LEARNABLE REFERENCE FINE-TUNING FOR ROLLOUT-FREE MATHEMATICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning can improve mathematical reasoning in large language models, but large-scale online rollouts are expensive. We study rollout-free fine-tuning from verified solutions alone, where the absence of negative examples leaves the contribution of each demonstration difficult to calibrate. We propose *Learnable Reference Fine-Tuning* (LRFT), which jointly trains a target policy and a reference policy on verified demonstrations. The reference models the demonstration distribution, and target-to-reference likelihood ratios modulate token-level target updates. Unlike target-only self-weighting or fixed-reference regularization, LRFT learns its update scale from the verified data. Across four checkpoints from two Qwen families and five mathematical-reasoning benchmarks, LRFT achieves the highest mean Pass@1 among all evaluated methods for every checkpoint and the best performance in most reported Pass@k comparisons. Relative to DFT, it also shows lower measured KL divergence from the base policy and less polarized gold-token probabilities. These results suggest that a demonstration-fitted reference provides an effective rollout-free learning signal while limiting distributional shift.

1 INTRODUCTION

Mathematical reasoning is a stringent post-training setting: a problem may admit many correct derivations with different intermediate steps and decompositions. An effective method should improve correctness without concentrating excessive probability on the few trajectories observed during training. This matters for sampling-based inference, where preserving probability mass beyond the modal completion can let repeated sampling recover solutions missed by greedy decoding.

Supervised fine-tuning (SFT) and reinforcement learning (RL) are widely used to post-train reasoning models (Wei et al., 2021; Ouyang et al., 2022; Chung et al., 2024). Reward-based RL can substantially improve reasoning, but online rollouts, reward evaluation, and self-play may be prohibitively expensive (Christiano et al., 2017; Schulman et al., 2015; 2017). We instead study the rollout-free regime, which learns from a fixed corpus of verified solutions (Levine et al., 2020). Although attractive when interaction is unavailable or costly, this regime leaves a central ambiguity: every training response is correct, but the data do not indicate how strongly each demonstration should update the target policy.

This ambiguity is consequential for strong pretrained models. Maximum likelihood can overspecialize the policy to the empirical demonstration distribution, while target-probability reweighting makes update magnitudes endogenous to the policy being optimized. Aggressive post-training can also cause substantial distributional drift or degrade capabilities retained by the base checkpoint (Kumar et al., 2022; Luo et al., 2023; Yue et al., 2025). A rollout-free objective should therefore extract a useful signal from positive demonstrations while controlling its effect on the model distribution.

Recent rollout-free methods reweight supervised objectives to approximate sparse-reward RL (Qin and Springenberg, 2025; Wu et al., 2025). Self-weighted methods derive update scales from the target policy; reference-based methods instead use a fixed or periodically updated policy as a stabilizing anchor (Zhu et al., 2025a; Zhu et al., 2025b). Neither explicitly learns a demonstration-conditioned update scale: self-weighting remains coupled to the target, whereas fixed or lagged references mainly impose proximity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose *Learnable Reference Fine-Tuning* (LRFT), which jointly trains a target policy and a learnable reference on verified demonstrations. As shown in Figure 1, the reference fits the positive demonstrations, and its detached likelihood forms a target-to-reference ratio that modulates target updates. The reference is thus not merely a fixed KL anchor; it provides a data-adaptive likelihood scale learned from the same verified support.

Because both policies share an initialization, deterministic joint training may keep their trajectories tightly coupled. LRFT instead mixes their token-level objectives using independently sampled Beta coefficients. This token-wise stochasticity gives the policies distinct aggregate update weights without extra labels, rollouts, or inference-time cost. In our ablation, stochastic mixing outperforms fixed, decaying, and two-stage alternatives in mean Pass@1.

Across four checkpoints, LRFT achieves the highest five-benchmark mean Pass@1 among all compared methods. It is also strongest in most Pass@k comparisons, while remaining substantially closer than DFT to the base policy in measured KL divergence and gold-token probability profiles (Yue et al., 2025; Gao et al., 2022).

Our contributions are:

- We introduce LRFT, a rollout-free objective that fits a reference to verified solutions and uses target-to-reference likelihood ratios to modulate target updates.
- We propose token-wise Beta mixing to induce distinct stochastic update patterns despite shared initialization.
- Across four Qwen checkpoints and five mathematical-reasoning benchmarks, LRFT achieves the highest mean Pass@1 for every checkpoint among the evaluated methods.

2 Related Work

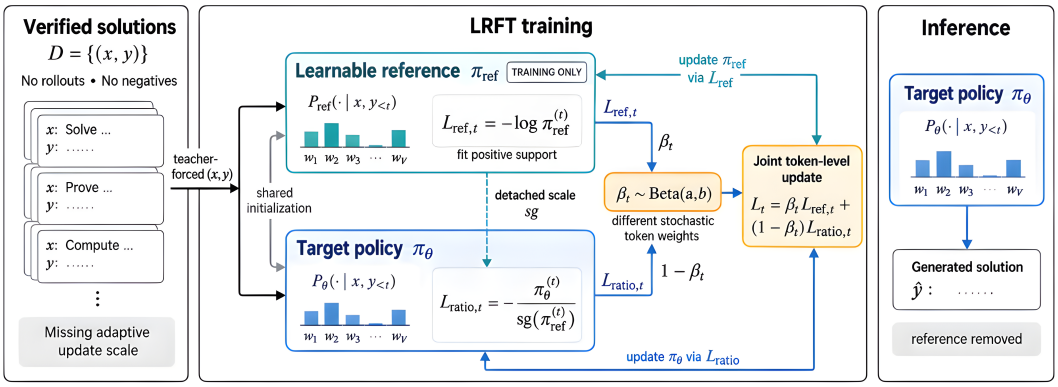

Figure 1: Overview of Learnable Reference Fine-Tuning (LRFT). Given verified solution pairs (x, y) without rollouts or negative examples, LRFT jointly trains a target policy π_θ and a learnable reference policy π_{ref} from a shared initialization. The reference policy is fitted to the verified demonstrations through a token-level maximum-likelihood objective, while the target policy is optimized using a target-to-reference likelihood-ratio surrogate, with the reference likelihood detached by a stop-gradient operator. To prevent the two policies from evolving synchronously, LRFT samples token-wise mixing coefficients $\beta_t \sim \text{Beta}(a, b)$ and forms a stochastic joint objective. The reference policy is used only during training and is removed at inference, where generation is performed solely by the learned target policy.

Instruction tuning and SFT optimize curated input–output pairs (Wei et al., 2021; Ouyang et al., 2022; Chung et al., 2024), whereas methods such as TRPO and PPO optimize rewards from preferences or verifiers (Christiano et al., 2017; Ouyang et al., 2022; Schulman et al., 2015; 2017). Other approaches include offline preference optimization (e.g., DPO) and rollout-based reasoning methods

such as GRPO and verifier-based RL (Cobbe et al., 2021; Lewkowycz et al., 2022; Lightman et al., 2023; Rafailov et al., 2023; Qwen Team, 2024; Guo et al., 2025). LRFT instead uses fixed verified demonstrations without preference labels or negative examples.

Offline learning from positive demonstrations. The cost of rollout-based RL has motivated offline alternatives. Offline RL learns from static behavior-policy data, using importance weighting to address distribution mismatch and trust regions to stabilize updates (Precup et al., 2000; Sutton et al., 1998; Levine et al., 2020; Schulman et al., 2015; 2017). This perspective motivates LRFT's likelihood ratio. We do not claim that the finite-data objective is an unbiased importance-sampling estimator: the learned reference is not the behavior policy that generated the data. The ratio instead provides a relative, demonstration-fitted update scale.

Recent work views positive-only fine-tuning through sparse-reward RL. Importance-weighted SFT (iw-SFT) derives a lower-bound perspective and reweights under distribution shift (Qin and Springenberg, 2025); Dynamic Fine-Tuning (DFT) modifies the SFT gradient through probability-based self-weighting (Wu et al., 2025). In self-weighted objectives, however, the update scale remains coupled to the target policy.

Reference-based stabilization. Reference-based methods stabilize post-training relative to another policy. Anchored SFT (ASFT) regularizes toward a fixed base policy (Zhu et al., 2025a), while Proximal SFT (PSFT) studies reference construction and update frequency (Zhu et al., 2025b). These fixed, lagged, or externally specified references mainly impose proximity. LRFT instead learns a reference from verified demonstrations and uses its likelihoods to scale target updates. Its optional base-policy KL term remains conceptually separate, retaining the conventional anchoring role (Table 1).

Table 1: Conceptual comparison of reference mechanisms among the closest rollout-free baselines discussed above. DFT uses the target policy as an effective self-scale rather than an explicit reference policy.

Method	Explicit reference?	Source of update scale	Primary role
DFT	No	Target-policy probability	Self-weighting
ASFT	Yes	Fixed base policy	Stable anchor
PSFT	Yes	Lagged or updated policy	Proximity anchor
LRFT	Yes	Demonstration-fitted policy	Target-update scale

3 METHOD

3.1 PROBLEM SETTING AND MOTIVATION

Let $\mathcal{D} = \{(x, y)\}$ contain positive demonstrations whose responses y are verified for inputs x. Training uses no online rollouts, self-play, or learned reward model. We seek to improve the sampling performance of $\pi_\theta(y \mid x)$ without unnecessarily concentrating on observed trajectories.

Without negative examples, the data do not determine how strongly each correct solution should update the policy. SFT fits the empirical distribution, target-probability reweighting makes the scale endogenous, and fixed references do not adapt it to demonstrations. LRFT learns π_{ref} from the positive data and uses it to define a data-adaptive target-update scale.

3.2 FROM REWARD MAXIMIZATION TO RATIO GUIDANCE

Let $r(x, y) \in \{0, 1\}$ denote a verifier-defined reward indicating whether y is correct for x. The target policy maximizes its success probability through the objective

$$J(\theta) = \mathbb{E}_x \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[r(x, y) \right]. \tag{1}$$

The dataset contains neither samples drawn from π_θ nor incorrect responses. Let π_{ref} be a reference policy with adequate support. At the population level, importance weighting gives

$$J(\theta) = \mathbb{E}_x \mathbb{E}_{y \sim \pi_{\text{ref}}(\cdot|x)} \left[r(x, y) \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} \right]. \tag{2}$$

Equation (2) is exact only when samples are drawn from a reference distribution with appropriate support. Because the finite training set is not sampled from the jointly learned reference policy, we use this identity only to motivate a relative likelihood scale, not as an empirical importance-sampling estimator. We fit π_{ref} on $\mathcal{D}$ and apply the resulting scale to the observed demonstrations.

As $\mathcal{D}$ contains only verified solutions, we set $r(x, y) = 1$ for all $(x, y) \in \mathcal{D}$ and define the following reference-guided offline surrogate:

$$\mathcal{L}_{\text{ratio}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\frac{\pi_\theta(y \mid x)}{\text{sg}(\pi_{\text{ref}}(y \mid x))} \right], \tag{3}$$

Training is teacher-forced: both policies score the same gold token y_t given $(x, y_{<t})$. The ratio is therefore a token-level surrogate, not a ratio of independently sampled trajectories. It compares target and reference probabilities for the verified next token. Detaching the denominator lets this comparison affect the target without allowing its loss to update the reference scale directly.

The ratio-guided form in (3) subsumes the gradient structures of familiar rollout-free objectives under particular choices of the denominator. SFT minimizes

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

DFT instead uses the self-weighted objective

$$\mathcal{L}_{\text{DFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}[\text{sg}(\pi_\theta(y \mid x)) \log \pi_\theta(y \mid x)].$$

When the denominator in (3) is fixed to one, its gradient is proportional to the DFT gradient; when it is set to $\text{sg}(\pi_\theta)$, its gradient coincides with the SFT gradient. These correspondences are statements about gradients under stop-gradient semantics rather than pointwise equalities between objective values. The resulting scale is therefore fixed or self-coupled in the two special cases, whereas LRFT learns the scale from verified demonstrations.

3.3 Learnable Reference Policy

Existing rollout-free objectives typically rely on either an implicit, target-derived scale or an externally specified reference. In self-weighted objectives, the target policy itself determines the reweighting signal; in fixed-reference objectives, the reference constrains divergence but does not adapt to the observed demonstrations.

In LRFT, the reference policy π_{ref} is trained to model the likelihood of verified demonstrations in $\mathcal{D}$ via maximum likelihood:

$$\mathcal{L}_{\text{ref}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \log \pi_{\text{ref}}(y \mid x). \tag{4}$$

Maximum likelihood learns a probability scale on the positive support. Unlike a frozen base policy, the reference adapts to regularities and heterogeneity in the demonstrations; unlike self-weighting, it has parameters distinct from the target. Thus, π_{ref} is jointly optimized with, but not defined by, π_θ: the reference tracks the demonstrations, while the target receives ratio-guided updates for downstream generation.

3.4 Joint Learning with a Reference Policy

Equations (4) and (3) assign complementary roles to the two policies: maximum likelihood fits the reference policy to positive demonstrations, whereas the target-policy gradient is modulated by the target-to-reference likelihood ratio.

Under shared initialization, fixed mixture weights can cause the two policies to evolve in near synchrony. For a token position t, define

$$\pi_{\text{ref}}^{(t)} := \pi_{\text{ref}}(y_t \mid x, y_{<t}), \qquad \pi_\theta^{(t)} := \pi_\theta(y_t \mid x, y_{<t}). \tag{5}$$

Near the shared initialization, $\pi_\theta^{(t)} \approx \pi_{\text{ref}}^{(t)}$, and the corresponding per-token gradients are nearly aligned:

$$\nabla \mathcal{L}_{\text{ref}} = \nabla \left(-\log \pi_{\text{ref}}^{(t)} \right) = -\frac{1}{\pi_{\text{ref}}^{(t)}} \nabla \pi_{\text{ref}}^{(t)},$$
$$\nabla \mathcal{L}_{\text{ratio}} = \nabla \left(-\frac{\pi_\theta^{(t)}}{\text{sg}(\pi_{\text{ref}}^{(t)})} \right) = -\frac{1}{\text{sg}(\pi_{\text{ref}}^{(t)})} \nabla \pi_\theta^{(t)}. \tag{6}$$

If the policies have identical parameters and optimizer states, this alignment may keep π_θ close to π_{ref} and thereby attenuate the relative-ratio signal.

3.5 Final Objective

To induce distinct update patterns, LRFT randomizes the mixture of the two losses at the token level. For every token position t, we independently sample $\beta_t \sim \text{Beta}(a, b)$ and optimize

$$\mathcal{L} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\sum_{t=1}^{|y|} \left(\beta_t \log \pi_{\text{ref}}^{(t)} + (1-\beta_t) \frac{\pi_\theta^{(t)}}{\text{sg}(\pi_{\text{ref}}^{(t)})} \right) \right]. \tag{7}$$

For implementation, the same first-order gradient can be obtained by writing the objective as a single fraction:

$$\mathcal{L} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\sum_{t=1}^{|y|} \frac{\beta_t \pi_{\text{ref}}^{(t)} + (1-\beta_t)\pi_\theta^{(t)}}{\text{sg}(\pi_{\text{ref}}^{(t)})} \right]. \tag{8}$$

Although (8) and (7) have different values, stop-gradient makes their first-order gradients identical: the two weighted components yield the same $\nabla \log \pi_{\text{ref}}^{(t)}$ and $\nabla \pi_\theta^{(t)} / \text{sg}(\pi_{\text{ref}}^{(t)})$ terms. Standard first-order optimizers therefore produce identical update directions and relative magnitudes.

The Beta mean $a/(a+b)$ controls the expected allocation between reference and target optimization, while concentration $a+b$ controls token-level variability. With the symmetric default $\text{Beta}(10, 10)$, neither policy is favored in expectation, yet their token weights differ within each mini-batch. Fixed mixing repeatedly applies the same deterministic proportion.

For additional stability, we optionally regularize the target policy toward the base policy π_{base}:

$$\mathcal{L}_{\text{KL}} = \lambda \, \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\frac{1}{|y|} \sum_{t=1}^{|y|} D_{\text{KL}} \left(\pi_\theta(\cdot \mid x, y_{<t}) \,\|\, \pi_{\text{base}}(\cdot \mid x, y_{<t}) \right) \right]. \tag{9}$$

The objective requires only teacher-forced passes, without sampled responses, reward-model queries, or preference comparisons. The reference is discarded after training, so inference architecture and cost match the target base model. The optional KL term adds an explicit base-policy anchor but does not replace the reference-guided ratio.

LRFT training step. For each mini-batch of verified demonstrations:
(1) Compute teacher-forced token probabilities under the target and reference policies.
(2) Sample the coefficients in (7) and construct the joint token-level objective.
(3) Add the regularizer in (9), when used.
(4) Back-propagate the resulting loss and jointly update both policies.

4 Experiments

We compare LRFT with SFT, DFT (Wu et al., 2025), and ASFT (Zhu et al., 2025a). We report Pass@1 and Pass@k, and analyze gold-token probabilities, target–reference likelihood ratios, and KL divergence from the base policy (Yue et al., 2025).

Table 2: Mean single-sample Pass@1 accuracy (%) computed from 16 stochastic completions per problem. Boldface denotes the best result within each base-model block.

Base Model	Method	MATH500	Minerva	Olympiad	AIME 2024	AMC 2023	Avg.
Qwen2.5-Math-1.5B	Base	31.66	8.51	15.88	4.16	19.38	15.92
	SFT	43.54	12.31	11.96	0.78	17.58	17.23
	DFT	64.00	23.74	26.47	5.57	38.98	31.75
	ASFT	63.75	24.06	26.73	4.84	38.52	31.58
	LRFT	**67.10**	**25.18**	**27.90**	**7.87**	**40.39**	**33.69**
Qwen2.5-Math-7B	Base	40.09	13.74	17.04	8.75	27.66	21.46
	SFT	52.91	19.13	17.84	2.49	25.04	23.48
	DFT	68.35	34.87	33.81	8.34	44.02	37.88
	ASFT	70.61	30.90	33.46	9.16	45.19	37.86
	LRFT	**75.69**	**37.60**	**37.53**	**16.10**	**56.33**	**44.65**
Qwen3-1.7B	Base	71.25	32.43	30.85	8.02	47.15	37.94
	SFT	36.13	10.59	9.21	0.73	13.40	14.01
	DFT	54.95	19.79	21.31	2.38	31.35	25.96
	ASFT	69.80	28.66	31.54	10.21	43.32	36.71
	LRFT	**74.31**	**33.88**	**35.58**	**15.26**	**51.33**	**42.07**
Qwen3-4B	Base	70.13	34.67	30.25	10.42	43.01	37.70
	SFT	54.05	27.89	19.78	1.45	22.27	25.09
	DFT	66.68	34.98	30.31	7.93	45.74	37.13
	ASFT	**83.91**	44.53	44.23	**24.07**	63.36	52.02
	LRFT	82.19	**47.64**	**44.34**	20.73	**67.30**	**52.44**

4.1 Experimental Setup

Models and data. We evaluate four checkpoints from two Qwen model families spanning several parameter scales: Qwen2.5-Math-1.5B, Qwen2.5-Math-7B (Qwen Team, 2024), Qwen3-1.7B, and Qwen3-4B (Qwen Team, 2025). Following (Wu et al., 2025), we construct the training set by randomly sampling 100,000 examples from NuminaMath-CoT (Li et al., 2024), which contains diverse mathematical problems paired with step-by-step solutions.

Baselines and training protocol. We compare LRFT with the base checkpoint, SFT, DFT (Wu et al., 2025), and ASFT (Zhu et al., 2025a). Our implementation is based on the Verl framework (Sheng et al., 2024). All methods are trained for one epoch using AdamW (Loshchilov and Hutter, 2017) with a global batch size of 256. We use a learning rate of 2×10^{-5} for models with at least 4B parameters and 5×10^{-5} otherwise. Unless stated otherwise, LRFT uses token-wise coefficients sampled from $\mathrm{Beta}(10, 10)$ throughout training. The target and reference policies share the same base initialization; the reference is retained only during training. All main results use $\lambda = 0$ and therefore exclude the optional KL regularizer.

Evaluation. We evaluate on five mathematical-reasoning benchmarks: MATH500 (Hendrycks et al., 2021), AIME 2024 (American Institute of Mathematics, 2024), AMC 2023 (Mathematical Association of America, 2023), Minerva Math (Lewkowycz et al., 2022), and OlympiadBench (He et al., 2024). For Table 2, we estimate Pass@1 using 16 stochastic completions per problem at temperature 1.0. For Figure 3, we separately sample 128 completions per problem on MATH500, Minerva Math, and OlympiadBench, and 1,024 on AIME 2024 and AMC 2023, and compute Pass@k using the standard unbiased estimator. We report k up to 128 on the former benchmarks and up to 1,024 on the latter, using powers of two.

4.2 Main Results

Aggregate performance. Table 2 shows that LRFT has the highest five-benchmark mean for every checkpoint. On Qwen2.5-Math-7B, it reaches 44.65% mean Pass@1, exceeding DFT and ASFT by 6.77 and 6.79 points, respectively. All methods use the same rollout-free positive demonstrations; they differ in how they scale and stabilize the learning signal.

Comparison with fixed-reference anchoring. ASFT is an informative reference-based baseline because it explicitly anchors updates to the base policy. LRFT has a higher mean on all four checkpoints. On Qwen3-4B, ASFT leads on MATH500 and AIME 2024, whereas LRFT leads on the other three benchmarks and by 0.42 points on average. This is consistent with a learned reference providing a task-adaptive positive-update scale beyond restricting drift.

Benchmark-level results. On Qwen2.5-Math-7B, LRFT reaches 16.10% on AIME 2024 versus 9.16% for ASFT, and 37.53% on OlympiadBench versus 33.81% for DFT. These gains span competition-style and broader reasoning benchmarks, although repeated-seed evaluation is needed to assess statistical significance.

Behavior on Qwen3 base checkpoints. On Qwen3-1.7B and Qwen3-4B, SFT lowers mean accuracy from 37.94% to 14.01% and from 37.70% to 25.09%, respectively. DFT also underperforms the Qwen3-1.7B base (25.96% versus 37.94%). LRFT instead exceeds both base models, reaching 42.07% and 52.44%, respectively. Reference-guided scaling may therefore help mitigate over-updating during positive-only fine-tuning.

4.3 Sampling Behavior and Distributional Sharpness

We next examine performance across sampling budgets and quantify the extent to which fine-tuning changes the token distribution relative to the base policy.

Token-level probability profiles. Higher Pass@k can result from a sharper output distribution that concentrates probability on a narrower set of continuations. Figure 2 shows a pronounced U-shaped distribution for DFT, with substantial mass near zero and one. LRFT is visually closer to Base and SFT and assigns more empirical mass to intermediate gold-token probabilities, consistent with less token-level polarization than DFT.

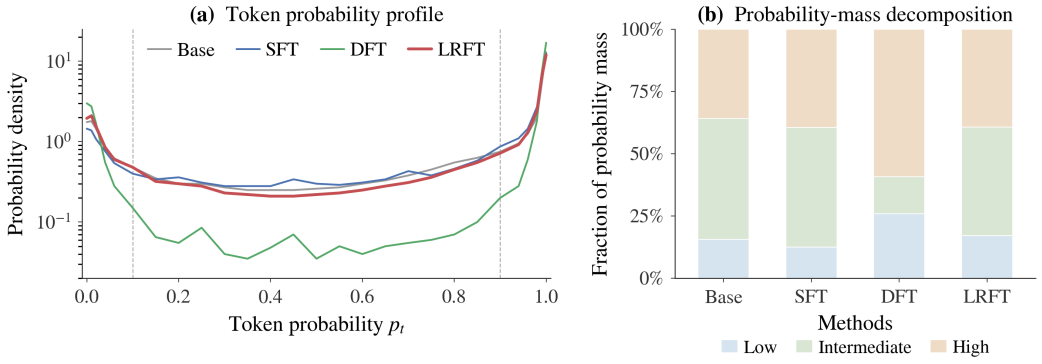

Figure 2: Empirical densities of gold-token probabilities on the training data. DFT concentrates more mass near zero and one, whereas LRFT is visually closer to Base and SFT.

Sampling performance across k. LRFT achieves the highest Pass@k in most plotted benchmark–budget pairs (Figure 3).

Distributional drift and sampling trade-off. The KL measurements complement the task metrics by quantifying deviation from the base policy under the reported evaluation protocol. LRFT has KL values of 0.0107 and 0.0103 in the two reported directions, compared with 0.4754 and 0.2962 for DFT (Table 3). Thus, on this evaluation, LRFT attains its performance gains without the large distributional displacement measured for DFT. The observation should not be interpreted as a universal guarantee of stability; rather, it provides empirical evidence that the learned-reference signal and positive-only objective can coexist with a small measured shift from the base policy.

Target–reference ratio dynamics. Figure 4 visualizes the distribution of $\log(\pi_\theta/\pi_{\text{ref}})$ during LRFT training. The median remains near zero while the distribution broadens. This behavior indi-

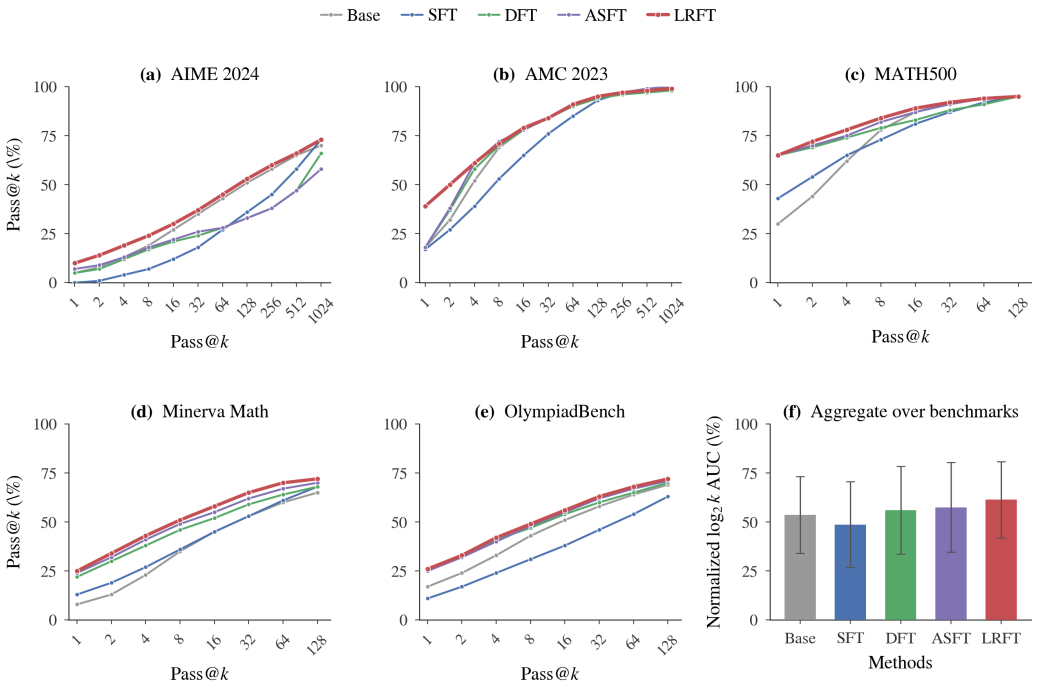

Figure 3: Pass@k on five mathematical-reasoning benchmarks and their mean for Base, SFT, DFT, ASFT, and LRFT.

cates that the target and reference policies do not collapse to an exactly identical likelihood model, yet their relative likelihoods remain centered rather than being dominated by a persistent global bias. Combined with the token-level probability profiles, this analysis provides a distributional view complementary to the accuracy results.

Table 3: Token-level KL divergence from the base policy on Qwen2.5-Math-1.5B, evaluated on teacher-forced training sequences and averaged first over response tokens and then over sequences. Lower values indicate smaller measured distributional shifts.

Method	$D_{\mathrm{KL}}(\pi_{\mathrm{base}} \| \pi)$	$D_{\mathrm{KL}}(\pi \| \pi_{\mathrm{base}})$
SFT	0.0072	0.0074
DFT	0.4754	0.2962
LRFT	0.0107	0.0103

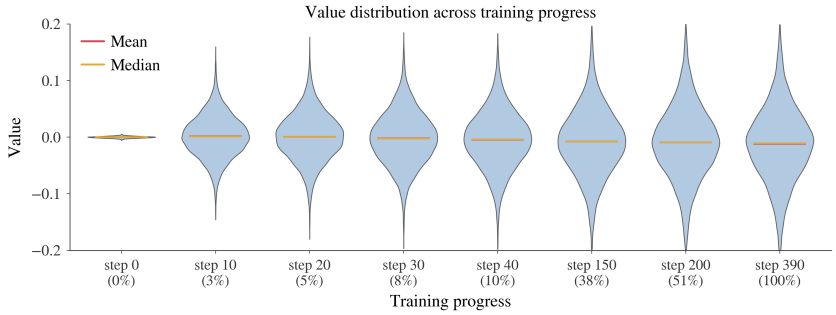

Figure 4: Distribution of $\log(\pi_\theta/\pi_{\mathrm{ref}})$ during training. The median remains near zero while the dispersion increases.

4.4 Ablation on β Mixing Strategies

LRFT uses β_t to stochastically allocate token-level weight between the reference likelihood term and the ratio-guided target term. This design is intended to prevent shared initialization and deterministic weighting from making the two policy updates nearly redundant. Table 4 compares the proposed sampling scheme with deterministic, stage-wise, and constant alternatives on Qwen2.5-Math-1.5B.

The two symmetric Beta choices isolate the effect of concentration while preserving an expected weight of 0.5 for each objective. $\mathrm{Beta}(10, 10)$ introduces more token-level variation than $\mathrm{Beta}(20, 20)$, whereas fixed 0.5 removes stochastic variation altogether. The log-decay, exponential-decay, and two-stage schedules instead impose a global training-time allocation that is shared by all tokens in a step. Consequently, the comparison tests not only how much weight is assigned to each loss, but also whether the allocation should vary across token positions.

Table 4: Ablation of β-mixing strategies on Qwen2.5-Math-1.5B. Values are mean single-sample Pass@1 (%) over 16 stochastic completions per problem; boldface marks the best value in each column.

β Strategy	MATH500	Minerva	Olympiad	AIME 2024	AMC 2023
$\mathrm{B}(10, 10)$	**67.10**	25.18	**27.90**	**7.87**	**40.39**
$\mathrm{B}(20, 20)$	65.13	**26.42**	26.81	7.19	40.16
Log decay	63.90	26.39	26.41	7.55	38.09
Exp decay	63.86	24.71	26.05	7.18	38.75
Two-stage	66.56	24.65	27.73	7.30	36.68
Fixed 0.5	61.96	23.60	23.30	4.84	22.50

Among the evaluated strategies, $\mathrm{Beta}(10, 10)$ achieves the highest mean Pass@1 across the five reported benchmarks, whereas fixed 0.5 mixing performs worst. Relative to fixed mixing, the stochastic strategy improves every reported benchmark in this ablation, with the largest absolute change on AMC 2023 (40.39% versus 22.50%). $\mathrm{Beta}(20, 20)$ is competitive on Minerva Math, but does not match the overall mean of the more variable $\mathrm{Beta}(10, 10)$ allocation. None of the deterministic or stage-wise alternatives matches the stochastic scheme across the full benchmark suite. These results support token-wise stochastic mixing as a useful component of LRFT, while not by themselves establishing a unique causal explanation for its effect.

Conclusion

We presented LRFT, a rollout-free fine-tuning method that learns a reference policy from verified demonstrations and uses target-to-reference likelihood ratios to modulate target-policy updates. Across four Qwen checkpoints and five mathematical-reasoning benchmarks, LRFT achieves the strongest mean Pass@1 among the compared methods while exhibiting less measured distributional drift and gold-token polarization than DFT. Future work should establish whether these properties extend to other model families, domains, and larger-scale training regimes.

Scope and Limitations

Our evaluation is restricted to mathematical reasoning and four checkpoints from two Qwen model families. It remains unknown whether the observed gains transfer to other architectures, model scales, training corpora, or application domains. Pass@k, token-probability profiles, and KL divergence characterize sampling performance and distributional change, but do not directly measure the semantic diversity or faithfulness of valid reasoning trajectories. We report point estimates without confidence intervals or repeated training runs; consequently, small differences should not be over-interpreted. LRFT also maintains two trainable policies and therefore increases memory consumption and optimizer state relative to SFT. Finally, although the mixing ablation is consistent with the utility of stochastic mixing, it does not establish gradient decorrelation as the causal mechanism.

Ethical Considerations

LRFT can reduce the computational burden associated with online-rollout-based post-training, but it does not guarantee reliable reasoning. Language-model-generated mathematical solutions may be fluent while still being incorrect. Outputs should therefore be independently verified before they are used in educational, scientific, financial, or other high-stakes contexts. The method is trained on publicly available mathematical datasets; users remain responsible for complying with the relevant licenses, provenance requirements, and applicable policies.

Reproducibility Statement

Sections 3 and 4 specify the training objective, model checkpoints, training-data source and size, optimizer, batch size, learning rates, benchmarks, decoding temperature, sampling budgets, and mixing schedules. The method is fully specified at the level of teacher-forced token probabilities: the reference is optimized by maximum likelihood, its likelihood is detached in the ratio-guided target loss, and independently sampled Beta coefficients mix the two token-level terms. The main tables and figures report the numerical results and the behavioral analyses used in the paper. Our implementation is based on the Verl framework; subject to organizational review, we will release the implementation, training configurations, prompts, evaluation scripts, and the precise answer-extraction protocol in a public repository upon acceptance.

AI Use Statement

Generative AI tools were used to assist with debugging experimental code, refining the visual presentation of figures and tables, improving the clarity and language of the manuscript, and providing feedback on the presentation and evaluation methodology. All AI-assisted suggestions, code modifications, and manuscript revisions were reviewed and validated by the authors. All experiments were executed by the authors, and all scientific decisions were made by the authors. The authors independently verified the data, analyses, figures, citations, and scientific claims, and take full responsibility for the content of this paper.

References

American Institute of Mathematics. AIME 2024 competition mathematical problems. 2024.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization. *arXiv preprint arXiv:2210.10760*, 2022.

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

He, C., Luo, R., Bai, Y., Hu, S., Leng Thai, Z., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., and Sun, M. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. *arXiv e-prints*, pp. arXiv:2402.14008, 2024.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring Mathematical Problem Solving With the MATH Dataset. *arXiv e-prints*, pp. arXiv:2103.03874, 2021.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

Li, J., Beeching, E., Tunstall, L., Lipkin, B., Soletskyi, R., Huang, S. C., Rasul, K., Yu, L., Jiang, A., Shen, Z., Qin, Z., Dong, B., Zhou, L., Fleureau, Y., Lample, G., and Polu, S. NuminaMath, 2024.

Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., and Cobbe, K. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.

Mathematical Association of America. AMC 2023 competition problems. 2023.

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

Precup, D., Sutton, R. S., and Singh, S. Eligibility traces for off-policy policy evaluation. 2000.

Qin, C. and Springenberg, J. T. Supervised fine tuning on curated data is reinforcement learning (and can be improved). *arXiv preprint arXiv:2507.12856*, 2025.

Qwen Team. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement. *arXiv e-prints*, pp. arXiv:2409.12122, 2024.

Qwen Team. Qwen3 Technical Report. *arXiv e-prints*, pp. arXiv:2505.09388, 2025.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal Policy Optimization Algorithms. *arXiv e-prints*, pp. arXiv:1707.06347, 2017.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024.

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Wu, Y., Zhou, Y., Ziheng, Z., Peng, Y., Ye, X., Hu, X., Zhu, W., Qi, L., Yang, M.-H., and Yang, X. On the Generalization of SFT: A Reinforcement Learning Perspective with Reward Rectification. *arXiv e-prints*, pp. arXiv:2508.05629, 2025.

Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., and Huang, G. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

Zhu, H., Su, J., Lai, P., Ma, R., Zhang, W., Yang, L., and Chen, G. Anchored Supervised Fine-Tuning. *arXiv e-prints*, pp. arXiv:2509.23753, 2025.

Zhu, W., Xie, R., Wang, R., Sun, X., Wang, D., and Liu, P. Proximal supervised fine-tuning. *arXiv preprint arXiv:2508.17784*, 2025.