RETHINKING SUPERVISED FINE-TUNING THROUGH POLICY PROJECTION: FROM UNIFORM IMITATION TO SELECTIVE ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised fine-tuning (SFT) has become a key recipe for adapting large language models (LLMs) to different domains through imitation of task-specific demonstrations. However, its deterministic objective applies indiscriminate learning to each reference token, which may lead to overfitting and ignore whether the supervision is compatible with the model's current predictive state. We first develop a structured policy projection framework that views fine-tuning as controlled movement in policy space and provides a principled design space by characterizing each token-level projection in terms of where the policy is directed with minimum shift and how strongly it contributes to optimization. This formulation reveals standard SFT as a rigid case with a degenerate projection target and uniform strength. Based on this, we further propose Modulated Fine-Tuning (MFT), which combines a compatibility-adaptive relaxed projection with an information-aware projection strength to modulate token-wise supervision according to the current policy, consolidating compatible and informative supervision while applying more conservative updates to strongly conflicting targets. Experiments across diverse domains show that MFT consistently outperforms SFT, improving average performance by +6.96 on mathematical reasoning and +10.8 on general domain tasks.

1 INTRODUCTION

Supervised fine-tuning (SFT) is a simple and effective paradigm for adapting pretrained language models to task-specific demonstrations (Wei et al., 2022; Ouyang et al., 2022). Nonetheless, increasing evidence shows that vanilla SFT may generalize poorly and induce over-specialization, distributional drift, or reduced generation diversity (Li et al., 2025; HAO et al., 2025; Klypa & Cherednichenko, 2026). More recent analyses further show that supervised data are intrinsically heterogeneous: conflicts between fine-tuning supervision and pretrained knowledge can substantially affect learning, and highly conflicting examples may even be detrimental to downstream performance (Zhong et al., 2025; Xue et al., 2026; Li et al., 2026). These observations point to a limitation beyond conventional overfitting: SFT does not distinguish supervision that is compatible with the current model from supervision that strongly contradicts it. Each demonstrated token is optimized under the same maximum-likelihood principle, even though the model may have very different prior preferences across tokens.

Recent efforts improve SFT by filtering or selecting more informative demonstrations (Deb et al., 2025; Pang et al., 2025; Zhou et al., 2026), reshaping token-level objectives or learning signals (Qin & Springenberg, 2025; Wu et al., 2026; Li et al., 2026; Sabbaghi et al., 2026), and constraining policy drift through entropy, proximal, self-distillation, or KL-based regularization (Li et al., 2024b; Zhu et al., 2026a; Wang et al., 2026b; Zhu et al., 2026b; Wang et al., 2026a). Most closely related to our work are studies that further connect SFT to reinforcement learning (RL) through importance weighting or implicit-reward interpretations (Qin & Springenberg, 2025; Wu et al., 2026; Gan et al., 2026). However, these approaches do not provide a unified characterization of both where token-level supervision moves the policy and how strongly each position influences the shared update.

In this work, we revisit SFT from the geometry of policy-space projection. This perspective is rooted in the classical theory of Bregman projections and mirror descent, which provides a general

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.

framework that views optimization as projection in a non-Euclidean geometry (Beck & Teboulle, 2003). In RL, this geometric view has been further connected to KL-regularized and proximal policy optimization (PPO) Schulman et al. (2017) through the theory of regularized Markov decision processes (Geist et al., 2019). More recently, Shenfeld et al. (2026) have applied the projection perspective to language-model fine-tuning and shown that on-policy RL can be interpreted as alternating information and moment projections, which explains its preference for KL-minimal solutions.

Motivated by the minimum-shift behavior in binary-reward RL (Shenfeld et al., 2026), we ask the following central **research question**:

Can supervised adaptation acquire new behavior while minimally disrupting the pretrained policy?

To study this question, we develop a structured policy projection framework for SFT. The framework views supervised learning as a collection of token-level projections whose effects are accumulated through a shared policy update, while constraining each projection to avoid unnecessary policy shift (Fig. 1). Importantly, each token-level projection is characterized by two complementary aspects: the *projection target*, which determines where the policy is directed in probability space, and the *projection strength*, which determines how much that projection influences the shared update. This formulation provides a structured way to understand both the displacement induced by supervision and its token-wise contribution to optimization. From this view, standard SFT exhibits a rigid structure: every demonstrated token is treated as the full projection target and assigned uniform strength, irrespective of its compatibility with the current policy.

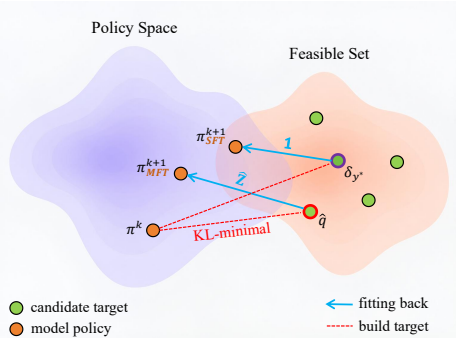


Figure 1: At the k-th update, the framework first selects a projection target from the feasible set, and then fits the model policy toward the selected target. SFT projects toward the hard reference target δ_{y^*} with uniform strength, while MFT selects the KL-minimal relaxed target $\hat{q}$ and modulates the fitting strength by $\hat{Z}$.

Building on this lens, we introduce **Modulated Fine-Tuning (MFT)**, a new fine-tuning algorithm that modulates both the projection target and projection strength according to the current policy. As illustrated in Fig. 2, MFT replaces the full target projection with a relaxed projection that adapts the policy displacement to model confidence while preserving the relative structure of non-target tokens, and assigns information-aware strength to regulate the influence of each supervised position. Together, this dual-mechanism enables selective adaptation rather than treating all demonstrated tokens uniformly, emphasizing informative, policy-compatible supervision while mitigating excessive fitting and meaningless policy shift on strongly conflicting targets. This allows the model to acquire new behavior patterns while reducing unnecessary distortion of useful pretrained structures.

Our key contributions are as follows. (i) We develop a **structured policy projection framework** that provides a principled **design space** for analyzing and constructing SFT by characterizing token-level learning into a minimum-shift projection target and its relative strength in the shared policy update. Within this framework, standard SFT emerges as a rigid special case with a degenerate target and uniform strength. (ii) Based on this analysis, we introduce **Modulated Fine-Tuning (MFT)**, which relaxes its fixed target and uniform-strength design through compatibility-adaptive projection and information-aware strength modulation. (iii) Extensive experiments across diverse tasks demonstrate consistent improvements over SFT, with relative gains of 29.2% and 28.1% on mathematical reasoning and general domain benchmarks, respectively. Ablation studies further verify the complementary benefits of target relaxation and strength modulation.

2 Related Work

Stabilizing Supervised Fine-Tuning. Recent work has increasingly focused on improving SFT through (i) *better control of token-level optimization* and (ii) *the distributional changes induced during fine-tuning*. In the first line of work, Li et al. (2026) study that the effectiveness of probability-based SFT objectives depends on the model's capability regime, while Sabbaghi et al. (2026) intro-

duce information-aware weighting to emphasize tokens with higher learning value. Several recent methods further adapt token updates to the model's predictive state. Gong et al. (2026) adapt token weights in long-CoT supervision according to gradient utility. Diao et al. (2026) use token-level entropy to identify confident conflicts and suppress destructive gradients. Wang et al. (2026c) derive an entropy-adaptive gating rule for allocating gradients across different confidence regimes. Lin et al. (2026) reduce the weight of hard tokens to mitigate general capability degradation. The second line of research controls the distributional change induced by SFT. Li et al. (2024b) introduce entropy-regularized distribution matching to preserve output diverse, while CurioSFT (Wang et al., 2026a) uses adaptive self-distillation to maintain exploration.

Connecting SFT and RL. A group of work has revisited SFT through the lens of RL. Qin & Springenberg (2025) connect curated SFT to sparse-reward RL through a lower-bound formulation and importance weighting. Wang et al. (2025) unify SFT and preference learning through their induced implicit rewards. DFT (Wu et al., 2026) rewrites SFT through a policy-gradient perspective and corrects its induced weighting, while ASFT (Zhu et al., 2026a) further adds a KL term to control its policy drift. Instead of reinterpretation and reward reweighting, recent work also imports optimization principles from RL more directly. PSFT (Zhu et al., 2026b) introduces trust-region constraints to stabilize supervised policy updates, whereas GFT (Gan et al., 2026) introduces group-based advantages with adaptive coefficient rectification. Ming et al. (2026) further brings token-level on-policy guidance into SFT through one-token rollouts. Beyond offline SFT, hybrid training methods, such as LUFFY (Yan et al., 2025) and CHORD (Zhang et al., 2026), integrate off-policy expert supervision into on-policy RL through adaptive weighting mechanisms that balance imitation and exploration.

Overall, existing methods improve SFT mainly through reweighting, distributional regularization, or RL-inspired reinterpretations, but do not jointly characterize the local projection target and its relative contribution to the shared policy update. In contrast, we formulate supervised adaptation under a minimum-shift policy projection framework that separates and modulates these two dimensions.

3 Preliminary: The Minimum Shift Principle

Previous work (Shenfeld et al., 2026) shows that single-step reinforcement learning with binary rewards exhibits a *minimum-shift bias*: among policies that satisfy the task objective, optimization favors solutions that remain closest in KL divergence to the current policy. This provides us with a perspective to understand why RL tends to forget less.

For an input x, let $\pi_k(\cdot \mid x)$ denote the current policy at iteration k, and let $R(x, y) \in \{0, 1\}$ be a binary reward. Denote the representable policy class as $\Pi \subseteq \Delta(\mathcal{Y})$. The set of task-optimal distributions is

$$\mathcal{C}(x) \coloneqq \left\{ q \in \Delta(\mathcal{Y}) : \mathbb{E}_{y \sim q(\cdot|x)}[R(x, y)] = 1 \right\}. \tag{1}$$

The corresponding policy update can be expressed through two alternating projections. First, among all task-valid distributions, the policy update selects the one closest to the current policy π_k:

$$q_k = \arg\min_{q \in \mathcal{C}(x)} D_{\mathrm{KL}}(q \,\|\, \pi_k). \tag{2}$$

The policy is then fitted to the resulting target by minimizing the KL divergence in the direction corresponding to maximum-likelihood fitting:

$$\pi_{k+1} = \arg\min_{\pi \in \Pi} D_{\mathrm{KL}}(q_k \,\|\, \pi). \tag{3}$$

In the idealized distribution-space setting, this alternating procedure theoretically converges to

$$\pi^\dagger = \arg\min_{\pi \in \mathcal{C}(x) \cap \Pi} D_{\mathrm{KL}}(\pi \,\|\, \pi_0). \tag{4}$$

Hence, among representable policies that solve the new task, binary-reward RL is implicitly biased toward a solution that induces minimal policy displacement from initialization. This minimum-shift behavior provides a useful principle for understanding how new behavior can be acquired while limiting unnecessary disruption of the pretrained policy.

4 METHOD

4.1 A STRUCTURED POLICY PROJECTION FRAMEWORK

For supervised learning, at each teacher-forced state, the supervision requirement can be expressed analogously to a binary-reward condition, with the reference token as the uniquely rewarded action. Motivated by the minimum-shift principle in Sec. 3, we require each local adaptation to satisfy its supervision constraint while avoiding unnecessary deviation from the current policy.

However, at the sequence level, each token induces a distinct local policy movement, while all such movements must be realized jointly through the same parameter update. Therefore, a local projection is not fully characterized by its destination alone, its relative influence on the shared update must also be specified. To obtain a structured modeling space for sequence-level supervised learning, we accordingly introduce two complementary design variables for each supervised position: a projection target, which specifies where the local policy should move, and a projection strength, which controls its relative contribution to the shared policy optimization. Together, these two variables define a structured policy projection framework that provides a design space for analyzing and improving supervised fine-tuning.

Given an input x and a demonstration $y^* = (y_1^*, \ldots, y_T^*)$, let $s_t = (y_{<t}^*, x)$ denote the teacher-forced state at token position t, where y_t^* is the reference token. For brevity, we write $\pi_\theta(y_t) = \pi_\theta(y_t \mid y_{<t}^*, x)$. Let $q \in \Delta(\mathcal{Y})$ be a candidate distribution, and $\mathcal{C}$ an admissible set induced by the learning signal.

The Policy Projection Framework. Within this design space, we instantiate the local projection target as the KL-minimal policy within the admissible set,

$$q_t = \arg\min_{q \in \mathcal{C}_t} D_{\mathrm{KL}}\left(q \,\|\, \pi_\theta(\cdot \mid s_t)\right). \tag{5}$$

Given the token-wise design variables q_t and Z_t, we realize their combined effect through the shared policy update,

$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\left[\sum\nolimits_{t=1}^{T} Z_t\, \mathbb{E}_{y_t\sim q_t}\left[\nabla_\theta \log \pi_\theta(y_t \mid s_t)\right]\right]. \tag{6}$$

Here, q_t determines the local policy displacement, whereas Z_t determines how strongly this projection influences the shared optimization. Both are constructed token-wise based on the predictive state from the frozen current policy, and thus treated as stop gradient during the shared update. Under teacher forcing, the KL of the induced product target decomposes additively across token positions, allowing the local targets to be constructed independently before being jointly fitted by the shared policy (see Appendix A.1). This framework provides a design space for us to further improve supervised fine-tuning in Sec. 4.2 and Sec. 4.3.

SFT as a Degenerate Projection. From the above lens, standard SFT can be recovered by imposing the hard token-level requirement

$$\mathcal{C}_t^{\mathrm{SFT}} = \{q \in \Delta(\mathcal{Y}) : \mathbb{E}_{y\sim q}[R_t(y)] = 1\}, \quad \text{where} \quad R_t(y) = \mathbb{1}[y = y_t^*] \tag{7}$$

Its KL-minimal target is therefore

$$\hat{q}_t^{\mathrm{SFT}} = \arg\min_{q \in \mathcal{C}_t^{\mathrm{SFT}}} D_{\mathrm{KL}}\left(q \,\|\, \pi_\theta(\cdot \mid s_t)\right) = \delta_{y_t^*}, \qquad Z_t^{\mathrm{SFT}} = 1, \tag{8}$$

where $\delta_{y_t^*}$ denotes the Dirac distribution concentrated at the reference token y_t^*. Substituting these choices back into the framework yields

$$\nabla_\theta \mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\left[\sum\nolimits_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t^* \mid s_t)\right]. \tag{9}$$

This view exposes the two rigid choices made by vanilla SFT: its feasible set collapses the projection target to the reference token, and every token is assigned the same projection strength. See Appendix A.2 for detail.

4.2 RELAXED PROJECTION TARGET

Standard SFT enforces a hard token-level supervision constraint that projects the local policy fully onto the reference token, regardless of its compatibility with the current policy. This rigid target can

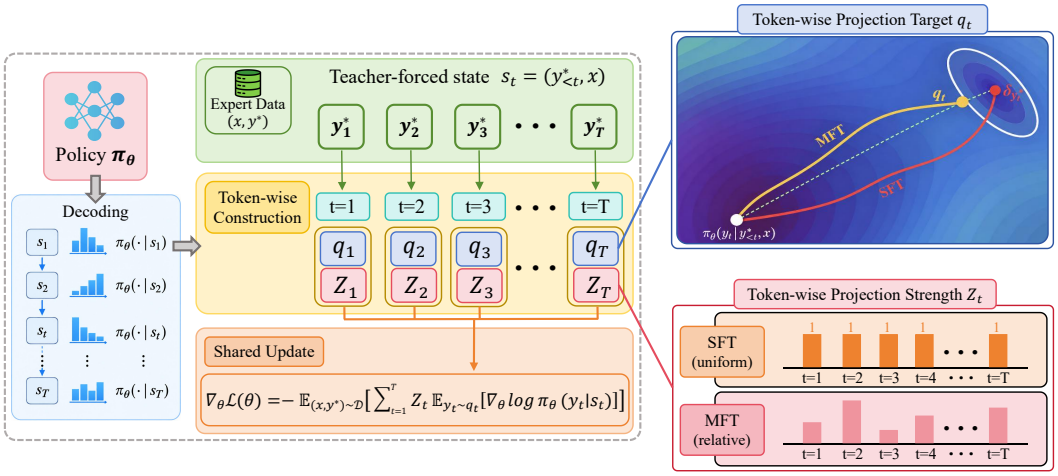


Figure 2: **Overview of our proposed MFT.** Given a teacher-forced state s_t, MFT constructs a token-wise projection target q_t and projection strength Z_t at each decoding state, which are aggregated through a shared sequence-level policy update. Standard SFT uses a full reference-token projection with uniform strength, whereas MFT adaptively relaxes the projection target according to policy compatibility and modulates its strength across supervised positions. Together, these two factors enable selective supervised adaptation while limiting unnecessary policy shift.

induce unnecessary policy displacement, especially when the demonstrated token strongly conflicts with the model's current prediction. Under our proposed framework in Eq.(6), we first focus on the design of the projection target q_t by relaxing the supervision requirement while retaining the minimum-shift construction, so that the resulting projection can adapt to the current policy rather than being fixed to $\delta_{y_t^*}$. Specifically, we introduce a mechanism to softly control the fraction of the remaining target-probability gap that the projection must close.

Relaxed Projection Policy Design. We introduce a parameter $w_t \in [0, 1]$ to parameterize the correction strength over the remaining probability gap towards the target y_t^*. We define the relaxed convex constraint set as follows:

$$\mathcal{C}_t(w_t) = \left\{ q \in \Delta(\mathcal{Y}) : \frac{q(y_t^*) - \pi_\theta(y_t^*)}{1 - \pi_\theta(y_t^*)} \geq w_t \right\}. \tag{10}$$

Applying the KL-minimal projection in our framework to this relaxed feasible set yields

$$\widehat{q}_t = \arg\min_{q \in \mathcal{C}_t(w_t)} D_{\mathrm{KL}}(q \,\|\, \pi_\theta) = (1 - w_t)\pi_\theta + w_t \delta_{y_t^*} \tag{11}$$

Importantly, this linear interpolation is not imposed as a heuristic smoothing rule, but emerges as the KL-minimal solution under the relaxed constraint. This closed form defines a simple one-dimensional projection path between the current policy and the hard SFT target $\delta_{y_t^*}$, with w_t specifying the location of the relaxed target along this path. This preserves the minimum-shift structure of our framework while introducing a controllable alternative to the degenerate SFT endpoint dynamically. The derivation is provided in Appendix A.3.

Remark 1 (Non-Target Structure Preservation)**.** *For the relaxed projection target $\hat{q}_t$ in Eq.(11), the conditional distribution over all non-target tokens is exactly preserved:*

$$\widehat{q}_t(y_t \mid y_t \neq y_t^*) = \pi_\theta(y_t \mid y_t \neq y_t^*), \qquad \forall\, y_t \neq y_t^*. \tag{12}$$

Moreover, the KL displacement reduces to the corresponding binary divergence,

$$D_{\mathrm{KL}}(\widehat{q}_t | \pi_\theta) = D_{\mathrm{KL}}(\mathrm{Bern}(\widehat{q}_t(y_t^*)) \| \mathrm{Bern}(\pi_\theta(y_t^*))) \,. \tag{13}$$

Thus, the relaxation in Eq.(10) modifies only the allocation of probability mass in $\hat{q}_t$ between the demonstrated token and its complement, while preserving the relative structure within the non-target space. The detailed proof can be found in Appendix A.4.

Compatibility-Guided Relaxation Schedule. Directly setting the correction strength from the raw target probability can be sensitive to state-dependent softmax normalization, making the same target

probability difficult to interpret consistently across heterogeneous prediction states. We therefore measure target compatibility relative to the local policy mode, which cancels the softmax normalization factor and yields the relative confidence ratio r_t:

$$r_t = \frac{\pi_\theta(y_t^*)}{\max_{y_t \in \mathcal{Y}} \pi_\theta(y_t)} \in (0, 1]. \tag{14}$$

The ratio r_t measures the alignment between the demonstrated token and the policy mode. Greater alignment (larger r_t) induces a more aggressive projection, whereas greater disagreement leads to a more conservative movement. With this ratio, we then set the parameter w_t in Eq.(10) to be $w_t := r_t^\lambda$, where $\lambda \in [0, 1]$ controls how sensitively this compatibility is translated into the relaxation level. Larger λ produces stronger separation between compatible and conflicting targets.

Remark 2 (Relative Confidence as Logit Margin)**.** *Let $z_t(y_t)$ denote the pre-softmax logit of token y_t at state $(y^*_{<t}, x)$, and define the margin between the policy mode and the demonstrated token as*

$$\Delta_t = \max_{y_t \in \mathcal{Y}} z_t(y_t) - z_t(y_t^*) \geq 0. \tag{15}$$

Then the relative confidence and the corresponding relaxation strength satisfy

$$w_t = \exp(-\lambda \Delta_t). \tag{16}$$

This ratio form eliminates the softmax normalization factor and depends solely on the logit margin between the demonstrated token and the policy mode. The proof and further discussions of the consequential properties are provided in Appendix A.5.

Remark 3 (KL-Regularized Interpretation of Relaxed Projection)**.** *$\widehat{q}_t$ is equivalently the solution of a KL-regularized binary-reward policy optimization problem:*

$$\widehat{q}_t = \arg\max_{q \in \Delta(\mathcal{Y})} \left\{ \mathbb{E}_{y_t \sim q} \left[\mathbf{1}(y_t = y_t^*) \right] - \beta_t D_{\mathrm{KL}}(q \| \pi_\theta) \right\}, \tag{17}$$

where

$$\beta_t^{-1} = \mathrm{logit}(\widehat{q}_t(y_t^*)) - \mathrm{logit}(\pi_\theta(y_t^*)). \tag{18}$$

While vanilla SFT can be reinterpreted as the zero-regularization limit:

$$q_t^{\mathrm{SFT}} = \delta_{y_t^*} = \lim_{\beta \to 0^+} \arg\max_{q \in \Delta(\mathcal{Y})} \left\{ \mathbb{E}_{y_t \sim q} \left[\mathbf{1}(y_t = y_t^*) \right] - \beta D_{\mathrm{KL}}(q \| \pi_\theta) \right\}. \tag{19}$$

This equivalence provides an optimization interpretation of the feasible-set relaxation introduced above. It shows that the relaxed projection can be viewed as an adaptive KL-regularized proximal update: it maximizes progress toward the demonstrated token under a finite policy-displacement budget. In particular, β_t^{-1} is exactly the additional logit preference that the relaxed projection assigns to the demonstrated token relative to every non-target token, while β_t characterizes the corresponding effective KL regularization against policy displacement. In contrast, vanilla SFT is recovered as the zero-regularization limit, which drives the projection directly to the degenerate target $\delta_{y_t^*}$. See Appendix A.6 for the detailed derivation.

4.3 Information-Aware Projection Strength

We now address the second design variable of the framework in Eq.(6), the projection strength Z_t, which controls the relative contribution of this local projection to the shared policy update. Unlike standard SFT, which assigns a uniform projection strength regardless of the model's predictive state, we make the strength reflect the supervised target's contribution to the model's predictive uncertainty. This is achieved by coupling the model's predictive confidence in the target token with its information distance from the current policy:

$$Z_t = \pi_\theta(y_t^*) D_{\mathrm{KL}} \left(\delta_{y_t^*} \parallel \pi_\theta \right). \tag{20}$$

To avoid the vanishing projection strength problem when the target is already highly probable, we therefore adjust the information distance through a smooth log-sum-exp transformation between a constant unit-scale floor and the KL distance (see Appendix A.7).

$$\widehat{Z}_t = \pi_\theta(y_t^*) \log \left(e + \exp D_{\mathrm{KL}} \left(\delta_{y_t^*} \parallel \pi_\theta \right) \right). \tag{21}$$

Combining the relaxed projection target in Eq.(11) with the information-aware strength in Eq.(21) completes MFT under the framework of Eq.(6). The two mechanisms act on distinct stages of adaptation: the former determines the token-level distribution to be projected toward, while the latter determines its relative contribution to the shared policy update. The complete formulation of the MFT policy update is provided in Appendix A.8.

Table 1: Performance comparison of fine-tuning methods on math reasoning and general tasks. The best overall performance is in **bold**, and the best performance among offline methods is underlined.

	Mathematical Reasoning Tasks						General Tasks			
	AIME26	**AIME25**	**Math500**	**Olympiad**	**Minerva**	**Avg.**	**MMLU**	**ARC-C**	**GPQA**	**Avg.**
Base Model	3.96	4.79	48.34	21.31	16.94	19.07	38.1	32.4	29.8	33.4
Supervised Fine-Tuning										
SFT	3.12	5.20	61.09	16.26	23.52	23.84	37.3	48.8	29.1	38.4
DFT	7.51	7.92	70.14	34.66	29.24	29.89	56.2	42.4	39.2	45.9
ASFT	7.91	7.08	69.63	34.36	27.22	29.24	52.6	39.1	34.5	42.1
PSFT	6.25	5.83	58.71	24.53	20.06	23.08	39.6	38.3	31.3	36.4
MFT (Ours)	9.18	9.79	71.03	35.17	28.81	30.80	58.4	**49.1**	40.2	49.2
Hybrid Training										
LUFFY	9.58	13.34	78.06	40.91	33.29	35.04	55.3	39.8	38.3	44.5
CHORD	11.25	13.11	79.14	42.11	35.99	36.32	58.8	40.9	41.4	47.0
Reinforcement Learning										
SFT + RL	8.95	12.50	77.90	40.96	33.02	34.67	58.4	39.6	43.5	47.2
DFT + RL	10.84	12.51	79.81	42.53	33.72	35.88	59.7	44.6	**43.8**	49.4
ASFT + RL	11.04	12.09	79.65	42.19	34.51	35.90	59.8	44.9	43.3	49.3
PSFT + RL	10.21	11.45	77.43	41.43	34.44	34.93	58.3	39.4	40.1	45.9
MFT + RL (Ours)	**12.08**	**13.95**	**80.74**	**43.70**	**38.22**	**37.74**	**60.3**	46.4	43.6	**50.1**

5 EXPERIMENTS

5.1 SETUP

Models, Datasets, and Evaluations. Our main experiments are conducted based on Qwen3-8B-Base (Yang et al., 2025). We use the NuminaMath-CoT (Li et al., 2024a) dataset for training by randomly selecting 20k instances for SFT and 10k instances for RL. (i) For math reasoning tasks, we test the models' performance on AIME 2026 (Zhang & Math-AI, 2026), AIME 2025 (Zhang & Math-AI, 2025), Math500 (Hendrycks et al., 2021b), OlympiadBench (He et al., 2024), and Minerva Math (Lewkowycz et al., 2022) to monitor the in-domain generalization performance. (ii) For general tasks, we assess the models' extent of knowledge retention and generalization ability on MMLU-Pro (Wang et al., 2024), ARC-Challenge (Clark et al., 2018), and GPQA-Diamond (Rein et al., 2024). Implementation details are provided in Appendix B.1.

Baselines. We compare MFT across three post-training settings. (i) Supervised fine-tuning uses only fixed expert demonstrations and does not require policy rollouts during training. In this setting, we compare against standard SFT and recent stabilized variants, including DFT (Wu et al., 2026), ASFT (Zhu et al., 2026a), and PSFT (Zhu et al., 2026b). (ii) Reinforcement learning instead trains from trajectories sampled by the current policy and optimized with reward feedback. Here, we apply GRPO (Shao et al., 2024) warm-started by each fine-tuning method to evaluate its effectiveness in initializing subsequent RL. (iii) Hybrid Training uses expert demonstrations together with on-policy rollouts within the RL stage, including LUFFY (Yan et al., 2025) and CHORD (Zhang et al., 2026).

5.2 MAIN RESULTS

Table 1 summarizes the results on mathematical reasoning and general domain tasks. We have the following observations. (1) On mathematical reasoning tasks, MFT substantially improves standard SFT, raising the average score from 23.84 to 30.80 (+6.96) and also surpassing recent competitive fine-tuning methods on average, showing that the proposed modulation consistently improves supervised adaptation on the target reasoning domain. (2) The gains are more pronounced on general domain benchmarks, where MFT improves the average from 38.4 to 49.2 over SFT (+10.8) and exceeds the strongest competing fine-tuning baseline by +3.3. Since these tasks lie outside the fine-tuning domain, the result suggests stronger transfer rather than merely better fitting of the demonstrations. This is consistent with MFT's selective adaptation design, which emphasizes informative, policy-compatible supervision while limiting excessive fitting and policy shift on strongly conflicting targets, thereby better preserving useful pretrained structure. (3) MFT also provides a stronger initialization for the subsequent reinforcement learning. Starting from MFT, GRPO reaches

Table 3: Performance comparison of fine-tuning methods on code generation tasks. The best overall performance is in **bold**, and the best performance among offline methods is underlined.

	HumanEval		MultiPL-E								
	HE	HE+	Python	C++	Java	PHP	TS	C#	Bash	JS	Avg.
Base	48.2	40.9	45.73	39.13	34.81	47.20	38.36	34.18	**25.32**	47.20	38.99
SFT	61.0	55.5	60.98	44.72	44.30	45.34	53.46	36.08	20.25	52.17	44.66
MFT	63.4	59.8	65.24	45.34	**46.84**	47.83	54.09	35.44	18.99	**54.66**	46.05
SFT + RL	64.0	59.1	57.93	47.83	43.04	49.69	54.72	36.08	22.78	53.42	45.69
MFT + RL	**66.5**	**61.0**	**65.85**	**48.45**	42.41	**50.93**	**55.35**	**37.34**	22.15	54.04	**47.07**

37.74 on mathematical reasoning and 50.1 on general tasks, improving over SFT + GRPO by +3.07 and +2.9, respectively, and achieving the best overall averages among the compared methods. This suggests that MFT places the policy closer to useful reasoning trajectories, enabling RL to focus on reinforcing and refining successful reasoning paths rather than correcting distortions introduced during supervised fine-tuning, and thus supporting a more effective transition from demonstration learning to reward-driven optimization.

5.3 Efficacy on Diverse Tasks

To assess the generalizability of our method, we extend our evaluation to other representative scenarios. (i) For **code generation**, we select 10k samples for SFT and 10k for RL from Magicoder (Wei et al., 2024). Model performance is assessed on three widely used code generation benchmarks: HumanEval (Chen et al., 2021), HumanEval+ (Liu et al., 2023), and MultiPL-E (Cassano et al., 2023). (ii) For **medical knowledge**, we sample 30k instances for SFT and 10k for RL from MedMCQA (Pal et al., 2022). We assess the models' performance on MedQA (Jin et al., 2021), MMLU-medical (Hendrycks et al., 2021a) and the MedMCQA test set. Details are provided in Appendix B.2.

Table 2: Performance comparison on medical tasks. Best overall results are in **bold** and best offline results are underlined.

	MedQA	MMLU-Med	MedMCQA	Avg.
Base	41.95	64.88	44.01	50.28
SFT	44.38	62.29	44.15	50.27
MFT	45.17	67.83	45.04	52.68
SFT + RL	46.35	67.83	44.56	52.91
MFT + RL	**47.45**	**69.33**	**45.54**	**54.11**

As shown in Tab. 3, MFT improves the overall performance on HumanEval and the multilingual MultiPL-E benchmarks, with gains spanning diverse programming languages. This suggests that MFT improves the transfer of learned coding capabilities rather than over-specializing to the fine-tuning distribution. Tab. 2 reveals that MFT effectively mitigates the performance degradation induced by vanilla SFT on knowledge-intensive datasets such as MMLU-Medical, achieving better results across all three medical tasks. Moreover, MFT provides a stronger initialization for subsequent RL in both domains, yielding stronger overall performance after RL.

5.4 Ablation Study

To understand the contribution of each design choice in our dual mechanisms, we conduct comprehensive ablation studies on the mathematical reasoning and general reasoning tasks. Detailed analyses are provided in Appendix B.3.

Effect of Dual Mechanisms. We first isolate the impact of the relaxed I-projection target and the information-calibrated projection strength. As demonstrated in Tab. 4, removing either mechanism degrades performance, while removing both degenerates the model to the vanilla SFT baseline. This confirms that dynamically modulating how far to update and how strongly to update are complementary, jointly improving in-domain reasoning performance and cross-domain generalization.

Sensitivity to Relaxation Schedule. We evaluate the effect of the hyperparameter λ, which controls the aggressiveness of the target relaxation $w_t = r_t^{\lambda}$. Tab. 5 shows that intermediate values of λ outperform both the unrelaxed extreme and the fully linear scaling. This highlights that our nonlinear schedule provides sufficient correction for imperfectly supported targets while suppressing overly aggressive updates when the reference token is strongly misaligned with the policy mode.

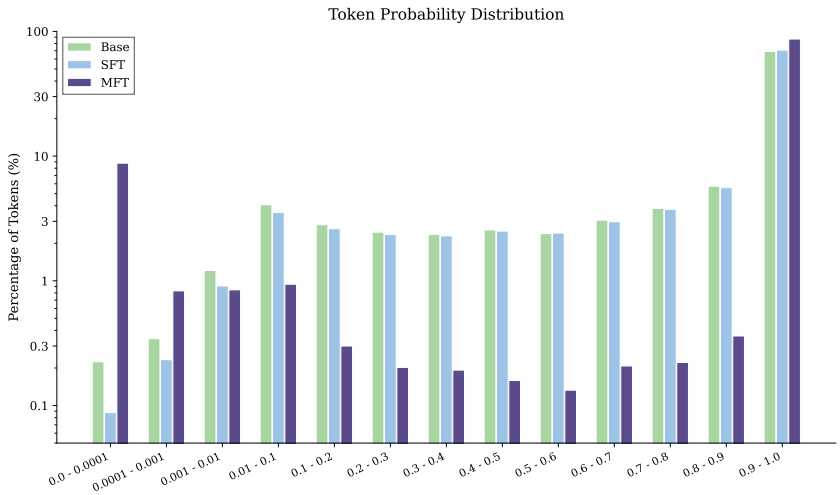


Figure 3: Comparison of token probability distributions on the training set before training and after fine-tuning between SFT and MFT.

Table 4: Ablation study of our dual mechanisms across mathematical and general reasoning tasks. The best overall performance is in **bold**.

	Mathematical Reasoning Tasks						General Tasks			
	AIME26	**AIME25**	**Math500**	**Olympiad**	**Minerva**	**Avg.**	**MMLU**	**ARC-C**	**GPQA**	**Avg.**
w/o. $\hat{q}$, $\hat{Z}$	0.41	0.41	33.76	8.16	9.82	10.51	23.0	**47.7**	25.0	31.9
w/o. $\hat{q}$	1.86	1.45	55.04	19.98	19.13	19.49	32.0	42.5	25.7	33.4
w/o. $\hat{Z}$	1.66	1.24	52.25	18.56	17.73	18.29	29.4	42.2	25.9	32.5
full	**3.32**	**1.86**	**55.41**	**20.34**	**20.38**	**20.26**	**34.9**	42.1	**26.7**	**34.6**

Design of the Projection Strength. Finally, we compare our proposed information-preserving calibration against alternative weighting heuristics. As shown in Tab. 6, our formulation achieves better overall accuracy. This validates the probability-information coupling underlying our projection strength, which preserves informative learning signals while avoiding excessively aggressive updates in highly conflicting regimes.

5.5 Token-Level Distributional Pattern Analysis

Fig. 3 further shows that MFT induces a substantially different adaptation pattern from vanilla SFT. While SFT broadly shifts supervised tokens toward higher probabilities, MFT yields a more polarized distribution, with fewer tokens remaining at intermediate confidence and more concentrated at either high or very low probability. This behavior is consistent with our selective projection mechanism: supervision that is already compatible with the current policy is reinforced toward higher confidence, whereas strongly conflicting or weakly supported tokens receive more conservative updates instead of being uniformly fitted. The resulting redistribution suggests that MFT does not increase confidence across all supervised tokens, but selectively reshapes their probabilities according to the policy state, helping explain its stronger generalization over standard SFT.

6 Conclusion

We present a structured policy projection framework that characterizes fine-tuning as a KL-minimal policy projection followed by a target-strength decomposition of the resulting update. Under this view, standard SFT emerges as a special case with a degenerate target and uniform projection strength, revealing limited flexibility in controlling policy displacement across heterogeneous token states. Building on this diagnosis, we proposed Modulated Fine-Tuning (MFT), which relaxes the projection target according to relative policy confidence and modulates the update strength through target-specific information. Experiments across several fields demonstrate that MFT exhibits not only better in-domain performance but also superior generalization capability.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing of the paper. We have not used generative AI tools for coming up idea, and auto research are not applicable to this work. Additionally, we used generative AI tools for checking and correcting errors. We have reviewed all AI-assisted work. LLM-generated sentence was verified by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The sources of both the training datasets and benchmarks have been provided in Sec. 5.1. We have also provided detailed training and evaluation parameters in Appendix B.1 to make the results reproducible.

REFERENCES

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q Feldman, et al. Multipl-e: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Transactions on Software Engineering*, 49(7):3675–3691, 2023.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde 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, Josh 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. *arXiv preprint arXiv:2107.03374*, 2021.

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

Rohan Deb, Kiran Koshy Thekumparampil, Kousha Kalantari, Gaurush Hiranandani, Shoham Sabach, and Branislav Kveton. FisherSFT: Data-efficient supervised fine-tuning of language models using information gain. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=e02oLEbehE.

Muxi Diao, Lele Yang, Wuxuan Gong, Yutong Zhang, Zhonghao Yan, Yufei Han, Kongming Liang, Weiran Xu, and Zhanyu Ma. Entropy-adaptive fine-tuning: Resolving confident conflicts to mitigate forgetting. *arXiv preprint arXiv:2601.02151*, 2026.

Wangjie Gan, Miao Pan, Linbo Xi, Wenqi Zhang, Jintao Chen, Jianwei Yin, and Xuhong Zhang. GFT: From imitation to reward fine-tuning with unbiased group advantages and dynamic coefficient rectification. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28909–28922, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1444. URL https://aclanthology.org/2026.findings-acl.1444/.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *International conference on machine learning*, pp. 2160–2169. PMLR, 2019.

Xuan Gong, Senmiao Wang, Hanbo Huang, Ruoyu Sun, and Shiyu Liang. VCORE: Variance-controlled optimization-based reweighting for chain-of-thought supervision. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28161–28180, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1298. URL https://aclanthology.org/2026.acl-long.1298/.

Yifan HAO, Xingyuan Pan, Hanning Zhang, Chenlu Ye, Rui Pan, and Tong Zhang. Understanding overadaptation in supervised fine-tuning: The role of ensemble methods. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=1xsW6tvMb3.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. 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)*, pp. 3828–3850, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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)*, 2021b. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

Roman Klypa and Oleksandr Cherednichenko. Diversity in large language models under supervised fine-tuning. *arXiv preprint arXiv:2605.00195*, 2026.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh 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 Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=IFXTZERXdM7.

Gaotang Li, Ruizhong Qiu, Xiusi Chen, Heng Ji, and Hanghang Tong. Beyond log likelihood: Probability-based objectives for supervised fine-tuning across the model capability continuum. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2hQBG2ZlFb.

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. *Hugging Face repository*, 13(9):9, 2024a.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Ruoyu Sun, and Zhi-Quan Luo. Entropic distribution matching for supervised fine-tuning of LLMs: Less overfitting and better diversity. In *NeurIPS 2024 Workshop on Fine-Tuning in Modern Machine Learning: Principles and Scalability*, 2024b. URL https://openreview.net/forum?id=dulz3WVhMR.

Ziniu Li, Congliang Chen, Tian Xu, Zeyu Qin, Jiancong Xiao, Zhi-Quan Luo, and Ruoyu Sun. Preserving diversity in supervised fine-tuning of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NQEe7B7bSw.

Jiacheng Lin, Zhongruo Wang, Kun Qian, Tian Wang, Arvind Srinivasan, Hansi Zeng, Ruochen Jiao, Xie Zhou, Jiri Gesi, Dakuo Wang, Yufan Guo, Kai Zhong, Weiqi Zhang, sujay sanghavi, Changyou Chen, Hyokun Yun, and Lihong Li. SFT doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ru93xpQFi1.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and LINGMING ZHANG. Is your code generated by chatGPT really correct? rigorous evaluation of large language models for code generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=1qvx610Cu7.

Rui Ming, Haoyuan Wu, Shoubo Hu, Zhuolun He, and Bei Yu. One-token rollout: Guiding supervised fine-tuning of llms with policy gradient. *arXiv preprint arXiv:2509.26313*, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei, Hao Cheng, Chen Qian, and Yang Liu. Token cleaning: Fine-grained data selection for LLM supervised fine-tuning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=tXkOUS3vLS.

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Mahdi Sabbaghi, George Pappas, Adel Javanmard, and Hamed Hassani. Infosft: Learn more and forget less with information-aware token weighting. *arXiv preprint arXiv:2605.14967*, 2026.

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

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

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7HNRYT4V44.

Bo Wang, Qinyuan Cheng, Runyu Peng, Rong Bao, Peiji Li, Qipeng Guo, Linyang Li, Zhiyuan Zeng, Yunhua Zhou, and Xipeng Qiu. Implicit reward as the bridge: A unified view of SFT and DPO connections. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=xUx2B2NHvj.

Hao Wang, Hao Gu, Hongming Piao, Kaixiong Gong, Yuxiao Ye, Xiangyu Yue, Sirui Han, Yike Guo, and Dapeng Wu. Learning while staying curious: Entropy-preserving supervised fine-tuning via adaptive self-distillation for large reasoning models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13567–13581, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.617. URL https://aclanthology.org/2026.acl-long.617/.

Xinyu Wang, Changzhi Sun, Yuanbin Wu, and Xiaoling Wang. Stabilizing llm supervised fine-tuning via explicit distributional control. *arXiv preprint arXiv:2605.04468*, 2026b.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-pro: A more robust and challenging multi-task language understanding benchmark. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=y10DM6R2r3.

Zecheng Wang, Deyuan Liu, Chunshan Li, Yupeng Zhang, Zhengyun Zhao, Dianhui Chu, Bingning Wang, and Dianbo Sui. Gradients must earn their influence: Unifying sft with generalized entropic objectives. *arXiv preprint arXiv:2602.11424*, 2026c.

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. URL https://openreview.net/forum?id=gEZrGCozdqR.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and LINGMING ZHANG. Magicoder: Empowering code generation with OSS-instruct. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=XUeoOBid3x.

Yongliang Wu, Yizhou Zhou, Zhou Ziheng, 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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Lv7PjbcaMi.

Chao Xue, Yao Wang, Mengqiao Liu, Di Liang, Xingsheng Han, Peiyang Liu, Xianjie Wu, Chenyao Lu, Lei Jiang, Yu Lu, Haibo Shi, Shuang Liang, Minlong Peng, and Flora D. Salim. Why supervised fine-tuning fails to learn: A systematic study of incomplete learning in large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30186–30213, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1393. URL https://aclanthology.org/2026.acl-long.1393/.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=vO8LLoNWWk.

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. *arXiv preprint arXiv:2505.09388*, 2025.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy RL meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=dCm9bBrk5d.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2025, 2025.

Yifan Zhang and Team Math-AI. American invitational mathematics examination (aime) 2026, 2026.

Qihuang Zhong, Liang Ding, Xiantao Cai, Juhua Liu, Bo Du, and Dacheng Tao. KaFT: Knowledge-aware fine-tuning for boosting LLMs' domain-specific question-answering performance. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 24085–24100, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1235. URL https://aclanthology.org/2025.findings-acl.1235/.

Yucheng Zhou, Junwei Sheng, Qianning Wang, and Jianbing Shen. Compatibility-aware dynamic fine-tuning for large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29997–30008, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1383. URL https://aclanthology.org/2026.acl-long.1383/.

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. Anchored supervised fine-tuning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=PORko7QT64.

Wenhong Zhu, Ruobing Xie, Rui Wang, Xingwu Sun, Di Wang, and Pengfei Liu. Proximal supervised fine-tuning. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=hQtwQqYikp.

A PROOFS

A.1 TEACHER-FORCED SEQUENCE EXTENSION

In the main framework, each projection target is constructed locally at a teacher-forced token state, while all token-wise projections are realized jointly through the shared sequence-level policy update. We therefore verify that, under teacher forcing, this local construction is consistent with the sequence-level objective by showing that the corresponding KL divergence decomposes exactly across token positions.

Teacher-Forced Sequence Extension. The minimum-shift targets can be constructed independently at each token position under state $s_t = (y^*_{<t}, x)$ and jointly realized through the shared sequence-level policy update.

$$D_{\mathrm{KL}}\left(\prod_{t=1}^{T} \widehat{q}_t \,\middle\|\, \prod_{t=1}^{T} \pi_\theta(\cdot \mid s_t)\right) = \sum_{t=1}^{T} D_{\mathrm{KL}}\left(\widehat{q}_t \,\|\, \pi_\theta(\cdot \mid s_t)\right). \tag{22}$$

Proof. Decompose the KL divergence:

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(\prod_{t=1}^{T}\widehat{q}_t \,\middle\|\, \prod_{t=1}^{T}\pi_\theta(\cdot \mid s_t)\right) \\
&= \mathbb{E}_{y_{1:T}\sim\widehat{q}}\left[\log \frac{\prod_{t=1}^{T}\widehat{q}_t(y_t \mid s_t)}{\prod_{t=1}^{T}\pi_\theta(y_t \mid s_t)}\right] \\
&= \sum_{t=1}^{T}\mathbb{E}_{y_t\sim\widehat{q}_t}\left[\log \frac{\widehat{q}_t(y_t \mid s_t)}{\pi_\theta(y_t \mid s_t)}\right] \\
&= \sum_{t=1}^{T} D_{\mathrm{KL}}\left(\widehat{q}_t \,\|\, \pi_\theta(\cdot \mid s_t)\right). \qquad (23)
\end{aligned}
$$

□

A.2 Derivation of SFT under the Framework.

We show that standard SFT is exactly a special case of the proposed framework. At each teacher-forced state $s_t = (y^*_{<t}, x)$, define the binary supervision signal

$$R_t(y_t) = \mathbb{1}[y_t = y^*_t], \qquad (24)$$

and impose the hard feasible set

$$\mathcal{C}^{\mathrm{SFT}}_t = \{q \in \Delta(\mathcal{Y}) : \mathbb{E}_{y_t\sim q}[R_t(y_t)] = 1\}. \qquad (25)$$

Since

$$\mathbb{E}_{y_t\sim q}[R_t(y_t)] = \sum_{y_t\in\mathcal{Y}} q(y_t)\mathbb{1}[y_t = y^*_t] = q(y^*_t), \qquad (26)$$

the constraint is equivalent to

$$q(y^*_t) = 1. \qquad (27)$$

This further implies

$$q(y_t) = 0, \qquad \forall y_t \neq y^*_t. \qquad (28)$$

Hence the feasible set contains a single distribution,

$$\mathcal{C}^{\mathrm{SFT}}_t = \{\delta_{y^*_t}\}. \qquad (29)$$

The corresponding minimum-shift projection is therefore trivially

$$\widehat{q}^{\mathrm{SFT}}_t = \arg\min_{q\in\mathcal{C}^{\mathrm{SFT}}_t} D_{\mathrm{KL}}\left(q \,\|\, \pi_\theta(\cdot \mid s_t)\right) = \delta_{y^*_t}. \qquad (30)$$

Standard SFT assigns equal influence to every supervised position, which corresponds to the uniform projection strength

$$Z^{\mathrm{SFT}}_t = 1. \qquad (31)$$

Substituting $\widehat{q}^{\mathrm{SFT}}_t = \delta_{y^*_t}$ and $Z^{\mathrm{SFT}}_t = 1$ into the general sequence-level update gives

$$
\begin{aligned}
\nabla_\theta \mathcal{L}_{\mathrm{SFT}} &= -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\left[\sum_{t=1}^{T}\mathbb{E}_{y_t\sim\delta_{y^*_t}}\left[\nabla_\theta \log \pi_\theta(y_t \mid s_t)\right]\right] \\
&= -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\left[\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(y^*_t \mid s_t)\right], \qquad (32)
\end{aligned}
$$

which is exactly the gradient of the standard teacher-forced negative log-likelihood objective

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\left[\sum_{t=1}^{T}\log \pi_\theta(y^*_t \mid s_t)\right]. \qquad (33)$$

A.3 Derivation of the Relaxed Projection

We provide the derivation of the resulting target under the relaxed feasible set as follows.

Proof. For $y \neq y_t^*$, define the normalized conditional distributions:

$$\tilde{q}(y_t) = \frac{q(y_t)}{1 - q(y_t^*)}, \qquad \tilde{\pi}_\theta(y_t) = \frac{\pi_\theta(y_t)}{1 - \pi_\theta(y_t^*)}. \tag{34}$$

By the chain rule for relative entropy, the objective decomposes into a binary divergence and a conditional divergence:

$$\begin{aligned}
D_{\mathrm{KL}}(q \,\|\, \pi_\theta) &= \sum_{y_t \in \mathcal{Y}} q(y_t) \log \frac{q(y_t)}{\pi_\theta(y_t)} \\
&= q(y_t^*) \log \frac{q(y_t^*)}{\pi_\theta(y_t^*)} + \sum_{y_t \neq y_t^*} q(y_t) \log \frac{q(y_t)}{\pi_\theta(y_t)} \\
&= q(y_t^*) \log \frac{q(y_t^*)}{\pi_\theta(y_t^*)} + \sum_{y_t \neq y_t^*} (1 - q(y_t^*)) \tilde{q}(y) \log \left(\frac{1 - q(y_t^*)}{1 - \pi_\theta(y_t^*)} \cdot \frac{\tilde{q}(y_t)}{\tilde{\pi}_\theta(y_t)} \right) \\
&= q(y_t^*) \log \frac{q(y_t^*)}{\pi_\theta(y_t^*)} + (1 - q(y_t^*)) \log \frac{1 - q(y_t^*)}{1 - \pi_\theta(y_t^*)} \underbrace{\sum_{y_t \neq y_t^*} \tilde{q}(y_t)}_{=1} \qquad (35) \\
&\quad + (1 - q(y_t^*)) \sum_{y_t \neq y_t^*} \tilde{q}(y_t) \log \frac{\tilde{q}(y_t)}{\tilde{\pi}_\theta(y_t)} \\
&= \underbrace{\left[q(y_t^*) \log \frac{q(y_t^*)}{\pi_\theta(y_t^*)} + (1 - q(y_t^*)) \log \frac{1 - q(y_t^*)}{1 - \pi_\theta(y_t^*)} \right]}_{= D_{\mathrm{KL}}(\mathrm{Bern}(q(y_t^*)) \,\|\, \mathrm{Bern}(\pi_\theta(y_t^*)))} \qquad (36) \\
&\quad + (1 - q(y_t^*)) \underbrace{\sum_{y_t \neq y_t^*} \tilde{q}(y_t) \log \frac{\tilde{q}(y_t)}{\tilde{\pi}_\theta(y_t)}}_{= D_{\mathrm{KL}}(\tilde{q} \,\|\, \tilde{\pi}_\theta)} \\
&= D_{\mathrm{KL}}\big(\mathrm{Bern}(q(y_t^*)) \,\|\, \mathrm{Bern}(\pi_\theta(y_t^*))\big) + (1 - q(y_t^*)) D_{\mathrm{KL}}(\tilde{q} \,\|\, \tilde{\pi}_\theta). \qquad (37)
\end{aligned}$$

For any fixed value of $q(y_t^*)$, the second term is strictly non-negative and is uniquely minimized when $\tilde{q} = \tilde{\pi}_\theta$. Therefore, the optimal distribution q on $\mathcal{Y}$ must satisfy:

$$q(y_t) = \begin{cases} q(y_t^*), & y_t = y_t^*, \\ (1 - q(y_t^*)) \dfrac{\pi_\theta(y_t)}{1 - \pi_\theta(y_t^*)}, & y_t \neq y_t^*, \end{cases} \tag{38}$$

which can be equivalently rewritten as a convex combination of $\delta_{y_t^*}$ and π_t:

$$q = \frac{q(y_t^*) - \pi_\theta(y_t^*)}{1 - \pi_\theta(y_t^*)} \delta_{y_t^*} + \left(1 - \frac{q(y_t^*) - \pi_\theta(y_t^*)}{1 - \pi_\theta(y_t^*)} \right) \pi_\theta. \tag{39}$$

Finally, the constraint requires:

$$\frac{q(y_t^*) - \pi_\theta(y_t^*)}{1 - \pi_\theta(y_t^*)} \geq w_t \implies q(y_t^*) \geq \pi_\theta(y_t^*) + w_t(1 - \pi_\theta(y_t^*)) \geq \pi_\theta(y_t^*). \tag{40}$$

Since the binary divergence $D_{\mathrm{KL}}\big(\mathrm{Bern}(q(y_t^*)) \,\|\, \mathrm{Bern}(\pi_\theta(y_t^*))\big)$ is strictly monotonically increasing for $q(y_t^*) \geq \pi_\theta(y_t^*))$, the minimum is uniquely attained at the lower boundary:

$$\widehat{q}_t(y_t^*) = \pi_\theta(y_t^*) + w_t(1 - \pi_\theta(y_t^*)). \tag{41}$$

Substituting $\widehat{q}_t(y_t^*)$ into the mixture formulation yields:

$$\widehat{q}_t = (1 - w_t)\pi_\theta + w_t \delta_{y_t^*}, \tag{42}$$

which concludes the proof. □

A.4 Derivation of the Structure Preservation

Remark 1 (Non-Target Structure Preservation). *For the relaxed projection target $\hat{q}_t$ in Eq.(11), the conditional distribution over all non-target tokens is exactly preserved:*

$$\widehat{q}_t(y_t \mid y_t \neq y_t^*) = \pi_\theta(y_t \mid y_t \neq y_t^*), \qquad \forall\, y_t \neq y_t^*. \tag{12}$$

Moreover, the KL displacement reduces to the corresponding binary divergence,

$$D_{\mathrm{KL}}(\widehat{q}_t|\pi_\theta) = D_{\mathrm{KL}}(\mathrm{Bern}(\widehat{q}_t(y_t^*))\|\mathrm{Bern}(\pi_\theta(y_t^*)))\,. \tag{13}$$

Proof. Given the relaxed projection target,

$$\widehat{q}_t = (1-w_t)\pi_\theta + w_t\delta_{y_t^*}. \tag{43}$$

Therefore, the target probability satisfies

$$\widehat{q}_t(y_t^*) = \pi_\theta(y_t^*) + w_t\big(1-\pi_\theta(y_t^*)\big), \tag{44}$$

while for every non-target token $y_t \neq y_t^*$,

$$\widehat{q}_t(y_t) = (1-w_t)\pi_\theta(y_t). \tag{45}$$

Thus,

$$1-\widehat{q}_t(y_t^*) = (1-w_t)\big(1-\pi_\theta(y_t^*)\big). \tag{46}$$

Hence, for $w_t \in [0,1)$, the conditional distribution over the non-target set satisfies

$$\begin{aligned}
\widehat{q}_t(y_t \mid y_t \neq y_t^*) &= \frac{\widehat{q}_t(y_t)}{1-\widehat{q}_t(y_t^*)} \\
&= \frac{(1-w_t)\pi_\theta(y_t)}{(1-w_t)\big(1-\pi_\theta(y_t^*)\big)} \\
&= \frac{\pi_\theta(y_t)}{1-\pi_\theta(y_t^*)} \\
&= \pi_\theta(y_t \mid y_t \neq y_t^*),
\end{aligned} \tag{47}$$

Consequently, for any two non-target tokens $y_t^{(i)}, y_t^{(j)} \neq y_t^*$,

$$\frac{\widehat{q}_t(y_t^{(i)})}{\widehat{q}_t(y_t^{(j)})} = \frac{(1-w_t)\pi_\theta(y_t^{(i)})}{(1-w_t)\pi_\theta(y_t^{(j)})} = \frac{\pi_\theta(y_t^{(i)})}{\pi_\theta(y_t^{(j)})}, \tag{48}$$

For the KL displacement,

$$\begin{aligned}
D_{\mathrm{KL}}(\widehat{q}_t\|\pi_\theta) &= \widehat{q}_t(y_t^*)\log\frac{\widehat{q}_t(y_t^*)}{\pi_\theta(y_t^*)} + \sum_{y_t\neq y_t^*}(1-w_t)\pi_\theta(y_t)\log\frac{(1-w_t)\pi_\theta(y_t)}{\pi_\theta(y_t)} \\
&= \widehat{q}_t(y_t^*)\log\frac{\widehat{q}_t(y_t^*)}{\pi_\theta(y_t^*)} + (1-w_t)\big(1-\pi_\theta(y_t^*)\big)\log(1-w_t) \\
&= \widehat{q}_t(y_t^*)\log\frac{\widehat{q}_t(y_t^*)}{\pi_\theta(y_t^*)} + \big(1-\widehat{q}_t(y_t^*)\big)\log\frac{1-\widehat{q}_t(y_t^*)}{1-\pi_\theta(y_t^*)} \\
&= D_{\mathrm{KL}}\left(\mathrm{Bern}(\widehat{q}_t(y_t^*))\,\|\,\mathrm{Bern}(\pi_\theta(y_t^*))\right)
\end{aligned} \tag{49}$$

□

A.5 Relative Confidence as Logit Margin

Remark 2 (Relative Confidence as Logit Margin). *Let $z_t(y_t)$ denote the pre-softmax logit of token y_t at state $(y_{<t}^*, x)$, and define the margin between the policy mode and the demonstrated token as*

$$\Delta_t = \max_{y_t\in\mathcal{Y}} z_t(y_t) - z_t(y_t^*) \geq 0. \tag{15}$$

Then the relative confidence and the corresponding relaxation strength satisfy

$$w_t = \exp(-\lambda\Delta_t). \tag{16}$$

Proof. Let

$$\pi_\theta(y_t \mid s_t) = \frac{\exp(z_t(y_t))}{\sum_{y' \in \mathcal{Y}} \exp(z_t(y'))}. \tag{50}$$

Denoting

$$z_t^{\max} = \max_{y_t \in \mathcal{Y}} z_t(y_t), \tag{51}$$

the probability of the policy mode is

$$\max_{y_t \in \mathcal{Y}} \pi_\theta(y_t) = \frac{\exp(z_t^{\max})}{\sum_{y' \in \mathcal{Y}} \exp(z_t(y'))}. \tag{52}$$

Therefore,

$$\begin{aligned} r_t &= \frac{\pi_\theta(y_t^*)}{\max_{y_t \in \mathcal{Y}} \pi_\theta(y_t)} \\ &= \frac{\exp(z_t(y_t^*))}{\exp(z_t^{\max})} \\ &= \exp(-\Delta_t). \end{aligned} \tag{53}$$

The softmax partition function cancels exactly. □

A.6 KL-Regularized Interpretation of Relaxed Projection

Remark 3 (KL-Regularized Interpretation of Relaxed Projection)**.** *$\widehat{q}_t$ is equivalently the solution of a KL-regularized binary-reward policy optimization problem:*

$$\widehat{q}_t = \arg\max_{q \in \Delta(\mathcal{Y})} \left\{ \mathbb{E}_{y_t \sim q} \left[\mathbf{1}(y_t = y_t^*) \right] - \beta_t D_{\mathrm{KL}}(q \| \pi_\theta) \right\}, \tag{17}$$

where

$$\beta_t^{-1} = \mathrm{logit}(\widehat{q}_t(y_t^*)) - \mathrm{logit}(\pi_\theta(y_t^*)). \tag{18}$$

While vanilla SFT can be reinterpreted as the zero-regularization limit:

$$q_t^{\mathrm{SFT}} = \delta_{y_t^*} = \lim_{\beta \to 0^+} \arg\max_{q \in \Delta(\mathcal{Y})} \left\{ \mathbb{E}_{y_t \sim q} \left[\mathbf{1}(y_t = y_t^*) \right] - \beta D_{\mathrm{KL}}(q \| \pi_\theta) \right\}. \tag{19}$$

Proof. Consider the KL-regularized binary-reward objective

$$\max_{q \in \Delta(\mathcal{Y})} \left\{ \mathbb{E}_{y_t \sim q} \left[\mathbf{1}(y_t = y_t^*) \right] - \beta_t D_{\mathrm{KL}} \left(q \,\|\, \pi_\theta \right) \right\}, \qquad \beta_t > 0. \tag{54}$$

Since the first term is linear in q and $D_{\mathrm{KL}}(q \| \pi_\theta)$ is strictly convex in q for $\pi_\theta(y_t) > 0$, the objective is strictly concave and therefore admits a unique optimum.

Expanding Eq.(54) gives

$$\max_{q \in \Delta(\mathcal{Y})} \left\{ \sum_{y_t \in \mathcal{Y}} q(y_t) \mathbf{1}(y_t = y_t^*) - \beta_t \sum_{y_t \in \mathcal{Y}} q(y_t) \log \frac{q(y_t)}{\pi_\theta(y_t)} \right\}. \tag{55}$$

Introducing a Lagrange multiplier α for the normalization constraint $\sum_{y_t} q(y_t) = 1$, we obtain

$$\mathcal{J}(q, \alpha) = \sum_{y_t \in \mathcal{Y}} q(y_t) \mathbf{1}(y_t = y_t^*) - \beta_t \sum_{y_t \in \mathcal{Y}} q(y_t) \log \frac{q(y_t)}{\pi_\theta(y_t)} + \alpha \left(\sum_{y_t \in \mathcal{Y}} q(y_t) - 1 \right).$$

The stationarity condition with respect to $q(y_t)$ is

$$\mathbf{1}(y_t = y_t^*) - \beta_t \left(\log \frac{q(y_t)}{\pi_\theta(y_t)} + 1 \right) + \alpha = 0. \tag{56}$$

Hence,

$$q(y_t) \propto \pi_\theta(y_t) \exp\left(\frac{\mathbf{1}(y_t = y_t^*)}{\beta_t} \right), \tag{57}$$

which yields the normalized solution

$$q(y_t) = \frac{\pi_\theta(y_t)\exp(\mathbf{1}(y_t = y_t^*)/\beta_t)}{\pi_\theta(y_t^*)e^{1/\beta_t} + 1 - \pi_\theta(y_t^*)}. \tag{58}$$

For the demonstrated token y_t^*, Eq.(58) gives

$$q(y_t^*) = \frac{\pi_\theta(y_t^*)e^{1/\beta_t}}{\pi_\theta(y_t^*)e^{1/\beta_t} + 1 - \pi_\theta(y_t^*)}. \tag{59}$$

Consequently,

$$1 - q(y_t^*) = \frac{1 - \pi_\theta(y_t^*)}{\pi_\theta(y_t^*)e^{1/\beta_t} + 1 - \pi_\theta(y_t^*)}, \tag{60}$$

and therefore

$$\frac{q(y_t^*)}{1 - q(y_t^*)} = e^{1/\beta_t}\frac{\pi_\theta(y_t^*)}{1 - \pi_\theta(y_t^*)}. \tag{61}$$

Taking logarithms on both sides yields

$$\beta_t^{-1} = \mathrm{logit}(q(y_t^*)) - \mathrm{logit}(\pi_\theta(y_t^*))\,. \tag{62}$$

We now match this solution to the relaxed projection target

$$\widehat{q}_t = (1 - w_t)\pi_\theta + w_t\delta_{y_t^*}. \tag{63}$$

Setting $q(y_t^*) = \widehat{q}_t(y_t^*)$ in Eq.(62) gives

$$\beta_t^{-1} = \mathrm{logit}(\widehat{q}_t(y_t^*)) - \mathrm{logit}(\pi_\theta(y_t^*))\,, \tag{64}$$

It remains to verify that matching the target-token probability identifies the entire distribution. For every $y_t \neq y_t^*$, Eq.(58) gives

$$q(y_t) = \frac{\pi_\theta(y_t)}{\pi_\theta(y_t^*)e^{1/\beta_t} + 1 - \pi_\theta(y_t^*)}. \tag{65}$$

Thus,

$$\frac{q(y_t)}{1 - q(y_t^*)} = \frac{\pi_\theta(y_t)}{1 - \pi_\theta(y_t^*)}, \qquad y_t \neq y_t^*. \tag{66}$$

Hence the conditional distribution over the non-target set is exactly preserved:

$$q(y_t \mid y_t \neq y_t^*) = \pi_\theta(y_t \mid y_t \neq y_t^*). \tag{67}$$

The relaxed projection in Eq.(63) satisfies the same property, since for every $y_t \neq y_t^*$,

$$\widehat{q}_t(y_t) = (1 - w_t)\pi_\theta(y_t), \tag{68}$$

while

$$1 - \widehat{q}_t(y_t^*) = (1 - w_t)\,(1 - \pi_\theta(y_t^*))\,. \tag{69}$$

Therefore,

$$\frac{\widehat{q}_t(y_t)}{1 - \widehat{q}_t(y_t^*)} = \frac{\pi_\theta(y_t)}{1 - \pi_\theta(y_t^*)}. \tag{70}$$

Since q and $\widehat{q}_t$ have both the same target-token probability and the same conditional distribution over all non-target tokens, they are identical:

$$q = \widehat{q}_t. \tag{71}$$

This proves the KL-regularized characterization of the relaxed projection.

Finally, we recover standard SFT as the zero-regularization limit. From Eq.(59),

$$\lim_{\beta\to 0^+} q(y_t^*) = \lim_{\beta\to 0^+} \frac{\pi_\theta(y_t^*)e^{1/\beta}}{\pi_\theta(y_t^*)e^{1/\beta} + 1 - \pi_\theta(y_t^*)} = 1. \tag{72}$$

For every $y_t \neq y_t^*$,

$$\lim_{\beta\to 0^+} q(y_t) = \lim_{\beta\to 0^+} \frac{\pi_\theta(y_t)}{\pi_\theta(y_t^*)e^{1/\beta} + 1 - \pi_\theta(y_t^*)} = 0. \tag{73}$$

Therefore,

$$\lim_{\beta\to 0^+} q = \delta_{y_t^*} = q_t^{\mathrm{SFT}}. \tag{74}$$

□

A.7 Properties of Information-Aware Projection Strength.

From the definition of KL divergence, we have

$$D_{\mathrm{KL}}(\delta_{y_t^*} \parallel \pi_\theta) = \sum_{y_t \in \mathcal{Y}} \delta_{y_t^*}(y_t) \log \frac{\delta_{y_t^*}(y_t)}{\pi_\theta(y_t)} \tag{75}$$

$$= \delta_{y_t^*}(y_t^*) \log \frac{\delta_{y_t^*}(y_t^*)}{\pi_\theta(y_t^*)} + \sum_{y_t \neq y_t^*} \delta_{y_t^*}(y_t) \log \frac{\delta_{y_t^*}(y_t)}{\pi_\theta(y_t)} \tag{76}$$

$$= 1 \cdot \log \frac{1}{\pi_\theta(y_t^*)} + \sum_{y_t \neq y_t^*} 0 \cdot \log \frac{0}{\pi_\theta(y_t)} \tag{77}$$

$$= -\log \pi_\theta(y_t^*). \tag{78}$$

Therefore, the unadjusted projection strength reduces to

$$Z_t = -\pi_\theta(y_t^*) \log \pi_\theta(y_t^*), \tag{79}$$

which corresponds to the reference token's contribution to predictive entropy.

However,

$$\lim_{\pi_\theta(y_t^*) \to 1} Z_t = \lim_{\pi_\theta(y_t^*) \to 1} (-\pi_\theta(y_t^*) \log \pi_\theta(y_t^*)) = 0 \tag{80}$$

To retain a non-vanishing signal in this regime, we calibrate the information distance through a smooth log-sum-exp transformation. The standard log-sum-exp inequality gives

$$\max\{a, b\} \leq \log(e^a + e^b) \leq \max\{a, b\} + \log 2. \tag{81}$$

Setting $a = 1$ and $b = d_t$ proves the first inequality. Denote $d_t = D_{\mathrm{KL}}\left(\delta_{y_t^*} \middle\| \pi_\theta\right) = -\log \pi_\theta(y_t^*) \to \infty$ as $\pi_\theta(y_t^*) \to 0$, we obtain

$$\log(e + e^{d_t}) = d_t + \log(1 + e^{1-d_t}) = d_t + o(1). \tag{82}$$

In contrast, for highly compatible targets,

$$\lim_{\pi_\theta(y_t^*) \to 1} \widehat{Z}_t = \log(e + 1) > 0. \tag{83}$$

A.8 Policy Update of MFT

To derive the complete formula of the shared policy update of MFT, we first need to derive a fundamental property that the score function evaluates to zero under its own distribution:

$$\begin{aligned}
\mathbb{E}_{y_t \sim \pi_\theta}\left[\nabla_\theta \log \pi_\theta(y_t | y_{<t}^*, x)\right] &= \sum_{y_t \in \mathcal{Y}} \pi_\theta(y_t | y_{<t}^*, x) \nabla_\theta \log \pi_\theta(y_t | y_{<t}^*, x) \\
&= \sum_{y_t \in \mathcal{Y}} \pi_\theta(y_t | y_{<t}^*, x) \frac{\nabla_\theta \pi_\theta(y_t | y_{<t}^*, x)}{\pi_\theta(y_t | y_{<t}^*, x)} \\
&= \sum_{y_t \in \mathcal{Y}} \nabla_\theta \pi_\theta(y_t | y_{<t}^*, x) \\
&= \nabla_\theta \left(\sum_{y_t \in \mathcal{Y}} \pi_\theta(y_t | y_{<t}^*, x) \right) \\
&= \nabla_\theta(1) \\
&= 0
\end{aligned} \tag{84}$$

Then we substitute the stop-gradient terms $\hat{q}_t$ and $\hat{Z}_t$ into our structured policy projection framework:

$$\begin{aligned}
\nabla_\theta \mathcal{L}_{MFT} &= -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\Big[\sum_{t=1}^{T} \hat{Z}_t \mathbb{E}_{y_t\sim\hat{q}_t}[\nabla_\theta \log \pi_\theta(y_t|y^*_{<t},x)]\Big] \\
&= -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\Big[\sum_{t=1}^{T} \hat{Z}_t \Big((1-w_t)\underbrace{\mathbb{E}_{y_t\sim\pi_\theta}[\nabla_\theta \log \pi_\theta(y_t|y^*_{<t},x)]}_{=0} \\
&\qquad + w_t \nabla_\theta \log \pi_\theta(y^*_t|y^*_{<t},x)\Big)\Big] \\
&= -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\Big[\sum_{t=1}^{T} \Big(\pi_\theta(y^*_t)\log\Big(e+\frac{1}{\pi_\theta(y^*_t)}\Big)\Big) w_t \Big(\frac{\nabla_\theta \pi_\theta(y^*_t)}{\pi_\theta(y^*_t)}\Big)\Big] \\
&= -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\Big[\sum_{t=1}^{T} \Big(\frac{\pi_\theta(y^*_t)}{\max_{y_t\in\mathcal{Y}}\pi_\theta(y_t)}\Big)^{\lambda} \log\Big(e+\frac{1}{\pi_\theta(y^*_t)}\Big)\nabla_\theta \pi_\theta(y^*_t)\Big]
\end{aligned} \tag{85}$$

Crucially, although the policy-anchoring component does not contribute directly to the expected score gradient, its role is not to rescale optimization. Instead, it determines the geometry of the local projection target before the shared update is formed. Through this term, $\widehat{q}_t$ remains anchored to the current policy and preserves its non-target structure, while $\hat{Z}_t$ separately controls how much the resulting projection contributes to the shared parameter update. Thus, target relaxation and projection strength act at different levels: the former specifies the distribution to be fitted, whereas the latter weights the influence of that distribution during optimization.

B Experiments

B.1 Main Experiments Settings

Implementation details. For supervised fine-tuning, we use max_response_length=2048, global_batch_size=512, AdamW optimizer with learning rate 5×10^{-5} and train for 1 epoch. For GRPO, we use max_response_length=2048, global_batch_size=128, kl_loss_coef 0.001, learning rate 3×10^{-6}, generate K = 8 rollouts per prompt before each policy update and train for 80 steps. And before RL, we apply supervised fine-tuning with learning rate 1×10^{-5} to do a slight warm-up. For evaluation, we adopt accuracy as the metric. To avoid high variance in results and ensure fair comparisons, we report avg@16 results on all benchmarks with a temperature of 1.0 and maximum generation length of 4096 tokens.

B.2 Diverse Tasks Settings

We conduct the following experiments based on Qwen3-1.7B-Base (Yang et al., 2025). **(i) Code generation task.** For supervised fine-tuning, training is conducted for one epoch with global_batch_size=512 and max_response_length=1024. And we employ AdamW optimizer with learning rate 5×10^{-5}. For GRPO, we use max_response_length=1024, global_batch_size=128, kl_loss_coef 0.001, learning rate 1×10^{-6}, generate K = 8 rollouts per prompt before each policy update and train for 80 steps. To ensure fair comparisons across models, we report pass@1 as the metric for all benchmarks. **(ii) Medical knowledge task.** For supervised fine-tuning, we train for three epochs with a learning rate of 2×10^{-5}, global_batch_size=512, and max_response_length=512. For GRPO, we use max_response_length=1024, global_batch_size=128, kl_loss_coef 0.001, learning rate 1×10^{-6}, generate K = 8 rollouts per prompt before each policy update and train for 80 steps. For evaluation, we adopt accuracy with a single sampling as the metric for all benchmarks.

B.3 Ablation Study

We conduct all ablation experiments based on Qwen3-1.7B-Base and use the same dataset and benchmarks as the main experiments.

Table 5: Sensitivity to the relaxation schedule λ on mathematical reasoning and general domain benchmarks. The best overall performance is in **bold**.

	Mathematical Reasoning Tasks						General Tasks			
	AIME26	AIME25	Math500	Olympiad	Minerva	Avg.	MMLU	ARC-C	GPQA	Avg.
$\lambda = 0$	1.86	1.45	55.04	19.98	19.13	19.49	32.0	**42.5**	25.7	33.4
$\lambda = 0.25$	1.86	1.24	54.96	20.16	19.56	19.56	32.7	40.2	26.4	33.1
$\lambda = 0.5$	**3.32**	1.86	**55.41**	**20.34**	**20.38**	**20.26**	**34.9**	42.1	**26.7**	**34.6**
$\lambda = 0.75$	1.65	1.03	53.78	19.18	19.54	19.04	33.6	39.7	25.5	32.9
$\lambda = 1$	2.69	**2.28**	54.46	19.89	19.95	19.85	34.2	40.5	26.5	33.7

Table 6: Ablation on projection strength design across mathematical reasoning and general domain benchmarks. The best overall performance is in **bold**.

	Mathematical Reasoning Tasks						General Tasks			
	AIME26	AIME25	Math500	Olympiad	Minerva	Avg.	MMLU	ARC-C	GPQA	Avg.
$Z = 1$	1.66	1.24	52.25	18.56	17.73	18.29	29.4	**42.2**	25.9	32.5
$Z = \pi_\theta(y_t^*)$	1.24	1.44	54.59	19.21	19.88	19.27	34.0	40.8	25.9	33.6
$Z = \pi_\theta(y_t^*)^{\frac{1}{2}}$	2.69	0.63	54.84	**20.59**	20.05	19.76	33.0	41.4	24.6	33.0
$Z = \min(c\pi_\theta(y_t^*), 1)$	0.83	1.45	**55.69**	20.08	19.71	19.55	30.9	41.9	26.0	32.9
$Z = -\pi_\theta(y_t^*)\log\pi_\theta(y_t^*)$	0.83	1.44	50.54	17.51	16.47	17.36	29.1	40.4	23.7	31.1
$Z = \pi_\theta(y_t^*)\log\left(e + \frac{1}{\pi_\theta(y_t^*)}\right)$	**3.32**	**1.86**	55.41	20.34	**20.38**	**20.26**	**34.9**	42.1	**26.7**	**34.6**

Sensitivity to Relaxation Schedule. We study the sensitivity of MFT to the relaxation hyperparameter λ by varying λ from the fully linear regime to increasingly weaker confidence modulation while keeping all other training configurations fixed in Tab. 5. The results show a clear non-monotonic trend: intermediate values of λ provide the most consistent performance across the two groups of tasks, whereas either extreme leads to a noticeable degradation on part of the evaluation suite. This suggests that the relaxation schedule should neither follow the raw confidence ratio too aggressively nor collapse toward an almost unmodulated projection. Instead, a moderate nonlinear schedule provides a better balance between sufficiently correcting low-confidence targets and avoiding overly strong displacement when the reference token is poorly aligned with the current policy.

Design of the Projection Strength. We further compare different functional forms of the projection strength while keeping the other setting fixed. Specifically, we consider the following alternatives:

- **Uniform strength:** $Z = 1$, which recovers the policy-agnostic weighting of standard SFT by assigning equal influence to all supervised tokens.
- **Linear confidence strength:** $Z = \pi_\theta(y_t^*)$, which increases the projection strength proportionally to the model probability of the reference token.
- **Smoothed confidence strength:** $Z = \pi_\theta(y_t^*)^{\frac{1}{2}}$, which retains confidence-dependent weighting while assigning relatively greater strength to low-probability targets than the linear schedule.
- **Clipped confidence strength:** $Z_t = \min(c\pi_\theta(y_t^*), 1)$, which increases with target confidence but caps the maximum influence of high-confidence tokens.
- **Entropy-contribution strength:** $Z_t = H_t^* = -\pi_\theta(y_t^*)\log\pi_\theta(y_t^*)$, which directly uses the reference token's contribution to the predictive entropy as its projection strength.
- **Ours:** $Z = \pi_\theta(y_t^*)\log(e + 1/\pi_\theta(y_t^*))$, which retains the uncertainty-sensitive structure of H_t^* while preventing the projection strength from vanishing for well-supported targets.

As shown in Tab. 6, these alternative formulations exhibit different trade-offs across mathematical reasoning and general domain tasks, whereas our proposed formulation provides the strongest overall performance. In particular, purely confidence-based formulations capture target-policy agreement but do not explicitly reflect target-specific predictive uncertainty, while the entropy contribution

provides such an uncertainty-aware signal but vanishes as $\pi_\theta(y_t^*) \to 1$. Our formulation preserves this information-sensitive structure while maintaining sufficient projection strength for supervision already well supported by the current policy.