REFERENCE-AWARE DYNAMIC FINE-TUNING FOR MULTI-SAMPLE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Dynamic Fine-Tuning (DFT) has recently shown that a simple detached reweighting of supervised fine-tuning can substantially improve average sampled accuracy on reasoning tasks. We identify a complementary limitation: the same DFT checkpoints that improve average-sample performance can degrade high-k pass rates, sometimes falling below the base model under multi-sample decoding. We trace this tension to a local distributional effect of the DFT update. By increasing the demonstrated-token probability, DFT also contracts high-probability non-reference alternatives through the softmax; this sharpening is beneficial when alternatives are local noise, but can reduce pass@k when alternatives correspond to viable reasoning continuations. Motivated by this mechanism, we propose Reference-Aware Dynamic Fine-Tuning (RA-DFT), an objective that uses the frozen base model's teacher-forced probability as a local confidence signal. RA-DFT attenuates DFT-style sharpening at low-reference-probability positions, where alternatives are more likely to support multi-sample coverage, while retaining strong sharpening at high-reference-probability positions. The method preserves the standard supervised fine-tuning pipeline and only changes the token-level loss weight. Across five models and six mathematical reasoning benchmarks, RA-DFT consistently improves over DFT across the evaluated pass@k range and improves majority voting, while preserving the average-sample gains that make DFT a strong baseline. Compared with a DFT+KL anchoring baseline, RA-DFT yields a better overall low-k/high-k balance, suggesting an advantage of reference-conditioned selective sharpening over global base anchoring.

1 INTRODUCTION

Large language models (DeepSeek-AI et al., 2025; OpenAI et al., 2023; Yang et al., 2024b; Jaech et al., 2024; Comanici et al., 2025; Team, 2024; Hui et al., 2024; Yang et al., 2024a) have achieved strong performance on instruction following and reasoning after large-scale pretraining and post-training (Chung et al., 2024; Wei et al., 2022; Zhou et al., 2023; Taori et al., 2023; Dubey et al., 2024; Qwen Team et al., 2024; Yang et al., 2025). For reasoning problems, however, model performance is often not determined by a single greedy or sampled completion alone. Modern reasoning systems frequently sample multiple candidate solutions and then aggregate, select, or verify them, as in pass@k evaluation, self-consistency, verifier-based selection, and rejection sampling (Chen et al., 2021; Cobbe et al., 2021; Wang et al., 2023; Lewkowycz et al., 2022; Lightman et al., 2024; Wang et al., 2024). This makes reasoning fine-tuning different from ordinary likelihood maximization: an effective objective should improve the reliability of individual samples while preserving enough distributional coverage for multi-sample decoding to find correct solutions.

Supervised fine-tuning (SFT) remains the standard positive-demonstration interface for adapting language models (Mishra et al., 2022; Zhang et al., 2026; Chung et al., 2024; Zhou et al., 2023). Dynamic Fine-Tuning (DFT) (Wu et al., 2026) recently improved this interface by reweighting each token-level cross-entropy term with the model's current probability of the demonstrated token, detached from the gradient. DFT substantially improves average sampled accuracy over SFT in mathematical reasoning settings (Wu et al., 2026). We therefore treat DFT as a strong reasoning fine-tuning baseline and ask a complementary question: does the sharpening that improves average sampled accuracy also preserve the coverage needed for high-k sampling?

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.

Figure 1 previews our main observation. On the reproduced DFT checkpoints used in this study, DFT improves average@16 and low-k performance, but its pass@64 can drop below the base model across the evaluated model families. This behavior matters because pass@k is the metric underlying many sampling-based reasoning procedures (Chen et al., 2021; Cobbe et al., 2021; Wang et al., 2023). The observation suggests that DFT changes not only the probability of the demonstrated trajectory, but also the shape of the local distribution around it: it makes individual samples more reliable, but may suppress alternatives that would have contributed to multi-sample success.

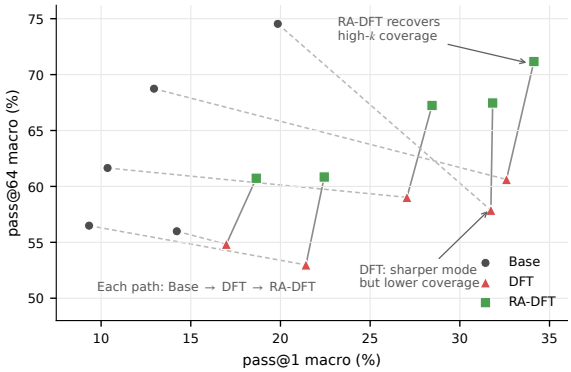


Figure 1: **Pass@1–pass@64 tension.** DFT improves low-k performance but can reduce high-k sampling performance; RA-DFT is designed to retain DFT's low-k benefit while recovering high-k coverage.

We analyze this tension at the token level. At a teacher-forced position, a DFT update increases the demonstrated-token logit, but through the softmax, it also pushes down non-demonstrated logits. A first-order expansion shows that this contraction contains an alternative-specific term whose magnitude grows with the probability of the alternative token. Thus, DFT-style sharpening acts most strongly on high-probability alternatives, precisely the alternatives that dominate sampling when the demonstrated token is not selected. This effect is not uniformly harmful: at some positions, alternatives are local noise, such as wrong digits, formatting artifacts, or tokenization variants. Suppressing such alternatives can improve the reliability of a sampled solution. At other positions, strong alternatives may include viable continuations, such as a different intermediate variable, an equivalent algebraic manipulation, or an alternative phrasing of a correct step. Suppressing these alternatives can reduce the benefit of drawing multiple samples. The issue is therefore not sharpening itself, but *non-selective sharpening*.

This motivates a reference-aware notion of selectivity. We use the frozen base model's teacher-forced probability of the demonstrated token as a local confidence signal. When the base model assigns a low probability to the demonstrated token, the local continuation is less constrained under the base policy, and strong alternatives are more likely to include viable reasoning branches. When the base model assigns a high probability to the demonstrated token, the continuation is more canonical, and alternatives are more likely to be noise. The base model is not used as a correctness oracle or a distillation target; it provides a signal for deciding where DFT-style sharpening should be attenuated.

Based on this principle, we propose **Reference-Aware Dynamic Fine-Tuning** (RA-DFT). RA-DFT preserves the positive-demonstration training interface of SFT and DFT, but replaces DFT's token weight with a reference-aware weight computed from the current model probability and the frozen base probability. The resulting objective remains a single teacher-forced reweighted cross-entropy loss. Unlike a global anchoring baseline such as KL regularization, RA-DFT does not simply constrain the model toward the base distribution everywhere. Instead, it changes where sharpening occurs: attenuate DFT-style updates where the base model is uncertain, and retain DFT-like sharpening where the base model is confident.

We validate RA-DFT with both mechanism diagnostics and benchmark evaluations. On held-out teacher-forced responses, DFT contracts the probability of the strongest non-reference alternative across model families. When positions are bucketed by the base reference probability, RA-DFT restores alternative mass more in low-reference buckets while remaining sharper than DFT+KL in high-reference buckets, supporting the distinction between global anchoring and reference-conditioned selective preservation. On six mathematical reasoning benchmarks, including Math500, Minerva Math, OlympiadBench, AIME 2024/2025, and AMC 2023 (Hendrycks et al., 2021; Lewkowycz et al., 2022; He et al., 2024), RA-DFT consistently improves over DFT across the evaluated pass@k range and also improves majority voting. Compared with DFT+KL, RA-DFT achieves stronger average@16 and pass@1 performance in every model family while maintaining competitive or better high-k performance, indicating that reference-conditioned selective sharpening yields a more favorable balance between low-k accuracy and high-k coverage than global base anchoring.

Our contributions are as follows:

- We identify a pass@1–pass@k tension in Dynamic Fine-Tuning: DFT can improve average sampled accuracy while reducing high-k sampling performance below the base model in our evaluated setting.
- We provide a local distributional analysis showing that DFT-style sharpening suppresses high-probability alternatives through the softmax, offering a mechanism for why average-sample gains need not translate into pass@k gains.
- We propose RA-DFT, a reference-aware reweighting of DFT that uses the frozen base model's teacher-forced probability as a local confidence signal to attenuate sharpening selectively.
- We evaluate RA-DFT across five models and six mathematical reasoning benchmarks, showing that it improves high-k performance over DFT while retaining the average-sample benefits.

2 METHOD

2.1 PRELIMINARIES

We consider the standard supervised fine-tuning setting where each training example provides a prompt and a single reference response. Let $\mathcal{D} = \{(x, y^\star)\}$, where x is a prompt and $y^\star = (y_1^\star, \ldots, y_T^\star)$ is the reference response. For a policy π_θ, define the teacher-forced probability of the reference token at position t as

$$p_t = \pi_\theta(y_t^\star \mid x, y_{<t}^\star). \tag{1}$$

Standard supervised fine-tuning (SFT) minimizes the token-level cross-entropy

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(x,y^\star)\sim\mathcal{D}} \sum_{t=1}^{T} \log p_t. \tag{2}$$

Dynamic Fine-Tuning (DFT) Wu et al. (2026) replaces the uniform token weight in SFT with the current reference-token probability:

$$\mathcal{L}_{\mathrm{DFT}} = -\mathbb{E}_{(x,y^\star)\sim\mathcal{D}} \sum_{t=1}^{T} \mathrm{sg}(p_t) \log p_t, \tag{3}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. Thus, DFT keeps the same supervised data interface as SFT while changing the per-token weighting rule.

2.2 THE PASS@1–PASS@k FAILURE MODE OF DFT

DFT was introduced as a stronger alternative to standard SFT, with substantial gains in average@16 accuracy over SFT Wu et al. (2026). We therefore treat DFT as a strong fine-tuning baseline for reasoning tasks. Our motivation comes from a complementary observation, previewed in Figure 1: the sharpening that improves average sampled accuracy does not necessarily improve high-k sampling. On our reproduced DFT checkpoints, average@16 improves, but pass@64 drops below the base model across the evaluated model families.

This observation suggests that DFT changes not only the probability of the reference trajectory, but also the distribution of nearby continuations. Average sampled accuracy measures the expected correctness of an individual sample averaged over repeated draws. Pass@k, in contrast, depends on whether at least one of k samples solves the problem, and therefore also depends on whether useful alternative continuations remain sampleable. A method can improve average@16 by making the distribution more peaked, yet hurt pass@k if it suppresses alternatives that would have led to correct sampled solutions.

The technical question is:

> *Can we retain DFT's average-sample benefit while avoiding the unnecessary suppression of alternatives needed for high-k decoding?*

2.3 DISTRIBUTIONAL MECHANISM AND DESIGN PRINCIPLE

We now analyze the local mechanism behind the pass@1–pass@k tension of DFT. Our goal here is not to give a full theory of sampled multi-step reasoning, but to isolate what a DFT update does to alternative tokens at a single teacher-forced position and use this to motivate RA-DFT.

Fix a teacher-forced position t, and let $a := y_t^\star \in \mathcal{V}$ be the demonstrated token. For each $j \in \mathcal{V}$, let

$$p_j := \pi_\theta(j \mid x, y_{<t}^\star),$$

so that $p_a = p_t$. Let $z = (z_j)_{j\in\mathcal{V}}$ be the corresponding pre-softmax logits. The DFT token loss at position t is

$$\mathcal{L}_t^{\mathrm{DFT}} = -\,\mathrm{sg}(p_a)\log p_a.$$

The following result, proved in Appendix B.1, characterizes the local logit effect of DFT:

$$\frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_a} = -\,\mathrm{sg}(p_a)(1-p_a), \qquad \frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_v} = \mathrm{sg}(p_a)p_v, \quad v \neq a. \tag{4}$$

Consequently, under a local gradient step on the logits,

$$z^+ = z - \eta \nabla_z \mathcal{L}_t^{\mathrm{DFT}},$$

the logit increments satisfy

$$\Delta z_a = \eta\ \mathrm{sg}(p_a)(1-p_a) \geq 0, \qquad \Delta z_v = -\eta\ \mathrm{sg}(p_a)p_v \leq 0, \quad v \neq a, \tag{5}$$

with strict inequalities whenever $0 < p_a < 1$ and $p_v > 0$. Thus, DFT not only increases the demonstrated-token logit, but also decreases all alternative logits through softmax competition.

We next translate the logit-direction result above into a first-order response of alternative probabilities under the same local logit step. For any alternative token $v \neq a$, let $\Delta p_v := p_v^+ - p_v$. The following proposition gives the induced first-order change; the proof is deferred to Appendix B.2.

Proposition 2.1 (Local response of alternative probabilities under DFT)**.** *For any $v \neq a$,*

$$\Delta p_v = \eta p_a p_v \left(\sum_{j\in\mathcal{V}} p_j^2 - p_a - p_v \right) + O(\eta^2). \tag{6}$$

Equation (6) shows that alternatives with larger p_v receive stronger downward pressure through the $-\eta p_a p_v^2$ term, although the overall first-order change in p_v also depends on the rest of the local distribution through $\sum_{j\in\mathcal{V}} p_j^2$. This suppression can be helpful or harmful depending on which alternatives are being suppressed. Removing noisy alternatives can improve the reliability of the dominant continuation, whereas removing alternatives that could support distinct valid continuations can reduce sampled-solution coverage under high-k decoding. The issue is therefore not sharpening per se, but *non-selective sharpening*.

To decide where DFT-style sharpening should be attenuated, we use the base model's teacher-forced probability as a local reference signal. Define

$$r_t = \pi_{\mathrm{base}}(y_t^\star \mid x, y_{<t}^\star), \tag{7}$$

where r_t is the probability assigned by the frozen base model to the demonstrated token at position t. We use r_t only as a local confidence signal, not as a correctness certificate. When r_t is low, the base model is uncertain about the demonstrated token, so preserving alternatives is more likely to matter for sampled-solution coverage. When r_t is high, the demonstrated token is already a high-confidence continuation under the base model, so stronger sharpening is less likely to remove useful alternatives and more likely to improve the reliability of the dominant continuation. This leads to the design principle behind RA-DFT: *attenuate DFT-style sharpening at low-reference positions, while retaining strong sharpening at high-reference positions*. Moreover, when the reference signal is strong, the update should not become arbitrarily small merely because the current model temporarily assigns low probability to the demonstrated token. This motivates a reference-aware recovery floor in the weight introduced next.

2.4 REFERENCE-AWARE DYNAMIC FINE-TUNING

We propose **Reference-Aware Dynamic Fine-Tuning** (RA-DFT). RA-DFT keeps the same teacher-forced reweighted cross-entropy form as DFT, but replaces the DFT weight $\mathrm{sg}(p_t)$ with a reference-aware token weight:

$$\mathcal{L}_{\text{RA-DFT}}(\theta) = -\mathbb{E}_{(x,y^\star)\sim\mathcal{D}} \sum_{t=1}^{T} \mathrm{sg}(w_t) \log p_t, \tag{8}$$

where

$$w_t = \begin{cases} p_t, & r_t < \varepsilon, \\ r_t^\alpha \max(r_t, p_t), & r_t \geq \varepsilon. \end{cases} \tag{9}$$

Here $\alpha > 0$ controls the strength of reference-aware attenuation, and ε determines when RA-DFT falls back to the DFT weight because the reference signal is too weak to trust; On the supported branch $r_t \geq \varepsilon$, the factor r_t^α modulates the update by reference confidence, assigning a smaller weight to lower r_t positions all else equal, while $\max(r_t, p_t)$ provides a recovery floor when the current model underestimates a reference-supported token. The stop-gradient is applied to the entire weight w_t, so RA-DFT remains a single reweighted cross-entropy objective.

The weight in Equation (9) addresses the DFT failure mode by changing *where* DFT-style sharpening occurs. For non-negligible reference probabilities, the weight has two regimes. If $p_t \geq r_t$, then $w_t = r_t^\alpha p_t$, so the DFT weight p_t is attenuated by the reference-confidence factor r_t^α. Relative to DFT, this reduces the magnitude of the weighted update and hence weakens the DFT-style contraction tendency captured by Equation (6), especially at low-r_t positions where preserving alternatives is more likely to matter for pass@k. If $p_t < r_t$, then $w_t = r_t^{\alpha+1}$. This decouples the weight from the currently underestimated probability p_t and prevents the learning signal from becoming arbitrarily small as $p_t \to 0$ on tokens that the base model considers reliable. In this way, RA-DFT is designed to attenuate the part of DFT-style sharpening that can contribute to excessive alternative collapse while retaining strong updates on reference-supported positions.

2.5 MECHANISM DIAGNOSTICS

We test the mechanism with the teacher-forced strongest alternative probability $a_t(\theta) = \max_{v \neq y_t^\star} \pi_\theta(v \mid x, y_{<t}^\star)$, which targets the alternatives most affected by the contraction term in Equation (6). We report retained mass relative to the base model, $C(\theta) = \mathbb{E}_t[a_t(\theta)]/\mathbb{E}_t[a_t(\theta_{\text{base}})]$. To test selectivity, we compute r_t once with the frozen base model, sort held-out positions by r_t, and split them into ten equal-frequency buckets $B0, \ldots, B9$. We call the low-r_t buckets open because the base model has not committed strongly to the demonstrated token, making alternatives more likely to represent different valid continuations. We call the high-r_t buckets canonical because the demonstrated token is already a high-confidence local continuation, making alternatives more likely to be local noise. For bucket b, we report $C_b(\theta) = \mathbb{E}[a_t(\theta) \mid t \in b]/\mathbb{E}[a_t(\theta_{\text{base}}) \mid t \in b]$. In the Qwen2.5-Math-1.5B diagnostic in Figure 2, B0–B2 are open, while B3+ are canonical because B3 already has median $r_t > 0.9$. This terminology is operational rather than a correctness oracle; Appendix D shows that low-r_t positions are more substitution-tolerant, whereas high-r_t positions are more substitution-sensitive.

Figure 2 supports the two predictions. Panel (a) shows a DFT-specific collapse fingerprint: DFT contracts strongest-alternative mass across all five model families, while SFT does not. Panel (b) shows that RA-DFT's recovery is selective: it preserves alternatives more in open buckets, where multiple valid continuations are plausible, similarly to DFT+KL; but remains much sharper than DFT+KL in canonical buckets, showing that it restores less alternative mass in high-r_t regions, where local substitutions are empirically more harmful in our validation in Appendix D. This is the distributional behavior the method is designed to produce. DFT+KL here acts as a diagnostic comparator for additive base anchoring.

2.6 IMPLICATIONS FOR PASS@1 AND PASS@k

The benchmark behavior is consistent with the mechanism. Pass@1 benefits from sharp canonical positions, because local mistakes and formatting noise are suppressed. Pass@k benefits from preserving alternatives at open positions because multiple samples can explore distinct valid continuations.

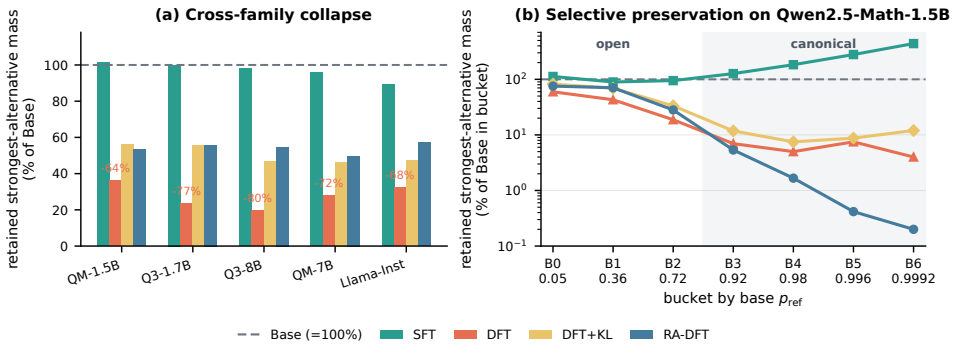


Figure 2: **Mechanism diagnostic for RA-DFT. (a)** Retained strongest alternative mass relative to the base model across five model families. SFT stays close to the base distribution, while DFT retains only 20–36% of the base strongest-alternative mass. DFT+KL and RA-DFT restore part of this mass. **(b)** Bucket-wise retained strongest alternative mass on Qwen2.5-Math-1.5B. Buckets are equal-frequency bins sorted by r_t; the numbers under B0–B6 are median r_t values. RA-DFT tracks DFT+KL more closely in open buckets, but remains substantially lower than DFT+KL in canonical buckets, distinguishing selective preservation from global anchoring. B7–B9 are not plotted individually because they all remain near zero in absolute mass.

DFT improves the former but can harm the latter. DFT+KL restores alternatives more broadly, which helps high-k relative to DFT but sacrifices some low-k sharpness. RA-DFT is designed to occupy the useful middle ground: DFT-like sharpening where the base model is confident, and attenuation of DFT where the base model leaves room for alternatives.

3 Experiments

3.1 Experimental setup

Models and baselines. We train all methods on the same NumiNa-CoT subset with 100000 instances (Li et al., 2024) following DFT Wu et al. (2026) and test five base models: Qwen2.5-Math-1.5B, Qwen2.5-Math-7B (Yang et al., 2024a), Qwen3-1.7B-Base, Qwen3-8B-Base Yang et al. (2025), and Llama-3.1-8B-Instruct (Dubey et al., 2024). We compare the base model, SFT, DFT (Wu et al., 2026), DFT+KL, and our RA-DFT. SFT provides the standard supervised baseline, while DFT is a stronger fine-tuning objective for improving average accuracy in this setting. DFT+KL is included as a natural anchoring baseline: it keeps the DFT loss but adds a KL penalty to the base model distribution, testing whether the pass@k regression can be fixed by a generic global constraint toward the base model. This baseline is important because it separates the effect of simply preserving the base distribution from the reference-conditioned selectivity introduced by RA-DFT. We evaluate on Math500 (Hendrycks et al., 2021), Minerva Math (Lewkowycz et al., 2022), OlympiadBench (He et al., 2024), AIME 2024/2025, and AMC 2023. Experiment details are provided in Appendix E.

3.2 Main results

We first compare average@16 and pass@16 results in Table 1. DFT substantially improves average@16 over SFT, confirming that it is a strong average-accuracy baseline. However, the improvement does not consistently translate into multi-sample success: the DFT rows often show much smaller gains in pass@16 than in average@16, and in several model families, DFT remains below the base model on pass@16 despite large average@16 improvements. RA-DFT outperforms the benchmark-averaged average@16 and pass@16 of DFT in every evaluated model. Specifically, the pass@16 margins over DFT are $+6.26$, $+9.07$, $+6.15$, $+8.78$, and $+4.45$ points on the average column, respectively.

Table 2 extends the comparison from $k = 16$ to the full range of evaluated pass@k. RA-DFT improves over DFT for every model and every $k \in \{1, 2, 4, 8, 16, 32, 64\}$. The improvement is already positive at pass@1, but it becomes more pronounced as k grows. At pass@64, RA-DFT improves over

Table 1: Average@16 and pass@16 results on six mathematical reasoning benchmarks. Each cell reports average@16 / pass@16 in percentage. Boldface marks the best value for each metric within each model family and benchmark.

	Math500	Minerva	Olympiad	AIME24	AIME25	AMC23	Avg.
Qwen2.5-Math-1.5B	21.55/81.50	6.64/43.15	13.12/51.64	2.92/**25.25**	1.46/13.34	18.91/70.35	10.77/47.54
w/SFT	43.36/82.39	16.80/50.88	13.69/45.61	0.62/12.52	0.62/4.21	16.72/65.93	15.30/43.59
w/DFT	63.78/84.76	**26.91**/51.04	28.32/53.62	4.17/15.03	**5.00**/18.81	34.53/72.10	27.12/49.23
w/DFT+KL	59.26/89.87	22.54/**58.72**	25.84/57.17	4.17/23.17	2.71/20.36	35.31/79.15	24.97/54.74
w/RA-DFT	**64.59/89.88**	26.26/57.45	**29.34/59.21**	**5.42**/22.63	4.38/**23.75**	**40.94/80.00**	**28.49/55.49**
Qwen2.5-Math-7B	30.98/87.02	6.94/44.75	9.36/48.78	6.67/**43.42**	0.42/20.76	20.94/76.49	12.55/53.54
w/SFT	54.73/88.32	22.33/61.12	19.57/54.86	1.25/18.31	1.04/10.70	25.62/73.50	20.76/51.13
w/DFT	70.03/84.32	**34.12**/55.08	34.87/56.01	8.54/18.69	4.58/21.18	42.19/77.06	32.39/52.06
w/DFT+KL	67.12/91.68	28.95/**64.01**	32.15/63.76	7.29/26.72	**5.21**/23.39	45.16/84.83	30.98/59.06
w/RA-DFT	**71.31/91.87**	31.94/63.17	**36.94/65.35**	**10.62**/33.33	**5.21/26.65**	**47.50/86.42**	**33.92/61.13**
Qwen3-1.7B-Base	25.45/79.62	7.40/40.54	9.58/43.31	1.88/12.12	0.42/8.40	12.81/59.65	9.59/40.61
w/SFT	35.24/75.95	11.58/45.30	8.34/38.06	0.21/8.12	0.83/2.87	13.75/59.07	11.66/38.23
w/DFT	54.24/79.89	20.50/46.25	21.08/46.23	2.71/11.55	1.67/12.47	26.25/60.93	21.07/42.89
w/DFT+KL	51.66/84.53	20.11/**53.75**	19.89/51.85	2.92/17.78	1.88/10.95	27.97/**67.48**	20.74/47.72
w/RA-DFT	**56.40/85.15**	**21.05**/51.87	**21.64/52.50**	**3.12/18.31**	**3.12/18.96**	**30.63**/67.45	**22.66/49.04**
Qwen3-8B-Base	43.88/**91.97**	16.84/**64.19**	20.39/63.10	4.38/**35.29**	3.75/**28.66**	27.03/**83.80**	19.38/**61.17**
w/SFT	52.98/87.20	23.53/62.05	20.42/54.80	2.29/14.70	0.62/12.38	23.75/72.07	20.60/50.53
w/DFT	**67.85**/83.99	31.20/51.98	33.44/56.05	6.25/14.52	3.54/12.37	**46.41**/78.29	31.45/49.53
w/DFT+KL	66.85/91.01	31.14/63.09	32.12/**63.37**	5.83/24.97	5.21/22.38	39.22/82.66	30.06/57.91
w/RA-DFT	**67.85**/90.55	**32.56**/63.87	**36.14**/62.85	**6.46**/24.62	**6.46**/25.90	42.03/82.03	**31.92**/58.31
Llama-3.1-8B-Instruct	36.50/77.66	17.62/47.14	10.35/39.10	3.12/**21.50**	0.21/5.34	16.72/62.82	14.09/42.26
w/SFT	27.35/73.76	9.72/40.66	7.04/34.91	0.42/6.20	0.21/2.82	9.69/57.35	9.07/35.95
w/DFT	44.98/76.68	19.67/45.74	13.41/40.54	3.33/18.14	**1.46**/11.97	18.28/62.29	16.85/42.56
w/DFT+KL	**47.46/81.41**	20.45/49.64	15.19/45.59	2.92/18.33	1.25/**13.70**	21.25/68.19	18.09/46.14
w/RA-DFT	46.86/81.29	**20.61/50.47**	**15.52/46.49**	**4.17**/19.66	1.04/13.34	**22.97/70.82**	**18.53/47.01**

DFT by $+8.22$, $+10.55$, $+7.86$, $+9.59$, and $+5.90$ points across the five models. Averaged across families, the gain over DFT is about $+6.9$ points at pass@16 and $+8.4$ points at pass@64. This trend is consistent with the mechanism in Section 2: attenuating unnecessary contraction of alternatives has a limited effect when only one sample is drawn, but becomes increasingly important when multiple samples can explore distinct continuations.

The comparison with DFT+KL in Table 1 and Table 2 provides an important control for the anchoring hypothesis introduced in Section 3.1. DFT+KL often improves high-k performance over DFT, which is expected because anchoring to the base model partially preserves alternatives. However, this anchoring is global and additive rather than reference-selective. As a result, DFT+KL loses a noticeable amount of low-k sharpness: RA-DFT has higher average@16 than DFT+KL in every family, with an average gain of about $+2.1$ points across families, and higher pass@1 in every family, with an average gain of about $+2.2$ points. Thus, DFT+KL partially addresses the high-k side of the problem, but it does not fully preserve the average-sample benefit that motivated DFT. RA-DFT is designed to do both: retain sharpening at reliable positions while preserving useful alternatives at uncertain positions.

3.3 Ablation study

We next ablate the three components of the RA-DFT weight: the reference-aware exponent α, the fallback threshold ε, and the combiner in the second factor. These ablations test whether the objective's high-k behavior follows from the proposed design rather than from incidental hyperparameter choices.

Table 3 shows three trends. First, RA-DFT is not overly sensitive to the exact choice of α: all tested values outperform the DFT baseline, and increasing α improves average@16 in this family. At the same time, high-k performance is not strictly monotonic, suggesting a trade-off between placing more weight on reference-supported positions and preserving alternative solution paths. Second, the fallback threshold ε is stable across a broad middle range. Very aggressive fallback, e.g., $\varepsilon = 0.5$, reduces the method toward DFT more often and substantially hurts high-k performance. Third, the

Table 2: Average pass@k and majority voting results over six math reasoning benchmarks. The left block reports pass@k; the right block reports majority voting, abbreviated as m@k. Boldface marks the best value within each model family and column, and the Δ row reports RA-DFT minus DFT.

	pass@k							majority voting					
	p@1	**p@2**	**p@4**	**p@8**	**p@16**	**p@32**	**p@64**	**m@1**	**m@4**	**m@8**	**m@16**	**m@32**	**m@64**
Qwen2.5-Math-1.5B	10.36	17.61	27.21	37.77	47.54	55.50	61.65	10.36	16.76	22.96	30.05	35.72	39.35
w/SFT	15.58	22.16	29.23	36.50	43.59	50.11	55.90	15.60	18.96	22.87	26.23	28.76	30.43
w/DFT	27.05	32.63	38.26	43.82	49.23	54.38	59.02	27.08	29.34	30.71	31.36	31.61	31.64
w/DFT+KL	25.47	33.74	41.43	48.47	54.74	60.22	64.79	25.46	30.85	34.76	37.25	38.75	39.55
w/RA-DFT	**28.45**	**36.00**	**43.04**	**49.46**	**55.49**	**61.47**	**67.24**	**28.45**	**33.06**	**36.29**	**38.27**	**39.22**	**39.64**
Δ RA-DFT–DFT	+1.40	+3.37	+4.78	+5.64	+6.26	+7.09	+8.22	+1.37	+3.72	+5.58	+6.91	+7.61	+8.00
Qwen2.5-Math-7B	12.95	21.44	32.13	43.37	53.54	61.94	68.74	12.93	20.43	27.82	35.84	42.36	**46.54**
w/SFT	21.11	28.77	36.52	44.03	51.13	58.07	65.18	21.15	25.09	29.42	32.96	35.29	36.76
w/DFT	32.60	37.78	42.80	47.56	52.06	56.47	60.62	32.61	34.46	35.62	36.21	36.52	36.63
w/DFT+KL	30.36	38.75	46.23	52.98	59.06	64.66	70.16	30.40	35.53	39.01	41.32	42.73	43.45
w/RA-DFT	**34.13**	**42.16**	**49.40**	**55.64**	**61.13**	**66.26**	**71.17**	**34.14**	**38.87**	**42.15**	**44.23**	**45.47**	46.10
Δ RA-DFT–DFT	+1.53	+4.38	+6.60	+8.08	+9.07	+9.79	+10.55	+1.53	+4.41	+6.53	+8.02	+8.95	+9.47
Qwen3-1.7B-Base	9.33	15.57	23.48	32.02	40.61	48.83	56.49	9.34	15.58	19.84	24.25	27.94	30.58
w/SFT	11.48	17.04	23.43	30.53	38.23	46.01	53.44	11.51	14.06	17.24	20.04	22.17	23.67
w/DFT	21.42	26.84	32.24	37.60	42.89	48.05	52.98	21.38	24.13	25.70	26.59	27.16	27.59
w/DFT+KL	20.23	27.21	34.25	41.18	47.72	53.86	59.92	20.25	24.63	27.86	30.36	31.92	32.81
w/RA-DFT	**22.44**	**29.33**	**36.11**	**42.69**	**49.04**	**55.16**	**60.84**	**22.45**	**26.76**	**29.99**	**32.31**	**33.57**	**34.20**
Δ RA-DFT–DFT	+1.02	+2.49	+3.87	+5.09	+6.15	+7.11	+7.86	+1.07	+2.63	+4.29	+5.72	+6.41	+6.61
Qwen3-8B-Base	19.85	30.98	42.73	**52.92**	**61.17**	**68.15**	**74.54**	19.83	27.90	36.36	**42.85**	**46.49**	**48.28**
w/SFT	20.56	28.03	35.72	43.32	50.53	57.57	64.80	20.56	24.29	28.58	32.08	34.79	36.62
w/DFT	31.73	36.50	40.92	45.18	49.53	53.89	57.87	31.70	33.79	34.76	35.22	35.54	35.83
w/DFT+KL	29.90	37.87	45.18	51.82	57.91	63.60	68.88	29.90	34.44	37.81	39.97	41.25	41.89
w/RA-DFT	**31.83**	**39.49**	**46.44**	52.73	58.31	63.15	67.46	**31.87**	**36.38**	**39.61**	41.80	42.81	43.18
Δ RA-DFT–DFT	+0.10	+2.99	+5.52	+7.55	+8.78	+9.26	+9.59	+0.17	+2.59	+4.85	+6.58	+7.27	+7.35
Llama-3.1-8B-Instruct	14.22	20.57	27.58	34.94	42.26	49.29	55.99	14.21	18.47	21.87	25.06	27.35	28.81
w/SFT	9.26	14.52	20.93	28.18	35.95	43.55	50.62	9.29	11.36	14.37	17.28	19.31	20.80
w/DFT	16.98	22.86	29.34	36.01	42.56	48.84	54.83	16.99	19.63	22.05	23.77	24.77	25.14
w/DFT+KL	18.50	25.19	32.19	39.27	46.14	52.54	58.57	18.48	21.98	25.19	27.39	28.76	29.42
w/RA-DFT	**18.65**	**25.46**	**32.62**	**39.86**	**47.01**	**54.00**	**60.73**	**18.72**	**22.25**	**25.49**	**27.76**	**29.26**	**30.12**
Δ RA-DFT–DFT	+1.67	+2.60	+3.28	+3.85	+4.45	+5.16	+5.90	+1.73	+2.62	+3.44	+3.99	+4.49	+4.98

combiner has the clearest effect on high-k results. Although mean, min, and no-floor variants achieve similar average@16, they consistently degrade pass@64 compared with the default max combiner. Removing the recovery floor entirely drops pass@64 by 5.8 points.

Overall, these ablations indicate that the improvement over DFT is not the result of a narrowly tuned hyperparameter setting. Rather, RA-DFT remains stronger than DFT across the tested variants, while the max-based recovery floor is especially important for preserving high-k performance. This supports the design in Section 2.4: the recovery term is not primarily needed to improve average sampled accuracy, but is load-bearing for maintaining diversity and recoverability at larger sampling budgets.

3.4 Analysis

Majority voting results. Pass@k measures whether at least one sample is correct, which is the relevant metric for verifier-based selection or rejection sampling. Majority voting asks a different question: whether the distribution places enough mass on the correct answer for plurality aggregation to succeed. Table 2 reports both views in the main text: the full pass@k curve and the full majority-voting curve, using m@k as a compact abbreviation for maj@k. The majority-voting columns show the same qualitative conclusion as pass@k. RA-DFT beats DFT in all five model families at every evaluated m@k, and it also beats DFT+KL in all five families at every evaluated m@k. This matters

Table 3: Ablation study for RA-DFT on Qwen2.5-Math-1.5B. We report average@16 and average pass@k over six mathematical reasoning benchmarks. Default settings are shaded.

Component	Variant	Avg@16	p@1	p@2	p@4	p@8	p@16	p@32	p@64
Selectivity α	$\alpha = 0$	28.34	27.58	35.05	42.06	48.61	54.93	60.92	65.87
	$\alpha = 0.5$	28.49	28.45	36.00	43.04	49.46	55.49	61.47	67.24
	$\alpha = 1$	28.59	28.67	36.35	43.52	50.00	55.72	60.94	65.20
	$\alpha = 2$	**29.26**	**28.98**	**36.74**	**43.92**	**50.48**	**56.70**	**62.98**	**69.82**
Fallback ε	$\varepsilon = 0$	27.76	28.07	36.29	43.90	50.66	56.70	62.04	66.67
	$\varepsilon = 0.01$	**28.99**	**28.87**	**36.81**	**44.29**	**51.17**	**57.46**	**63.03**	67.53
	$\varepsilon = 0.05$	28.49	28.45	36.00	43.04	49.46	55.49	61.47	67.24
	$\varepsilon = 0.1$	28.27	28.12	35.56	42.67	49.45	55.89	62.14	**67.84**
	$\varepsilon = 0.5$	28.26	28.04	34.66	40.99	46.91	52.30	57.46	62.86
Combiner	max	28.49	28.45	36.00	**43.04**	**49.46**	**55.49**	**61.47**	**67.24**
	mean	**28.79**	28.65	**36.13**	43.00	48.99	54.44	59.77	64.95
	min	28.69	**28.68**	35.49	41.80	47.42	52.79	58.43	64.31
	none	**28.79**	28.60	34.74	40.57	46.00	51.20	56.30	61.42

because pass@k and majority voting reward different distributional properties. Pass@k rewards any-correct coverage, while majority voting rewards stable consensus around the correct answer. Improving both suggests that RA-DFT is not merely adding random diversity. Instead, it preserves useful alternatives while keeping high-confidence positions sharp. This also clarifies why DFT+KL cannot fully solve the problem. A KL anchor can restore probability mass to the base distribution, which helps high-k sampling in some cases. But the anchor is not conditioned on whether a position is open or canonical. It may therefore preserve alternatives even where the demonstrated token is already a reliable continuation, weakening the low-k sharpening that DFT provides. The mechanism diagnostic in Section 2.5 shows the desired behavior more directly: RA-DFT restores alternative mass primarily in low-reference-probability positions while remaining sharp in high-reference-probability positions. The experiments here show the benchmark-level consequence of that selectivity: RA-DFT is consistently better than DFT on pass@k and majority voting, while yielding a more favorable overall balance than DFT+KL in our evaluated setting.

Saturated-base exception. One exception merits explicit discussion. On Qwen3-8B-Base, the base model already attains very strong pass@64 performance, and none of the fine-tuning methods, including RA-DFT, improves over the base model at the large sampling budget. This suggests a saturated-base regime, possibly due to the model's strong intrinsic mathematical capability or a favorable alignment between the base model and the evaluation benchmarks. Accordingly, we do not claim strict dominance over every base-anchored variant in such regimes. Importantly, however, this case is still consistent with our central claim: DFT suffers a sharp drop at pass@64, whereas RA-DFT recovers $+9.59$ points over DFT and remains stronger on low-k and majority-voting metrics. The evidence therefore supports the intended conclusion: RA-DFT mitigates DFT's high-k degradation while preserving its average-sample gains.

4 Conclusion

We studied a pass@1-pass@k tension in Dynamic Fine-Tuning: the same sharpening that improves average sampled accuracy can suppress high-probability alternatives and reduce high-k sampling performance. Our analysis shows that this behavior arises from the local softmax effect of the DFT update, which contracts competing continuations while increasing the demonstrated-token probability. We introduced RA-DFT, a reference-aware refinement of DFT that uses the frozen base model's teacher-forced probability as a local confidence signal. RA-DFT remains a single positive-demonstration reweighted cross-entropy objective. By attenuating DFT-style sharpening at low-reference-probability positions while retaining strong sharpening at high-reference-probability positions, RA-DFT yields a better overall balance between low-k accuracy and high-k coverage. Across five models and six mathematical reasoning benchmarks, RA-DFT improves over DFT across the evaluated pass@k range and improves majority voting while preserving DFT's average-sample gains. These results suggest that effective fine-tuning for reasoning should control not only how strongly the model imitates positive demonstrations, but also where distributional sharpening occurs.

AI USE STATEMENT

ChatGPT was used to provide feedback on the exposition, mathematical interpretation, and presentation of the methodology and experiments, and to assist with manuscript revision. The authors are responsible for the final content, including AI-assisted text and claims.

REFERENCES

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, 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, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(1), 2024.

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. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, and et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

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, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong

Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. ToRA: a tool-integrated reasoning agent for mathematical problem solving, 2024. URL https://arxiv.org/abs/2309.17452.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, August 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, Kai Dang, Yang Fan, Yichang Zhang, An Yang, Rui Men, Fei Huang, Bo Zheng, Yibo Miao, Shanghaoran Quan, Yunlong Feng, Xingzhang Ren, Xuancheng Ren, Jingren Zhou, and Junyang Lin. Qwen2.5-coder technical report. 2024. URL https://arxiv.org/abs/2409.12186.

OpenAI: Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang,

Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. OpenAI o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 2022.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q. Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in AI4Maths with 860k pairs of competition math problems and solutions, 2024. URL https://github.com/project-numina/aimo-progress-prize.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jian-Guang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. WizardMath: empowering mathematical reasoning for large language models via reinforced evol-instruct. In *The Thirteenth International Conference on Learning Representations*, 2025.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, 2022.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5: A party of foundation models, 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. 2023. URL https://crfm.stanford.edu/2023/03/13/alpaca.html.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2024. URL https://arxiv.org/abs/2312.08935.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, and Xu Yang. On the generalization of SFT: A reinforcement learning perspective with reward rectification. In *International Conference on Learning Representations*, 2026.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024a. URL https://arxiv.org/abs/2409.12122.

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, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen: An Yang, Baosong Yang, B Zhang, B Hui, B Zheng, B Yu, Chengpeng Li, D Liu, F Huang, H Wei, et al. Qwen2.5 technical report, 2024b. URL https://arxiv.org/abs/2412.15115.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, and Fei Wu. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LiLi Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

A Related work

Supervised fine-tuning and dynamic fine-tuning. Supervised fine-tuning (SFT) is the standard approach for adapting language models from expert demonstrations, due to its simplicity and stability (Chung et al., 2024; Wei et al., 2022; Zhou et al., 2023; Zhang et al., 2026; Mishra et al., 2022; Taori et al., 2023; Lightman et al., 2024). However, standard SFT assigns the same cross-entropy weight to every reference token, regardless of the model's current confidence or the role of the token in the output distribution. Dynamic Fine-Tuning (DFT) (Wu et al., 2026) addresses this by reweighting each token loss with the current model probability of the reference token. This simple modification improves average sampled accuracy and provides a strong demonstration-only alternative to standard SFT. Our work builds directly on DFT. We treat DFT as a strong baseline and study a complementary failure mode: although DFT improves the reliability of individual samples, its probability-weighted update can also sharpen the distribution non-selectively. In particular, increasing the demonstrated-token probability under softmax necessarily suppresses competing alternatives. We show that this effect can hurt high-k sampling, where success depends not only on the quality of one sample but also on whether useful alternative continuations remain sampleable.

Sampling-based evaluation and the pass@k trade-off. Many reasoning benchmarks evaluate models using pass@k, where a problem is solved if at least one of k sampled completions is correct (Chen et al., 2021; Cobbe et al., 2021; Luo et al., 2025; Wang et al., 2024; Shao et al., 2024; Gou et al., 2024). This metric captures a different desideratum from pass@1 or average sampled accuracy. Pass@1 rewards making the most likely sample correct, whereas pass@k also rewards preserving enough distributional coverage for multiple samples to explore distinct solution paths. Inference-time methods such as self-consistency exploit this property by sampling multiple reasoning chains and aggregating their answers (Wang et al., 2023). This distinction is central to our motivation. A fine-tuning method can improve average sampled accuracy by making the model distribution sharper, yet reduce pass@k if it removes alternatives that would have produced correct solutions under repeated sampling. RA-DFT is designed for this regime: it keeps the useful sharpening behavior of DFT at positions where the reference token is locally reliable, while attenuating sharpening where the base model is uncertain and alternative continuations may be valuable.

B METHOD DERIVATIONS

B.1 LOCAL LOGIT EFFECT OF DFT

Lemma B.1 (Local logit direction induced by DFT)**.** *Fix a teacher-forced position t, let $a := y_t^\star \in \mathcal{V}$ be the demonstrated token, and define*

$$p_j := \pi_\theta(j \mid x, y^\star_{<t}), \qquad j \in \mathcal{V}.$$

Let $z = (z_j)_{j\in\mathcal{V}}$ be the corresponding pre-softmax logits, so that $p = \mathrm{softmax}(z)$, and consider the DFT token loss

$$\mathcal{L}_t^{\mathrm{DFT}} = -\,\mathrm{sg}(p_a)\log p_a.$$

Then:

$$\frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_a} = -\,\mathrm{sg}(p_a)(1-p_a), \qquad \frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_v} = \mathrm{sg}(p_a)p_v, \quad v \neq a. \tag{10}$$

As a consequence, under the local gradient step

$$z^+ = z - \eta \nabla_z \mathcal{L}_t^{\mathrm{DFT}},$$

the logit increments satisfy

$$\Delta z_a = \eta\ \mathrm{sg}(p_a)(1-p_a) \geq 0, \qquad \Delta z_v = -\eta\ \mathrm{sg}(p_a)p_v \leq 0, \quad v \neq a, \tag{11}$$

with strict inequalities whenever $0 < p_a < 1$ and $p_v > 0$.

Proof. Since $\mathrm{sg}(p_a)$ is detached from the gradient, it is constant with respect to the local logits z. Therefore,

$$\frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_j} = -\,\mathrm{sg}(p_a)\frac{1}{p_a}\frac{\partial p_a}{\partial z_j}.$$

Using the softmax Jacobian,

$$\frac{\partial p_a}{\partial z_j} = p_a(\mathbf{1}\{j=a\} - p_j),$$

we obtain

$$\frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_j} = -\,\mathrm{sg}(p_a)(\mathbf{1}\{j=a\} - p_j).$$

Hence, for $j = a$,

$$\frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_a} = -\,\mathrm{sg}(p_a)(1-p_a),$$

while for any $v \neq a$,

$$\frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_v} = \mathrm{sg}(p_a)p_v.$$

This proves Equation (10).

Now apply one local gradient step on the logits:

$$z^+ = z - \eta \nabla_z \mathcal{L}_t^{\mathrm{DFT}}.$$

By definition,

$$\Delta z_j := z_j^+ - z_j = -\eta \frac{\partial \mathcal{L}_t^{\mathrm{DFT}}}{\partial z_j}.$$

Substituting Equation (10) gives

$$\Delta z_a = \eta \ \mathrm{sg}(p_a)(1 - p_a) \geq 0, \qquad \Delta z_v = -\eta \ \mathrm{sg}(p_a) p_v \leq 0, \quad v \neq a.$$

If $0 < p_a < 1$, then $1 - p_a > 0$, so $\Delta z_a > 0$. If $p_v > 0$, then $\Delta z_v < 0$. This proves Equation (11). □

B.2 Proof of Proposition 2.1

For clarity, we restate the proposition in the notation of Section 2.3. Fix a teacher-forced position t, let $a := y_t^\star \in \mathcal{V}$ be the demonstrated token, and define

$$p_j := \pi_\theta(j \mid x, y_{<t}^\star), \qquad j \in \mathcal{V},$$

with $p = \mathrm{softmax}(z)$, where $z = (z_j)_{j \in \mathcal{V}}$ denotes the local pre-softmax logits. Consider the DFT token loss

$$\mathcal{L}_t^{\mathrm{DFT}} = -\,\mathrm{sg}(p_a) \log p_a,$$

and the local logit step

$$z^+ = z - \eta \nabla_z \mathcal{L}_t^{\mathrm{DFT}}, \qquad \eta > 0.$$

For $v \neq a$, let $\Delta p_v := p_v^+ - p_v$. We show that

$$\Delta p_v = \eta p_a p_v \left(\sum_{j \in \mathcal{V}} p_j^2 - p_a - p_v \right) + O(\eta^2).$$

Proof. Fix $v \neq a$. Under the same local logit step

$$z^+ = z - \eta \nabla_z \mathcal{L}_t^{\mathrm{DFT}},$$

Lemma B.1 gives

$$\Delta z_a = \eta p_a (1 - p_a), \qquad \Delta z_j = -\eta p_a p_j, \quad j \neq a. \tag{12}$$

Now apply a first-order Taylor expansion of $p_v = \mathrm{softmax}(z)_v$:

$$\Delta p_v = \sum_{j \in \mathcal{V}} \frac{\partial p_v}{\partial z_j} \Delta z_j + O(\eta^2).$$

Using the softmax Jacobian,

$$\frac{\partial p_v}{\partial z_j} = p_v(\mathbf{1}\{j = v\} - p_j),$$

we obtain

$$\Delta p_v = p_v \left(\Delta z_v - \sum_{j \in \mathcal{V}} p_j \Delta z_j \right) + O(\eta^2). \tag{13}$$

Since $v \neq a$, Equation (12) gives

$$\Delta z_v = -\eta p_a p_v.$$

Moreover,

$$\begin{aligned} \sum_{j \in \mathcal{V}} p_j \Delta z_j &= p_a \Delta z_a + \sum_{j \neq a} p_j \Delta z_j \\ &= p_a \big(\eta p_a (1 - p_a) \big) + \sum_{j \neq a} p_j (-\eta p_a p_j) \end{aligned}$$

$$= \eta p_a^2 - \eta p_a \sum_{j \in \mathcal{V}} p_j^2.$$

Substituting these identities into Equation (13) yields

$$\Delta p_v = p_v \left(-\eta p_a p_v - \eta p_a^2 + \eta p_a \sum_{j \in \mathcal{V}} p_j^2 \right) + O(\eta^2)$$

$$= \eta p_a p_v \left(\sum_{j \in \mathcal{V}} p_j^2 - p_a - p_v \right) + O(\eta^2),$$

which proves the proposition. □

B.3 Gradient of the KL anchoring baseline

Let $r_i = \pi_{\text{base}}(i \mid x, y^\star_{<t})$ and $p_i = \pi_\theta(i \mid x, y^\star_{<t})$. For the token-level anchor $R_t = \mathrm{KL}(r \| p) = \sum_i r_i \log(r_i/p_i)$, the logit derivative is

$$\frac{\partial R_t}{\partial z_i} = p_i - r_i.$$

Thus, an additive KL term pulls the current distribution toward the base distribution at every position. A scalar KL coefficient controls only the overall strength of this pull; it does not by itself implement the explicit position-selective sharpening mechanism used by RA-DFT, which attenuates the DFT-style update according to the local reference confidence r_t. This is why DFT+KL is an important baseline, but not an explicitly confidence-gated, position-selective objective in the sense of RA-DFT.

C Mechanism diagnostics settings

The main mechanism diagnostic compares the base model and four trained checkpoints: SFT, DFT, DFT+KL, and RA-DFT. We evaluate them under teacher forcing on held-out NumiNa-CoT response tokens that were not used during training. For each token position t, we compute the base reference probability $r_t = \pi_{\text{base}}(y^\star_t \mid x, y^\star_{<t})$ once and assign positions to ten equal-frequency buckets sorted by r_t. This assignment is method-independent, so all checkpoints are compared on the same token populations within each bucket.

The local probe is the strongest-alternative probability

$$a_t(\theta) = \max_{v \neq y^\star_t} \pi_\theta(v \mid x, y^\star_{<t}).$$

This is the non-reference token most directly affected by the contraction term in Equation (6).

D Additional validation of the reference-probability axis

The main method uses the base reference probability r_t as an operational axis for deciding where alternative continuations should be preserved. This axis is not a correctness oracle. Rather, it is a statistical diagnostic: positions with low r_t are expected to be more tolerant to alternative continuations, while positions with high r_t are expected to be more sensitive to local substitutions. We include two representative validation checks for this interpretation. They are post-hoc diagnostics and are not used for training or model selection.

D.1 Paired-substitution validation

This check uses teacher-forced held-out responses, so the open/canonical assignment is based on the same quantity as in the main method. For every position t, we compute $r_t = \pi_{\text{base}}(y^\star_t \mid x, y^\star_{<t})$, sort positions by r_t, and form ten equal-frequency buckets. In this validation, the *open bucket* is the lowest-r_t bucket $B0$, and the *canonical bucket* is the highest-r_t bucket $B9$. We use the two extremes

because this check is intended to validate the interpretation of the axis, not to estimate a smooth curve over all buckets.

For each selected position, let

$$\tilde{y}_t = \arg\max_{v \neq y_t^\star} \pi_{\text{base}}(v \mid x, y_{<t}^\star)$$

be the base model's strongest non-reference token. We compare two matched prefixes for the same problem and the same position: the reference prefix $(x, y_{\leq t}^\star)$, and the substituted prefix $(x, y_{<t}^\star, \tilde{y}_t)$. From each prefix, the base model generates continuations at the same temperature 1.0 as in the main experiments. This pairing controls for problem identity, prefix length, and intervention location; the only local change is whether the prefix ends with the reference token or the strongest alternative.

For bucket b, we report the substitution effect

$$E_b = 100 \cdot \frac{\text{Acc}_{\text{ref},b} - \text{Acc}_{\text{sub},b}}{\text{Acc}_{\text{ref},b}}.$$

Positive values mean that substituting the strongest alternative reduces correctness. Negative values mean that the substituted prefix performs slightly better in that bucket, so the substitution provides no evidence of harm. If the r_t-axis captures the open/canonical distinction, E_b should be small or negative in B0 and substantially positive in B9.

Table 4: Paired-substitution validation on representative families. Open is the lowest-r_t bucket B0, and canonical is the highest-r_t bucket B9. Entries are substitution effects E_b in percent; positive values mean that the strongest-alternative substitution reduces correctness. The final column is measured in percentage points.

Family	Open B0	Canonical B9	Canonical–open
Qwen2.5-Math-1.5B	-2.2	$+21.2$	**+23.4**
Qwen3-1.7B-Base	-8.3	$+24.7$	**+33.0**

The pattern supports the intended interpretation. In the open bucket, replacing the reference token with the strongest alternative is not reliably harmful. In the canonical bucket, the same intervention substantially reduces final-answer correctness. Thus, low-r_t positions are statistically more tolerant to strong alternatives, whereas high-r_t positions are more sensitive to local substitutions.

D.2 On-policy canonical leakage

The paired-substitution check uses teacher-forced prefixes. We also ask whether fine-tuned models remain sharp at high-confidence positions when they generate normally. This diagnostic uses sampled responses rather than reference responses.

For a sampled response $\hat{y}$, consider a generated position t. We run the frozen base model on the sampled prefix $(x, \hat{y}_{<t})$ and define

$$u_t = \max_v \pi_{\text{base}}(v \mid x, \hat{y}_{<t}), \qquad b_t = \arg\max_v \pi_{\text{base}}(v \mid x, \hat{y}_{<t}).$$

Here u_t is the base model's confidence in its own preferred next token under the sampled prefix, and b_t is that preferred token. We call a sampled position *canonical* if $u_t \geq 0.9$. At such positions, the base model has a high-confidence local continuation. A *canonical leak* occurs when the sampled model emits a token different from b_t. The canonical leak rate is therefore

$$\frac{\#\{t : u_t \geq 0.9,\ \hat{y}_t \neq b_t\}}{\#\{t : u_t \geq 0.9\}}.$$

Lower canonical leak means that a model more often follows the base model's high-confidence local continuation during sampling. This is the on-policy analogue of the canonical sharpness measured by the teacher-forced diagnostic.

The result is consistent with the mechanism. Compared with DFT+KL, RA-DFT has roughly half the canonical leak rate in both representative families. This supports the claim that RA-DFT does not merely restore base-model alternatives globally; it keeps high-confidence local continuations sharper while preserving the open-position alternatives emphasized in the main diagnostic.

Table 5: On-policy canonical leak rate on representative families. Lower values indicate sharper behavior at high-confidence base positions. The final column reports how many times larger the DFT+KL leak rate is than the RA-DFT leak rate.

Family	Base	SFT	DFT	DFT+KL	RA-DFT	DFT+KL / RA-DFT
Qwen2.5-Math-1.5B	0.844%	2.170%	0.165%	0.196%	**0.099%**	1.98×
Qwen3-1.7B-Base	1.114%	1.864%	0.117%	0.135%	**0.062%**	2.18×

E Experiment details

E.1 Training details

We mainly follow the training settings of DFT (Wu et al., 2026). Specifically, all methods are trained for one epoch with matched optimization settings: global batch size 256 and maximum sequence length 2048. Qwen-family models use learning rate 5×10^{-5}, while Llama uses 2×10^{-5}. AdamW optimizer is used with a cosine learning rate schedule and warm up ratio of 0.1. For the DFT+KL baseline, we use a token-level forward KL anchor $D_{\mathrm{KL}}(\pi_{\mathrm{base}} \| \pi_\theta)$ on the same teacher-forced contexts. The base model is frozen and the KL coefficient is fixed to $\lambda = 0.05$ for all models. For our RA-DFT, we use fixed $\alpha = 0.5$ and $\varepsilon = 0.05$ in Equation (9) for all models. We use 8 GPUs for training in our experiment.

E.2 Evaluation details

Unless noted, reported averages are unweighted averages over the six benchmarks. We use a temperature of 1.0 and max generation length of 4096 tokens. Average@16 denotes the average accuracy over 16 independent generations. For pass@k, we draw $n = 64$ samples at temperature $T = 1.0$ and use the unbiased estimator

$$\widehat{\mathrm{pass@}k} = 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}},$$

where c is the number of correct generations among the n samples. We also report majority voting, denoted m@k, where the final answer is extracted from each of k samples and the plurality answer is selected.

E.3 Runtime comparison

On Qwen2.5-Math-1.5B, one full training epoch takes 18.6 minutes for SFT, 20.4 minutes for DFT, 29.2 minutes for DFT+KL, and 26.8 minutes for RA-DFT under the same hardware and settings. Normalized by DFT, the runtime is $0.92\times$ for SFT, $1.43\times$ for DFT+KL, and $1.32\times$ for RA-DFT. The additional cost of RA-DFT comes from using the frozen reference model to obtain base probabilities for the token weights; these probabilities can also be precomputed offline. Thus, RA-DFT adds moderate overhead over DFT, but is cheaper than the DFT+KL anchoring baseline in this setting.

F Ablation notes

The main-text ablation varies the selectivity exponent α, fallback threshold ε, and recovery combiner on Qwen2.5-Math-1.5B. The important takeaway is not that a single hyperparameter value is uniquely optimal, but that the design is stable across a range of settings. The max combiner is the most load-bearing component for high-k behavior: variants that weaken or remove the recovery floor achieve similar average@16 but noticeably lower pass@64. This supports the role of the recovery floor in maintaining reliable reference-supported updates without uniformly restoring alternatives.

G Limitations

Our study is intended as a focused investigation of the pass@1–pass@k trade-off in reasoning-oriented supervised fine-tuning, rather than as a claim of universal optimality for all tasks. In

particular, our experiments are conducted on mathematical reasoning benchmarks where sampling-based evaluation methods like pass@k and majority voting are widely used and practically relevant. RA-DFT also introduces modest additional computation compared with DFT because it uses base model's probabilities in the token weights.

H BROADER IMPACT

RA-DFT is a supervised fine-tuning objective for improving reasoning performance from positive demonstrations. Because it preserves the standard teacher-forced training pipeline, it may make reasoning fine-tuning more accessible in settings where additional supervision is expensive. The method does not introduce new data sources or deployment mechanisms; its risks are therefore similar to those of other fine-tuning methods that improve model reasoning capabilities. As with any capability-improving post-training method, RA-DFT should be applied with appropriate dataset governance, evaluation, and deployment safeguards for the intended application domain.