Bridging SFT and RL: Dynamic Policy Optimization for Robust Reasoning

Taojie Zhu[1,2,†] [*], **Dongyang Xu**[2,†,‡], **Ding Zou**[2,♠], **Sen Zhao**[3], **Qiaobo Hao**[2], **Zhiguo Yang**[2], **Yonghong He**[1,‡]

[1]Shenzhen International Graduate School, Tsinghua University
[2]Intelligent System Department, Zhongxing Telecom Equipment (ZTE)
[3]Academy of Advanced Interdisciplinary Studies, Chongqing University of Posts and Telecommunications

[†]Equal contribution. [‡]Corresponding author. [♠]Project Leader.
xu.dongyang2@zte.com.cn, heyh@sz.tsinghua.edu.cn

Abstract

Post-training paradigms for Large Language Models (LLMs), primarily Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL), face a fundamental dilemma: SFT provides stability (low variance) but suffers from high fitting bias, while RL enables exploration (low bias) but grapples with high gradient variance. Existing unified optimization strategies often employ naive loss weighting, overlooking the statistical conflict between these distinct gradient signals. In this paper, we provide a rigorous theoretical analysis of this bias-variance trade-off and propose **DYPO** (Dynamic Policy Optimization), a unified framework designed to structurally mitigate this conflict. DYPO integrates three core components: (1) a *Group Alignment Loss (GAL)* that leverages intrinsic group dynamics to significantly reduce RL gradient variance; (2) a *Multi-Teacher Distillation* mechanism that corrects SFT fitting bias via diverse reasoning paths; and (3) a *Dynamic Exploitation-Exploration Gating* mechanism that adaptively arbitrates between stable SFT and exploratory RL based on reward feedback. Theoretical analysis confirms that DYPO linearly reduces fitting bias and minimizes overall variance. Extensive experiments demonstrate that DYPO significantly outperforms traditional sequential pipelines, achieving an average improvement of 4.8% on complex reasoning benchmarks and 13.3% on out-of-distribution tasks. Our code is publicly available at https://github.com/Tocci-Zhu/DYPO.

1 Introduction

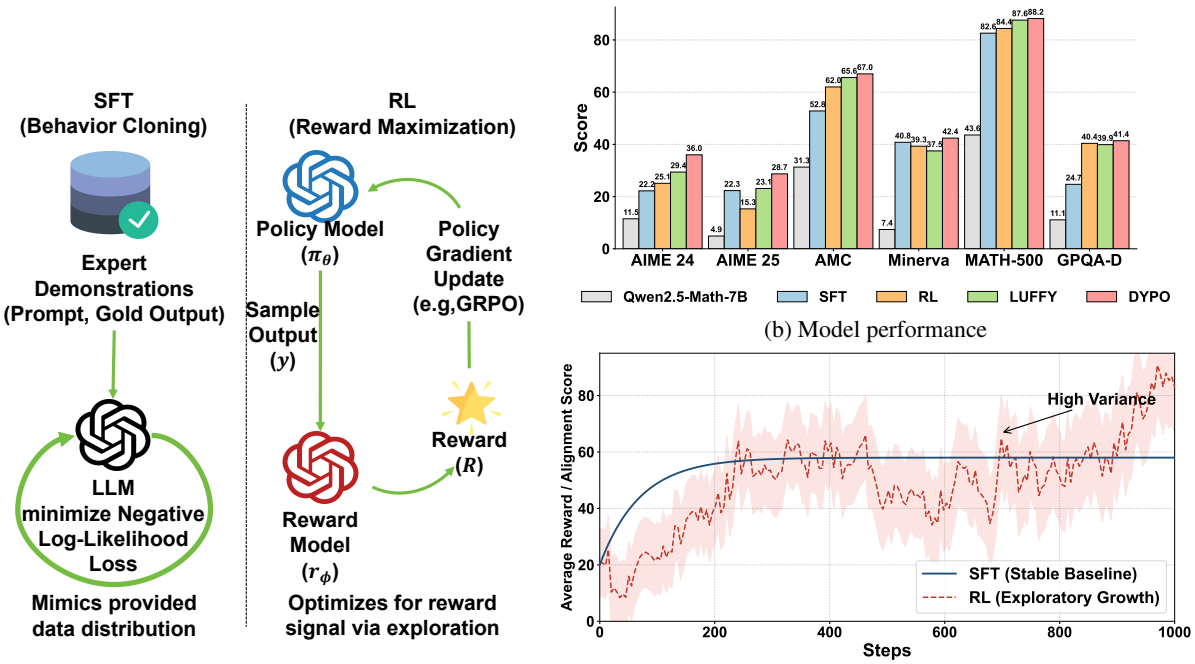
The reasoning capabilities of Large Language Models (LLMs) have become a central focus in artificial intelligence (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025). While reasoning-guidance techniques like Chain-of-Thought (CoT) prompting have significantly advanced model performance on multi-step tasks (Wei et al., 2022), traditional prompting methods relying on static templates struggle with scalability and dynamic adaptability. Consequently, the research focus has shifted toward the post-training stage to enhance robustness and generalization (Wang et al., 2023). Current mainstream post-training paradigms generally fall into two categories: i) Supervised Fine-Tuning (SFT): SFT offers efficient knowledge injection by learning from high-quality CoT corpora (Sanh et al., 2022; Wei et al., 2021). Its low-variance nature ensures stability and rapid fitting, but often at the cost of limited exploratory capacity and restricted Out-of-Distribution (OOD) generalization. ii) Reinforcement Learning (RL): Methods such as RLHF or RLVR allow models to autonomously explore the reasoning space via reward signals, substantially enhancing generalization (Ouyang et al., 2022; Ramamurthy et al., 2022; Schulman et al., 2017; Shao et al., 2024a). Unlike SFT, RL relies on the base model's intrinsic capabilities; consequently, weaker models often struggle to capture sparse reward signals in complex tasks.

To combine these strengths, researchers have widely adopted a "SFT-then-RL" training pipeline (Touvron et al., 2023; Yoshihara et al., 2025). However, this sequential approach suffers from *bias propagation*, where SFT-induced biases misguide subsequent RL exploration (Lv et al., 2025), alongside significant computational overhead.

Recent research has therefore shifted toward *unified optimization*, which combines SFT and RL objectives within a single training process (Yan et al., 2025; Fu et al., 2025; Zhang et al., 2025; Chen et al., 2025). Representative methods include SuperRL (Liu et al., 2025a), which adopts a binary switching strategy between supervision and reinforcement learning, and CHORD (Zhang et al., 2025), which harmonizes the two objectives through dynamic soft weighting. These approaches

[*]Work done during internship at ZTE.

Findings of the Association for Computational Linguistics: ACL 2026, pages 4468–4484
July 2-7, 2026 ©2026 Association for Computational Linguistics



(a) Data Flow & Training Mechanisms

(b) Model performance

(c) Learning Curves

Figure 1: **The SFT-RL Dilemma:** Balancing the high-bias stability of SFT against the high-variance exploration of RL.

highlight the growing interest in unified SFT-RL post-training, but they still apply a largely uniform optimization recipe across samples whose learning signals differ fundamentally in reliability.

Despite the growing interest in unified SFT-RL training, existing fusion strategies predominantly operate at a "surface level" via simple loss weighting (Lv et al., 2025). This approach overlooks two fundamental issues. First, it ignores the inherent statistical conflict between the gradient signals: SFT gradients are *high-bias* (fitting static data) but *low-variance* (Wu et al., 2025), whereas RL gradients are *low-bias* (reward-driven) but *high-variance* (due to sampling stochasticity) (Ramamurthy et al., 2022). Naively aggregating these conflicting vectors is sub-optimal, as RL's high variance destabilizes training while SFT's high bias constrains exploration. Second, this uniform approach fails to account for regime-dependent differences in sample difficulty. Specifically, trivial samples provide marginal optimization signals since model performance is already saturated; hard samples yield extremely sparse rewards, rendering RL highly inefficient; and only mid-difficulty samples simultaneously preserve reward discrimination and expose meaningful failure modes. Consequently, globally mixing SFT and RL objectives cannot fully resolve the multidimensional mismatch between stable but biased supervision and exploratory but high-variance policy optimization.

In this paper, we first provide a theoretical analysis formally defining this bias-variance trade-off in SFT-RL fusion. We then propose **DYPO** (DYnamic Policy Optimization), a unified framework that introduces structural solutions to concurrently mitigate both limitations.Unlike binary switching or soft weighting methods, DYPO performs instance-level routing based on rollout outcomes and assigns each regime to a distinct optimization objective.

Specifically, DYPO comprises three core components:

- **Dynamic Difficulty Grading**: A mechanism that dynamically categorizes queries based on group rollout outcomes. It effectively arbitrates the optimization pathway: routing complete failures (Hard) to stable SFT for knowledge injection, while directing inconsistent attempts (Mid) to low-bias RL for exploration.
- **Bias Correction (SFT)**: For 'Hard' samples, we employ a Multi-Teacher Distillation mechanism to correct the fitting bias inherent in SFT by aggregating diverse reasoning paths from different teacher models.
- **Variance Reduction (RL)**: For 'Mid' sam-

ples, we introduce a Group Alignment Loss (GAL) (Rafailov et al., 2023) that leverages intrinsic group dynamics.By effectively reinforcing winning samples while suppressing losing ones, GAL significantly reduces RL gradient variance compared to standard pairwise losses.

Theoretically, we prove that our Dynamic Difficulty Grading mechanism minimizes overall variance by strategically allocating queries based on reward feedback. For 'Hard' samples, the triggered multi-teacher strategy linearly reduces fitting bias; for 'Mid' samples, the GAL reduces gradient variance by orders of magnitude compared to GRPO. Experimentally, DYPO yields 5–10% performance gains on complex reasoning benchmarks.

2 Preliminaries

In this section, we formalize the reasoning trace generation problem and review the two foundational paradigms: SFT and RL. We specifically highlight their respective statistical challenges—fitting bias in SFT and gradient variance in RL—which motivate our proposed approach.

2.1 Problem Formulation

We model the reasoning task as a sequential decision-making process. Given an input prompt q sampled from a distribution $\mathcal{D}$, the LLM functions as a stochastic policy $\pi_\theta(\tau|q)$ parameterized by θ. Here, $\tau = (a_1, a_2, \ldots, a_T)$ represents a reasoning trajectory consisting of a sequence of tokens. The probability of generating a trajectory is factorized autoregressively:

$$\pi_\theta(\tau|q) = \prod_{t=1}^{T} \pi_\theta(a_t|q, a_{<t}) \tag{1}$$

Upon completion, the trajectory τ is evaluated by a reward function $R(q, \tau) \in \mathbb{R}$. The objective is to maximize the expected reward $J(\theta) = \mathbb{E}_{q\sim\mathcal{D}}\mathbb{E}_{\tau\sim\pi_\theta(\cdot|q)}[R(q, \tau)]$(Ouyang et al., 2022; Ziegler et al., 2019).

2.2 SFT and Fitting Bias

Standard SFT adapts the policy by minimizing the negative log-likelihood on a static dataset $\mathcal{D}_{\text{sft}}$ containing gold-standard pairs (q, τ^*)(Touvron et al., 2023; Wei et al., 2021):

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{(q,\tau^*)\sim\mathcal{D}_{\text{sft}}}[-\log \pi_\theta(\tau^*|q)] \tag{2}$$

While SFT provides stable supervision, it inherently suffers from **fitting bias**. Since the optimization is constrained to the fixed support of $\mathcal{D}_{\text{sft}}$, the model tends to overfit the specific distribution of the single teacher or dataset. This mimicry limits the model's ability to explore novel reasoning paths and often leads to sub-optimal local minima where the policy fails to generalize beyond the training examples.

2.3 GRPO and Gradient Variance

To enable exploration, we employ GRPO(Shao et al., 2024a). For each prompt q, GRPO samples a group of trajectories $G = \{\tau_1, \tau_2, \ldots, \tau_k\}$ and optimizes the policy using group-normalized advantages:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{q\sim\mathcal{D}}\Bigg[\frac{1}{k}\sum_{i=1}^{k}\Big(\text{CLIP}(\rho_i, \hat{A}_i, \epsilon)\Big) - \beta_{\text{KL}}\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\Bigg] \tag{3}$$

where ρ_i is the probability ratio and $\hat{A}_i$ is the advantage computed by standardizing rewards within the group G. Although GRPO provides an low-biased objective for reward maximization, it introduces high gradient variance. This instability arises from the stochastic nature of trajectory sampling and the reliance on a sparse reward signal. With a limited group size k, the Monte Carlo estimate of the gradient can be highly noisy, often destabilizing the training process in complex reasoning tasks.

3 Methodology

In this section, we present **DYPO**, a unified framework that dynamically balances exploration and stability by routing queries to the most suitable optimization pathway. The key intuition is that different queries expose learning signals of different reliability: easy queries are already saturated, hard queries lack usable reward signals, and only mid-difficulty queries preserve informative relative feedback for RL.

Formally, the unified objective of DYPO is constructed as a dynamic mixture of supervised and

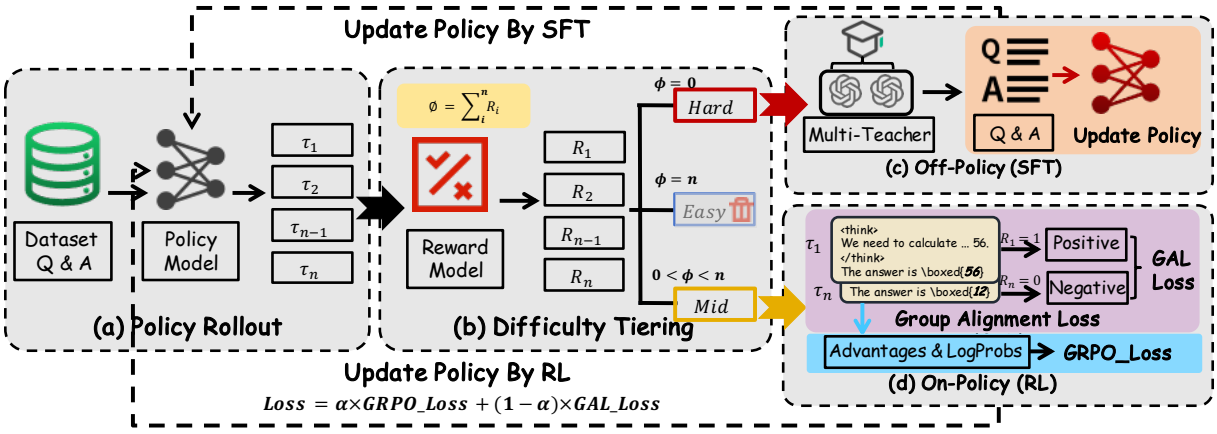


Figure 2: The overall framework of DYPO. The system employs a Dynamic Difficulty Grading mechanism to categorize queries into Easy, Hard, and Mid tiers based on group rollout outcomes, dispatching them to the most effective optimization pathway.

reinforcement learning objectives.

$$\mathcal{L}_{\text{DYPO}}(\theta) = \mathbb{E}_q \Big[\underbrace{\mathbb{I}_{\mathcal{H}}(q) \cdot \gamma \mathcal{L}_{\text{SFT}}}_{\text{Bias Mitigation (Sec. 3.2)}} + \underbrace{\mathbb{I}_{\mathcal{M}}(q) \cdot \big(\alpha \mathcal{L}_{\text{GRPO}} + (1-\alpha)\mathcal{L}_{\text{GAL}}\big)}_{\text{Variance Reduction (Sec. 3.3)}} \Big] \tag{4}$$

where $\mathbb{I}_{\mathcal{H}}$ and $\mathbb{I}_{\mathcal{M}}$ are indicator functions determined by a difficulty grading mechanism (Sec. 3.1). Easy samples have zero contribution to the training objective and are therefore omitted from Eq. (4) for simplicity. The coefficients γ and α control the strength of distillation and the bias-variance trade-off in RL, respectively. Through the structural separation of the learning process, Eq. (4) allows DYPO to better manage the bias-variance trade-off across different learning stages.

3.1 Dynamic Difficulty Grading

We propose a strategy to distinguish data samples based on their impact on training variance, termed Dynamic Difficulty Grading, to optimize the bias-variance trade-off. Specifically, we aim to filter out trivial instances yielding negligible gradients and overly complex outliers that induce high variance, thereby isolating the informative samples most conducive to robust optimization.

Specifically, given a query q, the policy π_θ generates a group of k trajectories $G = \{\tau_1, \ldots, \tau_k\}$. Let $R(\tau_i) \in \{0, 1\}$ denote the binary correctness reward. We categorize the training instance into three levels based on the reward distribution:

- **Easy ($\mathcal{E}$):** The model solves the problem consistently ($\forall \tau \in G, R(\tau) = 1$). These samples provide diminishing returns for gradient estimation and are **discarded** for efficiency.
- **Hard ($\mathcal{H}$):** The model fails completely ($\forall \tau \in G, R(\tau) = 0$). In this regime, valid reward signals are unavailable, causing standard RL gradients to fail. To bridge this gap, we adopt **Multi-Teacher Distillation**.
- **Mid ($\mathcal{M}$):** The group contains mixed results ($\exists \tau_i, \tau_j \in G, R(\tau_i) \neq R(\tau_j)$). This represents the critical learning frontier. We apply a hybrid objective of **GRPO** and **GAL** to leverage the relative feedback.

By accurately categorizing each sample into distinct difficulty levels, we integrate a refined sample stratification into our unified optimization framework. This ensures the model prioritizes the most effective learning signals during training. Subsequently, we detail how this framework leverages such stratification to effectively balance variance and bias.

3.2 Mitigating Supervisory Bias via Multi-Teacher Distillation

For instances falling into the Hard regime ($\mathbb{I}_{\mathcal{H}} = 1$), the model suffers from insufficient prior knowledge to formulate valid reasoning paths, making autonomous exploration prone to failure. To resolve the issue while mitigating the supervisory bias typically associated with single-source supervision, we introduce a **Multi-Teacher Distillation** strategy.

Rather than relying on a deterministic target from a single source, we maintain an ensemble of m teacher oracles. For each hard query, we uniformly sample a target trajectory τ_{tgt} from the

candidate set $\{\tau^{(1)}, \ldots, \tau^{(m)}\}$ derived from these teachers:

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{\tau_{\text{tgt}} \sim \mathcal{U}(\{\tau^{(1)}, \ldots, \tau^{(m)}\})} \left[-\log \pi_\theta(\tau_{\text{tgt}}|q) \right] \tag{5}$$

The theoretical validation for utilizing multiple teachers lies in the decomposition of supervisory bias. Let τ^* denote the ground-truth optimal reasoning path. A single teacher i provides a supervision signal $\tau^{(i)}$ which deviates from the truth according to the following decomposition:

$$\tau^{(i)} = \tau^* + \mathbf{b}_{\text{sys}} + \mathbf{b}_i \tag{6}$$

Here, $\mathbf{b}_{\text{sys}}$ represents the systematic bias common to all LLMs (e.g., limitation of language modality), while $\mathbf{b}_i$ represents the *idiosyncratic bias* specific to the i-th teacher model (e.g., preference for specific formatting or distinct hallucination patterns).

When relying on a single teacher ($m = 1$), the student model blindly inherits the full bias vector $\|\mathbf{b}_{\text{sys}} + \mathbf{b}_i\|$. However, under the *diversity assumption*—where different teachers exhibit uncorrelated bias directions (i.e., $\mathbb{E}[\mathbf{b}_i] \approx 0$)—the aggregation of m teachers significantly attenuates the idiosyncratic component. The effective bias of the multi-teacher ensemble is derived by averaging the individual error vectors:

$$\begin{aligned} \|\text{Bias}_{\text{multi}}\|^2 &= \left\| \mathbf{b}_{\text{sys}} + \frac{1}{m} \sum_{i=1}^{m} \mathbf{b}_i \right\|^2 \\ &= \|\mathbf{b}_{\text{sys}}\|^2 + \underbrace{\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{b}_i \right\|^2}_{\text{Idiosyncratic Term}} \end{aligned} \tag{7}$$

Assuming independence between teacher biases, the magnitude of the idiosyncratic bias reduces linearly with m:

$$\mathbb{E}\left[\left\| \frac{1}{m} \sum_{i=1}^{m} \mathbf{b}_i \right\|^2 \right] = \frac{1}{m} \bar{\sigma}_{\text{bias}}^2 \tag{8}$$

Consequently, we can formally establish that the multi-teacher objective strictly reduces the total supervisory bias compared to the single-teacher baseline ($m = 1$):

$$\mathbb{E}[\|\text{Bias}_{\text{multi}}\|^2] = \|\mathbf{b}_{\text{sys}}\|^2 + \frac{\bar{\sigma}_{\text{bias}}^2}{m} \tag{9}$$

$$\mathbb{E}[\|\text{Bias}_{\text{single}}\|^2] = \|\mathbf{b}_{\text{sys}}\|^2 + \bar{\sigma}_{\text{bias}}^2 \tag{10}$$

$$\mathbb{E}[\|\text{Bias}_{\text{multi}}\|^2] < \mathbb{E}[\|\text{Bias}_{\text{single}}\|^2] \tag{11}$$

In essence, aggregating supervision signals cancels out the idiosyncratic biases inherent to individual teachers, guiding the model toward the robust intersection of valid reasoning paths. By providing a stabilized policy prior with reduced bias, Multi-Teacher SFT enables effective exploration of the solution space, seamlessly bridging the gap between supervised likelihood maximization and expected reward maximization.

3.3 Variance-Reduced RL with Group Alignment

The Mid regime ($\mathbb{I}_{\mathcal{M}} = 1$) represents the critical learning frontier where the model exhibits capability but lacks consistency, producing a mixture of correct and incorrect responses. We identify this regime as the target scenario for RL intervention. While Reinforcement Learning is theoretically ideal for amplifying these correct signals to improve performance, standard RL algorithms often struggle with high variance in gradient estimates, which can severely impede convergence stability and speed. To address this bottleneck, we propose a novel optimization strategy: **Variance-Reduced RL with Group Alignment**.

Instability of GRPO Gradient. To motivate our approach, we first examine the gradient of GRPO. For a group of size k, the gradient is:

$$g_{\text{GRPO}} = \frac{1}{k} \sum_{i=1}^{k} \hat{A}_i \cdot \nabla_\theta \log \pi_\theta(\tau_i|q) \tag{12}$$

Let $\Sigma_s \triangleq \mathbb{E}[\|\nabla_\theta \log \pi_\theta\|^2]$ be the variance of the score function. Assuming normalized advantages ($\mathbb{E}[\hat{A}^2] \approx 1$), the variance of this estimator scales as:

$$\text{Var}(g_{\text{GRPO}}) \approx \frac{1}{k} \Sigma_s \tag{13}$$

While increasing k reduces variance, the unbounded nature of $\hat{A}_i$ induces high variance, leading to unstable updates during early exploration.

Group Alignment Loss (GAL). To mitigate this, we introduce GAL. As illustrated in Figure 3, the current policy first generates a group of rollouts, which are categorized into positive samples (successful trajectories) and negative samples (failed trajectories) based on correctness. The core intuition is to explicitly widen the gap between these two groups by "pulling" the policy towards correct reasoning paths while "pushing" it away from incorrect ones.

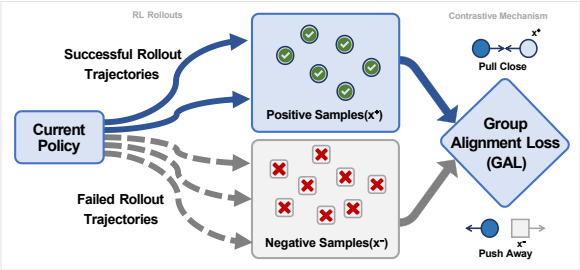


Figure 3: The Contrastive Mechanism in GAL.

Although GAL adopts a DPO-shaped contrastive form, it is not standard offline DPO. In DYPO, GAL is constructed from *on-policy* rollout groups sampled from the current policy, and its role is to serve as a variance-control term for GRPO rather than to align the model to a static preference dataset. Formally, we implement this by minimizing the following pairwise contrastive loss:

$$\mathcal{L}_{\text{GAL}}(\theta) = \mathbb{E}_{\substack{\tau_s,\tau_f \in G \\ R(\tau_s)>R(\tau_f)}} \left[-\log \sigma\big(\beta_{\text{GAL}} \cdot d(\tau_s, \tau_f)\big) \right] \tag{14}$$

where β_{GAL} is an inverse-temperature coefficient controlling the contrastive margin, and $d(\tau_s, \tau_f)$ represents the log-ratio difference between the successful trajectory τ_s and the failed trajectory τ_f, defined as:

$$d(\tau_s, \tau_f) = \log \frac{\pi_\theta(\tau_s|q)}{\pi_{\text{ref}}(\tau_s|q)} - \log \frac{\pi_\theta(\tau_f|q)}{\pi_{\text{ref}}(\tau_f|q)} \tag{15}$$

By applying the chain rule, the gradient of GAL is:

$$g_{\text{GAL}} = -\beta_{\text{GAL}} \underbrace{(1 - \sigma(\beta_{\text{GAL}} d))}_{\text{bounded weight } w_d} \cdot (\nabla_\theta \log \pi_s - \nabla_\theta \log \pi_f) \tag{16}$$

Unlike the unbounded $\hat{A}_i$ in GRPO, the weighting term w_d is strictly bounded in $(0, 1)$. Let $\eta = \mathbb{E}[(1-\sigma)^2]$ represent the *discrimination difficulty*. The variance of GAL (averaged over M pairs) is:

$$\text{Var}(g_{\text{GAL}}) \approx \frac{2\beta_{\text{GAL}}^2 \eta}{M} \Sigma_s \tag{17}$$

As the model learns to distinguish correct paths, $\sigma \to 1$ and $\eta \to 0$, causing $\text{Var}(g_{\text{GAL}}) \to 0$. Thus, GAL acts as a gradient variance reducer.

In the RL regime, we combine these objectives using a mixing coefficient $\alpha \in (0, 1)$:

$$g_{\text{mix}} = \alpha g_{\text{GRPO}} + (1 - \alpha) g_{\text{GAL}} \tag{18}$$

Assuming independence between the exploration noise of GRPO and the discrimination noise of GAL, the variance of the combined gradient is:

$$\begin{aligned} \text{Var}(g_{\text{mix}}) &\approx \alpha^2 \text{Var}(g_{\text{GRPO}}) + (1-\alpha)^2 \text{Var}(g_{\text{GAL}}) \\ &= \alpha^2 \left(\frac{\Sigma_s}{k} \right) + (1-\alpha)^2 \left(\frac{2\beta_{\text{GAL}}^2 \eta \Sigma_s}{M} \right) \end{aligned} \tag{19}$$

Since $\alpha < 1$ and $\eta \to 0$, it strictly follows that $\text{Var}(g_{\text{mix}}) < \text{Var}(g_{\text{GRPO}})$.

Summary. Analytically, we establish that the combined objective strictly bounds the gradient variance compared to GRPO (i.e., $\text{Var}(g_{\text{mix}}) < \text{Var}(g_{\text{GRPO}})$). Crucially, this stabilization is dynamic: as the model distinguishes successful trajectories τ_s from failed ones τ_f, the discrimination difficulty η decays to zero, naturally annealing the variance of GAL. This identifies GAL not merely as an auxiliary loss, but as an adaptive regularizer that actively dampens the high-variance noise of RL exploration.

4 Experiments

4.1 Setup

Dataset Construction. We align our data setup with LUFFY (Yan et al., 2025), utilizing the OpenR1-Math-220k (Face) subset with prompts primarily sourced from NuminaMath 1.5 (Jia et al., 2024). To facilitate multi-teacher distillation, we employ DeepSeek-R1 (Guo et al., 2025) and Qwen3-235B-A22B (Yang et al., 2025) to generate auxiliary reasoning traces. This ensemble strategy enriches the supervision signal and mitigates the policy's reliance on any single teacher's potentially biased reasoning patterns. All data will be open-sourced together with the code.

Implementation Details. Experiments were executed on a computing cluster equipped with 2 nodes, each containing 8 × NVIDIA A800 GPUs (80GB memory). To ensure fairness, we generate 8 trajectories (rollouts) per prompt for all trained models, with a maximum response length of 8,192 tokens. The learning rate is fixed at 1×10^{-6}. Our training pipeline is built upon the verl framework (Sheng et al., 2024). For the inference and rollout phases, we utilize vLLM (Kwon et al., 2023) to ensure high-throughput generation. All models were trained using bfloat16 precision to ensure numerical stability and efficiency.

Benchmarks. Our method is evaluated on five in-distribution (ID) benchmarks, includ-

Model	In-Distribution						Out-of-Distribution		
	AIME 24	AIME 25	AMC	MATH-500	Minerva	Avg	ARC-c	GPQA-D	Avg
Qwen2.5-Math-7B	11.5	4.9	31.3	43.6	7.4	19.7	18.2	11.1	14.6
Supervised Fine-Tuning									
SFT	22.2	22.3	52.8	82.6	40.8	44.1	75.2	24.7	50.0
Reinforcement Learning									
RL	25.1	15.3	62.0	84.4	39.3	45.2	**82.3**	40.4	61.4
SimpleRL-Zero	27.0	6.8	54.9	76.0	25.0	37.9	30.2	23.2	26.7
OpenReasoner-Zero	16.5	15.0	52.1	82.4	33.1	39.8	66.2	29.8	48.0
PRIME-Zero	17.0	12.8	54.0	81.4	39.0	40.8	73.3	18.2	45.8
Oat-Zero	33.4	11.9	61.2	78.0	34.6	43.8	70.1	23.7	46.9
SFT and RL									
SFT → RL	25.8	23.1	62.7	87.2	39.7	47.7	72.4	24.2	48.3
SuperRL	28.1	21.6	63.9	86.4	36.4	47.3	77.8	36.9	57.4
LUFFY	29.4	23.1	65.6	87.6	37.5	48.6	80.5	39.9	60.2
ReLIFT	28.3	22.9	65.1	87.4	37.1	48.2	74.9	40.9	57.9
SRFT	30.7	26.0	**69.8**	88.4	39.7	50.9	81.6	40.4	61.0
CHORD	31.2	24.4	66.8	**89.4**	39.3	50.2	81.1	40.4	60.8
DYPO	**36.0** (+10.2)	**28.7** (+5.6)	67.0 (+4.3)	89.2 (+2.0)	**42.4** (+2.7)	**52.5** (+4.8)	81.8 (+9.4)	**41.4** (+17.2)	**61.6** (+13.3)

Table 1: Overall performance on five competition-level mathematical reasoning benchmarks and two out-of-distribution benchmarks(Qwen2.5-Math-7B). Best results are **bolded** and second-best are underlined.

ing AIME 2024/2025, AMC (Li et al., 2024), MATH-500 (Hendrycks et al., 2021), and Minerva (Lewkowycz et al., 2022), as well as two out-of-distribution (OOD) tasks: ARC-c (Clark et al., 2018) and GPQA-Diamond (Rein et al., 2024). Performance is measured using pass@32 for the AIME/AMC subsets and pass@1 for the others. All inference is conducted with a temperature of 0.6 and option shuffling to prevent data leakage.

Baselines. We employ Qwen2.5-Math-7B (Yang et al., 2024) and Qwen3-4B-Base (Yang et al., 2025) as our base models and compare against four categories of baselines: (1) *Standard Supervised Baseline*, specifically the vanilla SFT; (2) *Zero-shot RL methods*, including SimpleRL-Zero (Zeng et al., 2025), OpenReasoner-Zero (Hu et al., 2025), PRIME-Zero (Cui et al., 2025), and Oat-Zero (Liu et al., 2025b); (3) *Post-SFT Optimization methods*, covering SFT → RL, LUFFY (Yan et al., 2025), ReLIFT (Ma et al., 2025),SRFT (Fu et al., 2025), SuperRL (Liu et al., 2025a) and CHORD (Zhang et al., 2025).

4.2 Main Results

4.2.1 Performance on Reasoning Benchmarks

As presented in Table 1 (Qwen2.5-Math-7B) and Table 2 (Qwen3-4B-Base), DYPO demonstrates consistent superiority across varying model architectures. On the Qwen2.5 benchmark, DYPO achieves an average in-distribution score of 52.5, setting a new state-of-the-art.

Comparison with SFT. DYPO significantly outperforms the SFT baseline, achieving a +8.4% average improvement on Qwen2.5-Math-7B and a substantial +18.8% on Qwen3-4B-Base.

Comparison with Zero-shot RL Methods. DYPO demonstrates superior stability over pure RL approaches like SimpleRL-Zero and Oat-Zero, surpassing the latter by a combined +19.4 points on the challenging AIME 24/25 benchmarks. While zero-shot methods often suffer from the high-variance "exploration trap," DYPO mitigates this by dynamically balancing the exploitation of priors with the exploration of new solutions, ensuring a more stable policy optimization process.

Comparison with Multi-stage Pipelines. DYPO maintains a clear lead over complex pipelines (SuperRL, LUFFY, ReLIFT, SRFT, CHORD), notably outperforming SRFT, by +4.8 points on AIME 25. This advantage is echoed in Qwen3 results, where DYPO outstrips the SFT→RL pipeline by +10.8%. Unlike the uniform optimization strategies in standard pipelines, DYPO employs a dynamic mechanism that mitigates gradient vanishing on both trivial and extremely hard samples, thereby maximizing sample utilization. Furthermore, the integration of multi-teacher distillation and GAL ensures a robust and stable training trajectory, avoiding the collapse often seen in complex pipelines.

Model	In-Distribution					Out-of-Distribution		
	AIME 24/25	AMC	MATH-500	Minerva	Avg	ARC-c	GPQA-D	Avg
Qwen3-4B-Base	9.3/5.3	40.0	66.8	27.9	29.9	49.4	14.1	31.8
SFT	33.3/27.3	62.9	73.8	43.0	48.1	73.8	28.8	51.3
RL	40.6/37.3	71.8	91.0	46.3	57.4	76.7	29.3	53.0
SFT → RL	43.3/39.3	75.4	77.4	44.9	56.1	77.4	27.8	52.6
DYPO	**59.3/44.0**	**86.0**	**94.6**	**50.4**	**66.9**	**92.5**	**44.4**	**68.5**

Table 2: Overall performance on mathematical reasoning benchmarks (Qwen3-4B-Base). Best results are **bolded** and second-best are underlined.

4.2.2 Generalization to Out-of-Distribution Tasks

Table 1 demonstrates that DYPO avoids the generalization degradation typically associated with in-domain optimization, achieving a top average OOD score of 61.6. On the PhD-level GPQA-Diamond benchmark, it outperforms both the standard SFT baseline (+16.7%). This indicates that DYPO transcends simple template memorization; by refining the reasoning policy rather than overfitting to surface-level patterns, it successfully transfers logical capabilities to diverse scientific domains.

4.3 Ablation Study

As shown in Table 3, we present an incremental analysis of the DYPO framework under different teacher strengths. The + *Multi-Teacher* variant serves as a data-matched supervised baseline, isolating the effect of stronger teacher supervision from the subsequent RL and routing components. Across all teacher settings (235B / 32B / 8B), performance improves monotonically as we add *+RL*, *+Dynamic Grading*, and *+GAL*, showing that DYPO is not merely distillation with stronger teachers. Even with the weaker 8B teacher, DYPO improves AIME 25 from 22.0 to 27.8 and GPQA-D from 30.8 to 39.4, demonstrating that the RL and routing components contribute substantial gains beyond supervision alone. Overall, Dynamic Difficulty Grading brings the largest jump on the hardest reasoning benchmarks, while GAL further stabilizes optimization and yields the best final performance across all teacher scales.

4.4 Offline Data Ratio, Reward and Entropy

To characterize the learning dynamics of DYPO, we monitor the Offline Data Ratio, Training Reward, and Policy Entropy across optimization steps.

Unlike static mixing (e.g., LUFFY), DYPO exhibits a self-evolving curriculum (Figure 4, left). The Offline Data Ratio transitions from full supervision (1.0 at $t = 0$) to a stable exploration-heavy state (≈ 0.35). This positioning suggests that DYPO treats offline demonstrations as a dynamic anchor: it autonomously de-leverages teacher signals as reasoning proficiency grows, yet retains a supervision floor to prevent distribution drift. The middle and right panels reveal the trade-off between convergence and diversity. While GRPO achieves rapid optimization, it suffers from premature mode collapse. In contrast, DYPO equilibrates reward maximization with policy stochasticity, maintaining robust entropy ($0.2 \sim 0.6$). This sustained diversity prevents the model from memorizing narrow reasoning templates, serving as the primary driver for its superior OOD generalization.

4.5 Empirical Analysis: Gradient Stability

A core theoretical contribution of DYPO is structurally resolving the bias-variance trade-off. We validate this empirically by analyzing the gradient norms of the policy network. As shown in Figure 5, standard GRPO (red) suffers from extreme volatility, implying a rugged landscape that complicates convergence. In contrast, DYPO (blue) maintains a significantly smoother trajectory. These results confirm that our offline component acts as an effective control variate, smoothing gradient estimates to allow for more aggressive learning rates.

5 Related Work

5.1 Post-Training Paradigms for LLM Reasoning

Enhancing the reasoning capabilities of LLMs has shifted from inference-time guidance (e.g., CoT prompting (Wei et al., 2022)) to robust post-training strategies (Wang et al., 2023). Current mainstream paradigms generally fall into two categories: SFT and RL. SFT effectively injects knowledge and stabilizes training by fitting high-quality demonstrations (Sanh et al., 2022; Wei et al., 2021), yet it suffers from high fitting bias and limited OOD generalization due to its reliance on static templates (Lv et al., 2025). Conversely, RL-based methods (e.g., RLHF or RLVR) encourage models to explore the reasoning space and maximize rewards, significantly boosting performance on complex tasks (Ouyang et al., 2022; Guo et al., 2025; Team et al., 2025). However, RL gradients are inherently high-variance and unstable, particularly when valid reward signals are sparse for weaker base models (Ramamurthy et al., 2022;

Model	AIME 24	AIME 25	AMC	GPQA-D
Qwen2.5-Math-7B	11.5	4.9	31.3	11.1
+ SFT	22.2	22.3	52.8	24.7
+ Multi-Teacher (235B / 32B / 8B)	26.6 / 26.8 / 24.5	23.3 / 23.5 / 22.0	61.4 / 61.8 / 59.5	33.3 / 32.3 / 30.8
+ RL (235B / 32B / 8B)	27.3 / 26.9 / 25.0	26.6 / 26.0 / 25.5	64.1 / 63.5 / 61.0	34.8 / 35.4 / 31.8
+ Dynamic Grading (235B / 32B / 8B)	33.3 / 31.5 / 31.8	28.7 / 27.9 / 28.1	63.6 / 62.0 / 62.4	36.4 / 35.4 / 34.3
+ GAL (DYPO) (235B / 32B / 8B)	**36.0 / 35.2 / 33.5**	**28.7 / 28.5 / 27.8**	**67.0 / 66.5 / 65.2**	**41.4 / 41.4 / 39.4**

Table 3: Ablation study under different teacher strengths.

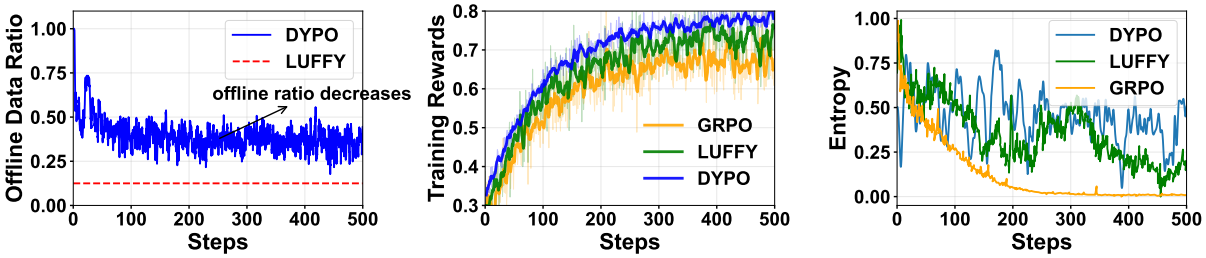


Figure 4: **Left:** Offline data ratio over steps.**Mid:** Training reward; **Right:** Policy entropy.

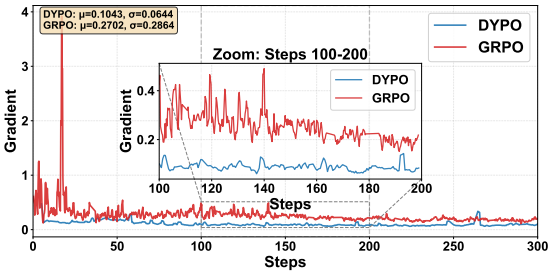


Figure 5: Gradient Norm Comparison.

Shao et al., 2024b). To combine these strengths, the traditional "SFT-then-RL" pipeline (Touvron et al., 2023; Yoshihara et al., 2025) is widely adopted but incurs multi-stage computational overhead and risks propagating SFT-induced biases into the exploration phase (Lv et al., 2025).

5.2 Unified Training and Optimization Trade-offs

To overcome the limitations of sequential pipelines, recent research focuses on unifying SFT and RL into a single-stage optimization process. Early attempts utilized simple loss weighting or fixed coefficients to balance stability and exploration (Fu et al., 2025; Yan et al., 2025). More advanced approaches employ dynamic scheduling or dual-control mechanisms (e.g., CHORD, HPT) to adjust the contribution of on-policy and off-policy data during training (Zhang et al., 2025). While recent theoretical works have explored the unified view of these objectives (Lv et al., 2025), a rigorous formalization of the gradient-level bias-variance trade-off in SFT-RL fusion remains underexplored. Existing unified methods largely operate at a "surface level" by re-weighting scalar losses, failing to structurally resolve the statistical conflict between the high-bias SFT vector and the high-variance RL vector. Unlike these approaches, our DYPO framework efficiently harmonizes SFT and RL via dynamic difficulty grading, while structurally reducing RL variance through GAL and mitigating SFT bias via multi-teacher distillation.

6 Conclusion

We address the inherent conflict between SFT fitting bias and RL gradient variance through DYPO, a unified framework that structurally mitigates this trade-off. Unlike static weighting approaches, DYPO employs Dynamic Difficulty Grading to adaptively route queries, leveraging Multi-Teacher Distillation to correct supervisory bias and GAL to suppress gradient variance.

Theoretically, we substantiate that DYPO achieves a linear rate of bias reduction while maintaining optimal variance control. Extensive empirical evaluations reveal that DYPO surpasses existing baselines, particularly in out-of-distribution scenarios. Beyond performance gains, the framework exhibits exceptional sample efficiency and architectural adaptability, proving effective across a spectrum of modern open-weights models. Ultimately, by dynamically governing the interplay between exploration and exploitation, DYPO establishes a unified and scalable paradigm for the next generation of reasoning-enhanced LLMs.

7 Limitations

Despite the promising performance of DYPO, we acknowledge certain limitations in our current study. First, our evaluation is primarily concentrated on logic-intensive domains, specifically mathematical reasoning tasks. While DYPO demonstrates superior stability in these objective-driven contexts, its efficacy on open-ended scenarios, such as creative writing or general chit-chat, remains to be fully explored. Second, regarding training efficiency, our method requires generating 8 trajectories per prompt to ensure robust dynamic estimation and fair comparison. This extensive online sampling inevitably introduces higher computational overhead and lower sample efficiency compared to offline baselines, representing a trade-off between optimization stability and training cost.

References

Liang Chen, Xueting Han, Li Shen, Jing Bai, and Kam-Fai Wong. 2025. Beyond two-stage training: Cooperative sft and rl for llm reasoning. *arXiv preprint arXiv:2509.06948*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, and 1 others. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, january 2025. *URL https://github.com/huggingface/open-r1*, page 9.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2025. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

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

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

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

LI Jia, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, and 1 others. 2024. Numinamath.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q. Jiang, Ziju Shen, and 1 others. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. `https://huggingface.co/datasets/Numinamath`. Hugging Face repository, 13:9.

Yihao Liu, Shuocheng Li, Lang Cao, Yuhang Xie, Mengyu Zhou, Haoyu Dong, Xiaojun Ma, Shi Han, and Dongmei Zhang. 2025a. Superrl: Reinforcement learning with supervision to boost language model reasoning. *Preprint*, arXiv:2506.01096.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025b. Understanding r1-zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*.

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Lixuan He, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, and 1 others. 2025. Towards a unified view of large language model post-training. *arXiv preprint arXiv:2509.04419*.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Bin Cui, and 1 others. 2025. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. *arXiv preprint arXiv:2506.07527*.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, and et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, and et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations (ICLR)*.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024a. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. ArXiv preprint arXiv:2307.09288.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023. Aligning large language models with human: A survey. *Preprint*, arXiv:2307.12966.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Gu, Aitor Lewkowycz, Yao Lu, Ambrose Slone, Quoc Le, and Barret Zoph. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems(NeurIPS)*.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. 2025. On the generalization of sft: A reinforcement learning perspective with reward rectification. *arXiv preprint arXiv:2508.05629*.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. 2025. Learning to reason under off-policy guidance. *Preprint*, arXiv:2504.14945.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Hiroshi Yoshihara, Taiki Yamaguchi, and Yuichi Inoue. 2025. A practical two-stage recipe for mathematical llms: Maximizing accuracy with sft and efficiency with reinforcement learning. *arXiv preprint arXiv:2507.08267*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and

Jingren Zhou. 2025. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. *Preprint*, arXiv:2508.11408.

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

A Mathematical Derivations and Theoretical Analysis

This appendix presents the rigorous mathematical formulations of the proposed loss functions and provides a theoretical analysis of their properties, specifically focusing on gradient variance comparisons and bias reduction mechanisms within the DYPO framework.

A.1 Loss Function Definitions

A.1.1 Conditional SFT Loss (Hard Regime)

In the **Hard** regime, where the policy fails to generate valid signals (i.e., all generated trajectories receive zero reward), exploration becomes inefficient. The framework strictly applies supervised fine-tuning using a Multi-Teacher strategy:

$$\mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{\tau_{\text{tgt}} \sim \mathcal{U}(\{\tau^{(1)},\dots,\tau^{(m)}\})} \left[-\log \pi_\theta(\tau_{\text{tgt}}|q) \right] \tag{20}$$

where $\{\tau^{(1)}, \dots, \tau^{(m)}\}$ are candidate solutions generated by m distinct teacher models, and $\mathcal{U}$ denotes the uniform distribution over these candidates.

A.1.2 GRPO Loss

The Group Relative Policy Optimization (GRPO) loss optimizes the policy via reinforcement learning by leveraging relative feedback within a group of sampled trajectories. Formally, for a query $q \sim \mathcal{D}$ (where $\mathcal{D}$ denotes the training data distribution), we sample a group of k trajectories $G = \{\tau_1, \dots, \tau_k\}$ from the current policy π_θ. The objective is defined as:

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\mathbb{E}_{q\sim\mathcal{D}} \Bigg[\frac{1}{k} \sum_{i=1}^{k} \Big(\text{CLIP}(\rho_i, \hat{A}_i, \epsilon) \Big) - \beta_{\text{KL}} \mathbb{D}_{\text{KL}}(\pi_\theta || \pi_{\text{ref}}) \Bigg] \tag{21}$$

where $\rho_i(\theta) = \frac{\pi_\theta(\tau_i|q)}{\pi_{\text{ref}}(\tau_i|q)}$ denotes the probability ratio between the current policy and the reference policy π_{ref}. The function $\text{CLIP}(\cdot)$ represents the standard clipping mechanism with hyperparameter ϵ to constrain policy updates. The term $\mathbb{D}_{\text{KL}}$ denotes the Kullback-Leibler divergence used to prevent mode collapse, weighted by the coefficient β_{KL}.

The advantage function $\hat{A}(\tau_i, G)$ is computed using group-based standardization:

$$\hat{A}(\tau_i, G) = \frac{R(\tau_i, q) - \mu_G}{\sigma_G + \xi} \tag{22}$$

where μ_G and σ_G represent the mean and standard deviation of rewards within group G, respectively, and ξ is a small constant added for numerical stability.

A.1.3 Group Alignment Loss (GAL)

To mitigate the high variance associated with pure RL in the **Mid** regime, we employ a contrastive objective. We leverage the intrinsic quality differences within the sampled group G by constructing pairwise comparisons. For a tuple (q, τ_s, τ_f) drawn from G where $R(\tau_s) > R(\tau_f)$ (implying $R(\tau_s) = 1$ and $R(\tau_f) = 0$ in binary settings):

$$\mathcal{L}_{\text{GAL}}(\theta) = \mathbb{E}_{\substack{\tau_s,\tau_f \in G \\ R(\tau_s)>R(\tau_f)}} \Bigg[-\log \sigma \Bigg(\beta_{\text{GAL}} \cdot \left(\log \frac{\pi_\theta(\tau_s|q)}{\pi_{\text{ref}}(\tau_s|q)} - \log \frac{\pi_\theta(\tau_f|q)}{\pi_{\text{ref}}(\tau_f|q)} \right) \Bigg) \Bigg] \tag{23}$$

Here, $\sigma(\cdot)$ denotes the sigmoid function, and β_{GAL} is the inverse temperature parameter controlling the discrimination margin.

A.1.4 Unified Objective and Difficulty Grading

The core of DYPO is the dynamic dispatching of queries based on the rollout outcome G. We define three mutually exclusive indicator functions based on the set of rewards $\{R(\tau)|\tau \in G\}$:

- **Indicator for Easy ($\mathbb{I}_{\mathcal{E}}$):** $\mathbb{I}(\forall \tau \in G, R(\tau) = 1)$. The loss is set to 0 to discard trivial samples.
- **Indicator for Hard ($\mathbb{I}_{\mathcal{H}}$):** $\mathbb{I}(\forall \tau \in G, R(\tau) = 0)$. This triggers the SFT fallback.
- **Indicator for Mid ($\mathbb{I}_{\mathcal{M}}$):** $\mathbb{I}(\exists \tau_i, \tau_j \in G, R(\tau_i) \neq R(\tau_j))$. This triggers the variance-reduced RL.

The final unified training objective is formulated as:

$$\mathcal{L}_{\text{DYPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D}}\Bigg[\underbrace{\mathbb{I}_{\mathcal{H}} \cdot \gamma\mathcal{L}_{\text{SFT}}}_{\text{Hard}} + \underbrace{\mathbb{I}_{\mathcal{M}} \cdot \big(\alpha\mathcal{L}_{\text{GRPO}} + (1-\alpha)\mathcal{L}_{\text{GAL}}\big)}_{\text{Mid}}\Bigg] \tag{24}$$

Note that when $\mathbb{I}_{\mathcal{E}} = 1$, the gradient contribution is effectively zero, implementing the sample discarding strategy.

Hyperparameters:

- $\alpha \in [0, 1]$: Weighting coefficient balancing the summation-based RL (GRPO) and contrastive alignment (GAL).
- $\gamma > 0$: Scaling factor for the supervised loss component.

A.2 Detailed Derivation of Multi-Teacher Bias Reduction

In this section, we provide the formal derivation for the bias reduction property of the Multi-Teacher Distillation strategy discussed in Section 3.2. We base our analysis on the bias decomposition formulation provided in the main text.

A.2.1 Definitions and Assumptions

Let $\tau^* \in \mathbb{R}^d$ be the optimal reasoning path (ground truth). The reasoning path generated by the i-th teacher, $\tau^{(i)}$, is modeled as:

$$\tau^{(i)} = \tau^* + \mathbf{b}_{\text{sys}} + \mathbf{b}_i \tag{25}$$

where $\mathbf{b}_{\text{sys}}$ is the systematic bias and $\mathbf{b}_i$ is the idiosyncratic bias.

To facilitate the derivation, we formalize the properties of the idiosyncratic bias term $\mathbf{b}_i$:

Assumption 1 (Zero-Mean Idiosyncratic Bias)**.** We assume that the idiosyncratic biases from different teachers are independent and centered around zero in the semantic space. That is, for any teacher i:

$$\mathbb{E}[\mathbf{b}_i] = \mathbf{0} \tag{26}$$

Assumption 2 (Variance Definition)**.** We define the magnitude of the idiosyncratic noise for a single teacher as $\bar{\sigma}^2_{\text{bias}}$. Formally, this is the expected squared Euclidean norm of the bias vector:

$$\mathbb{E}[\|\mathbf{b}_i\|^2] = \bar{\sigma}^2_{\text{bias}} \tag{27}$$

A.2.2 Derivation of Squared Bias

We compare the expected squared bias (estimation error) between the single-teacher baseline and the multi-teacher ensemble.

1. Single-Teacher SFT ($m = 1$). When supervision is provided by a single randomly selected teacher k, the bias is simply $\text{Bias}_{\text{single}} = \tau^{(k)} - \tau^* = \mathbf{b}_{\text{sys}} + \mathbf{b}_k$. The expected squared norm is:

$$\begin{aligned}\mathbb{E}[\|\text{Bias}_{\text{single}}\|^2] &= \mathbb{E}[\|\mathbf{b}_{\text{sys}} + \mathbf{b}_k\|^2] \\ &= \|\mathbf{b}_{\text{sys}}\|^2 + \mathbb{E}[\|\mathbf{b}_k\|^2] \\ &\quad + 2\mathbf{b}_{\text{sys}}^\top \underbrace{\mathbb{E}[\mathbf{b}_k]}_{=\mathbf{0}} \\ &= \|\mathbf{b}_{\text{sys}}\|^2 + \bar{\sigma}^2_{\text{bias}}\end{aligned} \tag{28}$$

2. Multi-Teacher SFT ($m > 1$). In the Multi-Teacher strategy, the effective supervision converges to the expectation over the sampled teachers, which is equivalent to the ensemble mean $\bar{\tau} = \frac{1}{m}\sum_{i=1}^m \tau^{(i)}$. The effective bias vector is:

$$\text{Bias}_{\text{multi}} = \left(\frac{1}{m}\sum_{i=1}^m \tau^{(i)}\right) - \tau^* = \mathbf{b}_{\text{sys}} + \frac{1}{m}\sum_{i=1}^m \mathbf{b}_i \tag{29}$$

The expected squared norm of the multi-teacher bias is:

$$\begin{aligned}\mathbb{E}[\|\text{Bias}_{\text{multi}}\|^2] &= \mathbb{E}\left[\left\|\mathbf{b}_{\text{sys}} + \frac{1}{m}\sum_{i=1}^m \mathbf{b}_i\right\|^2\right] \\ &= \|\mathbf{b}_{\text{sys}}\|^2 + \mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \mathbf{b}_i\right\|^2\right] \\ &\quad + 2\mathbf{b}_{\text{sys}}^\top \underbrace{\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^m \mathbf{b}_i\right]}_{=\mathbf{0}}\end{aligned} \tag{30}$$

We focus on the variance term (the second term). Due to the independence of $\mathbf{b}_i$, the cross-terms $\mathbb{E}[\mathbf{b}_i^\top \mathbf{b}_j]$ for $i \neq j$ are zero. Thus:

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i=1}^m \mathbf{b}_i\right\|^2\right] &= \frac{1}{m^2}\sum_{i=1}^m \mathbb{E}[\|\mathbf{b}_i\|^2] \\ &= \frac{1}{m^2} \cdot m \cdot \bar{\sigma}^2_{\text{bias}} \\ &= \frac{\bar{\sigma}^2_{\text{bias}}}{m}\end{aligned} \tag{31}$$

Substituting this back into Eq. (30), we obtain the final expression presented in the main text:

$$\mathbb{E}[\|\text{Bias}_{\text{multi}}\|^2] = \|\mathbf{b}_{\text{sys}}\|^2 + \frac{\bar{\sigma}^2_{\text{bias}}}{m} \tag{32}$$

A.2.3 Conclusion

By comparing the two results, we formally establish the reduction inequality:

$$\underbrace{\|\mathbf{b}_{\text{sys}}\|^2 + \frac{\bar{\sigma}^2_{\text{bias}}}{m}}_{\mathbb{E}[\|\text{Bias}_{\text{multi}}\|^2]} < \underbrace{\|\mathbf{b}_{\text{sys}}\|^2 + \bar{\sigma}^2_{\text{bias}}}_{\mathbb{E}[\|\text{Bias}_{\text{single}}\|^2]} \tag{33}$$

This confirms that increasing the ensemble size m strictly reduces the stochastic component of the supervisory bias.

A.3 Gradient Variance Analysis

We analyze the variance of the gradient estimators to theoretically justify the stability properties of the DYPO framework. We define the scalar variance of a gradient estimator g as $\text{Var}(g) = \mathbb{E}[\|g - \mathbb{E}[g]\|^2]$.

A.3.1 Variance of GRPO Loss

Consider the gradient of the GRPO loss for a single query q with group size k. In practical optimization, gradients are averaged over the group. The gradient is defined as:

$$g_{\text{GRPO}} = \frac{1}{k}\sum_{i=1}^{k} \nabla_\theta \log \pi_\theta(\tau_i|q) \cdot \hat{A}_i \tag{34}$$

Let $s_i = \nabla_\theta \log \pi_\theta(\tau_i|q)$ be the score function. Assuming sample independence and normalized advantages ($\mathbb{E}[\hat{A}_i^2] \approx 1$), the variance is:

$$\begin{aligned}\text{Var}(g_{\text{GRPO}}) &= \frac{1}{k^2}\sum_{i=1}^{k} \mathbb{E}[\|s_i\|^2] \cdot \mathbb{E}[\hat{A}_i^2] \\ &\approx \frac{1}{k^2} \cdot (k \cdot \Sigma_s) = \frac{\Sigma_s}{k}\end{aligned} \tag{35}$$

where $\Sigma_s = \mathbb{E}\left[\|\nabla_\theta \log \pi_\theta(\tau_i|q)\|^2\right]$. This shows $\text{Var}(g_{\text{GRPO}}) \propto 1/k$.

A.3.2 Variance of Group Alignment Loss

For the GAL objective, we construct M preference pairs. The gradient is averaged over these pairs:

$$g_{\text{GAL}} = \frac{1}{M}\sum_{j=1}^{M}(1 - \sigma(d_j)) \cdot \beta_{\text{GAL}}(s_{s,j} - s_{f,j}) \tag{36}$$

Assuming independence between pairs, the variance is bounded by:

$$\text{Var}(g_{\text{GAL}}) \approx \frac{2\beta^2_{\text{GAL}}\eta\Sigma_s}{M} \tag{37}$$

where $\eta = \mathbb{E}[(1 - \sigma(d))^2]$ represents the discrimination difficulty.

A.3.3 Variance of the Combined Gradient

In the RL regime (Mid), we combine these objectives using a mixing coefficient $\alpha \in (0, 1)$:

$$g_{\text{mix}} = \alpha g_{\text{GRPO}} + (1 - \alpha) g_{\text{GAL}} \tag{38}$$

Assuming independence between the exploration noise of GRPO and the discrimination noise of GAL, the variance of the combined gradient is:

$$\begin{aligned}\text{Var}(g_{\text{mix}}) &\approx \alpha^2 \text{Var}(g_{\text{GRPO}}) + (1 - \alpha)^2 \text{Var}(g_{\text{GAL}}) \\ &= \alpha^2 \left(\frac{\Sigma_s}{k}\right) + (1 - \alpha)^2 \left(\frac{2\beta^2_{\text{GAL}}\eta\Sigma_s}{M}\right)\end{aligned} \tag{39}$$

We observe that as the policy improves, the discrimination task becomes easier, causing $\eta \to 0$ (since $\sigma(d) \to 1$). Furthermore, since $\alpha < 1$ implies $\alpha^2 < 1$, the contribution of the GRPO term is strictly reduced. Therefore, under the condition that η is sufficiently small, it strictly follows that:

$$\text{Var}(g_{\text{mix}}) < \text{Var}(g_{\text{GRPO}}) \tag{40}$$

This inequality proves that the mixed objective yields a more stable gradient estimator than using GRPO alone, facilitating smoother convergence.

B Qualitative Analysis and Case Studies

In this section, we provide a qualitative analysis of our pipeline. We first illustrate our data construction strategy, which leverages the complementary strengths of multiple teacher models. We then present specific case studies for the SFT stage and the RL stage to demonstrate how DYPO enhances mathematical reasoning.

B.1 Data Construction: Leveraging Diversity

Our data construction method creates a high-quality, diverse dataset by distilling reasoning capabilities from multiple teacher models (e.g., DeepSeek-R1, Qwen3-235B-A22B). As illustrated in the example below, different teachers may approach the same problem via distinct but valid reasoning paths (e.g., algebraic vs. geometric). This

diversity prevents the student model from overfitting to a single reasoning pattern and improves generalization.

Problem Statement

Question: Given a rectangular billiard table with sides 1 and $\sqrt{2}$. A ball is shot from one of its corners at an angle of 45°. Will it ever fall into a pocket?

Our Data Construction Method

Teacher A (DeepSeek-R1) [Baseline]

Reasoning: The problem reduces to finding integers m, n such that $m \cdot 1 = n \cdot \sqrt{2}$. This implies $\frac{m}{n} = \sqrt{2}$. Since $\sqrt{2}$ is irrational, no such integers exist.

Answer: No

Teacher B (Qwen3-235B) [Complementary]

Reasoning Process: *1. Unfolding:* Mirroring implies trajectory line $y = x$. *2. Commensurability:* Ratio $1 : \sqrt{2}$ is irrational. *3. Conclusion:* Never hits a vertex.

Answer: No

B.2 Case Study I: Multi-Teacher Distillation (SFT)

Following the data construction methodology described above, the SFT stage trains the student model to recognize multiple solution paths. Here, we visualize the reasoning process for a constrained optimization problem. As shown below, **Teacher A** employs the method of Lagrange Multipliers (calculus-based), while **Teacher B** utilizes a coordinate transformation approach (algebraic/geometric). The student model benefits from exposure to both high-quality trajectories.

Prompt / Input

Find the greatest value M that the expression $7x+10y+z$ can take when x, y, z are real numbers satisfying $x^2 + 2x + \frac{1}{5}y^2 + 7z^2 = 6$. In which cases is this value achieved?

Trajectory 1: Teacher A (Lagrange Multipliers)

↪ *Internal Reasoning*

Okay, so I need to find the maximum value of the expression $7x + 10y + z$ given the constraint... Let's check the constraint equation $x^2+2x+(1/5)y^2+7z^2 = 6$. Using Lagrange multipliers:

$$\nabla f = \lambda \nabla g$$

Which gives us the system: 1) $7 = \lambda(2x + 2)$, 2) $10 = \lambda(2/5y)$, 3) $1 = \lambda(14z)$ ... [Middle reasoning omitted for brevity] ... Therefore, the maximum value is 55.

Final Response:

To find the greatest value M, we apply the method of Lagrange multipliers... The maximum value is achieved when:

$$x = -\frac{13}{62}, \quad y = \frac{175}{31}, \quad z = \frac{1}{62}$$

The value is $\boxed{55}$.

Trajectory 2: Teacher B (Coordinate Transformation)

↪ *Internal Reasoning*

Using Cauchy-Schwarz inequality... Let $u = x + 1$, $v = y/\sqrt{5}$, $w = z\sqrt{7}$. The constraint becomes $u^2 + v^2 + w^2 = 7$. ...

Final Response:

The solution is derived via coordinate transformation... Result is $\boxed{55}$.

B.3 Case Study II: Group Alignment via RL

While SFT establishes fundamental reasoning, the RL stage enables the exploration of complex solution paths. To ensure stability, the **Group Alignment Loss** reduces gradient variance by contrasting trajectories within a sampled group. We demonstrate this using a functional equation problem. Here, the Group Alignment Loss reinforces the rigorous application of the chain rule (Trajectory τ_1) while suppressing heuristic failures (Trajectory τ_2).

Prompt / Input

For arbitrary integers n, find the continuous function $f(x)$ which satisfies the following equation:

$$\lim_{h\to 0}\frac{1}{h}\int_{x-nh}^{x+nh} f(t)dt = 2f(nx)$$

Note that x can range over all real numbers and $f(1) = 1$.

Group Rollout Summary ($K = 8$):

ID	Final Answer	Reward	Status
τ_1	$\boxed{f(x) = x}$	**1.0**	**Successful**
τ_2	$\boxed{f(x) = 1}$	0.0	Failed
τ_3	$\boxed{f(x) = x}$	**1.0**	**Successful**
τ_4	$\boxed{f(x) = nx}$	0.0	Failed
τ_5	$\boxed{f(x) = x}$	**1.0**	**Successful**
τ_6	$\boxed{f(x) = 0}$	0.0	Failed
τ_7	$\boxed{f(x) = 1}$	0.0	Failed
τ_8	$\boxed{f(x) = x^n}$	0.0	Failed

Successful Trajectory (τ_1): Correct Chain Rule Application

↪ *Internal Reasoning*

Let the given equation be:

$$L = \lim_{h\to 0}\frac{1}{h}\int_{x-nh}^{x+nh} f(t)dt$$

This limit represents the derivative of the integral function with respect to h at $h = 0$. Let $F(t)$ be the antiderivative of $f(t)$. Then the integral is $F(x + nh) - F(x - nh)$. ...

$$= f(x + nh) \cdot n - f(x - nh) \cdot (-n)$$

As $h \to 0$, this becomes:

$$f(x) \cdot n - f(x) \cdot (-n) = 2nf(x)$$

... By continuity, $f(x) = x$ for all real x.

Final Response:
The function is $f(x) = x$. The answer is $\boxed{f(x) = x}$.

Failed Trajectory (τ_2): Missing Boundary Derivative

↪ *Internal Reasoning*

We need to evaluate the limit on the LHS... Let's assume the standard result where $\frac{1}{2h}\int_{x-h}^{x+h} f(t)dt \to f(x)$. ... Wait, let me reconsider. If I define $G(h) = \int_{x-nh}^{x+nh} f(t)dt$. $G'(h) = f(x + nh) - f(x - nh)$ (using standard Fundamental Theorem). At $h = 0$, $G'(0) = 0$. This implies $f(x) = 0$... or $f(x) = 1$.

Final Response:
The function satisfies $f(x) = 1$. The answer is $\boxed{f(x) = 1}$.

Analysis: The Failed Trajectory (τ_2) attempts to derive the limit but fails to apply the chain rule to the integration bounds ($G'(h) = f(x + nh) \cdot n - f(x - nh) \cdot (-n)$), leading to an erroneous constant solution.

C Additional Experimental Analysis

C.1 Single-Teacher vs. Multi-Teacher Ablation

To isolate the effect of supervision diversity, we compare the full DYPO framework using a single teacher against the same framework using two teachers, while keeping all other components unchanged. As shown in Table 4, multi-teacher supervision consistently improves both in-distribution and out-of-distribution performance.

Method	AIME 24	AMC	GPQA-D
DYPO (Single-Teacher)	32.8	64.2	37.5
DYPO (2-Teacher)	**36.0**	**67.0**	**41.4**

Table 4: Comparison between single-teacher and multi-teacher supervision in DYPO.

This result supports the role of multi-teacher distillation in reducing teacher-specific bias and providing a stronger prior for subsequent RL optimization.

C.2 Hyperparameter Sensitivity and Statistical Significance

We evaluate the sensitivity of DYPO to two key hyperparameters in the Mid regime: the mixing coefficient α and the inverse-temperature coefficient β_{GAL}. Results in Table 5 show that DYPO remains stable across a broad range of settings, with the default configuration ($\alpha = 0.5, \beta_{\text{GAL}} = 1$) achieving the best overall performance.

α	β_{GAL}	AIME 24	MATH-500
0.2	0.1	33.5 ± 1.4	87.6 ± 0.7
0.2	1	34.8 ± 0.9	88.5 ± 0.5
0.2	2	33.9 ± 1.2	87.9 ± 0.6
0.5	0.1	35.2 ± 0.8	88.9 ± 0.4
0.5	1	$\mathbf{36.0 \pm 1.1}$	$\mathbf{89.2 \pm 0.3}$
0.5	2	35.5 ± 1.0	88.7 ± 0.5
0.8	0.1	34.6 ± 1.3	88.3 ± 0.8
0.8	1	35.4 ± 1.1	88.8 ± 0.4
0.8	2	34.2 ± 1.5	88.1 ± 0.7

Table 5: Sensitivity analysis of α and β_{GAL}. Results are reported as mean $\pm$ standard deviation over multiple random seeds.

We further conduct paired significance tests against strong baselines using matched random seeds. The improvements of DYPO on the main benchmarks are statistically significant ($p < 0.05$).

D License and Artifacts Usage

We utilize the Qwen models and standard reasoning datasets (e.g., AIME, MATH), which are publicly available under the Apache 2.0 or MIT licenses. Our use of these artifacts for academic research and post-training optimization is strictly consistent with their intended usage policies. We release our code and the trained DYPO model checkpoints under the MIT License to promote reproducibility. This licensing is compatible with the original access conditions of the base models and datasets used in this work.