

# GFT: From Imitation to Reward Fine-Tuning with Unbiased Group Advantages and Dynamic Coefficient Rectification

Wangjie Gan<sup>1†</sup>, Miao Pan<sup>1†</sup>, Linbo Xi<sup>1†</sup>, Wenqi Zhang<sup>1\*</sup>, Jintao Chen<sup>1</sup>, Jianwei Yin<sup>1</sup>, Xuhong Zhang<sup>2\*</sup>

<sup>1</sup>Zhejiang University

<sup>2</sup>Innovation and Management Center, School of Software Technology (Ningbo), Zhejiang University  
{zhangwenqi, zhangxuhong}@zju.edu.cn

## Abstract

Large language models are typically post-trained using supervised fine-tuning (SFT) and reinforcement learning (RL), yet effectively unifying efficient knowledge injection with robust generalization remains challenging. In this work, we provide a training-dynamics analysis showing that SFT can be interpreted as a special case of policy gradient optimization with an extremely sparse implicit reward and unstable inverse-probability weighting, which together lead to single-path dependency, entropy collapse, and gradient explosion. Motivated by this diagnosis, we propose Group Fine-Tuning (GFT), a unified post-training framework that addresses these intrinsic limitations through two mechanisms: Group Advantage Learning, which constructs diverse response groups and derives normalized contrastive supervision to alleviate reward sparsity, and Dynamic Coefficient Rectification, which adaptively bounds inverse-probability weights to stabilize optimization while preserving efficient knowledge injection. Experiments demonstrate that GFT consistently surpasses SFT-based methods and yields policies that integrate more smoothly with subsequent RL training. Our code is publicly available at <https://github.com/ZJU-OmniAI/GFT>.

## 1 Introduction

The remarkable advancement of large language models has been driven to a great extent by two core post-training techniques: supervised fine-tuning (SFT) and reinforcement learning (RL) (Guo et al., 2025; Xu et al., 2025). A substantial body of prior work has investigated the respective strengths of these two paradigms. SFT leverages expert demonstration data to efficiently inject knowledge and skills, enabling models to rapidly acquire instruction-following abilities and

domain-specific competence (Chu et al., 2024; Chung et al., 2024). Meanwhile, RL guides models to explore and optimize within a broad policy space through reward signals, facilitating the learning of robust reasoning behaviors and generalizable strategies (Guo et al., 2025; Wang et al., 2024).

Despite the complementary strengths of SFT and RL, SFT training is highly sensitive to high-fidelity expert data (Zhou et al.; Gudibande et al., 2023) and often exhibits unstable optimization, which manifests in two salient failure modes in Figure 1. First, the strict imitation objective can overwrite and shift general-purpose representations acquired during pretraining, leading to catastrophic forgetting (Aw et al., 2023; Chu et al., 2024; Ruan et al., 2025; Luo et al., 2025) and degraded out-of-distribution generalization—consistent with the systematic regressions of SFT relative to the Base model in Figure 1(a). Second, SFT tends to overconstrain the policy to a narrow demonstration manifold, reducing policy entropy and solution diversity and thereby shrinking the exploration budget required by downstream RL (Chen et al., 2025a,b; Qin and Springenberg, 2025); as a result, Figure 1(b) shows a clear synergy break where RL alone (e.g., GRPO) delivers substantial gains, yet the common sequential pipeline (SFT+GRPO) yields consistently diminished improvements, i.e., “RL works, but its benefits are attenuated when preceded by SFT.”

To investigate the root causes of these challenges, we present a principled theoretical analysis from the perspective of training dynamics. We demonstrate that SFT can be interpreted as a special case of reinforcement learning, but one that suffers from two fundamental flaws: (1) It is constrained by **single-path dependency**, where the implicit reward  $r(x, y) = \mathbb{I}[y = y^*]$  restricts the learning signal to the exact expert trajectory, leading to **insufficient exploration** and **entropy collapse**. (2) It is vulnerable to **gradient explosion** during optimiza-

<sup>†</sup>Equal contribution.

<sup>\*</sup>Co-corresponding authors.

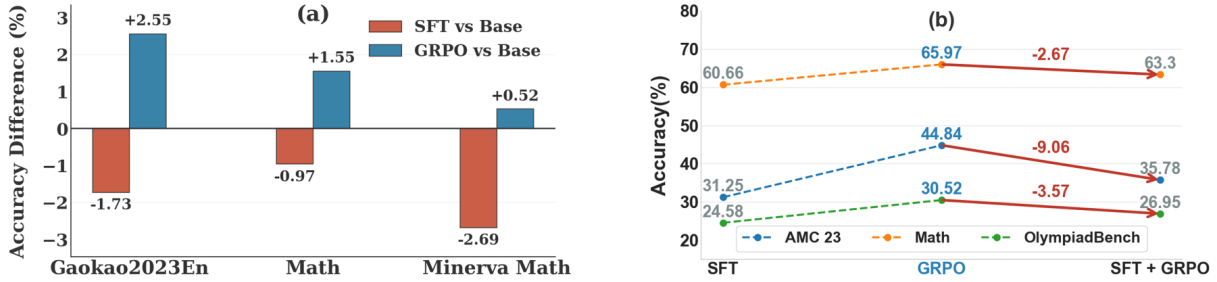


Figure 1: Performance of Qwen2.5-Math-1.5B on Numina-Math. (a) Accuracy changes relative to the base model: SFT consistently degrades performance, highlighting catastrophic forgetting. (b) Accuracy across different training pipelines: the SFT+GRPO pipeline exhibits poor synergy, underperforming GRPO alone.

tion. Since the gradient updates are scaled by an **unstable importance weight**  $w(y|x) = 1/\pi_\theta(y|x)$  (the reciprocal of the token probability), valid but unfamiliar expert tokens cause this weight to grow excessively large, triggering **gradient explosion** and driving the model toward **mechanical memorization** and **overfitting**. Together, these factors constitute the mathematical explanation for SFT’s limited generalization ability.

Motivated by these theoretical insights, we propose **Group Fine-Tuning (GFT)**, a unified post-training paradigm designed to directly mitigate these intrinsic deficiencies. GFT introduces two key mechanisms. *Group Advantage Learning* overcomes SFT’s single-path dependency by creating a diverse response group for each query, combining model-generated samples, expert demonstrations, and teacher outputs. By evaluating candidates according to their normalized within-group advantages, rather than rigidly imitating expert data, this approach produces learning signals that are comparable across diverse responses, thereby preserving essential exploration during early post-training. *Dynamic Coefficient Rectification* stabilizes optimization while preserving learning capacity through a clipping-like adaptive weighting scheme. By applying a dynamic threshold  $\tau$  to the **importance weight**  $w(y|x)$ , this mechanism suppresses gradient explosion for extreme samples while **preserving the effective gradient** for moderately low-probability tokens, enabling efficient injection of new knowledge into models.

We systematically evaluated GFT across multiple model families and math-reasoning benchmarks. Compared with standard SFT, strong SFT variants such as DFT (Wu et al., 2025) and ASFT (Zhu et al., 2025a), RL baselines such as GRPO, and component-wise ablations, GFT consistently outperforms all baselines on both standard and

competition-level tasks with substantially higher data efficiency. To further probe the post-training “synergy dilemma,” we use GFT as the initialization for subsequent RL and contrast it with the conventional “SFT→RL” pipeline; GFT provides a stronger cold start and more stable optimization, thereby significantly raising the attainable performance ceiling of RL. Finally, evaluations of catastrophic forgetting and output diversity show that GFT markedly mitigates the severe forgetting typical of SFT while achieving a practical unification of improved precision and preserved exploration.

#### Our main contributions include:

- From a training-dynamics perspective, we identify two causes of SFT’s weak generalization: (i) inherent single-path dependency, where each context is supervised by a single expert demonstration; and (ii) gradient explosion, which promotes mechanical memorization and catastrophic forgetting.
- We propose **GFT**, unifying unbiased group advantages and token-wise update stabilization into a single-stage post-training procedure by combining group advantage learning with dynamic importance weight rectification.
- Extensive experiments across multiple benchmarks show that GFT consistently outperforms standard SFT and strong SFT-based baselines, validating GFT as a foundational post-training paradigm for LLMs.

## 2 Preliminaries

In SFT learning process, the policy model  $\pi_\theta$  is trained to imitate expert demonstrations. Given an expert dataset  $\mathcal{D} = \{(x, y^*)\}$ , the gradient of the SFT objective with model parameters  $\theta$  is

$$\nabla_\theta \mathcal{L}_{\text{SFT}} = \mathbb{E}_{\mathcal{D}} \left[ -\nabla_\theta \log \pi_\theta(y^* | x) \right]. \quad (1)$$

This gradient increases the likelihood of the expert-provided response and does not explicitly

consider alternative outputs. However, in RL training process, the output  $y$  is generated by the current model  $\pi_\theta(\cdot | x)$  itself. The reward  $r(x, y)$  is then computed for this model-generated sample. The policy gradient takes the form

$$\nabla_\theta \mathcal{L}_{\text{RL}} = \mathbb{E}_{x,y} \left[ -\nabla_\theta \log \pi_\theta(y | x) r(x, y) \right]. \quad (2)$$

Notably, SFT can be viewed as a special case of RL. Specifically, if we interpret the SFT objective as maximizing a sparse reward that only provides non-zero feedback for the expert trajectory, its gradient can be rewritten as an on-policy expectation over  $\pi_\theta$  via importance sampling. This equivalent formulation is:

$$\nabla_\theta \mathcal{L} = -\mathbb{E}_{x,y} \left[ \frac{\mathbb{I}[y = y^*]}{\pi_\theta(y | x)} \nabla_\theta \log \pi_\theta(y | x) \right], \quad (3)$$

where the indicator  $\mathbb{I}[y = y^*]$  serves as a sparse reward that assigns a unit signal only when the sampled output exactly matches the expert demonstration, and zero otherwise. The term  $1/\pi_\theta(y | x)$  corresponds to an importance weight that corrects for sampling from the current policy  $\pi_\theta$  instead of the expert distribution. The detailed derivation is provided in Appendix A.

Eq. (3) exposes two intrinsic limitations of SFT from an RL perspective. First, **single-path dependency** arises as the sparse reward confines learning to a single expert trajectory, offering no comparative feedback over alternatives. Second, **gradient explosion** occurs because the importance weight  $1/\pi_\theta(y | x)$  grows excessively when the expert action probability is small, leading to highly unstable optimization behavior.

### 3 Method: Group Fine Tuning

To address the intrinsic limitations of SFT identified in Eq. (3), we propose two complementary mechanisms. **Group Advantage Learning** (GAL) constructs a group of multiple candidate trajectories for one query and evaluates each trajectory based on rule-consistent rewards, allowing the model to learn from diverse reasoning paths rather than treating only expert demonstrations as correct. **Dynamic Coefficient Rectification** (DCR) stabilizes training by clipping the weight  $1/\pi_\theta(y | x)$  for extremely low-probability tokens, preventing gradient explosion while preserving the original gradient for standard tokens to ensure efficient knowledge injection.

#### 3.1 Group Advantage Learning

To move beyond the limitations of single-path dependency, we expand the standard SFT dataset into a comprehensive hybrid response group  $\mathcal{G}_x = y_1, \dots, y_K$  for each query  $x$ . This group strategically integrates three complementary data sources: **Expert Demonstrations** ( $y_{\text{exp}}$ ) that provide ground truth to guarantee a valid optimization direction always exists; **Teacher Distillations** ( $y_{\text{demo}}$ ) from other powerful models, introducing diverse reasoning paradigms to break single-path dependency; and **Self-Generated Samples** ( $y_{\text{sample}}$ ) obtained from the model’s own rollouts, offering on-policy feedback to rectify intrinsic errors while reinforcing successful self-exploration. This design maintains high flexibility, allowing the composition to adapt based on data availability and training objectives. To effectively utilize the strengths of each data source within a unified learning framework, we assign a scalar reward  $R(y_k)$  to each response in group  $\mathcal{G}_x$ , then compute a standardized advantage score:

$$A(y_k) = \frac{R(y_k) - \mu(\mathcal{G}_x)}{\sigma_R(\mathcal{G}_x) + \epsilon}, \quad (4)$$

where  $\bar{R}(\mathcal{G}_x)$  and  $\sigma_R(\mathcal{G}_x)$  denote the mean and standard deviation of rewards within the group, and  $\epsilon > 0$  is a small constant that ensures numerical stability. This normalization centers and scales the rewards, creating a **relative, contrastive signal** within the group. Consequently, the reward mechanism guides the model to discern and prioritize high-quality responses, effectively unifying imitation, distillation, and self-improvement within a single, stable objective.

#### 3.2 Dynamic Coefficient Rectification

The theoretical analysis in Eq. (3) reveals that the inverse probability term  $1/\pi_\theta$  introduces an inherent instability into the SFT-style optimization. In practice, this instability arises in two common and complementary scenarios. First, when the model increases its exploration by rolling out uncertain or diverse responses, the predicted token probabilities  $\pi_t$  can become small, causing the corresponding update coefficients to grow excessively large. Second, even when fitting expert demonstrations or teacher-distilled responses, the model may initially assign low probability to valid but unfamiliar tokens, which similarly amplifies the inverse weighting term. Inspired by the gradient clipping technique prevalent in RL, we propose a simple

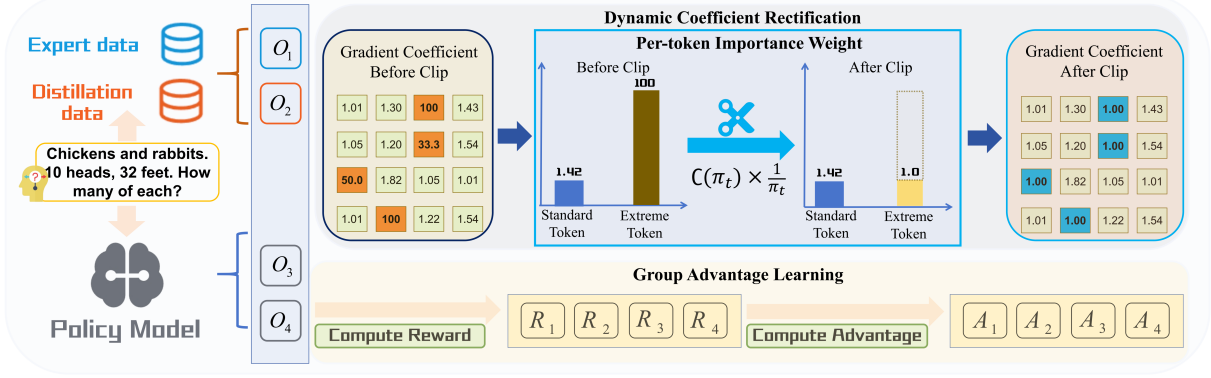


Figure 2: GFT comprises two components: (1) *Group Advantage Learning*, which computes standardized relative advantages ( $A_k$ ) from hybrid response groups (expert demonstrations, teacher outputs, and rollout samples); and (2) *Dynamic Coefficient Rectification*, which bounds importance weights via per-token gradient clipping.

rectification function to stabilize the training:

$$\mathcal{C}(\pi_t) = \begin{cases} \text{sg}(\pi_t) & \text{if } \pi_t < \tau \\ 1 & \text{if } \pi_t \geq \tau \end{cases} \quad (5)$$

Here,  $\tau$  is a confidence threshold, and  $\text{sg}(\cdot)$  denotes the stop-gradient operator. This design actively suppresses the explosive term  $1/\pi_t$  for low-confidence tokens ( $\pi_t < \tau$ ) by using  $\text{sg}(\pi_t)$  to yield a bounded effective coefficient, while leaving the gradient unchanged for confident predictions ( $\pi_t \geq \tau$ ). This ensures stable updates during exploration and preserves full learning strength for knowledge transfer, effectively resolving the instability inherent in the SFT objective.

### 3.3 Final GFT Objective

Combining Group Advantage Learning and Dynamic Coefficient Rectification, we derive the final training objective in its gradient form.

$$\nabla_{\theta} \mathcal{L} = \mathbb{E}_{y_k \in \mathcal{G}_x} \left[ A(y_k) \frac{\mathcal{C}(\pi)}{\pi_{\theta}(y_k|x)} \nabla \log \pi_{\theta}(y_k|x) \right]. \quad (6)$$

Eq. (6) presents the sequence-level gradient of GFT; the corresponding token-level formulation and loss definition are provided in Appendix B. This gradient directly resolves the two intrinsic limitations of SFT: group-wise advantage weighting introduces contrastive supervision across multiple trajectories, while dynamic coefficient rectification bounds the update magnitude for low-probability tokens to prevent gradient explosion.

## 4 Experiments

### 4.1 Experimental Setup

**Baselines and Models** We compare GFT against a diverse set of paradigms, ranging from standard

SFT and its recent stabilized variants—DFT (Wu et al., 2025), ASFT (Zhu et al., 2025a), and PSFT (Zhu et al., 2025b)—to the reinforcement learning baseline GRPO. We further compare with two recent hybrid approaches that integrate off-policy expert demonstrations into on-policy RL: LUFFY (Yan et al., 2025), which learns to reason under off-policy guidance, and CHORD (Zhang et al., 2025), which harmonizes SFT and RL via dynamic weighting. Following DFT (Wu et al., 2025), we evaluate five models covering diverse sizes, types and architectures: Qwen2.5-Math (1.5B, 7B) (Yang et al., 2024), LLaMA-3 (3.2-3B, 3.1-8B) (Dubey et al., 2024), and DeepSeekMath-7B-Base (Shao et al., 2024).

**Training Settings** Following prior works (Wu et al., 2025; Zhu et al., 2025a; Ming et al., 2025; Zhou et al., 2025), we utilize the NuminaMath CoT dataset (LI et al., 2024), selected for its extensive diversity ranging from high school exercises to international olympiads. For GFT, we construct a hybrid response group of size  $K = 8$  per query, comprising 1 expert demonstration, 3 teacher distillations from Qwen2.5-Math-72B, and 4 self-generated samples. Similarly, the GRPO baseline is configured to generate 8 outputs per query. To align total training volume, GFT and GRPO utilize a 10k subset (8 trajectories per query), whereas single-trajectory baselines (e.g., SFT) use 100k samples.

**Evaluation Settings** We conduct evaluations on a broad suite of 11 benchmarks: AMC23 (Mathematical Association of America, 2023), College Math (Hendrycks et al., 2020), Gaokao (Zhang et al., 2023), Math (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), TabMWP (Lu

et al., 2022), OlympiadBench (He et al., 2024), Mmlu Stem (Hendrycks et al., 2020), Sat Math (Zhong et al., 2024), Mawps (Koncel-Kedziorski et al., 2016), and Svamp (Patel et al., 2021). These benchmarks are carefully selected to cover a wide spectrum of difficulty levels and reasoning types, ensuring a holistic assessment of the model’s capabilities. We report the average Pass@1 accuracy across 16 decoding runs (Pass@16 Average) with a sampling temperature of 0.5 and a maximum generation length of 4096 tokens.

## 4.2 Main Results

Based on Table 1, GFT demonstrates strong data efficiency under a reduced training budget: with only 10k training examples, it matches or even surpasses a range of baselines trained with 100k examples. Crucially, mixing in distillation data yields only marginal changes for both SFT and GFT (i.e.,  $SFT(mix) \approx SFT$  and  $GFT(no\ mix) \approx GFT$ ), indicating that the gains are not primarily driven by additional distilled traces but by the proposed training mechanism. Notably, for smaller heterogeneous models like Llama-3.2-3B, GFT (NO MIX) surpasses mixing strategies, implying they are less robust to the distribution mismatch from the teacher’s distinct reasoning patterns. This superior performance is consistently observed across different model scales and model families, suggesting that the improvements are largely model-agnostic. In terms of the performance profile, GFT yields more uniform gains: whereas some methods exhibit uneven improvements or trade-offs across benchmarks, GFT tends to improve performance across diverse evaluations simultaneously. This suggests that GFT is not merely adapting to a specific question format, but is more reliably improving the quality of the underlying reasoning process. Meanwhile, GRPO can be close to GFT because both are largely driven by GAL that converts sparse (often near-binary) rewards into lower-variance, more informative signals; moreover, under our training setting without explicit KL regularization, GRPO’s implicit update stabilization can partially overlap with the effect of our DCR, effectively thereby narrowing down the apparent gap between them.

## 4.3 Ablation Studies

We validate the contributions of GAL and DCR via ablations on Qwen2.5-Math-1.5B, comparing the full GFT with variants that remove GAL, remove DCR, or remove both (equivalent to standard

SFT). We report results on Math500, Minerva, and Olympiad Bench to cover increasing difficulty and robustness requirements, and further inspect the optimization behavior of each variant using the learning-dynamics plot in Figure 3.

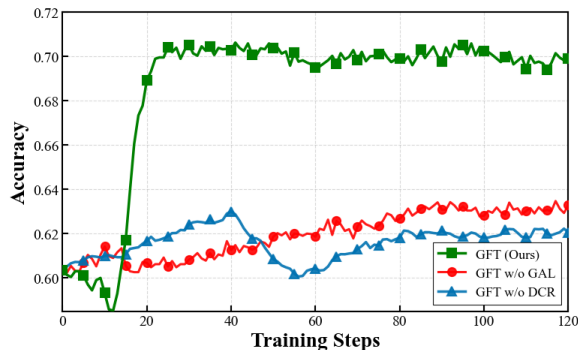


Figure 3: Learning dynamics on MATH-lighteval. Removing DCR causes severe volatility, while removing GAL results in slow convergence and a lower ceiling.

The results in Table 2 demonstrate the distinct contributions of each component. Removing GAL causes the sharpest decline on the hardest benchmark (Olympiad), validating that group-based contrastive feedback is vital for extracting signals in complex reasoning. In contrast, removing DCR primarily impacts robustness (Minerva), consistent with its role in rectifying gradient explosion. These performance patterns are further corroborated by the learning dynamics in Figure 3: the removal of DCR leads to severe training volatility, while removing GAL results in slow, suboptimal convergence. Ultimately, GFT synergizes both components to ensure efficient and stable optimization.

## 4.4 Compatibility with SFT and RL

We conduct a sequential-training compatibility study by combining SFT, GFT, and GRPO in different compositions (Figure 4). This design aims to diagnose the *synergy dilemma* in conventional post-training—where SFT may rigidify the policy and narrow the effective exploration manifold for downstream RL—and to evaluate whether GFT can both (i) serve as a stronger initializer for RL and (ii) improve the handoff from SFT to RL.

As shown in Figure 4, we design GFT to improve compatibility in two aspects. **(1) To improve RL exploration**, GAL prevents the cold-start policy from collapsing to a single expert-induced mode and maintains a multi-solution distribution via group-wise relative advantages. This broader support produces more diverse rollouts and

Table 1: Main results on seven math benchmarks. **SFT(mix)** indicates that the dataset is a mixture of expert datasets and distilled teacher datasets, while **GFT(no mix)** represents using only expert datasets without distilled data. **Bold** and **blue** denote the best intra-group and overall performance, respectively. Overall, GFT achieves the best average performance across diverse model scales.

Model	Method	AMC23	College Math	Gaokao2023En	Math	Minerva Math	TabMWP
Qwen2.5-Math-1.5B	Base Model	30.16	24.30	34.81	46.54	10.51	24.55
	+ SFT	31.25 (+1.09)	36.45 (+12.15)	48.86 (+14.05)	60.66 (+14.12)	23.99 (+13.48)	79.34 (+54.79)
	+ SFT(mix)	32.70 (+2.54)	36.35 (+12.05)	50.82 (+16.01)	60.41 (+13.87)	25.76 (+15.25)	80.15 (+55.60)
	+ GRPO	44.84 (+14.68)	35.58 (+11.28)	51.80 (+16.99)	65.97 (+19.43)	21.17 (+10.66)	76.94 (+52.39)
	+ ASFT	43.12 (+12.96)	29.40 (+5.10)	47.99 (+13.18)	60.35 (+13.81)	15.55 (+5.04)	65.06 (+40.51)
	+ PSFT	31.56 (+1.40)	33.77 (+9.47)	47.66 (+12.85)	59.51 (+12.97)	19.13 (+8.62)	71.61 (+47.06)
	+ DFT	36.40 (+6.24)	38.76 (+14.46)	52.75 (+17.94)	64.35 (+17.81)	23.75 (+13.24)	82.08 (+57.53)
	+ LUFFY	35.15 (+4.99)	36.61 (+12.31)	47.80 (+12.99)	59.80 (+13.26)	17.11 (+6.60)	76.97 (+52.42)
	+ CHORD	<b>47.50 (+17.34)</b>	38.70 (+14.40)	56.42 (+21.61)	68.10 (+21.56)	28.58 (+18.07)	83.09 (+58.54)
	+ GFT(no mix)	42.18 (+12.02)	39.37 (+15.07)	55.59 (+20.78)	68.13 (+21.59)	27.77 (+17.26)	82.21 (+57.66)
+ GFT (Ours)	<b>46.09 (+15.93)</b>	<b>40.51 (+16.21)</b>	<b>58.32 (+23.51)</b>	<b>70.50 (+23.96)</b>	<b>28.93 (+18.42)</b>	<b>85.24 (+60.69)</b>	
Qwen2.5-Math-7B	Base Model	42.66	34.31	49.50	59.10	19.20	85.32
	+ SFT	41.88 (-0.78)	38.31 (+4.00)	54.69 (+5.19)	67.16 (+8.06)	31.82 (+12.62)	87.67 (+2.35)
	+ SFT(mix)	43.06 (+0.40)	39.47 (+5.16)	56.83 (+7.33)	69.63 (+10.53)	32.45 (+13.25)	88.93 (+3.61)
	+ GRPO	55.63 (+12.97)	38.65 (+4.34)	61.63 (+12.13)	73.29 (+14.19)	32.60 (+13.40)	91.18 (+5.86)
	+ ASFT	52.81 (+10.15)	<b>40.76 (+6.45)</b>	61.55 (+12.05)	74.31 (+15.21)	32.47 (+13.27)	89.37 (+4.05)
	+ PSFT	41.56 (-1.10)	38.05 (+3.74)	56.74 (+7.24)	67.30 (+8.20)	34.86 (+15.66)	83.55 (-1.77)
	+ DFT	51.09 (+8.43)	39.31 (+5.00)	57.46 (+7.96)	70.42 (+11.32)	35.31 (+16.11)	86.94 (+1.62)
	+ GFT(no mix)	53.21 (+10.55)	38.74 (+4.43)	61.72 (+12.22)	74.78 (+15.68)	34.22 (+15.02)	92.90 (+7.58)
	+ GFT (Ours)	<b>56.09 (+13.43)</b>	40.24 (+5.93)	<b>63.47 (+13.97)</b>	<b>77.31 (+18.21)</b>	<b>39.86 (+20.66)</b>	<b>93.81 (+8.49)</b>
	DeepSeekMath-7B-Instruct	Base Model	16.09	27.56	38.00	42.73	19.44
+ SFT		20.93 (+4.84)	31.53 (+3.97)	43.13 (+5.13)	46.51 (+3.78)	18.71 (-0.73)	79.30 (+3.60)
+ SFT(mix)		21.57 (+5.48)	<b>32.28 (+4.72)</b>	42.43 (+4.43)	<b>47.52 (+4.79)</b>	20.00 (+0.56)	78.97 (+3.27)
+ GRPO		16.72 (+0.63)	27.59 (+0.03)	42.18 (+4.18)	43.39 (+0.66)	19.39 (-0.05)	77.74 (+2.04)
+ ASFT		15.52 (-0.57)	28.40 (+0.84)	39.03 (+1.03)	44.51 (+1.78)	15.38 (-4.06)	77.42 (+1.72)
+ PSFT		<b>25.78 (+9.69)</b>	30.05 (+2.49)	43.84 (+5.84)	45.36 (+2.63)	18.20 (-1.24)	78.91 (+3.21)
+ DFT		24.37 (+8.28)	30.68 (+3.12)	43.90 (+5.90)	47.01 (+4.28)	19.00 (-0.44)	79.88 (+4.18)
+ GFT(no mix)		18.43 (+2.34)	30.90 (+3.34)	42.56 (+4.56)	45.14 (+2.41)	19.38 (-0.06)	77.97 (+2.27)
+ GFT (Ours)		23.12 (+7.03)	30.98 (+3.42)	<b>48.15 (+10.15)</b>	44.79 (+2.06)	<b>20.42 (+0.98)</b>	<b>80.08 (+4.38)</b>
LLaMA-3.2-3B-Instruct		Base Model	23.78	25.40	38.06	44.63	14.83
	+ SFT	19.53 (-4.25)	26.12 (+0.72)	36.33 (-1.73)	43.66 (-0.97)	12.14 (-2.69)	68.40 (-0.72)
	+ SFT(mix)	21.09 (-2.69)	27.88 (+2.48)	37.76 (-0.30)	45.68 (+1.05)	14.15 (-0.68)	70.00 (+0.88)
	+ GRPO	23.25 (-0.53)	28.01 (+2.61)	40.61 (+2.55)	46.18 (+1.55)	18.44 (+3.61)	67.53 (-1.59)
	+ ASFT	18.44 (-5.34)	26.13 (+0.73)	37.49 (-0.57)	43.65 (-0.98)	11.38 (-3.45)	66.64 (-2.48)
	+ PSFT	24.37 (+0.59)	28.94 (+3.54)	40.73 (+2.67)	47.43 (+2.80)	15.41 (+0.58)	70.91 (+1.79)
	+ DFT	14.21 (-9.57)	26.63 (+1.23)	35.13 (-2.93)	41.45 (-3.18)	9.58 (-5.25)	67.39 (-1.73)
	+ GFT(no mix)	<b>27.66 (+3.88)</b>	<b>31.14 (+5.74)</b>	<b>43.88 (+5.82)</b>	<b>51.71 (+7.08)</b>	<b>21.29 (+6.46)</b>	72.87 (+3.75)
	+ GFT (Ours)	27.50 (+3.72)	30.15 (+4.75)	43.25 (+5.19)	49.60 (+4.97)	18.84 (+4.01)	<b>72.93 (+3.81)</b>

stronger advantage signals for GRPO, explaining why *GFT + GRPO* gains more than *SFT + GRPO* on harder benchmarks (Li et al., 2024). (2) To prevent distribution extremization and preserve exploration, DCR bounds per-token updates to avoid over-sharpening an SFT-initialized policy. Without this constraint, large steps can quickly drive the policy to a low-entropy, mode-concentrated distribution, reducing rollout diversity and weakening GRPO’s learning signal. By limiting update magnitude, DCR keeps the policy in a higher-entropy regime, matching the smoother dynamics and higher ceiling of *SFT + GFT + GRPO* in Figure 4. Notably, *GFT + GRPO* surpassing *SFT +*

*GRPO* does not mean GFT replaces SFT: SFT provides a reliable initialization point for alignment and formatting, while GFT improves RL compatibility by preserving support and stabilizing updates. Thus, *SFT + GFT + GRPO* works best as a staged pipeline: SFT sets the initialization point, GFT restores exploration capabilities without drifting, and GRPO leverages higher-quality trajectories to reach the top ceiling.

#### 4.5 Catastrophic Forgetting Analysis

Table 3 shows a clear contrast in catastrophic forgetting on general reasoning benchmarks. After domain training, **SFT** exhibits substantial degrada-

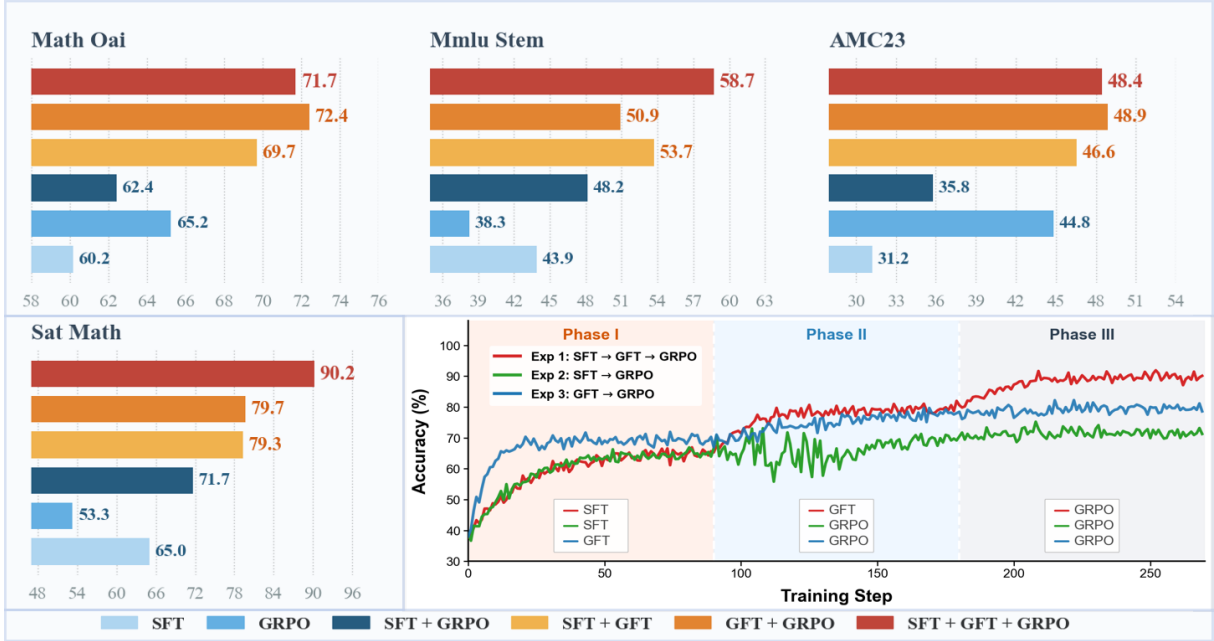


Figure 4: Performance comparison on Qwen2.5-Math-1.5B (Pass@16). Bottom-right: Sat-Math training dynamics. SFT+GFT+GRPO achieves top performance via stable optimization, demonstrating GFT’s high compatibility and effective synergy between SFT and GRPO.

Table 2: Ablation on **Qwen2.5-Math-1.5B**. GAL is important for complex reasoning (e.g., Olympiad) and DCR enhances performance by ensuring optimization stability, their synergy yields optimal results.

Method	AMC23	MATH	Olympiad
Base Model	30.16	46.54	23.39
GFT w/o (GAL + DCR)	31.25	60.66	24.58
GFT w/o GAL	35.78	63.91	26.63
GFT w/o DCR	42.81	65.97	27.82
<b>GFT (Ours)</b>	<b>46.09</b>	<b>70.50</b>	<b>30.52</b>

tion on MAWPS and SVAMP and also drops on MMLU-STEM, indicating severe forgetting. In contrast, **GRPO** largely preserves the base model’s prior capabilities, while **GFT** not only maintains comparable retention to GRPO but also improves MMLU-STEM. This ranking is further consistent with Figure 5, where the policy shift of SFT is the most pronounced, whereas GRPO and GFT remain significantly closer to the base policy.

To quantify forgetting more directly, we adopt the approach of (Shenfeld et al., 2025) and compute the *average KL divergence* between the trained model and the base model on the training dataset. Recent empirical studies further support the correlation between this KL-based drift and forgetting (Chu et al., 2024; Luo et al., 2025; Ruan et al., 2025). We therefore use the average KL diver-

Table 3: Performance of **LLaMA-3.2-3B-Instruct** on general reasoning benchmarks. While SFT induces substantial catastrophic forgetting, GFT largely preserves base performance.

Method	Mawps	Svamp	Mmlu stem
Base Model	96.06	86.36	41.03
+SFT	91.97 (- 4.09)	78.73 (- 7.63)	35.05 (-5.98)
+GRPO	94.60 (- 1.46)	<b>88.11 (+ 1.75)</b>	39.48 (-1.55)
<b>+GFT (Ours)</b>	<b>95.79 (- 0.27)</b>	84.65 (- 1.71)	<b>43.89 (+2.86)</b>

gence as a proxy for distributional drift, and hence forgetting. We analyze the training dynamics of Qwen2.5-Math-1.5B across different methods. As shown in Figure 5, all baselines converge to their peak performance approximately at step 100. At this stage, we observe a distinct contrast: **SFT** incurs the highest alignment tax with the largest KL divergence, whereas **GRPO** retains a *KL-minimal* solution; notably, **GFT** strikes a balance, stabilizing at a low KL level comparable to GRPO. We attribute this stability to our design: **GAL** reinforces high-quality output trajectories in a reward-driven manner, avoiding abrupt distributional shifts induced by pure cross-entropy trace fitting; meanwhile, **DCR** suppresses gradient explosions from “extreme tokens” (where  $\pi_\theta \approx 0$ ), preventing drastic policy drift. Together, these components enable efficient knowledge injection while retaining robust

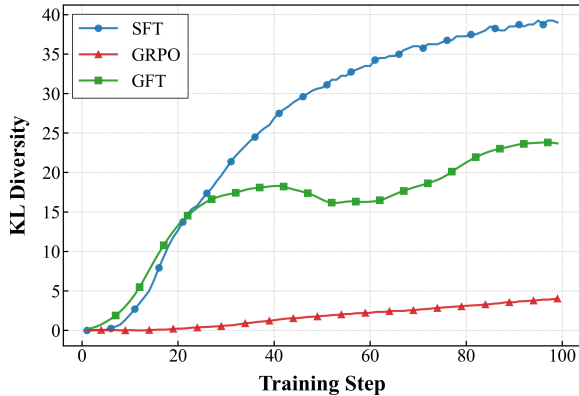


Figure 5: KL divergence quantifies distributional drift from the base model. SFT exhibits the highest divergence, while GFT maintains a significantly lower level, effectively mitigating catastrophic forgetting.

Table 4: Comparison of Pass@k ( $k = 128, 256$ ) performance between Distillation, GRPO and GFT. GFT consistently achieves the highest Pass@k scores, effectively enhancing response diversity.

Metric	Method	SAT Math	Minerva	TabMWP	Avg.
Pass@128	Base Model	39.69	9.71	24.17	24.52
	Distillation	66.67	22.98	79.32	56.32
	GRPO	52.95	19.89	76.77	49.87
	<b>GFT</b>	<b>72.58</b>	<b>28.59</b>	<b>85.31</b>	<b>62.16</b>
	Base Model	38.76	9.25	24.36	24.12
Pass@256	Distillation	67.20	21.84	79.28	56.11
	GRPO	51.90	19.77	75.82	49.16
	<b>GFT</b>	<b>73.33</b>	<b>27.17</b>	<b>85.23</b>	<b>61.91</b>

general-purpose reasoning.

#### 4.6 Diversity of GFT

Balancing solution diversity with correctness remains a challenge in post-training. While distillation preserves exploration by mimicking the teacher’s soft targets (Goyal et al., 2025), it often lacks explicit correctness incentives. Conversely, RL-style optimization (e.g., GRPO) tends to sharpen the policy toward specific high-reward trajectories, which effectively optimizes precision but may suppress the exploration space and reduce solution variety (Yue et al., 2025). To evaluate whether GFT can effectively reconcile this trade-off—maintaining intrinsic diversity while ensuring accuracy—we conduct a multi-sample evaluation using Pass@k as a proxy metric for solution coverage. Table 4 compares the diversity performance of **Distillation**, **GRPO**, and **GFT**.

GFT achieves the highest Pass@128 and

Pass@256 across benchmarks. Distillation improves exploration because soft targets from teacher train the student to match the teacher’s *output distribution*, but it does not use reward to distinguish correct reasoning. GRPO, in contrast, uses reward to *sharpen* the student distribution, which strengthens memory of rewarded (often correct) paths but also narrows exploration. GFT combines both signals by reward-evaluating trajectories from *both* the teacher distribution and the student’s own sampling distribution: it learns the teacher’s diverse modes (as in distillation) while using within-group advantages to explicitly compare student samples against teacher traces, pushing the student toward the teacher’s *high-reward diverse* modes. This teacher–student gap correction preserves diversity where it matters, leading to higher Pass@k.

#### 4.7 Hyperparameter Analysis

Table 5: Impact of group composition ratio ( $N_{demo} : N_{sample}$ ); **2:6** achieves the best accuracy, indicating richer contrast from self-samples with demo samples.

Ratio	Minerva Math	Olympiad	Sat Math	Avg.
8 : 0	15.11	22.48	36.92	24.84
6 : 2	29.53	29.60	71.68	43.60
4 : 4	28.93	30.52	69.93	43.13
<b>2 : 6</b>	<b>31.01</b>	<b>32.73</b>	<b>73.04</b>	<b>45.59</b>
0 : 8	23.31	28.61	40.60	30.84

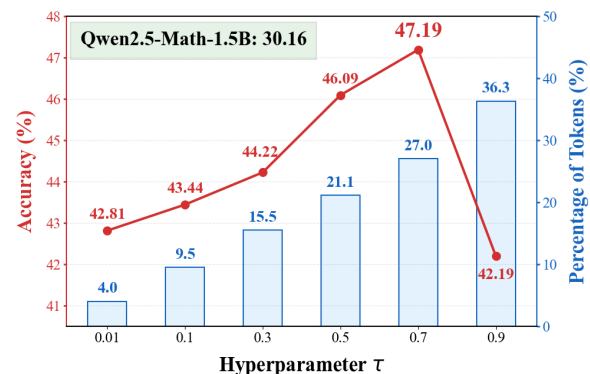


Figure 6: Effect of the clipping threshold  $\tau$ : larger  $\tau$  rectifies more tokens. Accuracy follows an inverted U-shape; insufficient clipping is unstable, while excessive clipping reduces learning efficiency.

To probe the impact of **group diversity** and **rectification strength**, we ablate the composition ratio ( $N_{demo} : N_{sample}$ ) and threshold  $\tau$  on Qwen2.5-Math-1.5B. With fixed  $K = 8$ , Table 5 identifies **2:6** as optimal, where minimal demonstrations *an-*

chor correctness while abundant self-samples provide richer *contrastive signals* for advantage learning. Regarding the clipping threshold  $\tau$ , Figure 6 reports accuracy together with the fraction of DCR-rectified tokens. As  $\tau$  increases, the rectification rate rises monotonically, indicating stronger clipping. Meanwhile, accuracy exhibits an **inverted U-shape**: small  $\tau$  yields insufficient clipping and unstable updates, whereas large  $\tau$  over-clips many tokens and attenuates informative gradients, harming learning efficiency. Consequently,  $\tau \approx 0.7$  achieves the best **stability–efficiency trade-off** in learning. Notably, GFT consistently outperforms the base model across the entire sweep of parameters, suggesting that DCR is robust to  $\tau$ .

## 5 Conclusion

In this work, we analyze SFT as a special case of RL. This perspective reveals two intrinsic limitations: single-path dependency that restricts exploration, and gradient explosion that causes instability. To address these, we propose **Group Fine-Tuning (GFT)**. This framework leverages Group Advantage Learning to enhance diversity via contrastive supervision and employs Dynamic Coefficient Rectification to stabilize optimization by preventing extreme weight updates. Experiments demonstrate that GFT effectively balances efficient knowledge injection with robust generalization, offering a principled paradigm for post-training.

## Limitations

Despite GFT’s effectiveness, we acknowledge three limitations. First, our evaluation focuses on mathematical reasoning with objective correctness; extending GFT to open-ended tasks with subjective rewards requires further exploration. Second, constructing response groups introduces marginal data preparation overhead compared to standard SFT, though this cost is significantly lower than online RL. Third, due to academic resource constraints, our experiments are limited to models up to 8B parameters; validating GFT on 70B+ models remains an important future direction.

## Acknowledgments

This work is supported by the Key R&D Program of Ningbo under Grant No.2024Z115.

## References

- Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning aligns llms to the human brain. *arXiv preprint arXiv:2312.00575*.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dasgupta, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Hase, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025a. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.
- Jierun Chen, Tiezheng Yu, Haoli Bai, Lewei Yao, Jiannan Wu, Kaican Li, Fei Mi, Chaofan Tao, Lei Zhu, Manyi Zhang, and 1 others. 2025b. The synergy dilemma of long-cot sft and rl: Investigating post-training techniques for reasoning vlms. *arXiv preprint arXiv:2507.07562*.
- Zhipeng Chen, Yingqian Min, Beichen Zhang, Jie Chen, Jinhao Jiang, Daixuan Cheng, Wayne Xin Zhao, Zheng Liu, Xu Miao, Yang Lu, Lei Fang, Zhongyuan Wang, and Ji-Rong Wen. 2025c. An empirical study on eliciting and improving r1-like reasoning models. *arXiv preprint arXiv:2503.04548*.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. volume 30.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2024. Sft memorizes, rl generalizes: A comparative study of foundation model post-training.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. 2025. Sft: A single-stage method with supervised and reinforcement fine-tuning for reasoning. *arXiv preprint arXiv:2506.19767*.

- Sachin Goyal, David Lopez-Paz, and Kartik Ahuja. 2025. Distilled pretraining: A modern lens of data, in-context learning and test-time scaling. *arXiv preprint arXiv:2509.01649*.
- Arnab Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and 1 others. 2024. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. pages 3828–3850.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- 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.
- Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. 2025. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *arXiv preprint arXiv:2507.00432*.
- Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. 2016. Mawps: A math word problem repository. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 1152–1157.
- 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. 35:3843–3857.
- Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. 2024. Numinamath.
- Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. 2024. Preserving diversity in supervised fine-tuning of large language models. *arXiv preprint arXiv:2408.16673*.
- Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. 2025. Uft: Unifying supervised and reinforcement fine-tuning. *arXiv preprint arXiv:2505.16984*.
- Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.
- Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*.
- Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. 2022. What matters in learning from offline human demonstrations for robot manipulation. In *CoRL*, pages 1678–1690.
- Mathematical Association of America. 2023. Amc 2023 competition problems.
- Rui Ming, Haoyuan Wu, Shoubo Hu, Zhuolun He, and Bei Yu. 2025. One-token rollout: Guiding supervised fine-tuning of llms with policy gradient. *arXiv preprint arXiv:2509.26313*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.
- Chongli Qin and Jost Tobias Springenberg. 2025. Supervised fine tuning on curated data is reinforcement learning (and can be improved). *arXiv preprint arXiv:2507.12856*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- Zhiwen Ruan, Yun Chen, Yutao Hou, Peng Li, Yang Liu, and Guanhua Chen. 2025. Unveiling overmemorization in finetuning llms for reasoning tasks. *arXiv preprint arXiv:2508.04117*.
- 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, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. 2025. RL’s razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Gokul Swamy, Sanjiban Choudhury, Wen Sun, Zhiwei Steven Wu, and J Andrew Bagnell. 2025. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *arXiv preprint arXiv:2503.01067*.
- Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.
- 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*.
- Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.
- 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. *arXiv preprint arXiv:2504.14945*.
- An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.
- Fei Yu, Anningzhe Gao, and Benyou Wang. 2024. Ovm, outcome-supervised value models for planning in mathematical reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 858–875.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.
- 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*.
- Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. 2023. Evaluating the performance of large language models on gaokao benchmark. *arXiv preprint arXiv:2305.12474*.
- Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.
- Chunting Zhou, Pengfei Liu, and Meta Ai. Lima: Less is more for alignment.
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.
- Xiangxin Zhou, Zichen Liu, Haonan Wang, Chao Du, Min Lin, Chongxuan Li, Liang Wang, and Tianyu Pang. 2025. Variational reasoning for language models. *arXiv preprint arXiv:2509.22637*.
- He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. 2025a. Anchored supervised fine-tuning. *arXiv preprint arXiv:2509.23753*.
- Wenhong Zhu, Ruobing Xie, Rui Wang, Xingwu Sun, Di Wang, and Pengfei Liu. 2025b. Proximal supervised fine-tuning. *arXiv preprint arXiv:2508.17784*.

## A Derivation: Viewing SFT as a Special Case of On-Policy RL

In this appendix, we provide a detailed derivation showing that supervised fine-tuning (SFT) can be interpreted as a special case of reinforcement learning (RL) with a sparse reward function. Specifically, we show that the gradient of the SFT objective can be rewritten as an on-policy expectation under the current policy via importance sampling.

### A.1 SFT Objective and Gradient

We consider a dataset of expert demonstrations  $\mathcal{D} = \{(x, y^*)\}$ , where  $x$  denotes the input and  $y^*$  is the expert-provided output. The standard SFT objective is defined as the negative log-likelihood:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x, y^*) \sim \mathcal{D}} [\log \pi_{\theta}(y^* | x)]. \quad (7)$$

Taking the gradient with respect to the model parameters  $\theta$ , we obtain

$$\nabla_{\theta} \mathcal{L}_{\text{SFT}}(\theta) = -\mathbb{E}_{(x, y^*) \sim \mathcal{D}} [\nabla_{\theta} \log \pi_{\theta}(y^* | x)]. \quad (8)$$

This expectation is taken over the expert data distribution rather than samples generated by the current policy.

### A.2 Importance Sampling Reformulation

We factorize the expert data distribution as

$$P(x, y^*) = P(x) P_{\text{expert}}(y^* | x), \quad (9)$$

and define the joint distribution induced by the current policy as

$$Q(x, y) = P(x) \pi_{\theta}(y | x). \quad (10)$$

Since both distributions share the same marginal  $P(x)$ , we can apply importance sampling to rewrite the expectation in Eq. (8) under  $Q(x, y)$ :

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{SFT}}(\theta) &= -\mathbb{E}_{(x, y) \sim Q} \left[ \nabla_{\theta} \log \pi_{\theta}(y | x) \cdot \frac{P_{\text{expert}}(y | x)}{\pi_{\theta}(y | x)} \right]. \quad (11) \end{aligned}$$

For deterministic expert demonstrations, the expert conditional distribution reduces to a Dirac delta:

$$P_{\text{expert}}(y | x) = \mathbb{I}[y = y^*]. \quad (12)$$

Substituting this into Eq. (11) yields

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{SFT}}(\theta) &= -\mathbb{E}_{(x, y) \sim Q} \left[ \frac{\mathbb{I}[y = y^*]}{\pi_{\theta}(y | x)} \nabla_{\theta} \log \pi_{\theta}(y | x) \right]. \quad (13) \end{aligned}$$

This recovers the equivalent on-policy formulation presented in the main text.

## A.3 Reinforcement Learning Interpretation

Equation (13) admits a direct reinforcement learning interpretation. In particular, it corresponds to an on-policy policy gradient with:

- **Policy:**  $\pi_{\theta}(y | x)$ ;
- **Reward function:**

$$r(x, y) = \mathbb{I}[y = y^*], \quad (14)$$

which provides a unit reward only when the sampled output exactly matches the expert demonstration;

- **Importance weight:**

$$w(x, y) = \frac{1}{\pi_{\theta}(y | x)}, \quad (15)$$

correcting for sampling from the model policy instead of the expert distribution.

Under this view, SFT can be regarded as a degenerate RL setting with an extremely sparse reward signal and high variance, where learning occurs only through trajectories that coincide exactly with expert demonstrations.

## A.4 Summary

In summary, the derivation proceeds by (i) expressing the SFT gradient as an expectation over expert data, (ii) applying importance sampling to rewrite it under the model policy, and (iii) specializing the expert distribution to a deterministic form. This establishes a formal equivalence between SFT and on-policy reinforcement learning with a sparse indicator reward, providing a unified perspective on supervised and reinforcement-based post-training.

## B Formulation of Group Fine-Tuning

In this appendix, we provide the explicit loss formulations and gradient expressions of Group Fine-Tuning (GFT), including both sequence-level and token-level forms. These formulations correspond to the gradient expression presented in Eq. (6) in the main text.

### B.1 Sequence-Level Objective

For each input query  $x$ , we construct a response group  $\mathcal{G}_x = \{y_1, \dots, y_K\}$ , where each response  $y_k$  is assigned a scalar reward  $R(y_k)$  and a standardized group advantage  $A(y_k)$  as defined in Eq. (4). We define the sequence-level GFT loss as

$$\mathcal{L}_{\text{GFT}}^{\text{seq}}(\theta) = -\mathbb{E}_x \left[ \sum_{y_k \in \mathcal{G}_x} A(y_k) \mathcal{C}(\pi_\theta(y_k | x)) \cdot \log \pi_\theta(y_k | x) \right]. \quad (16)$$

where  $\mathcal{C}(\cdot)$  is the dynamic coefficient rectification function defined in Eq. (5).

Taking the gradient of Eq. (16) yields the sequence-level policy gradient:

$$\nabla_\theta \mathcal{L}_{\text{GFT}}^{\text{seq}} = \mathbb{E}_x \left[ \sum_{y_k \in \mathcal{G}_x} A(y_k) \frac{\mathcal{C}(\pi_\theta(y_k | x))}{\pi_\theta(y_k | x)} \cdot \nabla_\theta \log \pi_\theta(y_k | x) \right]. \quad (17)$$

## B.2 Token-Level Decomposition

Each response sequence  $y_k = (y_{k,1}, \dots, y_{k,T_k})$  is generated autoregressively by the policy:

$$\pi_\theta(y_k | x) = \prod_{t=1}^{T_k} \pi_\theta(y_{k,t} | y_{k,<t}, x). \quad (18)$$

Accordingly, the sequence log-probability decomposes as

$$\log \pi_\theta(y_k | x) = \sum_{t=1}^{T_k} \log \pi_\theta(y_{k,t} | y_{k,<t}, x). \quad (19)$$

We use the shorthand

$$\pi_{k,t} \triangleq \pi_\theta(y_{k,t} | y_{k,<t}, x). \quad (20)$$

for the token-level prediction probability.

Substituting the above decomposition into Eq. (16), we obtain the token-level GFT loss:

$$\mathcal{L}_{\text{GFT}}^{\text{tok}}(\theta) = -\mathbb{E}_x \left[ \sum_{y_k \in \mathcal{G}_x} A(y_k) \sum_{t=1}^{T_k} \mathcal{C}(\pi_{k,t}) \log \pi_{k,t} \right]. \quad (21)$$

Taking the gradient yields the token-level policy gradient:

$$\nabla_\theta \mathcal{L}_{\text{GFT}}^{\text{tok}} = \mathbb{E}_x \left[ \sum_{y_k \in \mathcal{G}_x} A(y_k) \sum_{t=1}^{T_k} \frac{\mathcal{C}(\pi_{k,t})}{\pi_{k,t}} \nabla_\theta \log \pi_{k,t} \right]. \quad (22)$$

## B.3 Relation to SFT and RL Objectives

When the response group degenerates to a single expert demonstration ( $|\mathcal{G}_x| = 1$ ), the advantage is constant and Eq. (22) reduces to the standard SFT gradient. Conversely, when the group consists of diverse sampled trajectories with non-trivial advantage values, GFT recovers an on-policy reinforcement learning update with group-normalized advantage weighting and bounded importance coefficients.

This formulation establishes GFT as a strict generalization of SFT and a stabilized, contrastive variant of policy-gradient-based post-training.

## C Related Work

**Trade-off Between SFT and RL** Post-training paradigms typically navigate a trade-off between Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL). SFT is widely recognized for its efficiency in knowledge injection and ‘‘cold-starting’’ (Zhou et al., 2023; Chung et al., 2024); however, it is prone to mechanical memorization and often fails to generalize to out-of-distribution scenarios (Ouyang et al., 2022; Bai et al., 2022; Chu et al., 2024; Swamy et al., 2025; Huan et al., 2025). Conversely, RL excels at discovering robust strategies and optimizing long-term objectives (Christiano et al., 2017), yet it is computationally expensive and struggles to acquire complex reasoning skills from scratch without sufficient guidance (Schulman et al., 2017; Sheng et al., 2025; Mandlkar et al., 2022; Chen et al., 2025c).

**The Synergy Dilemma in Hybrid Post-Training** Standard hybrid approaches (e.g., SFT followed by RL) attempt to combine these complementary strengths but face a severe ‘‘synergy dilemma’’ (Ouyang et al., 2022; Rafailov et al., 2023). Recent studies conclude that this conflict arises from the fundamental training dynamics: the overfitting induced by SFT creates a rigid policy that severely constrains the exploration space required for subsequent RL (Chen et al., 2025a), while simultaneously leading to reasoning pattern mismatches that hinder effective policy alignment (Chen et al., 2025b). Although methods like interleaved updates (Liu et al., 2025) or preference optimization (Rafailov et al., 2023) offer partial solutions, they remain dependent on external feedback signals. In contrast, our work addresses this dilemma by transforming the rigid imitation objective into a **Group Advantage Learning** frame-

work, which explicitly preserves solution diversity and the exploration manifold by optimizing contrastive advantages derived from hybrid response groups.

**Single-Stage Hybrids: Mixing Imitation and Exploration** Several recent studies have attempted to unify SFT and RL by balancing imitation and exploration through modified objectives (Yuan et al., 2024). Single-stage hybrid methods, such as SRFT (Fu et al., 2025) and UFT (Liu et al., 2025), employ dynamic weighting mechanisms, interleaved updates, or dense verification signals (Wang et al., 2024; Yu et al., 2024) to mix supervised signals with reinforcement objectives. Similarly, frameworks like HybridFlow (Sheng et al., 2025) explore flexible combinations of offline and on-line data to bridge the gap. While approaches like CHORD (Zhu et al., 2025a) introduce anchor-based constraints to maintain stability, a common limitation across these methods is that they often treat SFT and RL as separate components to be linearly combined or alternated, rather than fusing them mathematically into a cohesive formulation derived from a unified training dynamic.

**Gradient-Level Stabilization and Its New Trade-offs** To address the instability inherent in post-training, other researchers have revisited the underlying gradient formulation. Theoretical analyses suggest a deeper equivalence between likelihood maximization and reinforcement learning (Swamy et al., 2025), prompting new rectification strategies. For instance, Wu et al. (2025) propose Dynamic Fine-Tuning (DFT), which counteracts gradient explosion by reweighting the loss with the model’s likelihood to cancel the inverse-probability term. However, this indiscriminate dampening creates a new dilemma: it suppresses the strong gradient signals required for injecting novel knowledge, potentially hindering adaptation to new domains. Alternatively, approaches like Proximal SFT (Zhu et al., 2025b) and Anchored SFT (Zhu et al., 2025a) introduce trust-region constraints to stabilize fine-tuning, yet such rigid regularizations may overly constrain the model’s plasticity. In the realm of Reinforcement Learning, stability is traditionally enforced via KL-divergence penalties (Ouyang et al., 2022) or clipping mechanisms (Schulman et al., 2017). More recently, group-based methods like GRPO (Shao et al., 2024) have emerged to mitigate gradient variance by normalizing advantages within generated groups, effectively removing the

reliance on unstable critic models, while system-level frameworks like HybridFlow (Sheng et al., 2025) attempt to stabilize training through flexible data scheduling.