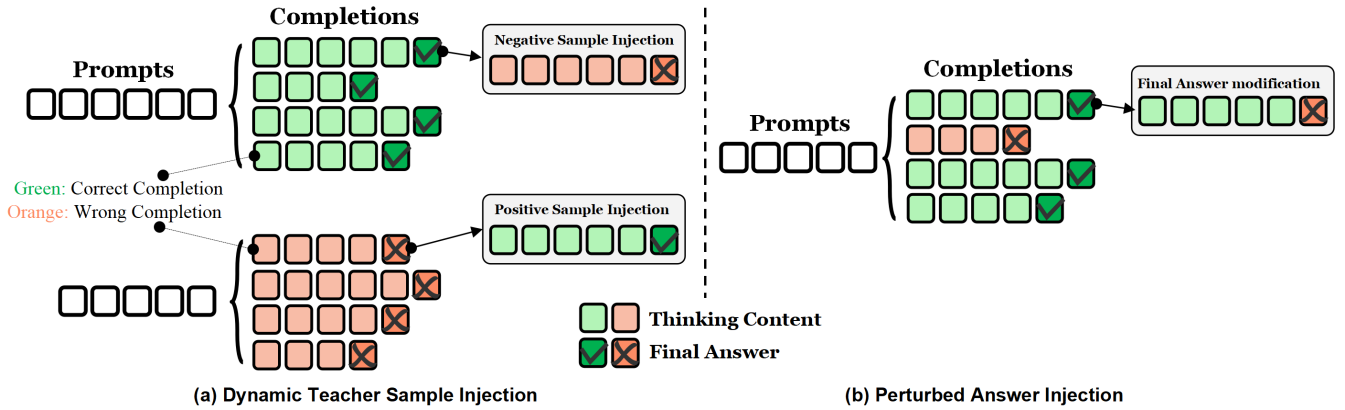

Figure 2: Sample Injection Strategies in TAPO. (a) *Dynamic Teacher Injection (DTI)* is triggered when all completions in a group are either correct (All-Correct) or incorrect (All-Wrong). A positive or negative teacher sample is injected respectively. (b) *Perturbed Answer Injection (PAI)* activates when correct and incorrect answers coexist. A correct completion is modified by flipping its final answer to introduce contrast. These injected samples preserve group normalization structure and restore gradient signal.

Unlike GRPO, TAPO decouples advantage computation for a subset of injected samples, allowing fixed-value intervention that bypasses noisy reward estimation.

Here, $r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|q,o_{i,<t})}$ denotes the importance sampling ratio evaluated over the current (possibly injected or perturbed) completion trajectory, and $C = |\{o_i \mid \texttt{is\_equivalent}(o_i, a)\}|$ denotes the number of completions equivalent to the reference answer a in the sampled group. The random score $s \sim \mathcal{U}(0,1)$ and threshold $\lambda \in (0,1)$ are used to probabilistically determine whether to apply PAI when $0 < C < G$. Injected or perturbed samples are assigned fixed advantage values (a_{teacher} or $a_{\text{perturbed}}$) without recomputing rewards, ensuring stability of group-wise normalization.

4.1 Dynamic Teacher Injection (DTI)

RL frameworks such as GRPO rely on reward variance within a sampled group of completions to compute effective gradient updates. However, groups where all completions are either correct or incorrect—termed *collapsed groups*—produce zero-variance rewards, yielding no training signal. Although some improvements, such as DAPO (Yu et al. 2025), address this by filtering out collapsed groups via dynamic sampling. While effective in high-signal regimes, this strategy could breaks down in early-stage training or on challenging datasets, where correct completions are scarce and valid groups are difficult to obtain. In such cases, the model enters a state of sample starvation, where no eligible group can be formed for learning.

To address this issue, we propose **Dynamic Teacher Injection (DTI)**, a targeted intervention that repairs collapsed groups without discarding them. This process is illustrated in Fig. 2(a). Given a query q and a group of G completions $o_1, \ldots, o_G \sim \pi_{\theta_{\text{old}}}$, we determine the number of completions semantically equivalent to the reference answer a:

$$C = |\{o_i \mid \texttt{is\_equivalent}(o_i, a)\}|. \tag{5}$$

When $C = 0$ or $C = G$, DTI injects a positive sample o^+ derived from the ground-truth solution or a negative sample o^- drawn from an unrelated query, respectively. Unlike previous methods that discard such groups or recompute noisy rewards, DTI retains the injected sample in the group and assigns it a fixed advantage a_{teacher} via *Advantage-Level Intervention (ALI)* (Sec. 4.4), ensuring stable optimization without altering group-level reward statistics. DTI enables full group utilization, restores gradient signal, and improves sample efficiency. As shown in Fig. 5(a), it also accelerates convergence and stabilizes training compared to GRPO and dynamic sampling baselines.

4.2 Perturbed Answer Injection (PAI)

Current reinforcement learning frameworks for language models typically rely on scalar rewards derived solely from final answer correctness. While effective for aligning surface-level outputs, such reward schemes penalize all incorrect answers uniformly, regardless of the value of their reasoning trajectories. Consequently, completions that follow sound logic but contain minor answer errors are treated on par with entirely wrong outputs. This lack of reasoning-sensitive reward granularity impedes the model's ability to internalize robust inference patterns.

To address this issue, we propose *Perturbed Answer Injection (PAI)*, which injects minimal yet structured contrastive perturbations into sampled groups. The core idea is to retain the original reasoning path while subtly altering the final answer, enabling the model to receive supervision signals that differentiate valid reasoning from mere correctness. Similar to DTI, PAI is coupled with Advantage-Level Intervention (ALI), assigning fixed advantage values to perturbed samples to reflect their partial correctness.

PAI is activated when the sampled group contains both correct and incorrect completions ($0 < C < G$), and a randomly drawn score $s \sim \mathcal{U}(0,1)$ falls below a predefined threshold λ. Under this condition, we randomly select one

correct completion o_i and modify its final answer, yielding a perturbed version $\tilde{o}_i$ that diverges from the reference answer a while preserving the original reasoning structure. The perturbed sample $\tilde{o}_i$ replaces one of the original completions and is assigned a fixed advantage:

$$\hat{A}_{i,t} = a_{\text{perturbed}}, \qquad \text{if } 0 < C < G \text{ and } s < \lambda, \text{(PAI activation)} \tag{6}$$

This assignment ensures that perturbations are applied only to mixed-validity groups under probabilistic gating. The PAI mechanism is illustrated in Fig. 2(b), and the fixed advantage $a_{\text{perturbed}}$ is defined via ALI (Sec. 4.4).

4.3 InfoLen-Aware Reward Shaping

In reinforcement learning training pipelines, it is common to truncate completions that exceed a predefined maximum length. By default, such overlong samples are assigned a hard punitive reward to discourage excessive responses. However, this may introduce reward noise, as valid but lengthy reasoning processes could be penalized purely for their output length, while short but semantically meaningless completions could receive disproportionately higher rewards. As a result, the model could misinterpret the penalty as a signal against legitimate reasoning patterns, which destabilizes training and impairs optimization.

To alleviate this issue, DAPO proposes *Overlong Filtering* and *Soft Overlong Punishment*, which either mask loss on truncated samples or penalize outputs proportionally to their excess length. While effective in improving stability, these strategies rely solely on the length dimension and risk suppressing completions that are long yet semantically informative. Building upon these observations, we introduce **InfoLen**, a fine-grained reward shaping method that jointly considers completion length and semantic redundancy, encouraging concise yet informative generation. We implement this idea via a continuous penalty function applied to the reward signal. Given a completion $y = \{y_1, y_2, \ldots, y_{|y|}\}$ and a task-specific reference length T_{ref}, we split y into a prefix $y_{[:T_{\text{ref}}]}$ and a suffix $y_{[T_{\text{ref}}:]}$. The semantic redundancy s is computed via cosine similarity between frozen sentence embeddings of the prefix and suffix:

$$s = \text{sim}(\phi(y_{[:T_{\text{ref}}]}), \phi(y_{[T_{\text{ref}}:]})). \tag{7}$$

We then define a length-aware penalty as:

$$\texttt{len\_penalty}(|y|) = \begin{cases} 0, & \text{if } |y| \le T_{\text{ref}}, \\ \frac{|y| - T_{\text{ref}}}{T_{\max} - T_{\text{ref}}}, & \text{if } T_{\text{ref}} < |y| \le T_{\max}, \\ 1, & \text{if } |y| > T_{\max}. \end{cases} \tag{8}$$

The total reward is shaped as:

$$r_{\text{total}} = r_{\text{base}} - \gamma \cdot (s \cdot \texttt{len\_penalty}(|y|)), \tag{9}$$

where γ is a hyperparameter controlling penalty strength. This design ensures that verbose yet informative completions are not overly penalized, while completions that are both excessively long and semantically repetitive receive stronger penalties. Our formulation jointly improves *conciseness* and *informativeness*, and enables *unified control* via a shared T_{ref}. InfoLen is model-agnostic and can be seamlessly integrated into GRPO, DAPO, and other RLHF-style training pipelines that operate on sampled completions. Empirically, InfoLen yields more stable training and slightly improved accuracy compared to other length penalty strategies (see Figure 5(b) and Table 2).

4.4 Advantage-Level Intervention (ALI)

In prior attempts, we injected contrastive completions such as DTI and PAI samples by directly modifying their reward values. However, this design introduced unintended side effects. Specifically, altering individual rewards perturbed the group-level statistics (mean and variance) used for advantage normalization in GRPO-style training, thereby destabilizing learning. As illustrated in Fig 3, reward perturbation inflated the variance and misaligned advantage distributions across the group.

To address this issue, we propose *Advantage-Level Intervention (ALI)*, which bypasses reward-level modification and directly intervenes at the advantage level. By assigning fixed advantage values to selected samples while preserving the original reward distribution, ALI maintains group-level normalization and enables more stable and precise gradient updates.

This strategy is adopted in the unified TAPO advantage rule defined in Eq. 4, where r_i is the scalar reward, and μ_G and σ_G are the group-wise mean and standard deviation computed over non-injected samples. This formulation ensures that injected samples contribute controlled optimization signals without distorting intra-group statistics. Meanwhile, uninjected samples follow the standard normalization path, preserving statistical integrity. As shown in Fig. 5(c), ALI consistently provides smoother and more stable training compared to reward-level intervention (RLI), validating its effectiveness in handling perturbed or contrastive samples across diverse group distributions.

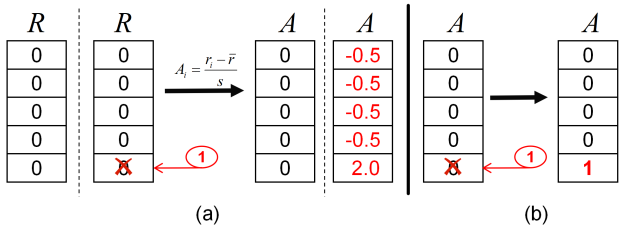

Figure 3: Comparison of reward-level vs. advantage-level intervention. (a) Injecting reward disrupts group statistics, resulting in unstable advantages. (b) ALI directly sets advantage without affecting other samples, preserving normalization consistency.

5 Experiments

In this section, we empirically validate that TAPO improves the reasoning ability of the base model, and uncover several insightful findings.

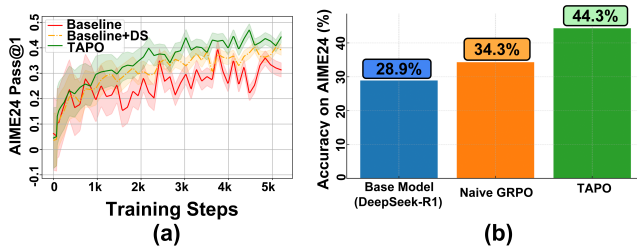

Figure 4: Evaluation on the AIME 2024 benchmark. (a) Accuracy comparison from base model to Naive GRPO and TAPO. (b) Training reward progression shows that TAPO achieves higher final reward and faster convergence than both the Naive GRPO baseline and the variant with Dynamic Sampling (DS).

5.1 Datasets and Evaluation Details

We train the model on a curated combination of high-quality math reasoning datasets, including AIME-2024 (of Problem Solving 2024), AMC (of Problem Solving 2023), OMNI-MATH (Gao et al. 2024), and MATH (Hendrycks et al. 2021). For evaluation, we use a diverse set of benchmarks spanning varying levels of difficulty: AIME-2024, AMC, MATH, MATH-500, MINERVA-MATH (Lewkowycz et al. 2022), OLYMPIADBENCH (He et al. 2024). Following standard practice, we report PASS@1 accuracy as the primary evaluation metric to assess performance.

5.2 Implementation Details

We implement TAPO based on the DeepSeek-R1-Distill-Qwen-1.5B backbone (DeepSeek-AI 2025), and train it using the Verl distributed reinforcement learning training library (Sheng et al. 2024). For TAPO-specific hyperparameters, we set the Perturbed Answer Injection (PAI) threshold λ to 0.4, and assign the fixed advantage values for injected samples as a_{teacher} with a value of ± 0.2 and $a_{\text{perturbed}}$ with a value of 0.1. The training batch size is set to 32, and the learning rate is set to 1×10^{-6}. The KL regularization (Vieillard et al. 2020) coefficient is set to 0.001, and the sampling temperature is fixed at 0.6. Each input query is sampled with 8 trajectories per iteration. To ensure robustness against randomness, we perform all experiments with three different random seeds and report the averaged outcomes. All experiments are conducted on four NVIDIA RTX 4090 GPUs (24GB VRAM each).

5.3 Main Results

Table 1 summarizes the performance of TAPO-1.5B and several strong baselines across six math reasoning benchmarks: AIME 2024, MATH-500, AMC-2023, Minerva Math, OlympiadBench, and MATH. Our TAPO-1.5B achieves the highest average accuracy of 62.87% among all models, outperforming both parameter-matched baselines and larger 7B-scale models. In particular, it surpasses the best prior 1.5B baseline (Still-1.5B, 57.42%) by 5.45 points and outperforms Qwen2.5-Math-7B-Instruct (7B, 50.70%) by a large margin. Compared to its backbone, DeepSeek-R1-Distill-Qwen-1.5B (28.9%), TAPO improves accuracy on AIME 2024 by 15.4 points, corresponding to over 50% relative gain. Even when starting from Naive GRPO (34.3%), TAPO further boosts performance by 10.0 points. These results demonstrate that our RL framework significantly enhances mathematical reasoning capability without increasing model size. As shown in Fig. 4(a), TAPO also accelerates training convergence and consistently achieves higher reward scores compared to GRPO and GRPO+DS variants, demonstrating improved optimization dynamics under our intervention strategies. This progression is visualized in Fig. 4(b), which highlights the step-wise improvements from the base model to GRPO and TAPO. TAPO also delivers consistent improvements across individual datasets. These results indicate that TAPO enhances both challenging competition-style tasks and stable academic benchmarks. Notably, all gains are achieved without increasing model size or using external datasets, but rather by applying targeted sample-level interventions that refine the learning signal and improve training stability.

5.4 Ablation Studies

Dynamic Teacher Injection (DTI). DTI mitigates the collapsed-group problem in GRPO-style training by repairing groups where all completions are either correct or incorrect, rather than discarding them. This intervention improves sample efficiency and stabilizes training in early-stage or low-signal regimes. As shown in Fig. 5, DTI yields smoother curve and explicitly accelerates convergence. On AIME-2024(Table 2), adding DTI to the GRPO baseline improves accuracy from 34.3% to 39.7% (+5.4), validating its effectiveness in restoring optimization in zero-variance scenarios.

PAI. Perturbed Answer Injection (PAI) is designed to enhance the model's sensitivity to the correctness of the reasoning process, rather than relying solely on the correctness of the final answer. Traditional training approaches tend to focus only on whether the final answer is correct, which can lead models to overlook the logical validity and coherence of the reasoning process. As a result, they may lack robustness when handling complex problems. PAI addresses this by injecting perturbed versions of answers but reserving correct reasoning content, guiding the model to generate more discriminative and informative thinking process. To validate the effectiveness of PAI, we evaluated model performance with and without PAI after applying Dynamic Teacher Injection (DTI). As shown in Table 2, adding PAI significantly improves accuracy on the AIME2024 benchmark, raising it from 39.7% to 42.8%, a gain of +3.1%. This result demonstrates that PAI enhances the model's sensitivity to process-level consistency, thereby further improving its ability to generate correct final answers.

InfoLen. Interestingly, although InfoLen operates with a softer shaping mechanism than hard truncation or length-only penalties, it achieves both higher accuracy and shorter output lengths. These results illustrate three key advantages of InfoLen over traditional length shaping strategies. First,

Model	AIME 2024	MATH 500	AMC 2023	Minerva Math	OlympiadBench	MATH	Avg.
TAPO-1.5B	**44.3**	**88.6**	**74.3**	31.6	**50.2**	**88.2**	**62.87**
Still-1.5B (Chen et al. 2025)	32.5	84.4	66.7	29.0	45.4	86.5	57.42
Qwen2.5-7B-SimpleRL (Zeng et al. 2025)	26.7	82.4	62.5	**39.7**	43.3	85.1	56.62
$\text{Model}_{\text{base}}$ + Naive GRPO	34.3	84.5	65	28.2	43.7	85.5	56.86
$\text{Model}_{\text{base}}$	28.9	82.8	62.9	26.5	43.3	86.2	55.08
Eurus-2-7B-PRIME (Cui et al. 2025)	26.7	79.2	57.8	38.6	42.1	83.8	54.70
rStar-Math-7B (Guan et al. 2025)	26.7	78.4	47.5	36.5	47.1	82.6	53.13
Qwen2.5-Math-7B-Instruct (Yang et al. 2024a)	13.3	79.8	50.6	34.6	40.7	85.2	50.70
O1-Preview (OpenAI 2024)	40.0	81.4	–	–	–	–	–

Table 1: Performance comparison across math benchmarks. All scores are percentages. $\text{Model}_{\text{base}}$ denotes the backbone DeepSeek-R1-Distill-Qwen-1.5B (Guo et al. 2025).

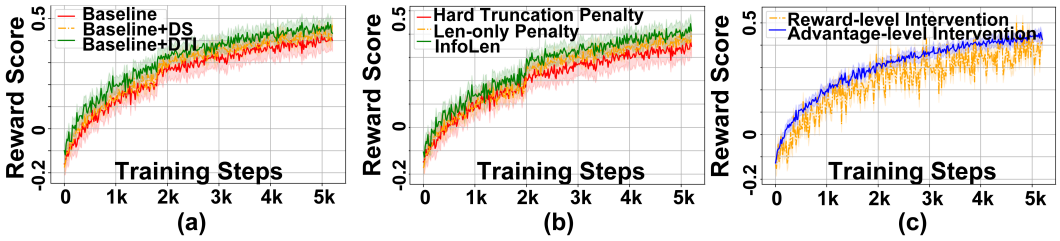

Figure 5: Training reward progression under different settings. (a) TAPO with DTI achieves faster convergence and higher final reward than the GRPO baseline and dynamic sampling (DS). (b) InfoLen reduces reward noise and yields smoother learning dynamics compared to hard truncation and length-only penalty. (c) Advantage-Level Intervention (ALI) offers more stable gradient supervision than reward-based intervention (RLI), especially in the presence of noisy or injected samples.

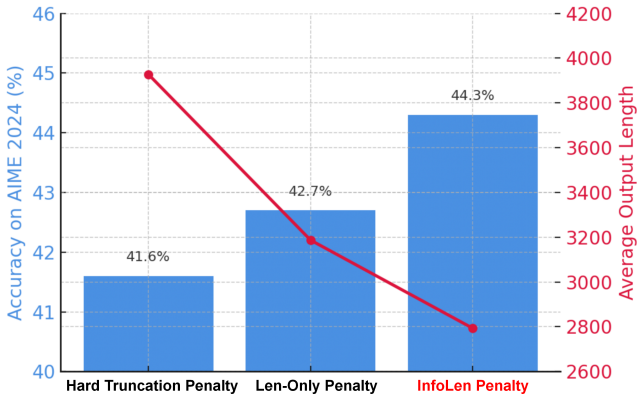

Figure 6: Impact of reward shaping strategies. InfoLen reduces redundancy while improving reasoning accuracy.

unlike *Length-only Penalty* which applies linear decay regardless of content, InfoLen penalizes completions based on both length and semantic redundancy. This encourages the model to avoid verbose but uninformative responses while retaining useful reasoning steps. As shown in Fig. 6, InfoLen yields shorter outputs than length-only shaping (2793 vs 3187 tokens) while further improving accuracy. Second, InfoLen promotes short yet informative completions. Since its penalty reflects semantic redundancy rather than length alone, the model learns to generate concise outputs without omitting important content, leading to more efficient inference patterns. Third, InfoLen improves training stability.

Model	$\text{AIME24}_{\text{pass@1}}$
DeepSeek-R1-Distill-Qwen-1.5B	28.9
Naive GRPO	34.3
+ DTI	39.7
+ PAI	42.8
+ InfoLen (**TAPO**)	**44.3**

Table 2: AIME2024 performance of progressive techniques applied to TAPO. Note that ALI is applied wherever DTI or PAI is used, and is not shown as a separate row in the ablation table.

Compared to hard truncation strategies that introduce noisy reward signals, InfoLen leads to more stable learning dynamics, as reflected by smoother reward progression during training (see Fig. 5(b)).

6 Conclusion

We propose **TAPO**, a reinforcement learning framework that enhances the reasoning capability of LLMs through targeted sample-level interventions, including Dynamic Teacher Injection (DTI) and Perturbed Answer Injection (PAI), along with InfoLen, a semantics-aware reward shaping strategy. Extensive experiments across six math benchmarks such as AIME 2024, AMC 2023, and MATH demonstrate that TAPO significantly improves both accuracy and training stability over the Naive GRPO baseline, and also outperforms strong 1.5B and 7B models without increasing model size.

Ethical Statement

This research adheres to the AAAI Code of Ethics. All experiments were conducted on publicly available datasets and simulated environments, without involving human subjects or personally identifiable information.

Acknowledgements

This work is part of the research project "Research on Hybrid Collaborative Intelligence and Artificial Intelligence–Driven Architectural Design and Robotic Construction Methods in High-Density Urban Environments (HIRA)", supported by the Shenzhen Science and Technology Innovation Commission (Grant No. 20231129094641002). The authors' affiliation institution, Tsinghua Shenzhen International Graduate School, also supports this research under the Scientific Research Start-up Funds (Grant No. 002023009C) and the Shenzhen Pengcheng Peacock Specific Program.

References

Aksitov, R.; Miryoosefi, S.; Li, Z.; Li, D.; Babayan, S.; Kopparapu, K.; Fisher, Z.; Guo, R.; Prakash, S.; Srinivasan, P.; et al. 2023. Rest meets react: Self-improvement for multi-step reasoning llm agent. *arXiv preprint arXiv:2312.10003*.

Bai, Y.; Jones, A.; Ndousse, K.; Askell, A.; Chen, A.; DasSarma, N.; Drain, D.; Fort, S.; Ganguli, D.; Henighan, T.; et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Cao, Y.; Zhao, H.; Cheng, Y.; Shu, T.; Chen, Y.; Liu, G.; Liang, G.; Zhao, J.; Yan, J.; and Li, Y. 2024. Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods. *arXiv preprint arXiv:2404.00282*.

Chaudhari, S.; Aggarwal, P.; Murahari, V.; Rajpurohit, T.; Kalyan, A.; et al. 2024. RLHF Deciphered: A Critical Analysis of Reinforcement Learning from Human Feedback for LLMs. *arXiv preprint arXiv:2404.08555*.

Chen, Z.; Min, Y.; Zhang, B.; Chen, J.; Jiang, J.; Cheng, D.; Zhao, W. X.; Liu, Z.; Miao, X.; Lu, Y.; et al. 2025. An Empirical Study on Eliciting and Improving R1-like Reasoning Models. *arXiv preprint arXiv:2503.04548*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 4299–4307.

Cui, G.; Yuan, L.; Wang, Z.; Wang, H.; Li, W.; He, B.; Fan, Y.; Yu, T.; Xu, Q.; Chen, W.; et al. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Dai, M.; Yang, C.; and Si, Q. 2025. S-GRPO: Early Exit via Reinforcement Learning in Reasoning Models. *arXiv preprint*, arXiv:2505.07686.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Fujimoto, S.; van Hoof, H.; and Meger, D. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning*, 1587–1596. PMLR.

Gao, B.; Song, F.; Yang, Z.; Cai, Z.; Miao, Y.; Dong, Q.; Li, L.; Ma, C.; Chen, L.; Xu, R.; et al. 2024. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Gao, L.; Madaan, A.; Zhou, S.; Alon, U.; Liu, P.; Yang, Y.; Callan, J.; and Neubig, G. 2022. PAL: Program-aided Language Models. *arXiv preprint arXiv:2211.10435*.

Guan, X.; Zhang, L. L.; Liu, Y.; Shang, N.; Sun, Y.; Zhu, Y.; Yang, F.; and Yang, M. 2025. rStar-Math: Small LLMs Can Master Math Reasoning with Self-Evolved Deep Thinking. *arXiv preprint arXiv:2501.04519*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z. L.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; et al. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; et al. 2022. Solving Quantitative Reasoning Problems with Language Models. *arXiv preprint arXiv:2206.14858*.

Li, R.; Fu, J.; Zhang, B.-W.; Huang, T.; Sun, Z.; Lyu, C.; Liu, G.; Jin, Z.; and Ge, L. 2023. TACO: Topics in Algorithmic COde generation dataset. *arXiv preprint*, arXiv:2312.14852.

Lin, Z.; Lin, M.; Xie, Y.; and Ji, R. 2025. CPPO: Accelerating the Training of Group Relative Policy Optimization-Based Reasoning Models. *arXiv preprint arXiv:2503.22342*.

Liu, T.; Chen, Z.; Fang, Z.; Luo, W.; Tian, M.; and Liu, Z. 2025. MathEval: A Comprehensive Benchmark for Evaluating Large Language Models on Mathematical Reasoning Capabilities. *Frontiers of Digital Education*, 2(16).

Liu, Z.; Sun, X.; and Zheng, Z. 2024. Enhancing LLM Safety via Constrained Direct Preference Optimization. *arXiv preprint arXiv:2403.02475*.

Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; and Roberts, A. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. *arXiv preprint arXiv:2301.13688*.

Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.-W.; Galley, M.; and Gao, J. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations (ICLR)*.

Ma, Y.; Gou, Z.; Hao, J.; Xu, R.; Wang, S.; Pan, L.; Yang, Y.; Cao, Y.; Sun, A.; Awadalla, H.; and Chen, W. 2024. SciAgent: Tool-augmented Language Models for Scientific Reasoning. *arXiv preprint arXiv:2402.11451*.

of Problem Solving, A. 2023. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php?title=AMC_Problems_and_Solutions. Accessed: 2025-04-20.

of Problem Solving, A. 2024. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions. Accessed: 2025-04-20.

OpenAI. 2024. Introducing OpenAI o1-preview. https://openai.com/index/introducing-openai-o1-preview/. Accessed: 2025-05-13.

OpenAI. 2024. OpenAI o1: Unreleased Reinforcement Learning Model. Internal report, not publicly available. Referencing CPPO discussion of OpenAI-o1.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Pal, A.; Karkhanis, D.; Dooley, S.; Roberts, M.; Naidu, S.; and White, C. 2024. Smaug: Fixing Failure Modes of Preference Optimisation with DPO-Positive. *arXiv preprint arXiv:2402.13228*.

Pan, S.; Lialin, V.; Muckatira, S.; and Rumshisky, A. 2023. Let's Reinforce Step by Step. *arXiv preprint arXiv:2311.05821*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint arXiv:2305.18290*.

Roziere, B.; Allal, L.; Izacard, G.; Ram, A.; and Lample, G. 2023. Code Llama: Open Foundation Models for Code. *arXiv preprint arXiv:2308.12950*.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. S.; Aly, R.; Chaffin, C.; Scao, T. L.; von Platen, P.; Patil, S.; Xu, Y.; et al. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. *arXiv preprint arXiv:2110.08207*.

Sanwal, M. 2025. Layered Chain-of-Thought Prompting for Multi-Agent LLM Systems: A Comprehensive Approach to Explainable Large Language Models. *arXiv preprint arXiv:2501.18645*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *arXiv preprint arXiv:2302.04761*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Seßler, K.; Rong, Y.; Gözlüklü, E.; and Kasneci, E. 2024. Benchmarking Large Language Models for Math Reasoning Tasks. In *arXiv preprint arXiv:2408.10839*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; and Song, J. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*.

Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2024. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv preprint arXiv:2409.19256*.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, 3008–3021.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 2000. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, 1057–1063.

Team, K.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Tong, W.; and Zhang, T. 2024. CODEJUDGE: Evaluating Code Generation with Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 20032–20051.

Uesato, J.; Kushman, N.; Kumar, R.; Song, F.; Siegel, N.; Wang, L.; Creswell, A.; Irving, G.; and Higgins, I. 2022. Solving Math Word Problems with Process- and Outcome-Based Feedback. *arXiv preprint arXiv:2211.14275*.

Vieillard, N.; Kozuno, T.; Scherrer, B.; Pietquin, O.; Munos, R.; and Geist, M. 2020. Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 12163–12174.

Wang, B.; Min, S.; Deng, X.; Shen, J.; Wu, Y.; Zettlemoyer, L.; and Sun, H. 2022a. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. *arXiv preprint arXiv:2212.10001*.

Wang, S.; Zhang, S.; Zhang, J.; Hu, R.; Li, X.; Zhang, T.; Li, J.; Wu, F.; Wang, G.; and Hovy, E. 2024. Reinforcement Learning Enhanced LLMs: A Survey. *arXiv preprint arXiv:2412.10400*.

Wang, X.; Hu, Z.; Lu, P.; Zhu, Y.; Zhang, J.; Subramaniam, S.; Loomba, A. R.; Zhang, S.; Sun, Y.; and Wang, W. 2023. SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models. *arXiv preprint arXiv:2307.10635*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022b. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022a. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 24824–24837.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36: 59008–59033.

Xu, J.; Ung, M.; Komeili, M.; Arora, K.; Boureau, Y.-L.; and Weston, J. 2022. Learning new skills after deployment: Improving open-domain internet-driven dialogue with human feedback. *arXiv preprint arXiv:2208.03270*.

Yan, X.; Song, Y.; Feng, X.; Yang, M.; Zhang, H.; Bou Ammar, H.; and Wang, J. 2025. Efficient Reinforcement Learning with Large Language Model Priors. In *International Conference on Learning Representations (ICLR)*.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; et al. 2024a. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yang, Z.; Li, P.; Yan, M.; Zhang, J.; Huang, F.; and Liu, Y. 2024b. React meets actre: When language agents enjoy training data autonomy. *arXiv preprint arXiv:2403.14589*.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Fan, T.; Liu, G.; Liu, L.; Liu, X.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yuan, Z.; Yuan, H.; Li, C.; Dong, G.; Lu, K.; Tan, C.; Zhou, C.; and Zhou, J. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv preprint arXiv:2308.01825*.

Zeng, W.; Huang, Y.; Liu, Q.; Liu, W.; He, K.; Ma, Z.; and He, J. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Zhang, J.; and Zuo, C. 2025. GRPO-LEAD: A Difficulty-Aware Reinforcement Learning Approach for Concise Mathematical Reasoning in Language Models. *arXiv preprint arXiv:2504.09696*.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic Chain of Thought Prompting in Large Language Models. *arXiv preprint arXiv:2210.03493*.

Zheng, J.; Cao, B.; Ma, Z.; Pan, R.; Lin, H.; Lu, Y.; Han, X.; and Sun, L. 2024. Beyond Correctness: Benchmarking Multi-dimensional Code Generation for Large Language Models. *arXiv preprint arXiv:2407.11470*.

Zhou, D.; Schärli, N.; Hou, L.; Wei, J.; Scales, N.; Wang, X.; Schuurmans, D.; Cui, C.; Bousquet, O.; Le, Q. V.; et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T. B.; Radford, A.; et al. 2019. Fine-Tuning Language Models from Human Preferences. *arXiv preprint arXiv:1909.08593*.