

GIFT: Reconciling Post-Training Objectives via Variational Finite-Temperature Gibbs Initialization

Zhengyang Zhao^{1,*}, Lu Ma^{1,*}, Yizhen Jiang^{1,2}, Xiaochen Ma¹, Zimo Meng¹,
Chengyu Shen¹, Lexiang Tang¹, Haoze Sun², Peng Pei², Wentao Zhang^{1,3,4,†}

¹Peking University ²Meituan ³Zhongguancun Academy

⁴Beijing Key Laboratory of Data Intelligence and Security (Peking University)
zhengyangzhao25@stu.pku.edu.cn, maluqaq@163.com

Abstract

The prevailing post-training paradigm for Large Reasoning Models (LRMs)—Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL)—suffers from an intrinsic optimization mismatch: the rigid likelihood maximization in SFT induces distributional collapse, thereby exhausting the exploration space necessary for subsequent RL. Motivated by the Gibbs optimum of KL-regularized RL, we derive a token-level variational surrogate that makes the SFT target structurally compatible with the subsequent RL stage, and propose Gibbs Initialization with Finite Temperature (GIFT). Standard SFT emerges as a degenerate zero-temperature limit of this surrogate, while a finite temperature preserves structural diversity. Our experiments demonstrate that GIFT outperforms standard SFT and surpasses competitive baselines when utilized for RL initialization. Our code is available at <https://github.com/zy1127/GIFT>.

1 Introduction

The training pipeline for Large Reasoning Models (LRMs) has converged into a standard two-stage paradigm. First, Supervised Fine-Tuning (SFT) establishes fundamental reasoning capabilities by leveraging expert demonstrations (Chung et al., 2022; Ouyang et al., 2022; Touvron et al., 2023). Second, Reinforcement Learning (RL) encourages the model to explore self-generated reasoning paths, allowing it to optimize for correctness beyond the constraints of static datasets (Shao et al., 2024a; Yu et al., 2025; Rafailov et al., 2024; Schulman et al., 2017). This hierarchical approach underpins current state-of-the-art reasoning systems (DeepSeek-AI et al., 2025b; Yang et al., 2025; Nvidia et al., 2024; Huang et al., 2026).

However, a fundamental theoretical discrepancy persists between these two phases. While SFT typically minimizes cross-entropy loss to imitate expert tokens, RL maximizes expected rewards through environmental exploration. Standard SFT aggressively suppresses the probability mass of non-target tokens, thereby forcing the policy distribution to collapse onto deterministic data points (Wu et al., 2025). This phenomenon, characterized as distributional collapse, erodes the structural priors inherited from pre-training and severely restricts the subsequent exploration space (Li et al., 2025). Consequently, as the model transitions to the RL stage, it encounters a critical bottleneck: the stochastic diversity required to discover high-reward trajectories has been substantially reduced by the rigid supervision of the preceding SFT phase (Gudiband et al., 2023; Kang et al., 2025).

Previous research has attempted to mitigate this issue by heuristically tuning data mixtures and inter-stage scheduling (Kang et al., 2025; Liu et al., 2025b), or by modifying the SFT objective to implicitly incorporate RL properties (Zhu et al., 2025b; Wu et al., 2025; Du et al., 2025; Rafailov et al., 2024). In contrast to these methods, we do not attempt to bypass the RL stage or rely on heuristic adjustments; instead, we aim to establish a mathematically principled coupling between the two phases. We posit that SFT should not be treated as an isolated imitation task but rather as a theoretically motivated initialization for subsequent RL within a coupled post-training framework.

Based on this insight, motivated by the Gibbs optimum of KL-regularized RL, we derive a token-level variational surrogate that makes the SFT target structurally compatible with the subsequent RL stage, and propose Gibbs Initialization with Finite Temperature (GIFT). GIFT ensures that while the model prioritizes expert trajectories, it preserves the structural diversity of the base model, offer-

*Equal contribution.

†Corresponding author.

ing a finite-temperature solution to the alignment-exploration trade-off. Within our framework, standard SFT is characterized as a degenerate zero-temperature limit that induces a rigid distributional collapse onto the expert data. In contrast, GIFT maintains a finite temperature, incorporating expert supervision as a continuous energy potential rather than a hard overwrite, thereby establishing a principled distributional bridge to subsequent RL.

Empirically, we conduct SFT followed by RL on a subset of the DeepMath-103k dataset (He et al., 2025). Experimental results demonstrate that GIFT consistently outperforms standard SFT and matches or surpasses other robust SFT variants when utilized as an initialization policy for RL across diverse mathematical reasoning benchmarks and out-of-distribution (OOD) tasks. Furthermore, our geometric and distributional analyses reveal that GIFT maintains objective consistency throughout the two-stage post-training process. By preserving the exploration landscape, GIFT facilitates accelerated convergence and superior final performance during the subsequent RL stage, enhancing the model’s reasoning capabilities.

The contributions of this work are twofold, spanning theoretical and practical dimensions. Theoretically, by leveraging the well-established Gibbs optimal policy from KL-regularized RL, we provide a novel reinterpretation of the post-training pipeline. Specifically, we show that standard SFT corresponds to the degenerate zero-temperature limit of this surrogate, which sharpens the policy and erodes the exploration space for subsequent RL. Practically, we propose GIFT to realize a principled finite-temperature initialization. Extensive experiments demonstrate that GIFT outperforms robust SFT baselines and matches competitive unified post-training methods on reasoning and OOD benchmarks, while maintaining superior objective consistency throughout the two-stage post-training pipeline.

2 Related Work

Post-training for Reasoning: SFT-then-RL Paradigm. Post-training for reasoning LLMs typically follows a two-stage paradigm: Supervised Fine-Tuning (SFT) and Reinforcement Learning with Verifiable Rewards (RLVR) (DeepSeek-AI et al., 2025a; Team et al., 2025; Liu et al., 2025b). SFT acts as a "cold-start" phase, establishing a strong initial policy by training on high-quality

reasoning chains. Subsequently, RLVR enhances problem-solving capabilities by allowing the model to explore the solution space beyond the SFT data, discovering novel and more robust reasoning paths. However, this transition is impeded by an intrinsic optimization conflict (Ouyang et al., 2022; Touvron et al., 2023). Specifically, SFT typically compels the model to fit deterministic one-hot labels, which aggressively suppresses the probability mass of alternative tokens. In contrast, RL necessitates a high-entropy policy distribution to facilitate exploration and the discovery of high-reward trajectories (Chen et al., 2025a; Zhang et al., 2024; Kumar et al., 2022; Niu et al., 2026). This mismatch often causes SFT to overfit specific patterns and induce a collapse of the probability manifold, thereby stifling the exploration capacity required for the subsequent RL stage (Kang et al., 2025; Vattikonda et al., 2025; Chen et al., 2025c).

Refining SFT and Bridging Post-Training Stages. To mitigate the distributional degradation inherent in standard SFT, numerous studies have refined its objective by modifying the training loss (e.g., dynamic re-weighting and clipping (Li et al., 2025; Wu et al., 2025; Zhu et al., 2025a,b)) or isolating loss calculations to critical high-entropy reasoning tokens (Wang et al., 2025; Jiang et al., 2025; Ruan et al., 2025). However, treating SFT as an isolated task often fails to address its overarching optimization mismatch with subsequent RL. To bridge this gap, alternative efforts seek to interleave the two stages (Ma et al., 2025; Zhang et al., 2025) or unify their objectives into a single framework (Yan et al., 2025; Ming et al., 2025; Liu et al., 2025a; Chen et al., 2025b). Despite their theoretical potential, these unified approaches introduce substantial computational complexity and have yet to supersede the empirically robust SFT-then-RL pipeline. In contrast, GIFT operates within this standard paradigm, proposing a lightweight, single-pass and mathematically principled initialization method designed to reconcile optimization objectives across the entire post-training process.

3 Method

We propose **Gibbs Initialization with Finite Temperature (GIFT)**, an SFT target designed to mathematically bridge the gap between initial fine-tuning and subsequent RL. Rather than treating SFT in isolation, we derive GIFT by additively decomposing the overarching post-training objective

across the two stages (Section 3.2). Concretely, GIFT is the closed-form optimizer of a per-token variational surrogate that up-weights the demonstration token while preserving the base distribution as a prior (Section 3.3). Parameterized by an inverse temperature β , it recovers standard one-hot SFT in the $\beta \rightarrow \infty$ degenerate limit. The training algorithm is summarized in Section 3.4.

3.1 Preliminaries

Supervised Fine-Tuning. Starting with a pre-trained base model π_{base} and a curated dataset of high-quality demonstrations $\mathcal{D} = \{(x, y^*)\}$, SFT optimizes the model parameters θ to maximize the log-likelihood of the target sequences. The objective is typically formulated as a cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{\text{SFT}}(\theta) &= -\mathbb{E}_{(x, y^*) \sim \mathcal{D}} \log \pi_{\theta}(y^* | x) \\ &= -\mathbb{E}_{(x, y^*) \sim \mathcal{D}} \sum_{t=1}^{|y^*|} [\log \pi_{\theta}(y_t^* | x, y_{<t}^*)].\end{aligned}\quad (1)$$

By minimizing this negative log-likelihood, the model learns to imitate the distribution of the expert data, providing a specialized foundation for subsequent alignment or reasoning-based stages.

Reinforcement Learning with Verifiable Rewards (RLVR). Following SFT, the model π_{sft} is further refined using RL algorithms—such as PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024b)—to maximize a KL-regularized reward objective:

$$\begin{aligned}J_{\text{RL}}(\theta) &= \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_{\theta}} \left[R(x, y) \right. \\ &\quad \left. - \lambda D_{\text{KL}}(\pi_{\theta}(\cdot | x) \parallel \pi_{\text{sft}}(\cdot | x)) \right],\end{aligned}\quad (2)$$

where $\lambda > 0$ is the KL regularization coefficient. The KL anchor at π_{sft} keeps the optimized policy close to the SFT initialization. This stage is critical for incentivizing advanced reasoning capabilities in LLMs, aiming to achieve superior performance that transcends the inherent limitations of static SFT datasets.

3.2 A Variational Bridge Between SFT and RL

A critical limitation of current post-training pipelines (Vattikonda et al., 2025; Kang et al., 2025; Ma et al., 2025) is the structural mismatch between the SFT outcome and the RL prerequisite. Standard SFT optimizes maximum likelihood, which

inherently drives the policy toward a highly peaked, degenerate distribution around the expert demonstrations. However, the subsequent RL phase relies on a KL divergence penalty $D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{sft}})$ to regularize exploration. If the anchor π_{sft} is overly collapsed, this KL penalty heavily penalizes any exploratory deviation, severely limiting the RL optimization.

Consequently, the SFT objective must not be treated as an isolated imitation task. Instead, it should be instantiated as an initialization that is structurally aligned with the KL-regularized form of the subsequent RL stage. We start by analyzing the target destination. In KL-regularized RL (Eq. 2), the inverse of the KL coefficient ($1/\lambda$) acts as an inverse temperature, dictating the strength of reward maximization. Viewing the entire pipeline as a unified process to reach an ultimate equilibrium, we govern its overarching goal with an *aggregate inverse temperature* $\bar{\eta}$. Following the autoregressive factorization with base prior $\pi_{\text{base}, t}$ and expert indicator $r_t(y) := \mathbb{I}[y = y_t^*]$, we seek a conditional distribution $q_t \in \Delta(\mathcal{V})$ that balances the expected reward against a KL penalty. This balance is formalized by the aggregate per-token variational free energy:

$$\mathcal{F}_{\bar{\eta}}(q_t) = \bar{\eta} \mathbb{E}_{y \sim q_t} [r_t(y)] - D_{\text{KL}}(q_t \parallel \pi_{\text{base}, t}). \quad (3)$$

We establish the mathematical bridge across this divide by decomposing the overarching objective. For any intermediate inverse temperature $0 \leq \beta \leq \bar{\eta}$, it yields an exact additive identity:

$$\mathcal{F}_{\bar{\eta}}(q_t) = \mathcal{F}_{\beta}(\pi_{\text{sft}}^*) + \mathcal{F}_{\bar{\eta}-\beta}^{(2)}(q_t; \pi_{\text{sft}}^*), \quad (4)$$

where π_{sft}^* is the optimal target at β , and the residual free energy is:

$$\mathcal{F}_{\bar{\eta}-\beta}^{(2)}(q_t; \pi_{\text{sft}}^*) := (\bar{\eta} - \beta) \mathbb{E}_{q_t} [r_t] - D_{\text{KL}}(q_t \parallel \pi_{\text{sft}}^*). \quad (5)$$

This decomposition resolves the structural disconnect. The residual term structurally mirrors the RL objective anchored at π_{sft}^* , operating under the remaining inverse temperature $(\bar{\eta} - \beta)$. Consequently, the first term $\mathcal{F}_{\beta}(\pi_{\text{sft}}^*)$ dictates that the SFT stage should not blindly maximize likelihood, but serve as a finite-temperature target that absorbs a portion (β) of the alignment pressure without triggering the prior collapse that limits RL. The exact algebraic validity of this two-stage additive decomposition is formally proven in Appendix C.

3.3 Deriving the SFT Target

Guided by the decomposition above, the SFT stage must act to maximize the stage-one objective $\mathcal{F}_\beta(q_t) = \beta \mathbb{E}_{q_t}[r_t] - D_{\text{KL}}(q_t \| \pi_{\text{base},t})$. The variational problem $\max_{q_t \in \Delta(\mathcal{V})} \mathcal{F}_\beta(q_t)$ admits a closed-form solution. By Donsker–Varadhan duality (Donsker and Varadhan, 1975), the supremum equals the log partition function:

$$\begin{aligned} \sup_{q_t \in \Delta(\mathcal{V})} \mathcal{F}_\beta(q_t) &= \log Z_t(\beta), \\ \text{where } Z_t(\beta) &:= \mathbb{E}_{\pi_{\text{base},t}}[e^{\beta r_t}], \end{aligned} \quad (6)$$

and is attained at the closed-form maximizer:

$$\pi_{\text{sft}}^*(y_t | x, y_{<t}^*) = \frac{1}{Z_t(\beta)} \pi_{\text{base},t}(y_t) e^{\beta r_t(y_t)}, \quad (7)$$

which we adopt as the formal SFT target. Substituting the indicator form of r_t , Eq. 7 reduces to the explicit two-case expression:

$$\pi_{\text{sft}}^*(y | x, y_{<t}^*) = \begin{cases} \frac{e^\beta \pi_{\text{base},t}(y_t^*)}{Z_t(\beta)} & y = y_t^*, \\ \frac{\pi_{\text{base},t}(y)}{Z_t(\beta)} & y \neq y_t^*, \end{cases} \quad (8)$$

with normalizer $Z_t(\beta) = (e^\beta - 1) \pi_{\text{base},t}(y_t^*) + 1$. On the logits level, this simply amounts to adding a constant bias β at the demonstration’s choice and re-normalizing.

This formulation fundamentally clarifies the failure mode of standard SFT. Taking the limit $\beta \rightarrow \infty$ sends the mass at y_t^* to one:

$$\lim_{\beta \rightarrow \infty} \pi_{\text{sft}}^*(\cdot | x, y_{<t}^*) = \delta_{y_t^*}, \quad (9)$$

recovering standard SFT as a degenerate zero-temperature limit. Crucially, as $\pi_{\text{sft}}^* \rightarrow \delta_{y_t^*}$, the residual penalty $D_{\text{KL}}(q_t \| \pi_{\text{sft}}^*)$ in Eq. 5 diverges for any q_t that places mass off y_t^* . The residual feasible region contracts strictly to $\{\delta_{y_t^*}\}$, restricting the search space available for the subsequent RL stage to refine the policy. Finite β prevents this premature policy collapse, maintaining support across the vocabulary and ensuring a smooth allocation of the aggregate inverse temperature. In practice, we treat β as a hyperparameter rather than tying it strictly to $\bar{\eta}$, accounting for trust-region clipping in modern RL (examined empirically in Section 4.4).

3.4 Proposed Algorithm: GIFT

Minimizing the KL divergence to our variational target π_{sft}^* (Eq. 7) is mathematically equivalent to

Algorithm 1 GIFT: Gibbs Initialization with Finite Temperature

Require: Data $\mathcal{D}$, base model π_{base} , inverse temperature β , learning rate α

- 1: Initialize $\pi_\theta \leftarrow \pi_{\text{base}}$
 - 2: **while** not converged **do**
 - 3: Sample $(x, y^*) \sim \mathcal{D}$; obtain base log-probs $\ell_{t,k} = \log \pi_{\text{base}}(k | x, y_{<t}^*)$
 - 4: Form GIFT target logits $\hat{z}_{t,k} = \ell_{t,k} + \beta \mathbb{I}[k = y_t^*]$, then $\pi_{\text{sft},t}^* = \text{Softmax}(\hat{z}_t)$
 - 5: $\mathcal{L} \leftarrow - \sum_{t,k} \pi_{\text{sft},t}^*(k) \log \pi_\theta(k | x, y_{<t}^*)$
 - 6: $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$
 - 7: **end while**
 - 8: **return** π_θ
-

optimizing the cross-entropy loss:

$$\begin{aligned} \mathcal{L}(\theta) &= - \mathbb{E}_{(x, y^*) \sim \mathcal{D}} \left[\sum_{t=1}^{|y^*|} \sum_{y_t \in \mathcal{V}} \pi_{\text{sft}}^*(y_t | x, y_{<t}^*) \log \pi_\theta(y_t | x, y_{<t}^*) \right]. \end{aligned} \quad (10)$$

The complete training procedure is outlined in Algorithm 1. Unlike standard soft-targets, GIFT employs an exponential tilt, $q(y) \propto \pi_{\text{base}}(y) \exp(\beta \mathbb{I}[y = y^*])$, which preserves the relative odds among non-oracle tokens (detailed in Appendix G).

4 Experiments

4.1 Experimental setup

Datasets. We utilize **DeepMath-103k** (He et al., 2025), a large-scale mathematical dataset specifically curated for high-difficulty reasoning tasks. It undergoes a rigorous decontamination process against multiple public benchmarks and provides verifiable ground-truth answers to facilitate rule-based rewards. Each problem is accompanied by high-quality solutions generated by DeepSeek-R1 (DeepSeek-AI et al., 2025a). From the full DeepMath-103k corpus, we randomly sample three disjoint subsets for our experiments: 10,000 samples for SFT, 10,000 samples for RL, and 1,000 samples as a validation set to monitor convergence during the SFT stage.

Evaluation. To comprehensively assess the reasoning capabilities, we employ a suite of 6 widely adopted benchmarks: mathematical reasoning

benchmarks **GSM8K** (Cobbe et al., 2021) for grade-school math; **Math500** (Hendrycks et al., 2021), a rigorous subset of the MATH dataset; **AIME24** & **AIME25** to test performance on high-difficulty mathematics competitions; and **HumanEval-plus**, **MBPP-plus** (Liu et al., 2023) for code reasoning. Furthermore, to verify generalizability, we evaluate on 4 OOD benchmarks: **GPQA** (Rein et al., 2023) for graduate-level reasoning, **MMLU-Pro** (Wang et al., 2024) and **MMLU-Redux** (Gema et al., 2025) for massive multitask language understanding, **ARC-Challenge** (Clark et al., 2018) for common-sense reasoning. We also validate GIFT on open-ended generation tasks (**MT-Bench** (Bai et al., 2024)), see Appendix E. All evaluations are conducted by VeRL (Sheng et al., 2024) with the vLLM backend. We set the maximum response length to 8,192 tokens and temperature $T = 0.6$ to evaluate the robustness of the policy. We report pass@1 performance for all benchmarks.

Baselines and Implementation. We evaluate various initialization strategies on two backbone models, Qwen2.5-7B and Llama-3.1-8B, categorizing baselines into three groups: (1) **Direct SFT/RL**, which includes standard SFT on the base model and applying RL directly without prior fine-tuning; (2) **Unified Paradigms**, featuring recent methods like LUFFY (Yan et al., 2025) and ReLIFT (Ma et al., 2025); and (3) **SFT-then-RL**, where we compare GIFT against Standard SFT and explicit ablations including SFT+Entropy (entropy regularization). To rigorously isolate the effect of our reward-calibrated Gibbs re-weighting, we also evaluate structurally similar target-softening approaches—SFT+Label Smoothing and SFT+KD—with their hyperparameters tuned to optimal capacity (see Appendix G). We also compare with recent SFT methods including DFT (Wu et al., 2025), ASFT (Zhu et al., 2025a), and PSFT (Zhu et al., 2025b). Notably, we omit the results of Unified Paradigms on Llama-3.1-8B due to their consistently poor performance on this weaker backbone. In the subsequent RL stage, we uniformly employ GRPO (Shao et al., 2024a) via the VeRL framework. Training is configured with a group size of $G = 8$, a learning rate of 1×10^{-6} , a PPO clip ratio of 0.2, and a maximum length of 8,192 tokens for long chain-of-thought reasoning. SFT is trained for 6 epochs and the subsequent RL stage runs for 1 epoch on $8 \times$ NVIDIA H200 GPUs. Detailed hyperparameters and settings are provided in Appendix H.

4.2 Main Results

We demonstrate the effectiveness of GIFT initialization from three aspects: significant improvements in reasoning tasks, robust OOD generalization, and superior exploration potential as evidenced by pass@k scaling prior to the RL training.

Reasoning Performance. As shown in Table 1, GIFT consistently achieves the highest average performance across all initialization strategies on both Qwen2.5-7B and Llama-3.1-8B. On Qwen2.5-7B, GIFT attains an average pass@1 of 59.55% (+4.49% over SFT), surpassing not only strong SFT variants like PSFT (56.33%) but also complex unified paradigms such as LUFFY (56.69%). Notably, on the challenging AIME benchmark, GIFT yields a substantial gain of 10% over Standard SFT (13.33% $\rightarrow$ 23.33%). On Llama-3.1-8B, GIFT similarly leads with an average of 39.07%, outperforming Standard SFT (34.12%) and PSFT (35.47%) by clear margins. In contrast, while simple regularization via SFT + Entropy yields moderate improvements, baselines such as DFT and ASFT exhibit severe degradation. Notably, even with the hyperparameter sweep in Appendix G, the best-tuned KD trails GIFT by 4.10%/4.88% on Qwen2.5-7B/Llama-3.1-8B, while Label Smoothing strictly degrades performance. This isolates the effect of GIFT’s reward-calibrated anchor, which explicitly boosts the oracle token while preserving the base model’s relative probability mass.

Generalization Capabilities. Table 2 details the OOD performance across different initialization strategies. GIFT demonstrates robust generalization capabilities, effectively mitigating the catastrophic forgetting of general knowledge typically induced by the rigid zero-temperature collapse of Standard SFT and other variants. On Qwen2.5-7B, GIFT achieves an average score of 64.10%, clearly outperforming Standard SFT (59.78%) and on par with complex Unified Paradigms like ReLIFT (64.55%). On Llama-3.1-8B, GIFT secures the highest average performance (55.24%), surpassing all baselines. These results confirm that by maintaining a finite temperature, GIFT successfully retains the structural priors essential for broad generalization, whereas standard zero-temperature SFT is prone to overfitting.

Exploration Potential. Pre-RL pass@k evaluation (Table 3) shows that GIFT scales more favorably with the sample budget. On Qwen2.5-7B, standard SFT holds a minor pass@1 advantage, but GIFT

Table 1: Average pass@1 performance of different initialization strategies for subsequent RL on six reasoning benchmarks. The best result in each column is bold.

Method Class	Strategy	GSM8K	AIME24	AIME25	MATH500	HE ⁺	MBPP ⁺	Average
<i>Qwen2.5-7B</i>								
<i>Direct SFT/RL</i>	Direct SFT	90.25	15.00	16.66	74.30	73.20	61.90	55.22
	Direct RL	90.50	12.22	6.67	77.13	73.20	63.80	53.92
<i>Unified Paradigms</i>	LUFFY	92.70	15.00	15.83	80.20	73.20	63.20	56.69
	ReLIFT	91.84	14.06	17.50	79.72	72.60	61.90	56.27
<i>SFT-then-RL</i>	SFT	91.79	13.33	14.44	78.67	72.00	60.10	55.06
	SFT + Entropy	90.90	18.33	11.67	79.90	73.80	61.90	56.08
	SFT + Label Smoothing	90.14	13.33	10.00	73.60	72.00	59.00	53.01
	SFT + KD	91.85	13.33	15.56	78.83	72.60	60.50	55.45
	DFT	86.50	6.67	8.89	65.87	73.20	64.30	50.91
	ASFT	90.65	12.22	7.78	74.07	73.80	63.00	53.59
	PSFT	91.94	15.83	16.67	81.33	73.20	59.00	56.33
	GIFT	92.06	23.33	17.78	82.00	76.20	65.90	59.55
<i>Llama3.1-8B</i>								
<i>Direct SFT/RL</i>	Direct SFT	75.59	6.67	0.00	39.50	39.00	43.10	33.98
	Direct RL	49.18	0.00	3.33	22.27	34.10	48.70	26.26
<i>SFT-then-RL</i>	SFT	73.39	0.00	3.33	39.20	43.30	45.50	34.12
	SFT + Entropy	70.24	6.67	3.33	36.30	40.90	48.40	34.31
	SFT + Label Smoothing	67.90	0.00	0.00	33.40	36.60	47.10	30.83
	SFT + KD	73.48	0.00	3.33	39.52	43.00	45.80	34.19
	DFT	64.75	0.00	3.33	29.47	40.90	49.50	31.33
	ASFT	60.78	3.33	0.00	24.33	39.60	47.10	29.19
	PSFT	74.53	3.33	6.67	39.40	44.50	44.40	35.47
	GIFT	81.33	6.67	6.67	44.45	44.50	50.80	39.07

extends the lead to +5.16% by pass@8. On Llama-3.1-8B—which exhibits more rapid SFT overfitting—GIFT closes its initial deficit and performs on par with SFT at pass@8. While Llama’s pre-RL pass@ k saturates early, the broader output distribution retained by GIFT becomes useful during subsequent RL (Table 1), where it supports the downstream gains. Detailed breakdowns are in Appendix F.

4.3 Geometric and Distributional Consistency

To validate whether the finite-temperature initialization fosters a more continuous and stable optimization trajectory during post-training, we analyze the geometric and distributional properties of the model updates. We evaluate the model’s representational and distributional shifts across two sequential stages: (1) the transition from the base model to the initialized policy ($Base \rightarrow SFT$), and (2) the subsequent evolution from the initialized policy to the final RL model ($SFT \rightarrow RL$). We quantify geometric consistency using cosine similarity and L2 distance of the last transformer layer, while measuring distributional consistency via KL divergence and Top- K token overlap (estimated on the held-out training subset). While smaller representational shifts are not unconditionally beneficial in

post-training, they serve as useful proxy metrics in our context to indicate the preservation of structural priors.

Geometric Consistency. As reported in Table 4, GIFT significantly enhances geometric consistency throughout the post-training process. GIFT consistently exhibits higher cosine similarity and lower L2 distance compared to standard SFT across stages and models. This indicates that GIFT effectively absorbs task-specific supervision while adhering to the base model’s semantic manifold, reducing the destructive parameter updates caused by standard SFT, thereby providing a more efficient exploration for the subsequent RL stage.

Distributional Consistency. As shown in Table 4, GIFT reduces KL divergence across the entire post-training pipeline in most cases. Complementing this, the top- k token overlap analysis in Figure 2 reveals that GIFT mitigates the drastic distributional shifts characteristic of standard SFT. Moreover, we can observe that the preservation of base priors during SFT stage is effectively leveraged in the subsequent RL stage, manifesting as higher top- k overlap that facilitates more efficient exploration. These proxy metrics suggest that GIFT mitigates drastic distributional shifts, maintaining optimization coherence without forcing a rigid collapse.

Table 2: Average pass@1 performance of different initialization strategies for subsequent RL on four out-of-distribution benchmarks. The best result in each column is bold.

Method Class	Strategy	GPQA	MMLU-Redux	ARC-Challenge	MMLU-Pro	Average
<i>Qwen2.5-7B</i>						
<i>Direct SFT/RL</i>	Direct SFT	19.70	69.03	82.68	50.23	55.41
	Direct RL	32.32	70.50	73.85	54.33	57.75
<i>Unified Paradigms</i>	LUFFY	34.85	73.30	89.85	59.08	64.27
	ReLIFT	37.88	73.67	89.85	56.82	64.55
<i>SFT-then-RL</i>	SFT	27.78	71.13	85.54	54.67	59.78
	SFT + Entropy	33.33	72.20	87.29	55.60	62.11
	DFT	17.17	57.77	84.85	39.75	49.89
	ASFT	25.25	68.27	86.90	49.98	57.60
	PSFT	34.34	73.79	87.77	59.05	63.74
	GIFT	39.09	73.27	88.01	56.01	64.10
<i>Llama3.1-8B</i>						
<i>Direct SFT/RL</i>	Direct SFT	21.21	61.50	77.82	40.05	50.14
	Direct RL	23.74	52.13	64.59	29.65	42.53
<i>SFT-then-RL</i>	SFT	23.74	55.97	74.49	33.67	46.97
	SFT + Entropy	27.78	63.37	80.25	41.87	53.32
	DFT	24.24	54.27	71.46	35.59	46.39
	ASFT	13.13	47.23	66.42	29.40	39.05
	PSFT	12.12	44.67	46.54	30.14	33.37
	GIFT	30.81	63.17	82.94	44.05	55.24

Table 3: Average pass@ k performance (%) on four mathematical reasoning benchmarks in SFT stage prior to RL.

Model	Strategy	$k=1$	$k=2$	$k=4$	$k=8$
Llama-3.1-8B	SFT	27.83	34.35	39.67	44.25
	GIFT	27.60	34.53	39.62	44.87
Qwen2.5-7B	SFT	46.64	52.16	55.90	58.02
	GIFT	46.06	53.92	58.48	63.18

Table 4: Consistency Metrics across Training Stages.

Strategy	Stage 1 Base $\rightarrow$ SFT			Stage 2 SFT $\rightarrow$ RL		
	Cos $\uparrow$	L2 $\downarrow$	KL $\downarrow$	Cos $\uparrow$	L2 $\downarrow$	KL $\downarrow$
<i>Llama-3.1-8B</i>						
SFT	0.8112	87.33	0.5893	0.9310	53.85	0.2316
GIFT	0.8207	85.05	0.5909	0.9422	49.08	0.2010
<i>Qwen2.5-7B</i>						
SFT	0.9201	117.18	0.3995	0.9764	62.61	0.0426
GIFT	0.9253	113.95	0.3907	0.9869	47.34	0.0327

4.4 Analysis of Inverse Temperature

The inverse temperature β regulates the trade-off between expert imitation and prior preservation. As shown in Figure 1, post-RL performance consistently peaks at a finite β . This optimal value correlates closely with a backbone’s resilience to SFT overfitting: Llama-3.1-8B overfits early (Figure 1a) and peaks at a lower $\beta \approx 10$, whereas Qwen2.5-7B maintains a stable validation loss (Figure 1b) and

Table 5: Post-RL residual drift $\hat{\eta}_{\text{final}}$ evaluated on a held-out training subset. All policies underwent the same RL phase but differ in their stage-one initialization (Standard SFT vs. GIFT at varying β).

Model	SFT	$\beta=5$	$\beta=10$	$\beta=15$	$\beta=20$	$\beta=25$
Qwen2.5-7B	0.030	—	0.014	0.013	0.009	0.019
Llama-3.1-8B	0.044	0.059	0.016	0.047	0.040	—

tolerates a higher $\beta \approx 20$ before degrading toward the standard SFT baseline ($\beta \rightarrow \infty$).

To understand how the alignment effort is distributed across the two stages, we estimate the total parameter drift from the base model using $\hat{\eta}_{\text{final}}(\beta)$, defined as the within-prompt difference in length-normalized log-likelihood ratios—relative to the base prior—between successful and failed rollouts on a held-out training subset. As shown in Table 5, this cumulative drift exhibits a distinct U-shape, reaching its minimum near the same β that maximizes downstream accuracy. This reflects the joint optimization dynamics of the pipeline. When β is excessively large, the SFT stage aggressively overfits the demonstrations, introducing a large initial drift into the model. The RL stage inherits this collapsed state, keeping the final cumulative drift artificially inflated. Conversely, when β is too small, the SFT stage offers insufficient guid-

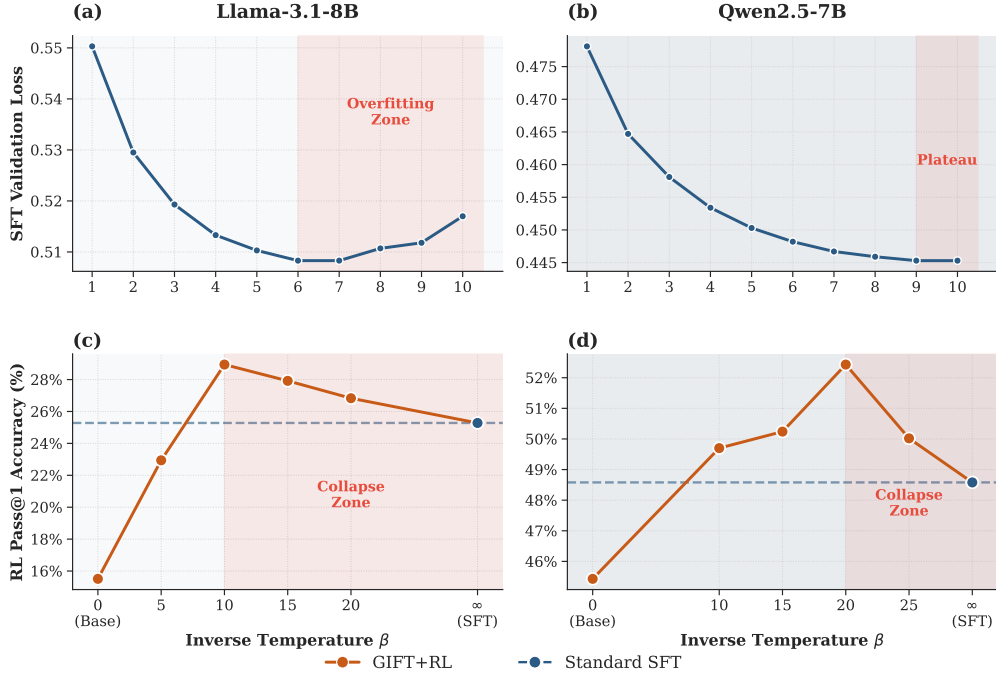


Figure 1: **Correlation between SFT validation dynamics and optimal RL initialization.** (a) and (b) show the SFT validation loss per epoch for Llama-3.1-8B and Qwen2.5-7B, respectively. (c) and (d) illustrate their corresponding post-RL pass@1 accuracy on mathematical reasoning benchmarks across different inverse temperatures β .

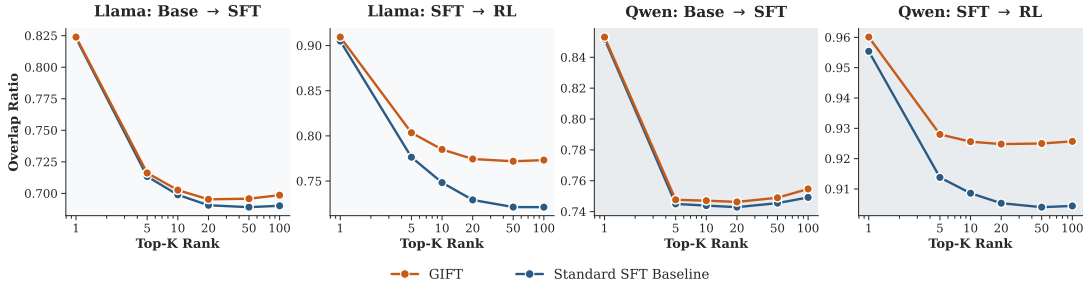


Figure 2: **Top-K Token Overlap Analysis.** Results for Llama-3.1-8B (Left) and Qwen2.5-7B (Right). GIFT exhibits consistently higher token overlap than the SFT baseline.

ance, forcing the RL stage to absorb the alignment burden and independently drive up the drift. The minimum of $\hat{\eta}_{\text{final}}$ thus marks a state of maximum efficiency, where the SFT target better conditions the policy without breaking the underlying prior.

Consequently, the ideal β admits a two-step localization: the onset of SFT validation overfitting (Figure 1a,b) provides a pre-RL upper bound that already excludes the high- β regime, and within this range $\hat{\eta}_{\text{final}}$ further pinpoints the operating point where alignment effort is most evenly absorbed across the two stages. In our main experiments, we fix $\beta = 20$ for Qwen2.5-7B and $\beta = 10$ for Llama-3.1-8B based on a small pilot sweep on each backbone (Figure 1c,d), and keep these values identical across all reported benchmarks.

5 Conclusion

In this work, we identify the zero-temperature distributional collapse induced by standard SFT and propose GIFT as a principled initialization for subsequent RL. Experiments validate its effectiveness across reasoning and out-of-distribution tasks. By reconciling the traditional two-stage paradigm, our work offers a theoretically motivated pathway toward constructing robust and generalizable reasoning systems.

6 Limitations

While our framework demonstrates strong performance on models such as Llama-3.1-8B and Qwen2.5-7B, evaluating its scaling properties

on massive-scale LLMs is currently constrained by substantial computational requirements. Our current formulation is particularly well suited to sparse, verifiable-reward post-training, where demonstrations provide high-quality trajectories and the subsequent RL stage benefits from retaining base-model alternatives. Furthermore, our empirical validation primarily focuses on standard benchmarks within specific linguistic boundaries. Extending this policy optimization approach to highly specialized domains or comprehensive multilingual settings remains a natural avenue for future work.

Acknowledgments

This work is supported by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM113), National Natural Science Foundation of China (92470121, 62402016), National Key R&D Program of China (2024YFA1014003), Zhongguancun Academy (C20250204, C20250602), Beijing Major Science and Technology Project (Z251100008125043, Z251100008425023), and High-performance Computing Platform of Peking University.

References

- Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. 2024. [Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7421–7454. Association for Computational Linguistics.
- Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. 2025a. [Retaining by doing: The role of on-policy data in mitigating forgetting](#). *Preprint*, arXiv:2510.18874.
- Huayu Chen, Kaiwen Zheng, Qinsheng Zhang, Ganqu Cui, Yin Cui, Haotian Ye, Tsung-Yi Lin, Ming-Yu Liu, Jun Zhu, and Haoxiang Wang. 2025b. [Bridging supervised learning and reinforcement learning in math reasoning](#). *Preprint*, arXiv:2505.18116.
- Jack Chen, Fazhong Liu, Naruto Liu, Yuhan Luo, Erqu Qin, Harry Zheng, Tian Dong, Haojin Zhu, Yan Meng, and Xiao Wang. 2025c. [Step-wise adaptive integration of supervised fine-tuning and reinforcement learning for task-specific llms](#). *Preprint*, arXiv:2505.13026.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416.
- 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](#). *Preprint*, arXiv:1803.05457.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Monroe D Donsker and SR Srinivasa Varadhan. 1975. [Asymptotic evaluation of certain markov process expectations for large time, i](#). *Communications on Pure and Applied Mathematics*, 28(1):1–47.
- Yuhao Du, Zhuo Li, Pengyu Cheng, Zhihong Chen, Yuejiao Xie, Xiang Wan, and Anningzhe Gao. 2025. [Simplify rlhf as reward-weighted sft: A variational method](#). *Preprint*, arXiv:2502.11026.
- Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. [Are we done with mmlu?](#) *Preprint*, arXiv:2406.04127.
- Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. [The false promise of imitating proprietary llms](#). *Preprint*, arXiv:2305.15717.
- Zhiwei He, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xingyu Chen, Yue Wang, Linfeng Song, Dian Yu, Zhenwen Liang, Wenxuan Wang, and 1 others. 2025. Deepmath-103k: A large-scale, challenging, decontaminated, and verifiable mathematical dataset for advancing reasoning. *arXiv preprint arXiv:2504.11456*.

- 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](#). *Preprint*, arXiv:2103.03874.
- Ailin Huang, Ang Li, Aobo Kong, Bin Wang, Binxing Jiao, Bo Dong, Bojun Wang, Boyu Chen, Brian Li, Buyun Ma, and 1 others. 2026. Step 3.5 flash: Open frontier-level intelligence with 11b active parameters. *arXiv preprint arXiv:2602.10604*.
- Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. 2025. [Rethinking entropy regularization in large reasoning models](#). *Preprint*, arXiv:2509.25133.
- Feiyang Kang, Michael Kuchnik, Karthik Padthe, Marin Vlastelica, Ruoxi Jia, Carole-Jean Wu, and Newsha Ardalani. 2025. [Quagmires in sft-rl post-training: When high sft scores mislead and what to use instead](#). *Preprint*, arXiv:2510.01624.
- Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. [Fine-tuning can distort pretrained features and underperform out-of-distribution](#). *Preprint*, arXiv:2202.10054.
- Gaotang Li, Ruizhong Qiu, Xiusi Chen, Heng Ji, and Hanghang Tong. 2025. [Beyond log likelihood: Probability-based objectives for supervised fine-tuning across the model capability continuum](#). *Preprint*, arXiv:2510.00526.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. [Is your code generated by chat-GPT really correct? rigorous evaluation of large language models for code generation](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. 2025a. [Uft: Unifying supervised and reinforcement fine-tuning](#). *Preprint*, arXiv:2505.16984.
- Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025b. [Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy](#). *Preprint*, arXiv:2506.13284.
- Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Bin Cui, and Wentao Zhang. 2025. [Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions](#). *Preprint*, arXiv:2506.07527.
- Rui Ming, Haoyuan Wu, Shoubo Hu, Zhuolun He, and Bei Yu. 2025. [One-token rollout: Guiding supervised fine-tuning of llms with policy gradient](#). *Preprint*, arXiv:2509.26313.
- Xueyan Niu, Bo Bai, Wei Han, and Weixi Zhang. 2026. [On the non-decoupling of supervised fine-tuning and reinforcement learning in post-training](#). *Preprint*, arXiv:2601.07389.
- Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, and 64 others. 2024. [Nemotron-4 340b technical report](#). *Preprint*, arXiv:2406.11704.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. [Direct preference optimization: Your language model is secretly a reward model](#). *Preprint*, arXiv:2305.18290.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. [Gpqa: A graduate-level google-proof q&a benchmark](#). *Preprint*, arXiv:2311.12022.
- Zhiwen Ruan, Yixia Li, He Zhu, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. 2025. [Enhancing large language model reasoning via selective critical token fine-tuning](#). *Preprint*, arXiv:2510.10974.
- 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, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Dheeraj Vattikonda, Santhoshi Ravichandran, Emiliano Penaloza, Hadi Nekoei, Megh Thakkar, Thibault Le Sellier de Chezelles, Nicolas Gontier, Miguel Muñoz-Mármol, Sahar Omid Shayan, Stefania Raimondo, Xue Liu, Alexandre Drouin, Laurent Charlin, Alexandre Piché, Alexandre Lacoste, and Massimo Caccia. 2025. [How to train your llm web agent: A statistical diagnosis](#). *Preprint*, arXiv:2507.04103.
- Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. 2025. [Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning](#). *Preprint*, arXiv:2506.01939.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. [Mmlu-pro: A more robust and challenging multi-task language understanding benchmark](#). *Preprint*, arXiv:2406.01574.
- 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](#). *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](#). *Preprint*, arXiv:2505.09388.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. [Dapo: An open-source llm reinforcement learning system at scale](#). *Preprint*, arXiv:2503.14476.
- Shiyue Zhang, Shijie Wu, Ozan Irsoy, Steven Lu, Mohit Bansal, Mark Dredze, and David Rosenberg. 2024. [Mixce: Training autoregressive language models by mixing forward and reverse cross-entropies](#). *Preprint*, arXiv:2305.16958.
- 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.
- He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. 2025a. [Anchored supervised fine-tuning](#). *Preprint*, arXiv:2509.23753.
- Wenhong Zhu, Ruobing Xie, Rui Wang, Xingwu Sun, Di Wang, and Pengfei Liu. 2025b. [Proximal supervised fine-tuning](#). *Preprint*, arXiv:2508.17784.

A Closed-Form Reference for KL-Regularized RL

We record here the standard variational closed form of the KL-regularized RL maximizer used as a reference throughout the main text. For an arbitrary anchor π_{ref} and KL coefficient $\lambda > 0$, the objective $J(\pi) = \mathbb{E}[R(x, y) - \lambda D_{\text{KL}}(\pi(\cdot|x) \parallel \pi_{\text{ref}}(\cdot|x))]$ is uniquely maximized at the Gibbs distribution

$$\pi^*(y|x) \propto \pi_{\text{ref}}(y|x) e^{R(x,y)/\lambda}, \quad (11)$$

with inverse temperature $1/\lambda$ on the reward. We invoke this expression at two points in the main text: (i) Eq. 2, with $\pi_{\text{ref}} = \pi_{\text{sft}}$, where it characterizes the equilibrium of the standard RLVR stage; and (ii) Section 3.2, where the residual term in Eq. 5 is read as a KL-regularized RL problem anchored at π_{sft}^* with $\lambda = 1/(\bar{\eta} - \beta)$; the corresponding identification of $\bar{\eta} - \beta$ as the inverse temperature contributed by the RL stage is made formal in Appendix C. While the above characterizes the sequence-level RL equilibrium, the GIFT target itself is derived natively at the token level from a per-token variational free energy, as detailed next in Appendix B.

B Derivation of the GIFT Token-Level Target

This section provides the detailed derivation of the closed-form maximizer of the per-token variational free energy used in Section 3.3. Throughout, we drop the conditioning on $(x, y_{<t}^*)$ and write $\pi_{\text{base}}(\cdot)$ for $\pi_{\text{base},t}(\cdot)$, $r(y) := r_t(y) = \mathbb{I}[y = y_t^*]$ for the per-token oracle indicator carried by the SFT demonstration (x, y^*) , and q for the candidate token-level distribution. The variational problem is

$$\max_{q \in \Delta(\mathcal{V})} \mathcal{F}_\beta(q) = \beta \mathbb{E}_q[r] - D_{\text{KL}}(q \parallel \pi_{\text{base}}), \quad (12)$$

with $\Delta(\mathcal{V})$ the simplex over the vocabulary $\mathcal{V}$.

Remark on the indicator form of r . Setting $r(y) = \mathbb{I}[y = y_t^*]$ is the formal encoding of the supervision actually carried by an SFT pair (x, y^*) : at position t , only the demonstration token is identified, with no token-level credit available for any alternative $y \neq y_t^*$. The derivation below applies verbatim to an arbitrary bounded $r : \mathcal{V} \rightarrow \mathbb{R}$, in which case $\pi_{\text{sft}}^*(y) \propto \pi_{\text{base}}(y) e^{\beta r(y)}$ is the corresponding maximizer; the indicator specialization is a structural necessity strictly imposed by standard SFT datasets: because the data only provides the single demonstrated sequence, we inherently

lack the explicit supervisory signal to evaluate any alternative tokens.

Donsker–Varadhan route. The Donsker–Varadhan variational characterization of the log-partition function states that

$$\log \mathbb{E}_{\pi_{\text{base}}} [e^{\beta r}] = \sup_{q \in \Delta(\mathcal{V})} \mathcal{F}_\beta(q), \quad (13)$$

with the supremum attained at the Gibbs distribution

$$\begin{aligned} \pi_{\text{sft}}^*(y) &= \frac{1}{Z_t(\beta)} \pi_{\text{base}}(y) e^{\beta r(y)}, \\ Z_t(\beta) &= \mathbb{E}_{\pi_{\text{base}}} [e^{\beta r}]. \end{aligned} \quad (14)$$

Lagrangian route. The same maximizer follows from a direct Lagrangian argument. Introducing a multiplier ν for the normalization constraint $\sum_y q(y) = 1$, the Lagrangian is

$$\begin{aligned} \mathcal{L}(q, \nu) &= \beta \sum_y q(y) r(y) - \sum_y q(y) \log \frac{q(y)}{\pi_{\text{base}}(y)} \\ &\quad - \nu \left(\sum_y q(y) - 1 \right). \end{aligned}$$

Differentiating with respect to $q(y)$ and setting the result to zero yields

$$\beta r(y) - \log q(y) + \log \pi_{\text{base}}(y) - 1 - \nu = 0,$$

so $q(y) \propto \pi_{\text{base}}(y) e^{\beta r(y)}$, which after normalization recovers Eq. 14. Strict concavity of $-D_{\text{KL}}(\cdot \parallel \pi_{\text{base}})$ on the relative interior of $\Delta(\mathcal{V})$, together with the linearity of $\beta \mathbb{E}_q[r]$, ensures that the maximizer is unique whenever π_{base} has full support on $\mathcal{V}$.

Specialization to the per-token oracle. Substituting $r(y) = \mathbb{I}[y = y_t^*]$, the partition function simplifies to

$$Z_t(\beta) = (e^\beta - 1) \pi_{\text{base}}(y_t^*) + 1,$$

and π_{sft}^* admits the explicit two-case form

$$\pi_{\text{sft}}^*(y) = \begin{cases} \frac{e^\beta \pi_{\text{base}}(y_t^*)}{Z_t(\beta)} & \text{if } y = y_t^*, \\ \frac{\pi_{\text{base}}(y)}{Z_t(\beta)} & \text{if } y \neq y_t^*, \end{cases}$$

matching Eq. 8 of the main text. At the level of logits, this is equivalent to adding a constant β to the base logit at y_t^* and re-applying softmax, which is exactly the target constructed in Algorithm 1. The cross-entropy loss in Algorithm 1 therefore fits $\pi_\theta(\cdot \mid x, y_{<t}^*)$ to the closed-form maximizer of Eq. 12.

The $\beta \rightarrow \infty$ limit. For finite β , π_{sft}^* retains support on the entire vocabulary $\mathcal{V}$. As $\beta \rightarrow \infty$, $\pi_{\text{sft}}^*(y_t^*) = e^\beta \pi_{\text{base}}(y_t^*) / Z_t(\beta) \rightarrow 1$ and $\pi_{\text{sft}}^*(y) \rightarrow 0$ pointwise for every $y \neq y_t^*$, so $\pi_{\text{sft}}^* \rightarrow \delta_{y_t^*}$ in distribution (Eq. 9). One-hot SFT thus appears as the degenerate zero-temperature limit of the GIFT target, rather than as a separate construction.

C Two-Stage Decomposition Identity

We establish the additive identity stated in Eq. 4 of the main text. Adopting the notation of Appendix B and writing π_{sft}^* for the token-level target in Eq. 14, we show that for any $0 \leq \beta \leq \bar{\eta}$ and any $q \in \Delta(\mathcal{V})$,

$$\mathcal{F}_{\bar{\eta}}(q) = \mathcal{F}_\beta(\pi_{\text{sft}}^*) + \mathcal{F}_{\bar{\eta}-\beta}^{(2)}(q; \pi_{\text{sft}}^*), \quad (15)$$

where $\mathcal{F}_\beta(q) = \beta \mathbb{E}_q[r] - D_{\text{KL}}(q \| \pi_{\text{base}})$ and $\mathcal{F}_{\bar{\eta}-\beta}^{(2)}(q; \pi_{\text{sft}}^*) = (\bar{\eta} - \beta) \mathbb{E}_q[r] - D_{\text{KL}}(q \| \pi_{\text{sft}}^*)$.

Algebraic derivation. Using the closed form $\pi_{\text{sft}}^*(y) = \pi_{\text{base}}(y) e^{\beta r(y)} / Z_t(\beta)$, we expand the residual KL term:

$$\begin{aligned} D_{\text{KL}}(q \| \pi_{\text{sft}}^*) &= \mathbb{E}_q \left[\log \frac{q(y)}{\pi_{\text{sft}}^*(y)} \right] \\ &= \mathbb{E}_q[\log q(y)] - \mathbb{E}_q[\log \pi_{\text{base}}(y)] \\ &\quad - \beta \mathbb{E}_q[r(y)] + \log Z_t(\beta) \\ &= D_{\text{KL}}(q \| \pi_{\text{base}}) - \beta \mathbb{E}_q[r] + \log Z_t(\beta). \end{aligned}$$

Substituting into $\mathcal{F}_{\bar{\eta}-\beta}^{(2)}(q; \pi_{\text{sft}}^*)$ and rearranging,

$$\begin{aligned} \mathcal{F}_{\bar{\eta}-\beta}^{(2)}(q; \pi_{\text{sft}}^*) &= (\bar{\eta} - \beta) \mathbb{E}_q[r] - D_{\text{KL}}(q \| \pi_{\text{sft}}^*) \\ &= \bar{\eta} \mathbb{E}_q[r] - D_{\text{KL}}(q \| \pi_{\text{base}}) - \log Z_t(\beta) \\ &= \mathcal{F}_{\bar{\eta}}(q) - \log Z_t(\beta). \end{aligned}$$

By the Donsker–Varadhan duality (Appendix B), $\log Z_t(\beta) = \mathcal{F}_\beta(\pi_{\text{sft}}^*)$, since the supremum of $\mathcal{F}_\beta$ is attained at π_{sft}^* . Substituting back yields Eq. 15.

Two-stage suprema match. Maximizing both sides of Eq. 15 over q , the left side gives $\log Z_t(\bar{\eta})$ by Donsker–Varadhan applied at inverse temperature $\bar{\eta}$. On the right side, $\mathcal{F}_\beta(\pi_{\text{sft}}^*)$ is constant in q and equals $\log Z_t(\beta)$, while $\sup_q \mathcal{F}_{\bar{\eta}-\beta}^{(2)}(q; \pi_{\text{sft}}^*) = \log \mathbb{E}_{\pi_{\text{sft}}^*}[e^{(\bar{\eta}-\beta)r}]$ by the same duality applied to the residual problem with reference π_{sft}^* . A direct computation using $\pi_{\text{sft}}^*(y) e^{(\bar{\eta}-\beta)r(y)} = \pi_{\text{base}}(y) e^{\bar{\eta}r(y)} / Z_t(\beta)$ gives

$$\log \mathbb{E}_{\pi_{\text{sft}}^*}[e^{(\bar{\eta}-\beta)r}] = \log Z_t(\bar{\eta}) - \log Z_t(\beta),$$

so the two suprema agree: the SFT-then-RL pipeline (stage one maximizing $\mathcal{F}_\beta$, stage two maximizing $\mathcal{F}_{\bar{\eta}-\beta}^{(2)}$ with reference π_{sft}^*) attains the same value as the single-stage problem at the aggregate inverse temperature $\bar{\eta}$.

Structural correspondence with KL-regularized

RL. Dividing the residual term in Eq. 5 by $(\bar{\eta} - \beta) > 0$ does not change its argmax over q , so $\mathcal{F}_{\bar{\eta}-\beta}^{(2)}(q; \pi_{\text{sft}}^*)$ shares its maximizer with the per-token KL-regularized objective

$$\mathbb{E}_q[r] - \frac{1}{\bar{\eta} - \beta} D_{\text{KL}}(q \| \pi_{\text{sft}}^*).$$

Identifying this with the standard form of Appendix A gives the structural correspondence $\lambda = 1/(\bar{\eta} - \beta)$, so $\bar{\eta} - \beta$ plays the role of the inverse temperature contributed by the RL stage.

We emphasize that this is a structural correspondence in form: the decomposition does not require the trajectory-level RL reward $R(x, y)$ in Eq. 2 to coincide with the per-token oracle indicator r used here. Any token-level credit assignment of R (e.g., per-token GRPO advantages) realizes the residual stage in the same form, with the empirical analog of $\bar{\eta} - \beta$ shaped by the actual KL coefficient and the implicit trust-region effects of clipping and finite-step updates. We accordingly treat $\bar{\eta}$ as a structural surrogate that organizes the two stages, and select β as a hyperparameter rather than tying it to a closed-form expression of λ (Section 4.4).

Degenerate limit $\beta \rightarrow \infty$. As $\beta \rightarrow \infty$, $\pi_{\text{sft}}^* \rightarrow \delta_{y_t^*}$ (Appendix B). The residual term $D_{\text{KL}}(q \| \pi_{\text{sft}}^*)$ in Eq. 5 then diverges for every q that places nonzero mass off y_t^* , so the residual feasible region on which $\mathcal{F}_{\bar{\eta}-\beta}^{(2)}$ is finite contracts strictly to $\{\delta_{y_t^*}\}$. The residual stage is then unable to refine the policy in any meaningful direction, recovering the prior collapse mode discussed in Section 3.2. For finite β , $D_{\text{KL}}(q \| \pi_{\text{sft}}^*)$ remains finite for every q supported on $\mathcal{V}$, leaving the residual stage well-posed and providing room for the RL stage to absorb the remaining inverse temperature $\bar{\eta} - \beta$.

D Training Dynamics.

We report the accuracy on mathematical reasoning benchmarks during the RL training in Figure 3. GIFT initialization demonstrates superior sample efficiency and asymptotic performance compared to standard SFT initialization. On Llama-3.1-8B, GIFT maintains a consistent performance advantage over the SFT baseline throughout the entire

optimization process. Meanwhile, on Qwen2.5-7B, we observe a crossover phenomenon: although GIFT begins with slightly lower performance, it rapidly overtakes the SFT baseline within the first 25% steps. This trajectory indicates that GIFT effectively avoids the premature convergence often induced by rigid supervision, trading off negligible initial exploitation for substantially greater exploration potential and superior long-term gains.

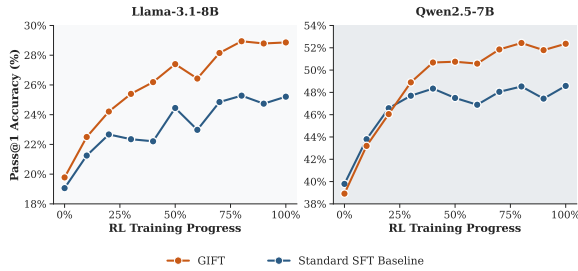


Figure 3: Average pass@1 accuracy across the RL training progress.

E Generalizability on Open-ended Generation Tasks

While our primary evaluation focuses on verifiable mathematics and reasoning, we extend our experiments to the **MT-Bench** benchmark (Bai et al., 2024) to verify whether the token-level GIFT target maintains generalizability in unconstrained, open-ended conversational settings. As shown in Table 7, GIFT consistently achieves the highest overall scores (*Total*) across both Qwen2.5-7B and Llama-3.1-8B models. Specifically, it excels in categories such as STEM and Extraction, while avoiding the severe performance degradation often observed in standard SFT variants (e.g., SFT+KD or Label Smoothing) when faced with diverse user prompts. These results demonstrate that GIFT successfully preserves the generative priors of the base model, ensuring robust open-ended dialogue capabilities while simultaneously achieving peak reasoning potential.

F Detailed Performance Prior to RL Stage

Table 6 details the pass@ k performance across individual mathematical reasoning benchmarks. We observe a clear trade-off between strict mode-seeking and exploration: while standard SFT remains highly competitive at $k = 1$ due to its rigid optimization of the target trajectories, it exhibits

severe exploration bottlenecks at larger k values. In contrast, GIFT exhibits robust scaling behavior. As the sample budget scales to $k = 8$, GIFT consistently overtakes the baseline, confirming that a finite-temperature initialization successfully maintains the diversity necessary for broader exploration.

G Hyperparameter Sweep for Baselines

GIFT belongs to the broader family of base-anchored soft-target objectives, but differs in target geometry. Knowledge distillation and label smoothing form convex mixtures $q = (1 - \alpha)\mathbb{I}[y = y_t^*] + \alpha p$ with p a base or uniform distribution, whereas GIFT applies an exponential tilt $q \propto \pi_{\text{base}}(y) e^{\beta \mathbb{I}[y=y_t^*]}$ that preserves the base relative probabilities among non-oracle tokens while multiplying the oracle-token odds by e^β . This form is matched to the Gibbs solution of the KL-regularized surrogate, and the sweep below confirms that the gains arise from this geometry rather than generic target softening.

To isolate the structural advantages of GIFT and address the concern that its gains might simply stem from generic target softening, we conduct a comprehensive hyperparameter sweep for Knowledge Distillation (KD) and Label Smoothing (LS) on both Qwen2.5-7B and Llama-3.1-8B. Formally, these baselines modify the standard one-hot target into the following convex mixtures:

$$q^{\text{KD}}(y) = (1 - \alpha)\mathbb{I}[y = y_t^*] + \alpha\pi_{\text{base}}(y), \quad (16)$$

$$q^{\text{LS}}(y) = (1 - \mu)\mathbb{I}[y = y_t^*] + \frac{\mu}{|\mathcal{V}|}, \quad (17)$$

where $\mathcal{V}$ is the vocabulary. We vary the distillation weight $\alpha \in \{0.01, 0.03, 0.05, 0.1\}$ and the smoothing parameter $\mu \in \{0.01, 0.03, 0.05, 0.1\}$. The results are summarized in Table 8.

H Implementation Details

In this section, we provide the detailed experimental configuration to facilitate reproducibility. All experiments are conducted using the VeRL framework (Sheng et al., 2024) on a cluster of 8×NVIDIA H200 GPUs.

H.1 Supervised Fine-Tuning

Training Configuration. We optimize the base models using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.01. The global batch size is set to 128. We employ a constant

Table 6: Detailed pass@ k performance (%) on individual mathematical reasoning benchmarks prior to RL.

Benchmark	Strategy	Llama-3.1-8B				Qwen2.5-7B			
		$k = 1$	$k = 2$	$k = 4$	$k = 8$	$k = 1$	$k = 2$	$k = 4$	$k = 8$
GSM8K	SFT	73.84	86.20	92.97	96.11	90.14	94.58	96.35	97.35
	GIFT	73.36	86.65	92.60	96.35	89.94	94.71	96.73	97.80
AIME 24	SFT	0.97	1.94	3.89	7.78	11.25	17.02	21.86	23.33
	GIFT	0.83	1.94	3.33	8.33	10.00	19.88	24.29	30.00
AIME 25	SFT	0.55	1.11	2.22	4.44	11.25	14.88	17.62	20.00
	GIFT	0.42	1.11	2.50	5.00	10.00	18.21	24.71	33.33
MATH-500	SFT	35.94	48.15	59.58	68.67	73.92	82.14	87.77	91.40
	GIFT	35.80	48.42	60.06	69.80	74.30	82.86	88.19	91.60

Table 7: Comprehensive performance on the MT-Bench benchmark across various conversational and reasoning categories.

Strategy	Total	Writing	Roleplay	Reasoning	Math	Coding	Extract	STEM	Humanities
<i>Qwen2.5-7B</i>									
SFT	5.79	4.0	7.7	4.1	5.3	5.3	5.0	6.7	8.2
SFT + Entropy	6.12	4.4	7.7	6.7	4.7	5.0	5.7	7.0	7.8
SFT + Label Smoothing	6.00	4.7	6.6	4.0	6.4	5.2	5.6	6.8	8.7
SFT + KD	6.12	5.5	7.2	4.5	5.8	5.7	5.6	7.1	7.6
DFT	5.60	4.1	5.8	4.0	6.3	5.1	5.8	5.9	7.8
ASFT	5.85	3.7	6.9	4.3	6.1	5.6	5.6	6.5	8.1
PSFT	5.64	3.7	6.7	4.1	4.1	5.7	5.2	7.0	8.6
GIFT	6.34	5.2	6.9	4.8	6.4	5.2	6.5	7.6	8.1
<i>Llama3.1-8B</i>									
SFT	4.46	4.9	5.4	3.5	2.5	3.4	3.8	4.7	7.5
SFT + Entropy	3.94	4.7	4.3	2.8	2.4	2.5	3.7	4.6	6.5
SFT + Label Smoothing	3.99	5.0	4.5	3.0	2.8	2.1	4.5	3.8	6.2
SFT + KD	3.30	2.8	3.3	2.2	4.0	3.3	3.0	4.1	3.7
DFT	5.05	5.0	5.9	3.5	4.2	3.6	5.5	5.6	7.1
ASFT	4.89	5.2	5.6	3.8	3.6	4.4	5.2	4.5	6.8
PSFT	4.19	5.3	4.4	3.0	3.7	2.3	2.7	5.6	6.5
GIFT	5.17	6.4	5.9	4.3	4.1	3.3	4.2	6.0	7.2

learning rate of 1×10^{-5} with no warmup steps. To accommodate long chain-of-thought reasoning, we set the maximum sequence length to 8,192 tokens. Following standard practices for generative fine-tuning, we apply right-side truncation for input sequences. For the explicit SFT ablation variants, we set the regularization coefficient to $\mu = 0.01$ for entropy regularization. For all other baseline methods, we strictly adopt the default hyperparameter settings recommended in their respective original papers to ensure a fair comparison. To initialize the subsequent RL stage, we uniformly select the SFT checkpoint from the 6th epoch across all backbones. As analyzed in Section 4.3 (Figure 1), Llama-3.1-8B reaches its lowest validation loss at this epoch before overfitting. For Qwen2.5-7B, the performance differences after epoch 6 are marginal. Thus, we uniformly adopt the 6th epoch to establish a standardized setting and avoid post-hoc selection bias based on RL downstream dynamics. Finally,

to ensure a fair comparison, the direct SFT baseline is trained on a combined dataset comprising both SFT and RL data to match the total data size configuration.

H.2 Reinforcement Learning

Following SFT, we apply the Group Relative Policy Optimization (GRPO) algorithm (Shao et al., 2024b) to refine the policy. During the exploration phase, we sample $G = 8$ outputs per prompt with a temperature of $T = 1.0$ to encourage diversity. The policy is optimized using a constant learning rate of 1×10^{-6} and a global batch size of 128 (mini-batch size of 64). We set the clip ratio to 0.2 and disable advantage normalization (norm_adv_by_std=False). To prevent reward hacking, we restrict training to 1 epoch.

Table 8: Hyperparameter sweep for Knowledge Distillation (KD) and Label Smoothing (LS) prior to RL. Standard SFT and GIFT are included as reference bounds. The best result among the swept variants is underlined, while the overall best is bolded.

Method	Hyperparameter	GSM8K	AIME24	AIME25	MATH500	HE ⁺	MBPP ⁺	Average
<i>Qwen2.5-7B</i>								
Standard SFT	-	91.79	13.33	14.44	78.67	72.00	60.10	55.06
SFT + KD (α)	$\alpha = 0.01$	<u>91.85</u>	<u>13.33</u>	<u>15.56</u>	<u>78.83</u>	72.60	60.50	<u>55.45</u>
	$\alpha = 0.03$	91.62	13.33	13.33	78.37	73.00	61.20	55.14
	$\alpha = 0.05$	91.55	10.00	11.11	78.14	73.20	61.80	54.30
	$\alpha = 0.10$	91.51	10.00	10.00	78.00	<u>73.80</u>	<u>62.40</u>	54.29
SFT + LS (μ)	$\mu = 0.01$	<u>90.14</u>	<u>13.33</u>	<u>10.00</u>	<u>73.60</u>	72.00	<u>59.00</u>	<u>53.01</u>
	$\mu = 0.03$	88.47	10.00	8.89	71.23	71.50	<u>58.40</u>	51.42
	$\mu = 0.05$	86.18	6.67	8.89	68.52	70.80	57.20	49.71
	$\mu = 0.10$	82.15	3.33	6.67	62.37	69.50	55.80	46.64
GIFT (Ours)	-	92.06	23.33	17.78	82.00	76.20	65.90	59.55
<i>Llama-3.1-8B</i>								
Standard SFT	-	73.39	0.00	3.33	39.20	43.30	45.50	34.12
SFT + KD (α)	$\alpha = 0.01$	<u>73.48</u>	0.00	<u>3.33</u>	<u>39.52</u>	<u>43.00</u>	45.80	<u>34.19</u>
	$\alpha = 0.03$	70.12	<u>3.33</u>	0.00	35.18	41.50	47.10	32.87
	$\alpha = 0.05$	64.76	<u>3.33</u>	0.00	30.14	39.40	48.20	30.97
	$\alpha = 0.10$	59.28	3.33	0.00	25.00	37.80	<u>49.50</u>	29.15
SFT + LS (μ)	$\mu = 0.01$	<u>67.90</u>	0.00	0.00	<u>33.40</u>	<u>36.60</u>	<u>47.10</u>	<u>30.83</u>
	$\mu = 0.03$	<u>64.23</u>	0.00	0.00	30.08	35.20	45.50	29.17
	$\mu = 0.05$	58.54	0.00	0.00	26.77	33.80	43.20	27.05
	$\mu = 0.10$	49.26	0.00	0.00	20.46	31.00	40.50	23.54
GIFT (Ours)	-	81.33	6.67	6.67	44.45	44.50	50.80	39.07

H.3 Evaluation and Reward Mechanism

Evaluation. We report the average pass@1 accuracy across four runs for all benchmarks. We merge FSDP checkpoints and utilize vLLM for high-throughput inference with a temperature of $T = 0.6$ and a max token limit of 8,192. A single β value is applied uniformly across all evaluated benchmarks without per-task tuning.

Reward Function. We employ a cascaded binary reward function that sequentially utilizes OpenMathInstruct extraction rules and the MathVerify parser. To ensure stability, the verification process is guarded by a 10-second thread-based timeout, assigning a reward of 1.0 if either method validates the solution against the ground truth.

I Use of AI Assistants

We utilized AI assistants exclusively for grammatical polishing and LaTeX formatting assistance.