

LLM Fine-tuning with an Unconstrained Fine-tuned Teacher

Jianhao Zhang Ou Wu* Yi Du

Hangzhou Institute for Advanced Study,

University of Chinese Academy of Sciences, Hangzhou, China,

zhangjianhao24@mailsucas.ac.cn, wuou@ucas.ac.cn, duyicnic.cn

Abstract

Fine-tuning large language models (LLMs) on a single target task requires balancing task adaptation with retention of prior capabilities. Existing teachers mainly guide preservation or transfer, but do not identify which target-task samples should be emphasized when fine-tuning must retain prior capabilities. We propose the **Unconstrained Fine-tuned Teacher (UFT)**, shifting the teacher role from “knowledge transfer” to “value discovery”. UFT is obtained by unconstrained fine-tuning on the target task and then compared with the source model to identify high-value samples. These samples are upweighted during student fine-tuning and receive value-weighted UFT-guided distillation, allowing the student to prioritize task-critical signals while preserving prior capabilities. Experiments on Magicoder and MetaMathQA show that UFT improves the trade-off between target-task performance and prior-capability retention, while converging faster and more stably. By turning teacher behavior into sample-value discovery, UFT establishes a new teacher role for data-efficient fine-tuning. Code at: <https://github.com/zjh-vinky/UFT>.

1 Introduction

Fine-tuning LLMs on a single target task must balance target-task adaptation with preserving prior capabilities (Wu et al., 2025a; Wang et al., 2025b; Lin et al., 2026). Although parameter-efficient methods like LoRA (Hu et al., 2022) can reduce capability degradation by constraining updates, they may underfit task-specific nuances (He, 2024; Shen et al., 2025). Knowledge distillation offers an alternative by transferring teacher knowledge to students (Hinton et al., 2015). However, existing teacher roles mainly preserve source-model behavior or transfer external capabilities, rather than discover which target-task samples carry value for adaptation while preserving prior capabilities.

*Corresponding author.

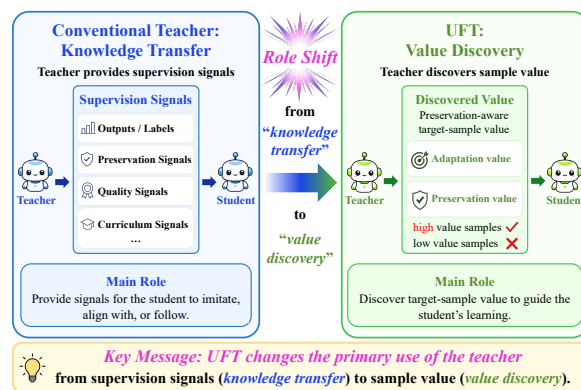


Figure 1: From Knowledge Transfer to Value Discovery.

Recent advances have diversified teacher functions in fine-tuning. Old-knowledge teachers prevent forgetting through regularization (Wu et al., 2025b), generalist teachers transfer universal skills via synthetic data generation (Yin et al., 2025), and structure teachers model relational knowledge (Truong et al., 2025). While useful for teacher-guided adaptation, these approaches neither assess target-task sample value nor indicate which samples deserve emphasis. This can make learning inefficient, because reliable task-specific samples that do not undermine prior capabilities are not prioritized over noisy or potentially disruptive ones.

The data weighting literature has addressed a related issue. FLOW (Sanyal et al., 2025) weights samples by pretrained-model losses, emphasizing easy samples to mitigate forgetting. Sow et al. (2025) uses dynamic loss-based reweighting, and ADAPT (Zhao et al., 2026) performs online reweighting from similarity-derived quality signals. However, they rely on generic proxies such as loss or quality heuristics, lacking a task-specialized expert signal that reflects target-task-only adaptation. They identify samples favored by proxies rather than those worth emphasizing for target-task adaptation under prior-capability preservation.

Consider fine-tuning a model on mathematical

reasoning (Wei et al., 2022; Luo et al., 2025). Standard distillation may uniformly transfer generalist-teacher knowledge, yet fail to identify examples containing reliable task-specific reasoning patterns. Indiscriminate learning from all target-task samples may overfit noisy reasoning traces and degrade prior capabilities. Existing data weighting methods may prioritize samples by proxy signals like loss, but still miss examples with reliable task-specific value. Thus, sample utilization remains inefficient because training cannot distinguish task-critical examples from difficult, noisy, or disruptive ones.

A fundamental limitation is the lack of teacher signals for *target-task value judgment*. Existing distillation largely follows “blind transfer”, assigning uniform weight to samples regardless of whether they are informative or redundant for the target task (Gou et al., 2021). When updates involve high-variance or noisy samples, convergence slows and generalization degrades (Mirzadeh et al., 2020; Agarwal et al., 2024). We argue that knowing what to emphasize is as critical as knowing how to preserve, yet existing teacher roles do not gauge target-task sample value under source-model preservation constraints, as illustrated in Fig. 1.

To address this, we propose the **Unconstrained Fine-tuned Teacher (UFT)**, shifting the teacher role from knowledge transfer to value discovery. UFT is obtained through unconstrained target-task fine-tuning, allowing the teacher to reveal the value of target samples beyond their raw supervision labels. By relating this target-task expert to source-reference behavior, UFT discovers samples with reliable adaptation value and limited conflict with confident source behavior. These discovered values guide value-weighted UFT supervision, enabling the student to focus on task-critical signals without unnecessary deviation from source-model behavior. Thus, UFT improves target-task adaptation while preserving source-model behavior.

Our contributions are summarized as follows:

1. We identify a gap in the teacher-role spectrum: existing teachers mainly preserve or transfer knowledge, while target-sample value discovery for adaptation with prior-capability preservation remains underexplored.
2. We propose UFT, an unconstrained fine-tuned teacher that shifts the teacher role from knowledge transfer to value discovery. By coupling UFT reliability with source-relative behavioral divergence, UFT discovers preservation-

aware utilities without sample-value annotations.

3. We establish a value-guided fine-tuning pipeline that converts discovered sample values into dual-teacher supervision. Results across three models on Magicoder and MetaMathQA show faster convergence and stronger target-task performance while preserving prior capabilities.

2 Related Work

2.1 Data Weighting in LLM Fine-tuning

Assigning differential importance to training data has become important for efficient LLM adaptation. Recent methods estimate data weights from proxy signals such as loss or confidence (Sanyal et al., 2025; Sow et al., 2025), validation signals or training dynamics (Li et al., 2025b; Pan et al., 2025; Zhao et al., 2026), and even smaller-model judgments (Afzali et al., 2026; Liu et al., 2026a). Although these strategies differ, they mostly rely on indirect signals to assess data importance. Thus, estimated weights may reflect superficial difficulty, transient optimization states, or proxy-specific biases, rather than which samples best support target-task adaptation while preserving prior capabilities.

Zhou et al. (2025) study model-centric data selection by comparing a base model with its fine-tuned counterpart and using loss to estimate per-sample learning potential, yielding learnability-based selection. This signal measures how much the model improves on each sample after fine-tuning. UFT shifts the criterion from improvement magnitude to sample value, jointly considering target-task evidence reliability and compatibility with confident source-model behavior. The unconstrained fine-tuned teacher provides task-specific evidence, while source-relative divergence evaluates source compatibility. The resulting utility identifies and emphasizes high-value samples during fine-tuning. Appendix D.8 further compares UFT with DavIR.

2.2 Knowledge Distillation for LLMs

Knowledge distillation transfers capabilities from teacher to student and was originally proposed for model compression (Hinton et al., 2015). In continual learning, model-based distillation mitigates forgetting by encouraging the adapted model to preserve source-model behavior (Schwarz et al., 2018). However, these approaches treat the teacher mainly as a static repository of prior knowledge, offering

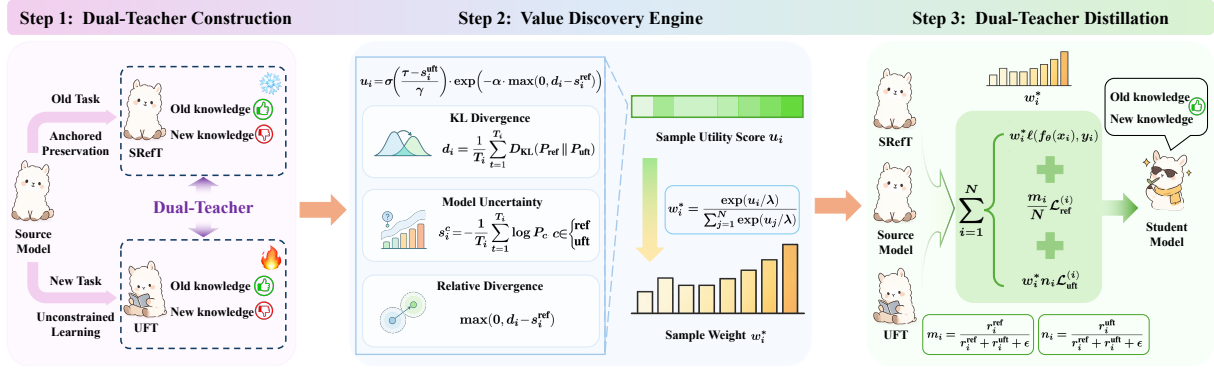


Figure 2: Overview of the UFT framework. It constructs an unconstrained teacher for value discovery, identifying samples useful for new-task learning and compatible with source-model behavior preservation. Stable student fine-tuning is then guided by dual teachers, with UFT supervising adaptation and SRefT anchoring source knowledge.

little guidance on which target-task samples should receive stronger supervision during fine-tuning.

Recent advances have diversified teacher roles. Some methods directly use learned teachers as on-policy supervisors over student-generated trajectories (Agarwal et al., 2024; Gu et al., 2024; Jang et al., 2026; Ko et al., 2026), but do not identify which samples should be emphasized under preservation constraints. Others use teachers as output refiners (Song et al., 2025; Adarsh et al., 2025), focusing on output quality rather than learning value. Teachers can also serve as training-free scorers of sample quality (Wang et al., 2025a; Xia et al., 2025), though such signals are largely static and insensitive to model-specific adaptation behavior.

Our approach reconceptualizes the teacher as a **value discovery mechanism**. UFT is constructed through unconstrained target-task fine-tuning, exposing task-induced behavioral drift and yielding a task-specialized teacher. Its contrast with the source model, together with teacher reliability, quantifies sample value under source-model preservation constraints, enabling *value-aware distillation* rather than uniform transfer across samples.

2.3 Model-Centric Selective Learning

Recent work uses model behavior to assess training value. RHO-1 (Lin et al., 2024) uses reference-relative excess loss to identify valuable tokens, while IFD (Li et al., 2024) estimates instruction value from model-perceived difficulty. Forgetting (Taheri et al., 2026) uses changes in model loss to distinguish tokens to learn or forget, and TOSS (Li et al., 2026) uses cross-model loss differences to identify safety-sensitive tokens. Select and Distill (Yu et al., 2024b) uses discrepancies between pretrained and fine-tuned teachers to selec-

tively transfer complementary knowledge. In these methods, model behavior provides a selection criterion, while teachers remain sources of transferable knowledge. Target-sample value discovery is not treated as the teacher’s primary function.

Our approach reconceptualizes the unconstrained fine-tuned teacher as a value discovery mechanism. UFT exploits task-specialized teacher behavior to identify target samples that warrant greater emphasis beyond their raw supervision. The discovered values guide both sample weighting and dual-teacher supervision, enabling the teacher to determine what the student should emphasize and how it should learn.

3 Methodology

3.1 Problem Formulation and Notation

Let $\theta \in \mathbb{R}^d$ denote model parameters. We fine-tune on a target-task dataset $\mathcal{D}_{\text{tgt}} = \{(x_i, y_i)\}_{i=1}^N$ with task loss $\mathcal{L}_{\text{task}}$. The source model M_{ref} with parameters θ_{ref} is the pre-adaptation checkpoint and serves as the Source Reference Teacher (SRefT), providing reference behavior for preservation.

Sample weights are $\mathbf{w} \in \Delta^{N-1}$, where $\Delta^{N-1} = \{\mathbf{w} : w_i \geq 0, \sum_i w_i = 1\}$. They control each sample’s contribution to task learning and UFT-guided target-task supervision. Preservation of source-model behavior is handled separately and is not solely governed by these utility-based weights.

Teacher Role Gap. Existing teacher roles in knowledge distillation, including old-knowledge guardians, generalist teachers, and structure teachers, focus on preserving capabilities or transferring content. However, they lack a teacher specialized in assessing which target-task samples provide reliable task-specific learning signals while showing

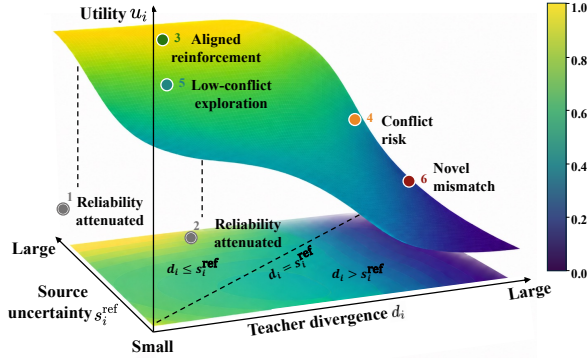


Figure 3: Sample utility regimes under UFT. UFT reliability and source-relative divergence divide samples into six representative regimes, with the boundary $d_i = s_i^{\text{ref}}$ separating tolerated from penalized divergence.

limited conflict with source-model behavior.

3.2 Constructing the Unconstrained Fine-tuned Teacher

To discover samples worth emphasizing while preserving prior capabilities, we require a teacher that explores target-task behavior without preservation constraints. We construct the Unconstrained Fine-tuned Teacher (UFT) M_{uft} with parameters θ_{uft} by unconstrained fine-tuning of M_{ref} on $\mathcal{D}_{\text{tgt}}$. Specifically, we initialize $\theta_{\text{uft}}^{(0)} = \theta_{\text{ref}}$ and optimize

$$\theta_{\text{uft}} = \arg \min_{\theta} \sum_{(x,y) \in \mathcal{D}_{\text{tgt}}} \ell(f_{\theta}(x), y). \quad (1)$$

This exposes task-induced behavioral drift from unconstrained target-task fine-tuning, providing contrast to identify reliable task-specific signals and potential conflicts with source-model behavior. The M_{uft} serves as a value-discovery teacher, providing target-task evidence for reliably supported samples. Since UFT reliability alone cannot assess source-behavior conflict, we next contrast UFT with source-model reference behavior.

For each sample $(x_i, y_i) \in \mathcal{D}_{\text{tgt}}$, the teacher divergence on the target response is defined as

$$d_i = \frac{1}{T_i} \sum_{t=1}^{T_i} D_{\text{KL}}(P_{\text{ref}}(\cdot | x_i, y_{i,<t}) \| P_{\text{uft}}(\cdot | x_i, y_{i,<t})), \quad (2)$$

where T_i is the response length. We compute per-sample uncertainty proxies as $s_i^c = -\frac{1}{T_i} \sum_{t=1}^{T_i} \log P_c(y_{i,t} | x_i, y_{i,<t})$ for $c \in \{\text{ref}, \text{uft}\}$. Smaller scores indicate higher confidence.

3.3 Value Discovery Mechanism

The divergence d_i between SRefT and UFT measures behavioral deviation induced by unconstrained target-task fine-tuning on sample i , but

is insufficient alone. A large divergence may indicate either a valuable task-specific signal or conflict with source-model behavior. Thus, we evaluate divergence relative to source-model uncertainty. Low uncertainty implies confident source behavior and limited drift tolerance, whereas high uncertainty allows more room for target-task adaptation.

This motivates a **relative divergence** measure that penalizes UFT-SRefT divergence only when it exceeds the source-model uncertainty budget:

$$u_i = \sigma\left(\frac{\tau - s_i^{\text{uft}}}{\gamma}\right) \cdot \exp(-\alpha \cdot \max(0, d_i - s_i^{\text{ref}})), \quad (3)$$

where $\sigma(\cdot)$ is the sigmoid function, τ is the UFT uncertainty threshold below which UFT supervision is treated as reliable, $\gamma > 0$ controls the softness of the reliability gate, and $\alpha > 0$ controls sensitivity to relative divergence. This formulation unifies six intuitive cases into two principles: (1) higher UFT reliability increases sample utility, and (2) divergence is tolerated only when it does not exceed source-model uncertainty, as illustrated in Fig. 3.

A detailed analysis is provided in Appendix A.

Weight Optimization. Given sample utilities $\{u_i\}_{i=1}^N$, we optimize weights to maximize expected utility while preventing weight collapse through entropy regularization

$$\max_{\mathbf{w} \in \Delta^{N-1}} \sum_{i=1}^N w_i \cdot u_i - \lambda \cdot \sum_{i=1}^N w_i \log w_i, \quad (4)$$

where $\lambda > 0$ controls the entropy regularization. This objective has the closed-form solution

$$w_i^* = \frac{\exp(u_i/\lambda)}{\sum_{j=1}^N \exp(u_j/\lambda)}. \quad (5)$$

The obtained weights w_i^* are then used to emphasize high-value samples in task learning and UFT-guided target-task supervision. The derivation is provided in Appendix B.1.

3.4 Value-guided Dual-Teacher Distillation

Given the optimized weights $\mathbf{w}^*$, we perform fine-tuning with dual-teacher distillation that balances preservation of source-model behavior and target-task adaptation. The student model receives guidance from both SRefT and UFT, with each teacher serving a distinct purpose.

SRefT provides **stability regularization** via knowledge distillation:

$$\mathcal{L}_{\text{ref}}^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} D_{\text{KL}}(P_{\text{ref}}(\cdot | x_i, y_{i,<t}) \| P_{\theta}(\cdot | x_i, y_{i,<t})). \quad (6)$$

Algorithm 1 LLM Fine-tuning with UFT

Require: $M_{\text{ref}}, \mathcal{D}_{\text{tgt}}, \tau, \gamma, \alpha, \lambda, \beta$ **Ensure:** Fine-tuned model M_θ

- 1: $\theta_{\text{uft}} \leftarrow \text{FineTune}(\theta_{\text{ref}}, \mathcal{D}_{\text{tgt}})$
 - 2: Compute d_i, s_i^{ref} , and s_i^{uft} for $x_i \in \mathcal{D}_{\text{tgt}}$
 - 3: Compute sample value u_i by (3)
 - 4: Compute sample weights $w_i^* = \frac{\exp(u_i/\lambda)}{\sum_j \exp(u_j/\lambda)}$
 - 5: Compute teacher weights m_i and n_i by (8)
 - 6: Train θ with weighted task loss and dual-teacher distillation from SRefT and UFT
 - 7: **return** M_θ
-

This term discourages unnecessary deviation from source-model behavior.

UFT provides **target-task guidance** on target-task samples:

$$\mathcal{L}_{\text{uft}}^{(i)} = \frac{1}{T_i} \sum_{t=1}^{T_i} D_{\text{KL}}(P_{\text{uft}}(\cdot|x_i, y_i, <t) \| P_\theta(\cdot|x_i, y_i, <t)). \quad (7)$$

This term provides a per-sample supervision signal from the unconstrained target-task expert.

Confidence-Adaptive Distillation. A key insight is that samples should rely differently on the two teachers. High SRefT confidence (s_i^{ref} is small) indicates that the target response is well supported by the source model, so deviations from source-model behavior should be discouraged. High SRefT uncertainty (s_i^{ref} is large) indicates weaker source guidance, allowing greater reliance on task-specific supervision if reliable. In parallel, high UFT confidence (s_i^{uft} is small) indicates reliable supervision from the unconstrained target-task teacher, making the sample suitable for UFT-guided learning.

To dynamically balance teacher guidance, we convert uncertainty scores into reliability scores $r_i^c = \exp(-s_i^c)$ for $c \in \{\text{ref}, \text{uft}\}$, and define

$$m_i = \frac{r_i^{\text{ref}}}{r_i^{\text{ref}} + r_i^{\text{uft}} + \epsilon}, \quad n_i = \frac{r_i^{\text{uft}}}{r_i^{\text{ref}} + r_i^{\text{uft}} + \epsilon}, \quad (8)$$

where ϵ is a constant for numerical stability. These coefficients control the student’s relative reliance on the two teachers for each sample. When SRefT is more confident than UFT ($r_i^{\text{ref}} \gg r_i^{\text{uft}}$), $m_i \approx 1$ and the student mainly follows SRefT to preserve source-model behavior. Conversely, when UFT is more confident than SRefT ($r_i^{\text{uft}} \gg r_i^{\text{ref}}$), $n_i \approx 1$, emphasizing UFT-guided target-task adaptation. When both teachers are uncertain, neither provides

dominant reliable guidance, and the reliance is determined by their stabilized relative reliabilities.

The total training objective combines task learning with confidence-adaptive distillation:

$$\min_{\theta} \sum_{i=1}^N \left[w_i^* \ell(f_\theta(x_i), y_i) + \beta \left(\frac{m_i}{N} \mathcal{L}_{\text{ref}}^{(i)} + w_i^* n_i \mathcal{L}_{\text{uft}}^{(i)} \right) \right]. \quad (9)$$

where β controls global distillation strength, and m_i and n_i are coupled reliability coefficients to balance source-behavior preservation and UFT-guided target-task supervision at the sample level. The uniformly normalized SRefT term provides a broad preservation anchor, while utility-weighted task and UFT terms concentrate target adaptation on high-value samples. By incorporating teacher reliability, m_i and n_i implement confidence-adaptive distillation, preventing over-regularization and under-utilization of target-task expertise, and enabling efficient high-value learning while discouraging unnecessary source-behavior deviation. The complete procedure is summarized in Algorithm 1.

4 Theoretical Analysis

We provide a theoretical analysis of UFT. First, validation-based checkpoint selection is justified within the finite UFT trajectory. Second, the resulting UFT weights remain reliable for downstream student fine-tuning under moderate initial utility approximation error and dynamic utility drift. Proofs are given in Appendix B.

Theorem 1 (Validation-based UFT checkpoint selection). *Let $\{\theta_t\}_{t=1}^T$ be checkpoints along the unconstrained fine-tuning trajectory, and let $\ell(\theta; z)$ denote the per-sample validation loss. For each checkpoint, define $\hat{\mathcal{L}}_{\text{val}}(\theta_t) = n_{\text{val}}^{-1} \sum_{j=1}^{n_{\text{val}}} \ell(\theta_t; z_j)$ and $\mathcal{L}_{\text{pop}}(\theta_t) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(\theta_t; z)]$, where $\{z_j\}_{j=1}^{n_{\text{val}}}$ are independent validation samples from the target-task population $\mathcal{D}$. Assume that the validation set is independent of the training trajectory and that $\ell(\theta_t; z) \in [0, B]$ almost surely over $z \sim \mathcal{D}$ for all $t \in \{1, \dots, T\}$, where $B > 0$. Let $\delta \in (0, 1)$, $t_{\text{val}} = \arg \min_{1 \leq t \leq T} \hat{\mathcal{L}}_{\text{val}}(\theta_t)$, and $t_{\text{pop}} = \arg \min_{1 \leq t \leq T} \mathcal{L}_{\text{pop}}(\theta_t)$. Then, with probability at least $1 - \delta$,*

$$\mathcal{L}_{\text{pop}}(\theta_{t_{\text{val}}}) \leq \mathcal{L}_{\text{pop}}(\theta_{t_{\text{pop}}}) + 2B \sqrt{\frac{\log(2T/\delta)}{2n_{\text{val}}}}. \quad (10)$$

Theorem 1 justifies validation-based selection of the UFT checkpoint for computing static utilities $\mathbf{u}$. We next show that, under bounded utility

drift, fixed UFT weights stay close to ideal dynamic weights during local student fine-tuning.

Proposition 1 (Local stability of fixed UFT weighting). *Let $\mathbf{u} = (u_1, \dots, u_N)$ be the static utilities computed by UFT before downstream student training, and define the fixed UFT weights as*

$$w_i^{uft} = \frac{\exp(u_i/\lambda)}{\sum_{j=1}^N \exp(u_j/\lambda)}. \quad (11)$$

Let $\{\theta_k\}_{k=0}^K$ be the student-training trajectory, and let $v_i(\theta_k)$ denote sample i 's ideal dynamic utility at checkpoint θ_k , with $\mathbf{v}(\theta_k) = (v_1(\theta_k), \dots, v_N(\theta_k))$. The corresponding dynamic reference weights are

$$w_i^{dyn}(\theta_k) = \frac{\exp(v_i(\theta_k)/\lambda)}{\sum_{j=1}^N \exp(v_j(\theta_k)/\lambda)}. \quad (12)$$

Suppose UFT utilities approximate initial dynamic utilities, with $|u_i - v_i(\theta_0)| \leq \epsilon$ for all i , and dynamic utilities have bounded drift, with $|v_i(\theta_k) - v_i(\theta_0)| \leq \Delta_K$ for all i and $k \leq K$. Let $\eta_K = \epsilon + \Delta_K$. Then $\|\mathbf{u} - \mathbf{v}(\theta_k)\|_\infty \leq \eta_K$ for all $k \leq K$. Moreover, for all i and $k \leq K$,

$$\exp\left(-\frac{2\eta_K}{\lambda}\right) \leq \frac{w_i^{uft}}{w_i^{dyn}(\theta_k)} \leq \exp\left(\frac{2\eta_K}{\lambda}\right). \quad (13)$$

As a corollary, for all $k \leq K$,

$$TV(\mathbf{w}^{uft}, \mathbf{w}^{dyn}(\theta_k)) \leq 1 - \exp\left(-\frac{2\eta_K}{\lambda}\right). \quad (14)$$

Here, TV denotes total variation distance between sample-weight distributions, ϵ the initial UFT-utility approximation error, and Δ_K the dynamic-utility drift over the first K student-training steps.

5 Experiments

In this section, we systematically evaluate UFT through four questions: **Q1:** Does UFT better balance target-task adaptation and prior-capability retention than strong baselines? **Q2:** Does the unconstrained fine-tuned teacher provide reliable target-task evidence for identifying high-value samples? **Q3:** Does each key component of UFT contribute effectively, and how robust is UFT under different hyperparameter settings? **Q4:** Does UFT improve training convergence efficiency?

5.1 Experiment Setup

Datasets & Evaluation Metrics. We conduct experiments on Magicoder (Wei et al., 2024) and MetaMathQA (Yu et al., 2024a) to evaluate UFT. Following recent studies (Wang et al., 2025b; Lin et al., 2026; Sanyal et al., 2025; Liu et al., 2026b; Yuan et al., 2026), we decouple evaluation into target performance and general capability, with benchmark roles adjusted across scenarios. For Magicoder, HumanEval (HE) (Chen et al., 2021) is the target benchmark, while MMLU (Hendrycks et al., 2021), HellaSwag (HASW) (Zellers et al., 2019), GSM8K (Cobbe et al., 2021), and OpenBookQA (OBQA) (Mihaylov et al., 2018) assess general capability. For MetaMathQA, GSM8K is the target benchmark, while HE is moved to the general-capability suite. We report general capability (mean of four benchmarks), target performance, and Overall (arithmetic mean of the two).

Baselines & Implementation Details. We compare UFT with a set of representative baselines, including standard fine-tuning (FT), LoRA (Hu et al., 2022), FLOW (Sanyal et al., 2025), DFT (Wu et al., 2026), TALR (Lin et al., 2026), and ADAPT (Zhao et al., 2026). All methods are evaluated on three model architectures of varying scales: Gemma3-4B-pt (Team et al., 2025), Llama3.2-3B (Grattafiori et al., 2024), and Qwen3.5-9B-Base (Team, 2026). All methods use AdamW with a cosine learning-rate schedule. For each target dataset, we hold out 1,000 examples for validation and use the remaining target-training split for all methods. For UFT, we set $\tau = Q_{60}(s^{uft})$, $\gamma = 0.2$, $\alpha = 0.8$, $\lambda = 0.2$, $\beta = 2$. We report mean and standard deviation over three random seeds. All experiments are conducted on 4 NVIDIA A100 GPUs. Detailed implementation is provided in Appendix C.

5.2 Main Results (Q1)

Table 1 reports Magicoder results across three models. On LLaMA3.2, Standard FT improves the target average but reduces the general capability average from 42.97% to 37.47%. UFT achieves the best target, general, and overall averages among fine-tuned methods, reaching 46.73%, 42.65%, and 44.69%, respectively, and surpassing the strongest baseline's overall average of 42.56%. On Gemma and Qwen, UFT shows the same pattern, with the best overall averages of 52.90% and 69.61%. This pattern is consistent with value discovery, which upweights samples with stronger target signals and

Model	Method	General Capability $\uparrow$					Target $\uparrow$	Overall $\uparrow$
		HASW	OBQA	MMLU	GSM8K	Avg.	HE	Avg.
LLaMA3.2-3B	Pre-trained	55.83 $\pm$ 0.00	31.23 $\pm$ 0.00	56.92 $\pm$ 0.00	27.89 $\pm$ 0.00	42.97 $\pm$ 0.00	28.16 $\pm$ 0.00	35.57 $\pm$ 0.00
	Standard FT	52.12 $\pm$ 0.47	28.39 $\pm$ 0.37	48.61 $\pm$ 0.43	20.74 $\pm$ 0.34	37.47 $\pm$ 0.29	45.13 $\pm$ 0.48	41.30 $\pm$ 0.30
	LoRA ICLR 22	53.38 $\pm$ 0.39	29.46 $\pm$ 0.34	51.98 $\pm$ 0.35	21.17 $\pm$ 0.39	39.00 $\pm$ 0.26	41.92 $\pm$ 0.40	40.46 $\pm$ 0.25
	FLOW ICML 25	54.63 $\pm$ 0.34	30.06 $\pm$ 0.30	55.83 $\pm$ 0.31	25.12 $\pm$ 0.23	41.41 $\pm$ 0.22	42.34 $\pm$ 0.34	41.88 $\pm$ 0.21
	DFT ICLR 26	53.36 $\pm$ 0.37	29.71 $\pm$ 0.21	51.97 $\pm$ 0.33	25.34 $\pm$ 0.24	40.10 $\pm$ 0.21	45.01 $\pm$ 0.36	42.56 $\pm$ 0.23
	TALR ICLR 26	54.84 $\pm$ 0.35	30.20 $\pm$ 0.34	54.01 $\pm$ 0.30	26.79 $\pm$ 0.34	41.46 $\pm$ 0.24	43.54 $\pm$ 0.35	42.50 $\pm$ 0.22
	ADAPT ICLR 26	53.76 $\pm$ 0.41	29.62 $\pm$ 0.29	52.87 $\pm$ 0.36	22.24 $\pm$ 0.29	39.62 $\pm$ 0.23	45.27 $\pm$ 0.39	42.45 $\pm$ 0.24
	UFT (Ours)	55.29 $\pm$ 0.31	30.96 $\pm$ 0.26	56.22 $\pm$ 0.28	28.12 $\pm$ 0.31	42.65 $\pm$ 0.22	46.73 $\pm$ 0.40	44.69 $\pm$ 0.23
Gemma3-4B	Pre-trained	57.41 $\pm$ 0.00	34.62 $\pm$ 0.00	60.24 $\pm$ 0.00	36.73 $\pm$ 0.00	47.25 $\pm$ 0.00	35.77 $\pm$ 0.00	41.51 $\pm$ 0.00
	Standard FT	53.06 $\pm$ 0.45	31.24 $\pm$ 0.38	53.21 $\pm$ 0.40	23.79 $\pm$ 0.28	40.33 $\pm$ 0.27	58.17 $\pm$ 0.49	49.25 $\pm$ 0.29
	LoRA ICLR 22	55.81 $\pm$ 0.36	31.82 $\pm$ 0.28	56.74 $\pm$ 0.32	31.61 $\pm$ 0.42	44.00 $\pm$ 0.25	54.69 $\pm$ 0.51	49.35 $\pm$ 0.29
	FLOW ICML 25	55.23 $\pm$ 0.34	32.38 $\pm$ 0.30	58.17 $\pm$ 0.29	34.16 $\pm$ 0.37	44.99 $\pm$ 0.23	55.86 $\pm$ 0.44	50.43 $\pm$ 0.26
	DFT ICLR 26	54.65 $\pm$ 0.38	31.80 $\pm$ 0.23	56.68 $\pm$ 0.34	34.39 $\pm$ 0.38	44.38 $\pm$ 0.24	58.76 $\pm$ 0.32	51.57 $\pm$ 0.21
	TALR ICLR 26	56.62 $\pm$ 0.32	32.24 $\pm$ 0.34	59.79 $\pm$ 0.27	35.07 $\pm$ 0.44	45.93 $\pm$ 0.25	56.08 $\pm$ 0.51	51.01 $\pm$ 0.29
	ADAPT ICLR 26	55.66 $\pm$ 0.39	31.43 $\pm$ 0.35	56.77 $\pm$ 0.33	32.02 $\pm$ 0.39	43.97 $\pm$ 0.26	58.39 $\pm$ 0.52	51.18 $\pm$ 0.30
	UFT (Ours)	56.93 $\pm$ 0.30	32.56 $\pm$ 0.29	59.28 $\pm$ 0.28	36.27 $\pm$ 0.46	46.26 $\pm$ 0.24	59.53 $\pm$ 0.34	52.90 $\pm$ 0.22
Qwen3.5-9B	Pre-trained	59.67 $\pm$ 0.00	34.17 $\pm$ 0.00	72.39 $\pm$ 0.00	86.23 $\pm$ 0.00	63.12 $\pm$ 0.00	61.05 $\pm$ 0.00	62.09 $\pm$ 0.00
	Standard FT	54.21 $\pm$ 0.41	31.06 $\pm$ 0.33	66.22 $\pm$ 0.36	78.57 $\pm$ 0.31	57.52 $\pm$ 0.25	75.82 $\pm$ 0.41	66.67 $\pm$ 0.25
	LoRA ICLR 22	56.73 $\pm$ 0.33	32.24 $\pm$ 0.32	69.44 $\pm$ 0.29	81.87 $\pm$ 0.24	60.07 $\pm$ 0.21	70.33 $\pm$ 0.24	65.20 $\pm$ 0.17
	FLOW ICML 25	57.21 $\pm$ 0.30	31.97 $\pm$ 0.31	70.86 $\pm$ 0.26	82.45 $\pm$ 0.37	60.62 $\pm$ 0.22	72.34 $\pm$ 0.35	66.48 $\pm$ 0.22
	DFT ICLR 26	56.62 $\pm$ 0.35	31.88 $\pm$ 0.34	70.05 $\pm$ 0.31	82.32 $\pm$ 0.44	60.22 $\pm$ 0.26	75.98 $\pm$ 0.32	68.10 $\pm$ 0.22
	TALR ICLR 26	57.44 $\pm$ 0.28	32.46 $\pm$ 0.32	70.94 $\pm$ 0.25	83.04 $\pm$ 0.41	60.97 $\pm$ 0.23	72.81 $\pm$ 0.46	66.89 $\pm$ 0.27
	ADAPT ICLR 26	56.48 $\pm$ 0.36	31.78 $\pm$ 0.34	68.52 $\pm$ 0.30	81.94 $\pm$ 0.43	59.68 $\pm$ 0.26	76.23 $\pm$ 0.41	67.96 $\pm$ 0.25
	UFT (Ours)	58.60 $\pm$ 0.27	32.69 $\pm$ 0.31	70.91 $\pm$ 0.25	84.96 $\pm$ 0.39	61.79 $\pm$ 0.22	77.42 $\pm$ 0.34	69.61 $\pm$ 0.21

Table 1: Performance comparison on the Magicoder dataset (%). Pre-trained and Standard FT are excluded from the ranking. The **best results** are highlighted in red, and the **second-best** results are highlighted in blue.

lower preservation cost. The dual-teacher design further balances target adaptation and source retention, enabling stronger adaptation with less general capability degradation. Appendix D.1 provides detailed MetaMathQA results, and Appendix D.2 visualizes this trade-off from a Pareto perspective.

5.3 Unconstrained Fine-Tuned Teacher (Q2)

The unconstrained fine-tuned teacher is central to UFT because it provides target-specialized evidence for value discovery. We assess its reliability as a value-discovery teacher from three aspects.

Unconstrained Drift Behavior. Table 2 compares the source model M_{ref} and the unconstrained teacher M_{uft} on target-task and general-capability benchmarks. M_{uft} substantially improves target-task performance over M_{ref} , showing that unconstrained fine-tuning exposes task-specific behavioral changes. Meanwhile, its degradation on general benchmarks indicates that target-task adaptation can introduce retention-sensitive drift when no preservation constraint is imposed. This contrast provides the signal that UFT exploits for value discovery: samples strongly supported by M_{uft} are emphasized only when their divergence remains compatible with confident source-model behavior.

Model	HASW	OBQA	MMLU	GSM8K	HE
M_{ref}	55.83	31.23	56.92	27.89	28.16
M_{uft}	52.80	29.21	50.60	22.10	45.50

Table 2: Performance comparison between the M_{ref} and the M_{uft} on Magicoder with LLaMA3.2.

Checkpoint Suitability. Fig. 4(a) and Fig. 4(b) show the target-task and general-capability dynamics during unconstrained fine-tuning on Magicoder with LLaMA3.2-3B. As training proceeds, M_{uft} first improves on the target task but later no longer provides better target-task evidence, while general capability continues to degrade, reflecting increasing source-model drift. To examine how the UFT training stage affects value discovery, Fig. 4(c) compares student performance using utilities derived from UFT-1, UFT-2, and UFT-3. UFT-1 is selected too early and remains insufficiently adapted to the target task, yielding weak utility signals. UFT-3 is trained too long and introduces overly aggressive drift from M_{ref} , weakening retention-aware valuation. In contrast, UFT-2 achieves the best downstream trade-off between target-task performance and general-capability preservation.

Value-Discovery Analysis. We examine whether

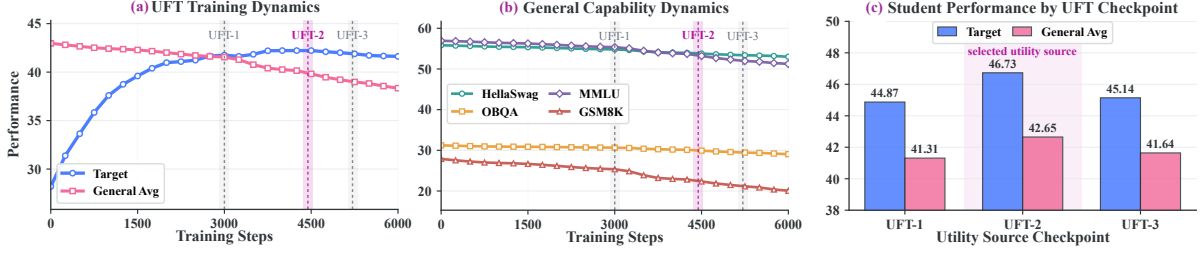


Figure 4: (a) Training dynamics of the unconstrained fine-tuned teacher. (b) General-capability dynamics during unconstrained fine-tuning. (c) Student performance using UFT checkpoints from different training stages.

Weighting Signal	General $\uparrow$	Target $\uparrow$	Overall $\uparrow$
Source-loss $\downarrow$	42.90	41.72	42.31
UFT-confidence $\uparrow$	39.96	45.18	42.57
Divergence $\downarrow$	42.10	41.88	41.99
Source-uncertainty $\uparrow$	40.92	43.74	42.33
Uniform-UFT	40.53	44.28	42.41
Shuffled-UFT	39.78	43.30	41.54
Inverse-UFT $\downarrow$	38.74	40.08	39.41
UFT	42.65	46.73	44.69

Table 3: Effectiveness analysis of UFT value discovery on LLaMA3.2-3B. Signal-side $\uparrow/\downarrow$ mean larger/smaller raw values receive larger weights. Metric-side $\uparrow$ means higher is better.

Weighting Signal	General $\uparrow$	Target $\uparrow$	Overall $\uparrow$
Source-loss $\downarrow$	46.12	62.84	54.48
UFT-confidence $\uparrow$	42.86	68.92	55.89
Divergence $\downarrow$	45.51	63.57	54.54
Source-uncertainty $\uparrow$	44.38	66.44	55.41
Uniform-UFT	44.07	67.36	55.72
Shuffled-UFT	42.95	65.34	54.15
Inverse-UFT $\downarrow$	41.55	60.72	51.14
UFT	45.82	69.45	57.63

Table 4: Effectiveness analysis of UFT value discovery on Gemma3-4B. Signal-side $\uparrow/\downarrow$ mean larger/smaller raw values receive larger weights. Metric-side $\uparrow$ means higher is better.

UFT’s utility scores capture target-sample value. All variants use the same training setting, and UFT-based variants share the same UFT teacher. Table 3 and 4 show that conservative signals such as Source-loss preserve general capability but under-adapt, while UFT-confidence improves target performance yet fails to control retention-sensitive drift. Uniform-UFT keeps the same UFT-guided supervision but removes value-based sample differentiation, showing the gain comes from value discovery rather than teacher guidance alone. Shuffled-UFT and Inverse-UFT show that utility ordering is essential for prioritizing samples that jointly support adaptation and retention. These results support UFT’s value-discovery role through utilities coupling target-task reliability and source-relative compatibility.

5.4 Ablation Study and Hyperparameter (Q3)

Ablation Study. We conduct ablations to examine how UFT balances adaptation and preservation. Table 5 shows that removing value-based sample differentiation causes the largest target and overall drops, indicating that value discovery must guide sample emphasis. Removing relative divergence mainly hurts general capability, showing that UFT-

side target evidence alone is insufficient without source-relative preservation control. The asymmetric effects of removing SRefT and UFT distillation further align with their preservation and adaptation roles, while removing confidence-adaptive distillation weakens sample-level teacher balancing. Overall, UFT’s main gain comes from preservation-aware value discovery, with dual-teacher distillation providing supervision for stable fine-tuning.

Method	General	Target	Overall
UFT	42.65	46.73	44.69
- w/o Utility weighting	40.53	44.28	42.41
- w/o SRefT distillation	40.19	46.84	43.52
- w/o UFT distillation	42.86	44.96	43.91
- w/o Confidence adaptive	41.55	45.52	43.54
- w/o Relative divergence	40.37	45.34	42.86

Table 5: Ablation study of UFT components.

Hyperparameter Analysis. We analyze UFT’s sensitivity to five hyperparameters. Fig. 5 shows how UFT coordinates value discovery, utility-guided weighting, and dual-teacher guidance. For the reliability gate, restrictive or sharp settings discard useful target signals, while permissive or smooth settings blur reliable and noisy UFT ev-

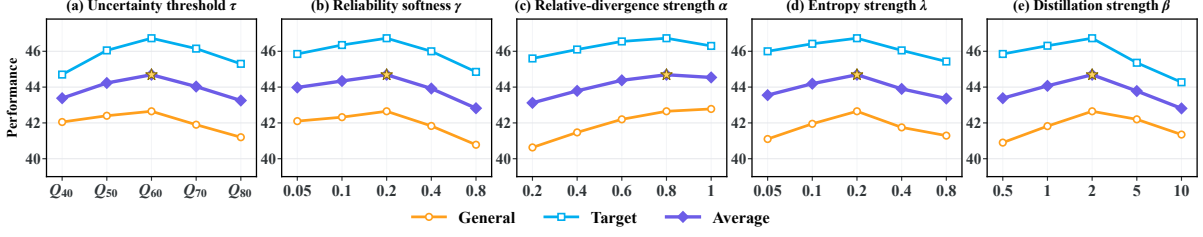


Figure 5: Hyperparameter sensitivity of UFT on Magicoder with LLaMA3.2-3B.

idence. Larger α shifts valuation toward source compatibility, improving retention but hurting target performance when task-induced drift is overpenalized. λ controls utility-weight concentration, with small values causing brittle emphasis and large values approaching uniform weighting. β controls dual-teacher guidance. Overall, stable trends around the selected setting suggest that UFT’s gains arise from coordinated value discovery rather than a narrow hyperparameter choice.

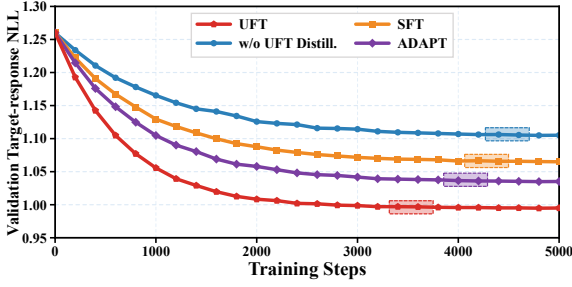


Figure 6: Convergence analysis on the target task.

5.5 Convergence and Runtime Analysis (Q4)

Convergence Analysis. We analyze convergence using a common validation target-response NLL, since different methods optimize different training objectives and their training losses are not directly comparable. We compute NLL every 200 training steps and define convergence as fluctuation within 2% over a 400-step window. Fig. 6 compares UFT with SFT, ADAPT, and UFT w/o UFT Distill. UFT reaches a lower, more stable NLL earlier, while removing UFT-guided distillation slows descent and raises the late-stage plateau. This aligns with Appendix B.3, where UFT-guided distillation provides the curvature term $\beta G_{\text{uft}} e_t$ for local contraction toward UFT-compatible target behavior.

Runtime Evaluation. Fig. 7 reports the cost and performance trade-off on Magicoder with Llama3.2, with UFT measured under its stage-wise pipeline from offline value discovery to final fine-tuning. UFT remains moderately costly relative to its gains. By using discovered sample values to

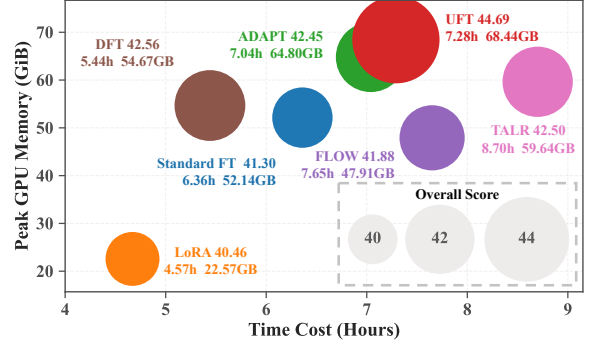


Figure 7: Runtime–performance trade-off. Time cost and peak GPU memory are measured on four GPUs, and bubble size denotes the overall score.

guide dual-teacher training, UFT achieves the best retention and target-task performance. Some methods incur lower overhead, but they yield smaller overall gains, leaving UFT practically valuable. Appendix D.5 details the teacher-construction cost.

5.6 Additional Robustness Analysis

We report instruction-tuning results on Alpaca-GPT4 (Peng et al., 2023) across 11 tasks in Appendix D.7, demonstrating UFT’s generalization. Appendix D.8 compares UFT with a DavIR (Zhou et al., 2025) variant, validating the effectiveness of UFT’s value discovery mechanism.

6 Conclusion

We introduced UFT, a fine-tuning framework that shifts the teacher role from knowledge transfer to preservation-aware value discovery in LLM adaptation. By exposing task-induced drift relative to the source model, UFT discovers high-value target-task samples with reliable adaptation value and limited source-behavior conflict. These discovered values guide dual-teacher supervision, where SRefT anchors source behavior and UFT provides value-weighted target-task guidance. Experiments on Magicoder and MetaMathQA show UFT improves the adaptation-retention trade-off.

Limitations

Although UFT improves the adaptation–retention trade-off, it has several limitations. First, the unconstrained fine-tuned teacher provides task-specialized evidence for value discovery, but constructing it and performing teacher-based valuation add computation and memory compared with lightweight methods. Second, UFT uses a one-shot utility estimation strategy. This avoids repeated on-line scoring and keeps student training efficient, but it also means that sample weights are not updated as the student evolves. Future work could explore dynamic utility refreshing to better capture changing sample value during training. Third, teacher-forced uncertainty and source-relative divergence provide annotation-free value signals, but their reliability may decrease for open-ended tasks or multi-answer data. Extending UFT to multilingual, multimodal, long-context, and continual multi-task settings, together with developing lower-cost navigator construction for improved robustness across tasks and fine-tuning recipes, remains an important direction for future work.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62476191 and 42550187, with computing resources provided by our research group. Generative AI tools were used to assist with language polishing during manuscript preparation.

References

- Shivam Adarsh, Kumar Shridhar, Caglar Gulcehre, Nicholas Monath, and Mrinmaya Sachan. 2025. SIKeD: Self-guided iterative knowledge distillation for mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9868–9880. Association for Computational Linguistics.
- Amirabbas Afzali, Myeongho Jeon, and Maria Brbic. 2026. When weak llms speak with confidence, preference alignment gets stronger. In *International Conference on Learning Representations*, volume 2026, pages 129320–129351.
- Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pages 21246–21263.
- Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. 2024. LoRA learns less and forgets less. *Transactions on Machine Learning Research*.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.
- 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, and 39 others. 2021. [Evaluating large language models trained on code](#). *Preprint*, arXiv:2107.03374.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have solved question answering? try arc, the ai2 reasoning challenge](#). *Preprint*, arXiv:1803.05457.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. [Training verifiers to solve math word problems](#). *Preprint*, arXiv:2110.14168.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2023. [A framework for few-shot language model evaluation](#).
- Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International journal of computer vision*, 129(6):1789–1819.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pages 32694–32717.
- Guangzeng Han and Xiaolei Huang. 2026. What makes good instruction-tuning data? an in-context learning perspective. In *Proceedings of the 64th Annual*

- Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1013–1027. Association for Computational Linguistics.
- Pengfei He. 2024. [Parameter efficient instruction tuning: An empirical study](#). *Preprint*, arXiv:2411.16775.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). *Preprint*, arXiv:1503.02531.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Ijun Jang, Jewon Yeom, Juan Yeo, Hyunggyu Lim, and Taesup Kim. 2026. Stable on-policy distillation through adaptive target reformulation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 42217–42227. Association for Computational Linguistics.
- Jongwoo Ko, Sara Abdali, Young Jin Kim, Tianyi Chen, and Pashmina Cameron. 2026. [Scaling reasoning efficiently via relaxed on-policy distillation](#). *Preprint*, arXiv:2603.11137.
- Ming Li, Han Chen, Chenguang Wang, Dang Nguyen, Dianqi Li, and Tianyi Zhou. 2025a. RuleR: Improving llm controllability by rule-based data recycling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 926–943.
- Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7602–7635.
- Weicheng Li, Lixin Zou, Min Tang, Qing Yu, Wanli Li, and Chenliang Li. 2025b. Meta-lora: Memory-efficient sample reweighting for fine-tuning large language models. In *Proceedings of the 31st international conference on computational linguistics*, pages 8504–8517.
- Yanping Li, Zhening Liu, Zijian Li, Zehong Lin, and Jun Zhang. 2026. Token-level data selection for safe llm fine-tuning. In *International Conference on Learning Representations*, volume 2026, pages 57317–57345.
- 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. 2026. Sft doesn’t always hurt general capabilities: Revisiting domain-specific fine-tuning in llms. In *International Conference on Learning Representations*, volume 2026, pages 46954–46992.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252.
- Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, and Weizhu Chen. 2024. Not all tokens are what you need for pretraining. In *Advances in Neural Information Processing Systems*, volume 37, pages 29029–29063. Curran Associates, Inc.
- Yajiao Liu, Congliang Chen, Junchi Yang, and Ruoyu Sun. 2026a. Rethinking data mixture for large language models: A comprehensive survey and new perspectives. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 275–289, Rabat, Morocco. Association for Computational Linguistics.
- Zeming Liu, Yibai Liu, Shihang Wang, Zheming Song, Junzhe Wang, Jingjing Liu, Qingjie Liu, Guangxu Chen, and Yunhong Wang. 2026b. Learn more, forget less: A gradient-aware data selection approach for llm. *Signal Processing*, page 110611.
- Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jian-Guang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. 2025. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. In *International Conference on Learning Representations*, volume 2025, pages 49573–49609.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2381–2391.
- Paramita Mirza, Lucas Weber, and Fabian K  ch. 2025. Stratified selective sampling for instruction tuning with dedicated scoring strategy. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19949–19974.
- Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198.

- Rui Pan, Dylan Zhang, Hanning Zhang, Xingyuan Pan, Minrui Xu, Jipeng Zhang, Renjie Pi, Xiaoyu Wang, and Tong Zhang. 2025. ScaleBiO: Scalable bilevel optimization for LLM data reweighting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31959–31982, Vienna, Austria. Association for Computational Linguistics.
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. [Instruction tuning with gpt-4](#). *Preprint*, arXiv:2304.03277.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.
- Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. 2025. Upweighting easy samples in fine-tuning mitigates forgetting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 52922–52957. PMLR.
- Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. 2018. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4528–4537. PMLR.
- Ao Shen, Zhiqian Lai, Qiang Wang, Xiongwei Li, Lizhi Zhang, Dongsheng Li, and Jiaxin Li. 2025. Accurate and efficient fine-tuning of quantized large language models through optimal balance in adaptation. *Transactions of the Association for Computational Linguistics*, 13:861–877.
- Yuncheng Song, Liang Ding, Changtong Zan, and Shujian Huang. 2025. Self-evolution knowledge distillation for LLM-based machine translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10298–10308, Abu Dhabi, UAE.
- Daouda Sow, Herbert Woisetschl ger, Saikiran Bulusu, Shiqiang Wang, Hans Arno Jacobsen, and Yingbin Liang. 2025. Dynamic loss-based sample reweighting for improved large language model pretraining. In *International Conference on Learning Representations*, volume 2025, pages 89490–89520.
- Ali Taheri, Alireza Taban, Qizhou Wang, Shanshan Ye, Abdolreza Mirzaei, Tongliang Liu, and Bo Han. 2026. Forgetting: A new mechanism towards better large language model fine-tuning. *Transactions on Machine Learning Research*. J2C Certification.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram  , Morgane Rivi  re, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).
- Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. 2025. Emo: Embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7594–7606. Association for Computational Linguistics.
- Shaobo Wang, Xiangqi Jin, Ziming Wang, Jize Wang, Jiajun Zhang, Kaixin Li, Zichen Wen, Zhong Li, Conghui He, Xuming Hu, and Linfeng Zhang. 2025a. Data whisperer: Efficient data selection for task-specific LLM fine-tuning via few-shot in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23287–23305. Association for Computational Linguistics.
- Yihan Wang, Andrew Bai, Nanyun Peng, and Cho-Jui Hsieh. 2025b. On the loss of context awareness in general instruction fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, pages 8610–8635. Curran Associates, Inc.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.
- Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. 2024. Magicoder: Empowering code generation with OSS-instruct. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52632–52657. PMLR.
- Chao-Chung Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung (Vivian) Chen, Shao-Hua Sun, and Hung-yi Lee. 2025a. Mitigating forgetting in llm fine-tuning via low-perplexity token learning. In *Advances in Neural Information Processing Systems*, volume 38, pages 1708–1744. Curran Associates, Inc.
- Junhong Wu, Yang Zhao, Yangyifan Xu, Bing Liu, and Chengqing Zong. 2025b. Boosting llm translation skills without general ability loss via rationale distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12217–12236.
- Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi,

- Ming-Hsuan Yang, and xu yang. 2026. On the generalization of sft: A reinforcement learning perspective with reward rectification. In *International Conference on Learning Representations*, volume 2026, pages 27550–27571.
- Mingxuan Xia, Haobo Wang, Yixuan Li, Zewei Yu, Jindong Wang, Junbo Zhao, and Runze Wu. 2025. Prompt candidates, then distill: A teacher-student framework for llm-driven data annotation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2750–2770.
- Fan Yin, Zifeng Wang, I-Hung Hsu, Jun Yan, Ke Jiang, Yanfei Chen, Jindong Gu, Long Le, Kai-Wei Chang, Chen-Yu Lee, Hamid Palangi, and Tomas Pfister. 2025. Magnet: Multi-turn tool-use data synthesis and distillation via graph translation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32600–32616, Vienna, Austria. Association for Computational Linguistics.
- Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024a. Meta-math: Bootstrap your own mathematical questions for large language models. In *International Conference on Learning Representations*, volume 2024, pages 45040–45061.
- Yu-Chu Yu, Chi-Pin Huang, Jr-Jen Chen, Kai-Po Chang, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. 2024b. Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models. In *European Conference on Computer Vision*, pages 219–236. Springer.
- Xiaosong Yuan, Chen Shen, Shaotian Yan, kaiyuan liu, Xiaofeng Zhang, Sinan Fan, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and Jieping Ye. 2026. Differential fine-tuning large language models towards better diverse reasoning abilities. In *International Conference on Learning Representations*, volume 2026, pages 110433–110463.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.
- Wanru Zhao, Yihong Chen, Yuzhi Tang, Wentao Ma, Shengchao Hu, Xu Hu, Alex Iacob, Abhinav Mehrotra, and Nic Lane. 2026. Rethinking data curation in llm training: Online reweighting offers better generalization than offline methods. In *International Conference on Learning Representations*, volume 2026, pages 112899–112923.
- Haotian Zhou, Tingkai Liu, Qianli Ma, Yufeng Zhang, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. 2025. DavIR: Data selection via implicit reward for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9220–9237, Vienna, Austria. Association for Computational Linguistics.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.

A Six Utility Regimes Analysis

In this section, we analyze the six utility regimes induced by Eq. (3). The utility is

$$u_i = \sigma\left(\frac{\tau - s_i^{\text{uft}}}{\gamma}\right) \cdot \exp[-\alpha \max(0, d_i - s_i^{\text{ref}})], \quad (15)$$

where smaller uncertainty indicates higher confidence. The first factor is a soft UFT-reliability gate, and the second penalizes UFT-SRefT divergence only when it exceeds the source-model uncertainty budget; thus, $d_i = s_i^{\text{ref}}$ separates tolerated from penalized divergence. For clarity, denote

$$\begin{aligned} r_i^{\text{uft}} &= \sigma\left(\frac{\tau - s_i^{\text{uft}}}{\gamma}\right) \\ p_i^{\text{rel}} &= \exp[-\alpha \max(0, d_i - s_i^{\text{ref}})], \end{aligned} \quad (16)$$

so that $u_i = r_i^{\text{uft}} p_i^{\text{rel}}$. A small s_i^{uft} gives a large reliability factor, while a large s_i^{uft} attenuates the sample regardless of the relative-divergence term.

Case 1: UFT unreliable, source confident. When $s_i^{\text{uft}} > \tau$ and s_i^{ref} is small, the UFT-reliability factor r_i^{uft} is small. The sample lies in a region where the source model is confident, but UFT does not provide reliable target-task evidence. Therefore, its utility is reliability-attenuated:

$$u_i = r_i^{\text{uft}} p_i^{\text{rel}}, \quad r_i^{\text{uft}} < \frac{1}{2}. \quad (17)$$

If $d_i > s_i^{\text{ref}}$, the relative-divergence penalty further suppresses the sample. This regime avoids using uncertain UFT evidence to perturb confident source-model behavior.

Case 2: UFT unreliable, source uncertain. When $s_i^{\text{uft}} > \tau$ and s_i^{ref} is large, the source model provides weaker guidance and allows a larger divergence budget. However, the utility remains attenuated because UFT itself is not reliable:

$$u_i = r_i^{\text{uft}} p_i^{\text{rel}}, \quad r_i^{\text{uft}} < \frac{1}{2}. \quad (18)$$

Thus, source uncertainty alone is insufficient for high utility. A sample is emphasized only when the unconstrained teacher also provides reliable target-task evidence.

Case 3: UFT reliable, source confident, divergence within budget. When $s_i^{\text{uft}} \leq \tau$, s_i^{ref} is small, and $d_i \leq s_i^{\text{ref}}$, the relative-divergence penalty is inactive:

$$u_i = r_i^{\text{uft}}. \quad (19)$$

Since UFT is reliable and the source model is confident, satisfying $d_i \leq s_i^{\text{ref}}$ means that UFT agrees with strongly anchored source behavior within a tight budget. This is an aligned reinforcement regime, where the sample supports target adaptation without threatening preserved behavior.

Case 4: UFT reliable, source confident, divergence exceeds budget. When $s_i^{\text{uft}} \leq \tau$, s_i^{ref} is small, and $d_i > s_i^{\text{ref}}$, the sample enters a high-conflict regime:

$$u_i = r_i^{\text{uft}} \exp[-\alpha(d_i - s_i^{\text{ref}})]. \quad (20)$$

Although UFT provides reliable target-task evidence, its deviation exceeds the small budget allowed by confident source behavior. The exponential term therefore softly suppresses the sample in proportion to the excess divergence.

Case 5: UFT reliable, source uncertain, divergence within budget. When $s_i^{\text{uft}} \leq \tau$, s_i^{ref} is large, and $d_i \leq s_i^{\text{ref}}$, the penalty is again inactive:

$$u_i = r_i^{\text{uft}}. \quad (21)$$

Unlike Case 3, the source model is uncertain, so the larger uncertainty budget permits more target-task exploration. This is a low-conflict exploration regime, where reliable UFT evidence is useful and remains compatible with the source model's limited confidence.

Case 6: UFT reliable, source uncertain, divergence exceeds budget. When $s_i^{\text{uft}} \leq \tau$, s_i^{ref} is large, and $d_i > s_i^{\text{ref}}$, the utility becomes

$$u_i = r_i^{\text{uft}} \exp[-\alpha(d_i - s_i^{\text{ref}})]. \quad (22)$$

This regime corresponds to novel mismatch. The source model is uncertain and therefore permits a larger divergence budget than in confident regions. However, once UFT-SRefT divergence exceeds even this relaxed budget, the sample is still down-weighted. Thus, source uncertainty relaxes preservation pressure but does not remove the relative-divergence penalty.

B Theoretical Analysis

B.1 Closed-form solution of the weight optimization.

In this section, we derive the closed-form optimizer of the entropy-regularized weighting objective used

in Eq. (4). Consider the optimization problem

$$\begin{aligned} \max_{\mathbf{w} \in \Delta^{N-1}} \sum_{i=1}^N w_i u_i - \lambda \sum_{i=1}^N w_i \log w_i, \\ \Delta^{N-1} := \left\{ \mathbf{w} \in \mathbb{R}_{\geq 0}^N : \sum_{i=1}^N w_i = 1 \right\}, \end{aligned} \quad (23)$$

where $\lambda > 0$. Since the first term is linear in $\mathbf{w}$ and the entropy term $-\sum_{i=1}^N w_i \log w_i$ is strictly concave on the probability simplex, the objective in Eq. (23) is strictly concave over Δ^{N-1} . Hence, the problem admits a unique global maximizer.

The unique optimizer can be characterized by the first-order optimality conditions under the simplex constraint. We therefore consider the Lagrangian

$$\begin{aligned} \mathcal{L}(\mathbf{w}, \mu) = \sum_{i=1}^N w_i u_i - \lambda \sum_{i=1}^N w_i \log w_i \\ + \mu \left(\sum_{i=1}^N w_i - 1 \right), \end{aligned} \quad (24)$$

where μ is the Lagrange multiplier associated with the simplex constraint. The first-order stationarity condition with respect to w_i gives

$$\frac{\partial \mathcal{L}}{\partial w_i} = u_i - \lambda(\log w_i + 1) + \mu. \quad (25)$$

Eq. (25) vanishes at the optimum, which yields

$$u_i - \lambda(\log w_i + 1) + \mu = 0, \quad (26)$$

which can be rearranged as

$$\log w_i = \frac{u_i + \mu}{\lambda} - 1. \quad (27)$$

The remaining constant C is fixed by the simplex constraint $\sum_{i=1}^N w_i = 1$, which gives

$$\begin{aligned} w_i &= \exp\left(\frac{u_i + \mu}{\lambda} - 1\right) \\ &= \exp\left(\frac{u_i}{\lambda}\right) \exp\left(\frac{\mu}{\lambda} - 1\right). \end{aligned} \quad (28)$$

Since the second factor is independent of i , we denote

$$C := \exp\left(\frac{\mu}{\lambda} - 1\right), \quad (29)$$

so that

$$w_i = C \exp(u_i/\lambda). \quad (30)$$

Enforcing the simplex constraint $\sum_{i=1}^N w_i = 1$, we obtain

$$C = \frac{1}{\sum_{j=1}^N \exp(u_j/\lambda)}. \quad (31)$$

Substituting Eq. (31) into Eq. (30) yields the unique optimum

$$w_i^* = \frac{\exp(u_i/\lambda)}{\sum_{j=1}^N \exp(u_j/\lambda)}. \quad (32)$$

Therefore, the entropy-regularized utility maximization problem admits the softmax solution in closed form.

B.2 Proof of the Theoretical Justification

We provide detailed proofs for the near-optimality of validation-based early stopping and the conditional stability of utility estimates under the early-stopped UFT checkpoint.

B.2.1 Proof of Theorem 1

Notation. Let $\{\theta_t\}_{t=1}^T$ denote the sequence of UFT checkpoints along the unconstrained fine-tuning trajectory. For each checkpoint θ_t , define the population risk

$$\mathcal{L}_{\text{pop}}(\theta_t) := \mathbb{E}_{z \sim \mathcal{D}}[\ell(\theta_t; z)], \quad (33)$$

and the empirical validation risk

$$\hat{\mathcal{L}}_{\text{val}}(\theta_t) := \frac{1}{n_{\text{val}}} \sum_{j=1}^{n_{\text{val}}} \ell(\theta_t; z_j^{\text{val}}), \quad (34)$$

where $\{z_j^{\text{val}}\}_{j=1}^{n_{\text{val}}}$ are independent validation samples from the target-task population $\mathcal{D}$. Let

$$\begin{aligned} t_{\text{val}} &= \arg \min_{1 \leq t \leq T} \hat{\mathcal{L}}_{\text{val}}(\theta_t) \\ t_{\text{pop}} &= \arg \min_{1 \leq t \leq T} \mathcal{L}_{\text{pop}}(\theta_t). \end{aligned} \quad (35)$$

Proof. Condition on the UFT training trajectory $\{\theta_t\}_{t=1}^T$. Since the validation set is independent of this trajectory, the validation losses are independent bounded random variables for each fixed checkpoint. For any θ_t , Hoeffding's inequality gives

$$\begin{aligned} \Pr\left(\left|\hat{\mathcal{L}}_{\text{val}}(\theta_t) - \mathcal{L}_{\text{pop}}(\theta_t)\right| \geq \varepsilon\right) \\ \leq 2 \exp\left(-\frac{2n_{\text{val}}\varepsilon^2}{B^2}\right). \end{aligned} \quad (36)$$

Applying the union bound over the finite checkpoint set yields

$$\begin{aligned} \Pr\left(\sup_{1 \leq t \leq T} \left|\hat{\mathcal{L}}_{\text{val}}(\theta_t) - \mathcal{L}_{\text{pop}}(\theta_t)\right| \geq \varepsilon\right) \\ \leq 2T \exp\left(-\frac{2n_{\text{val}}\varepsilon^2}{B^2}\right). \end{aligned} \quad (37)$$

Thus, with probability at least $1 - \delta$,

$$\sup_{1 \leq t \leq T} |\hat{\mathcal{L}}_{\text{val}}(\theta_t) - \mathcal{L}_{\text{pop}}(\theta_t)| \leq B \sqrt{\frac{\log(2T/\delta)}{2n_{\text{val}}}}. \quad (38)$$

On the event in Eq. (38), we have

$$\mathcal{L}_{\text{pop}}(\theta_{t_{\text{val}}}) \leq \hat{\mathcal{L}}_{\text{val}}(\theta_{t_{\text{val}}}) + B \sqrt{\frac{\log(2T/\delta)}{2n_{\text{val}}}}. \quad (39)$$

Since t_{val} minimizes the empirical validation risk,

$$\hat{\mathcal{L}}_{\text{val}}(\theta_{t_{\text{val}}}) \leq \hat{\mathcal{L}}_{\text{val}}(\theta_{t_{\text{pop}}}). \quad (40)$$

Applying Eq. (38) again to $\theta_{t_{\text{pop}}}$ gives

$$\hat{\mathcal{L}}_{\text{val}}(\theta_{t_{\text{pop}}}) \leq \mathcal{L}_{\text{pop}}(\theta_{t_{\text{pop}}}) + B \sqrt{\frac{\log(2T/\delta)}{2n_{\text{val}}}}. \quad (41)$$

Combining Eqs. (39)–(41), we obtain

$$\mathcal{L}_{\text{pop}}(\theta_{t_{\text{val}}}) \leq \mathcal{L}_{\text{pop}}(\theta_{t_{\text{pop}}}) + 2B \sqrt{\frac{\log(2T/\delta)}{2n_{\text{val}}}}. \quad (42)$$

B.2.2 Proof of Proposition 1

Proof. For any sample i and any $k \leq K$, the triangle inequality gives

$$|u_i - v_i(\theta_k)| \leq |u_i - v_i(\theta_0)| + |v_i(\theta_k) - v_i(\theta_0)| \leq \epsilon + \Delta_K = \eta_K. \quad (43)$$

Therefore $\|\mathbf{u} - \mathbf{v}(\theta_k)\|_{\infty} \leq \eta_K$. Let

$$\begin{aligned} Z_{\text{uft}} &= \sum_{j=1}^N \exp(u_j/\lambda) \\ Z_{\text{dyn}}(\theta_k) &= \sum_{j=1}^N \exp(v_j(\theta_k)/\lambda). \end{aligned} \quad (44)$$

Since $|u_j - v_j(\theta_k)| \leq \eta_K$ for all j , we have

$$\begin{aligned} \exp\left(-\frac{\eta_K}{\lambda}\right) \exp(v_j(\theta_k)/\lambda) &\leq \exp(u_j/\lambda) \\ &\leq \exp\left(\frac{\eta_K}{\lambda}\right) \exp(v_j(\theta_k)/\lambda). \end{aligned} \quad (45)$$

Summing over j yields

$$\begin{aligned} \exp\left(-\frac{\eta_K}{\lambda}\right) Z_{\text{dyn}}(\theta_k) &\leq Z_{\text{uft}} \\ &\leq \exp\left(\frac{\eta_K}{\lambda}\right) Z_{\text{dyn}}(\theta_k). \end{aligned} \quad (46)$$

Equivalently,

$$\exp\left(-\frac{\eta_K}{\lambda}\right) \leq \frac{Z_{\text{dyn}}(\theta_k)}{Z_{\text{uft}}} \leq \exp\left(\frac{\eta_K}{\lambda}\right). \quad (47)$$

For each sample i , the weight ratio satisfies

$$\begin{aligned} \frac{w_i^{\text{uft}}}{w_i^{\text{dyn}}(\theta_k)} &= \frac{\exp(u_i/\lambda)}{Z_{\text{uft}}} \cdot \frac{Z_{\text{dyn}}(\theta_k)}{\exp(v_i(\theta_k)/\lambda)} \\ &= \exp\left(\frac{u_i - v_i(\theta_k)}{\lambda}\right) \frac{Z_{\text{dyn}}(\theta_k)}{Z_{\text{uft}}}. \end{aligned} \quad (48)$$

Using $|u_i - v_i(\theta_k)| \leq \eta_K$ together with the normalization-bound above gives

$$\exp\left(-\frac{2\eta_K}{\lambda}\right) \leq \frac{w_i^{\text{uft}}}{w_i^{\text{dyn}}(\theta_k)} \leq \exp\left(\frac{2\eta_K}{\lambda}\right). \quad (49)$$

It remains to prove the total-variation bound. Let $a = \exp\left(-\frac{2\eta_K}{\lambda}\right)$. The ratio bound implies $w_i^{\text{uft}} \geq a w_i^{\text{dyn}}(\theta_k)$ and $w_i^{\text{dyn}}(\theta_k) \geq a w_i^{\text{uft}}$ for all i . Hence,

$$\min(w_i^{\text{uft}}, w_i^{\text{dyn}}(\theta_k)) \geq a w_i^{\text{dyn}}(\theta_k). \quad (50)$$

Summing over i gives

$$\sum_{i=1}^N \min(w_i^{\text{uft}}, w_i^{\text{dyn}}(\theta_k)) \geq a. \quad (51)$$

Using

$$\text{TV}(\mathbf{p}, \mathbf{q}) = 1 - \sum_{i=1}^N \min(p_i, q_i), \quad (52)$$

we obtain

$$\text{TV}(\mathbf{w}^{\text{uft}}, \mathbf{w}^{\text{dyn}}(\theta_k)) \leq 1 - \exp\left(-\frac{2\eta_K}{\lambda}\right). \quad (53)$$

B.3 Local Prediction-Space Contraction of UFT-Guided Distillation

We analyze the local dynamics underlying the faster convergence observed with UFT. Value-weighted UFT-guided distillation introduces a teacher-anchored contraction along directions where UFT supervision is reliable and compatible with source preservation.

Let $z_i(\theta)$ denote the teacher-forced student logits on sample i , and let z_i^{uft} denote the corresponding UFT logits. For notational simplicity, we suppress the token index; the same argument applies after summing over response positions. Define $a_i = w_i^* n_i$, where w_i^* is the utility-induced sample weight and n_i is the reliability coefficient assigned to UFT guidance. The value-weighted UFT distillation term is

$$\mathcal{L}_{\text{vd}}(\theta) = \sum_{i=1}^N a_i D_{\text{KL}}(P_{\text{uft}}(\cdot|x_i) \| P_{\theta}(\cdot|x_i)). \quad (54)$$

We analyze this term around a local UFT-compatible reference point θ_* , where the student predictions match the UFT predictions on the considered teacher-forced contexts. This reference characterizes the target-behavior component induced by UFT guidance, while the full student objective also includes the SRefT anchor for source preservation. Equivalently, θ_* may be viewed as a local minimizer of $\mathcal{L}_{\text{vd}}$, in which case the same contraction argument applies with $G_{\text{uft}} = \nabla^2 \mathcal{L}_{\text{vd}}(\theta_*)$. We use the matched-prediction form below to make the UFT-induced curvature explicit.

Let $e_t = \theta_t - \theta_*$. Assume that the logit map admits the local expansion

$$\begin{aligned} z_i(\theta) &= z_i^{\text{uft}} + J_i(\theta - \theta_*) + r_i(\theta), \\ \|r_i(\theta)\|_2 &= O(\|\theta - \theta_*\|_2^2), \end{aligned} \quad (55)$$

where $J_i = \nabla_{\theta} z_i(\theta_*)$ and $\|J_i\|_2$ is locally bounded. For the forward KL term, a second-order expansion around z_i^{uft} gives

$$\begin{aligned} D_{\text{KL}}(P_{\text{uft}}(\cdot|x_i) \| P_{\theta}(\cdot|x_i)) \\ = \frac{1}{2} (z_i(\theta) - z_i^{\text{uft}})^{\top} H_i (z_i(\theta) - z_i^{\text{uft}}) \\ + O(\|z_i(\theta) - z_i^{\text{uft}}\|_2^3). \end{aligned} \quad (56)$$

where

$$H_i = \nabla_z^2 D_{\text{KL}}(P_{\text{uft}}(\cdot|x_i) \| \text{softmax}(z)) \Big|_{z=z_i^{\text{uft}}} \quad (57)$$

is the categorical Fisher matrix evaluated at the UFT prediction. It is positive semidefinite and inherits the softmax shift-invariance null direction. Through the Jacobians $\{J_i\}$, this direction induces a parameter-space kernel of G_{uft} , which is automatically removed by the spectral projection below.

Since $\nabla \mathcal{L}_{\text{vd}}(\theta_*) = 0$ at the UFT-compatible reference, substituting the local expansion of $z_i(\theta)$ into the KL approximation yields

$$\mathcal{L}_{\text{vd}}(\theta) = \frac{1}{2} e^{\top} G_{\text{uft}} e + O(\|e\|_2^3), \quad (58)$$

where

$$G_{\text{uft}} = \sum_{i=1}^N a_i J_i^{\top} H_i J_i. \quad (59)$$

Thus, locally,

$$\nabla \mathcal{L}_{\text{vd}}(\theta_t) = G_{\text{uft}} e_t + O(\|e_t\|_2^2). \quad (60)$$

Now consider the full UFT student objective. Its local gradient update can be written as

$$e_{t+1} = e_t - \eta (\beta G_{\text{uft}} e_t + b_t), \quad (61)$$

where b_t collects the task-loss gradient, the SRefT-anchor gradient, and higher-order local residuals.

Let P be the spectral projection onto an invariant subspace of G_{uft} whose eigenvalues lie in $[\mu_{\text{uft}}, M_{\text{uft}}]$ with $0 < \mu_{\text{uft}} \leq M_{\text{uft}}$. Since G_{uft} is symmetric positive semidefinite, this spectral projection is well defined. By construction, $G_{\text{uft}} P = P G_{\text{uft}} P$, so the projected update decouples from the orthogonal complement. Suppose the remaining projected terms satisfy $\|P b_t\|_2 \leq \nu \|P e_t\|_2 + \delta_t$. Projecting the update onto P gives

$$P e_{t+1} = (I - \eta \beta P G_{\text{uft}} P) P e_t - \eta P b_t. \quad (62)$$

For $0 < \eta \leq 1/(\beta M_{\text{uft}})$, we have

$$\|I - \eta \beta P G_{\text{uft}} P\|_2 \leq 1 - \eta \beta \mu_{\text{uft}}. \quad (63)$$

Therefore,

$$\|P e_{t+1}\|_2 \leq (1 - \eta \beta \mu_{\text{uft}}) \|P e_t\|_2 + \eta \|P b_t\|_2. \quad (64)$$

Using the bound on $P b_t$ gives

$$\|P e_{t+1}\|_2 \leq (1 - \eta(\beta \mu_{\text{uft}} - \nu)) \|P e_t\|_2 + \eta \delta_t. \quad (65)$$

Hence, when $\beta \mu_{\text{uft}} > \nu$ and δ_t is controlled, the projected error contracts toward UFT-compatible target behavior.

If $\delta_t \leq \delta$ uniformly and $\rho = 1 - \eta(\beta \mu_{\text{uft}} - \nu) < 1$, the recurrence further gives

$$\limsup_{t \rightarrow \infty} \|P e_t\|_2 \leq \frac{\delta}{\beta \mu_{\text{uft}} - \nu}. \quad (66)$$

Thus the projected error enters a bounded neighborhood of the UFT-compatible reference. This behavior is consistent with the dual-teacher design: UFT pulls the student toward target-task behavior on the projected subspace, while SRefT anchors source-model behavior and contributes to the steady-state balance through δ_t .

This local contraction stems from UFT-guided distillation: $a_i = w_i^* n_i$ concentrates teacher-anchored curvature on samples with high discovered utility and reliable UFT guidance, while the relative-divergence term suppresses samples with strong source-relative conflict, reducing the conflicting SRefT-anchor component in b_t . Thus, UFT strengthens teacher-guided contraction on preservation-compatible high-value samples, explaining the faster convergence observed in Fig. 6.

C Experimental Details

In this section, we provide detailed descriptions of the experimental setup for reproducibility. Section C.1 introduces baseline implementations and configurations. Sections C.2 and C.3 describe the benchmarks and evaluation details used in our experiments. Section C.4 presents implementation details, including hyperparameter settings and training configurations. Section C.5 provides the prompt templates used for training. Section C.6 describes unconstrained fine-tuned teacher training and the validation setting for checkpoint selection.

C.1 Baseline Details

We compare UFT against strong baselines spanning full fine-tuning, parameter-efficient fine-tuning, and weighting-based fine-tuning.

- **Standard Fine-Tuning (Standard FT):** This baseline updates all model parameters using the standard cross-entropy loss. It typically suffers from severe catastrophic forgetting and thus serves as a lower-bound reference for general capability preservation.
- **LoRA (Hu et al., 2022):** LoRA freezes the backbone model and injects trainable low-rank matrices into Transformer layers. By restricting the update space, LoRA reduces training cost and can mitigate forgetting, but may underfit task-specific behaviors compared with full-parameter adaptation (Biderman et al., 2024).
- **FLOW (Sanyal et al., 2025):** FLOW is a data-weighting method that assigns sample weights according to pretrained-model loss signals. It emphasizes easier or more stable samples to reduce forgetting during fine-tuning.
- **DFT (Wu et al., 2026):** DFT is a dynamic fine-tuning method that improves SFT generalization from a reinforcement-learning perspective. It analyzes the implicit reward structure of SFT and dynamically rescales the token-level objective by token probability to stabilize gradient updates.
- **TALR (Lin et al., 2026):** TALR is a token-adaptive loss reweighting method motivated by a revisiting of domain-specific SFT. It shows that smaller learning rates can mitigate general-capability degradation, and further reweights token-level losses to better balance domain-specific adaptation and general capabilities.

Datasets	Aspect	Nums
Magicoder (Wei et al., 2024)	Code Generation	111K
MetaMathQA (Yu et al., 2024a)	Math Reasoning	395K

Table 6: Summary of training datasets.

- **ADAPT (Zhao et al., 2026):** ADAPT is an online data-reweighting framework that treats data curation as dynamic loss weighting during training rather than offline filtering or resampling. It assigns adaptive per-sample learning rates guided by similarity-based quality signals, adjusting sample importance without changing the number of training samples.

C.2 Training Data Details

In our experiments, we use two target-training datasets: Magicoder (Wei et al., 2024), a decontaminated code-generation dataset, and MetaMathQA (Yu et al., 2024a), a large-scale mathematical-reasoning dataset curated from real-world mathematical exam problems. Their source papers explicitly control target-benchmark contamination through HumanEval decontamination for Magicoder and GSM8K training-split construction for MetaMathQA. Detailed dataset statistics are summarized in Table 6.

C.3 Evaluation Details

We evaluate models from two perspectives: general capability retention and target-task performance. Following recent studies (Wang et al., 2025b; Lin et al., 2026; Sanyal et al., 2025; Liu et al., 2026b; Yuan et al., 2026), we decouple these two aspects to assess task adaptation and preservation of pre-existing abilities, with benchmark roles adjusted by fine-tuning scenario. For Magicoder (Wei et al., 2024), HumanEval (Chen et al., 2021) serves as the target-task benchmark for code generation, while HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), MMLU (Hendrycks et al., 2021), and GSM8K (Cobbe et al., 2021) evaluate general capability retention. For MetaMathQA (Yu et al., 2024a), GSM8K is the target-task benchmark for mathematical reasoning, while HellaSwag, OpenBookQA, MMLU, and HumanEval evaluate retention; here, HumanEval measures retained out-of-domain code generation after math fine-tuning. Table 7 summarizes the settings. All benchmarks

Datasets	Aspects	Metrics	Nums	Evaluate
HellaSwag (Zellers et al., 2019)	Commonsense Reasoning	Accuracy	10,003	0-shot
OpenBookQA (Mihaylov et al., 2018)	Commonsense Reasoning	Accuracy	500	0-shot
MMLU (Hendrycks et al., 2021)	Multi Task Understanding	Accuracy	14,042	5-shot
GSM8K (Cobbe et al., 2021)	Math Reasoning	Exact Match	1,319	5-shot
HumanEval (Chen et al., 2021)	Code Generation	PASS@1	164	0-shot

Table 7: Evaluation benchmarks and detailed settings.

are evaluated with lm-evaluation-harness (Gao et al., 2023) under standard prompting setups. General Capability Avg. is the arithmetic mean over the four general benchmarks, and Overall Avg. is the arithmetic mean of General Capability Avg. and target-task performance.

- **HellaSwag:** HellaSwag (Zellers et al., 2019) is a commonsense natural language inference benchmark. Given an event description or incomplete context, a model must choose the most plausible continuation from multiple candidate endings. The dataset is created through Adversarial Filtering, where a series of discriminators iteratively select adversarial machine-generated negative options, making the benchmark easy for humans but much harder for SOTA models. Models are typically evaluated using **Accuracy**.
- **OpenBookQA:** Open Book Question Answering (Mihaylov et al., 2018) is a question-answering benchmark inspired by open-book exams, designed to evaluate a model’s subject understanding. It contains elementary-school-level multiple-choice science questions and tests both comprehension of a core collection of science facts and the ability to apply such knowledge to novel situations. Models are typically evaluated using the **Accuracy** metric.
- **MMLU:** Massive Multitask Language Understanding (Hendrycks et al., 2021) assesses models’ broad knowledge and reasoning abilities across diverse domains. It contains multiple-choice questions from 57 subjects spanning STEM, humanities, social sciences, and professional knowledge, with four or five answer choices per question. For each subject, the model selects the correct answer using its knowledge and reasoning ability, and performance is typically measured by **Accuracy**.
- **GSM8K:** Grade School Math 8K (Cobbe et al., 2021) is a benchmark for assessing models’ math-

ematical reasoning ability. It contains grade-school-level math word problems, each paired with a step-by-step solution. Given a natural-language problem, the model must produce a solution that includes the correct numerical answer. The questions cover both basic arithmetic and more complex multi-step reasoning. Models are typically evaluated using the **EM** metric.

- **HumanEval:** HumanEval (Chen et al., 2021) is a benchmark for assessing the functional correctness of model-generated code. It contains 164 Python programming tasks and is commonly evaluated with **PASS@1**. Given a problem description and a function signature, the model must generate Python code that implements the required function. A solution is correct only if it passes all provided unit tests.

We report three evaluation metrics in our experiments, which are defined as follows.

- **Accuracy (ACC):** Accuracy measures the proportion of correct predictions over all evaluation examples. Unlike Exact Match, it allows multiple acceptable answers per example. Formally, it is defined as

$$\text{ACC} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i \in Y_i), \quad (67)$$

where N denotes the total number of evaluation examples, $\hat{y}_i$ is the model prediction for the i -th example, and Y_i denotes the set of acceptable ground-truth answers for that example.

- **Exact Match (EM):** Exact Match measures whether the predicted answer exactly matches the reference answer after normalization. Compared with Accuracy, it is stricter because it requires an exact match between the predicted output and the gold answer string. Formally, it is defined as

$$\text{EM} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i), \quad (68)$$

where $\hat{y}_i$ and y_i denote the predicted and reference answers for the i -th example, respectively.

- **PASS@ k :** PASS@ k is a standard metric for code generation. For each problem, the model generates k candidate solutions, and the metric estimates the probability that at least one solution passes all associated unit tests. Formally, it is defined as

$$\text{PASS@}k = \begin{cases} 1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}, & \text{if } n - c \geq k, \\ 1, & \text{if } n - c < k, \end{cases} \quad (69)$$

where n denotes the total number of generated solutions, c denotes the number of correct solutions among them, and k is the number of sampled solutions used for evaluation.

C.4 Implementation Details

All methods are implemented under a unified fine-tuning setup and optimized with AdamW using a cosine learning-rate scheduler. For Magicoder, we use a global batch size of 64, maximum sequence length 4096, and 3 training epochs; for MetaMathQA, we use a global batch size of 128, maximum sequence length 1024, and 2 epochs. The warmup ratio is 0.03 in both settings. For each target dataset, we hold out 1,000 examples for validation and use the remaining target-training split for fine-tuning. We evaluate all methods on three model families, including Gemma3-4B-pt (Team et al., 2025), Llama3.2-3B (Grattafiori et al., 2024), and Qwen3.5-9B-Base (Team, 2026). The compared methods include Standard FT, LoRA (Hu et al., 2022), FLOW (Sanyal et al., 2025), DFT (Wu et al., 2026), TALR (Lin et al., 2026), ADAPT (Zhao et al., 2026), and UFT. We use official repositories for all baselines except TALR, whose code is unavailable. For UFT, we set the UFT uncertainty threshold as $\tau = Q_{60}(s^{\text{uft}})$, reliability softness $\gamma = 0.2$, relative-divergence strength $\alpha = 0.8$, entropy strength $\lambda = 0.2$, and distillation strength $\beta = 2$. We tune the learning rate for all methods in $[5 \times 10^{-6}, 5 \times 10^{-4}]$ and report the best results. For UFT, the selected learning rates on Magicoder are 1×10^{-5} , 3×10^{-5} , and 1×10^{-4} for LLaMA3.2-3B, Gemma3-4B, and Qwen3.5-9B, respectively. On MetaMathQA, the selected learning rates are 4×10^{-5} , 5×10^{-5} , and 7×10^{-5} for LLaMA3.2-3B, Gemma3-4B, and Qwen3.5-9B, respectively.

We report the mean and standard deviation over three random seeds. All experiments are conducted on $4 \times \text{A100 GPUs}$.

C.5 Prompt Templates

To facilitate reproducibility, we provide the exact instruction templates used for fine-tuning. Unless otherwise specified, all fine-tuning templates follow the standard Alpaca-style instruction-response format. For evaluation, we use the task-specific prompts and configurations provided by lm-evaluation-harness (Gao et al., 2023).

For Magicoder, we utilize a specialized template with a coding-focused system preamble:

Magicoder Instruction Template

```
You are a proficient coding assistant. Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:
{instruction}

### Response:
```

For MetaMathQA, we use the standard instruction template:

MetaMathQA Instruction Template

```
Below is an instruction that describes a task. Write a response that appropriately completes the request.

### Instruction:
{instruction}

### Response:
```

C.6 Unconstrained Fine-tuned Teacher Training and Validation Setting

For each target dataset, we construct the unconstrained fine-tuned teacher by fine-tuning the source model on the target-task training split without preservation regularization. We hold out 1,000 examples as the validation split and use the remaining examples for UFT training. The unconstrained teacher follows the same target-task fine-tuning recipe as the student training stage, including the optimizer, scheduler, batch size, maximum sequence length, warmup ratio, and number of

Model	Method	General Capability $\uparrow$					Target $\uparrow$	Overall $\uparrow$
		HASW	OBQA	MMLU	HE	Avg.	GSM8K	Avg.
LLaMA3.2-3B	Pre-trained	55.83 $\pm$ 0.00	31.23 $\pm$ 0.00	56.92 $\pm$ 0.00	28.16 $\pm$ 0.00	43.04 $\pm$ 0.00	27.89 $\pm$ 0.00	35.47 $\pm$ 0.00
	Standard FT	50.02 $\pm$ 0.46	28.40 $\pm$ 0.34	48.49 $\pm$ 0.39	15.27 $\pm$ 0.49	35.55 $\pm$ 0.23	65.94 $\pm$ 0.31	50.75 $\pm$ 0.22
	LoRA ICLR 22	52.06 $\pm$ 0.39	28.60 $\pm$ 0.34	52.34 $\pm$ 0.35	25.31 $\pm$ 0.42	39.58 $\pm$ 0.20	61.02 $\pm$ 0.33	50.30 $\pm$ 0.21
	FLOW ICML 25	53.41 $\pm$ 0.36	30.34 $\pm$ 0.40	55.86 $\pm$ 0.34	26.97 $\pm$ 0.35	41.65 $\pm$ 0.19	63.73 $\pm$ 0.24	52.69 $\pm$ 0.18
	DFT ICLR 26	52.89 $\pm$ 0.40	29.21 $\pm$ 0.27	53.82 $\pm$ 0.33	26.83 $\pm$ 0.41	40.69 $\pm$ 0.21	65.89 $\pm$ 0.21	53.29 $\pm$ 0.17
	TALR ICLR 26	53.52 $\pm$ 0.38	30.04 $\pm$ 0.44	55.94 $\pm$ 0.36	27.32 $\pm$ 0.34	41.71 $\pm$ 0.22	63.68 $\pm$ 0.26	52.70 $\pm$ 0.20
	ADAPT ICLR 26	53.04 $\pm$ 0.43	29.42 $\pm$ 0.30	54.26 $\pm$ 0.37	24.98 $\pm$ 0.39	40.43 $\pm$ 0.20	65.96 $\pm$ 0.25	53.20 $\pm$ 0.19
	UFT (Ours)	55.37 $\pm$ 0.34	30.46 $\pm$ 0.33	55.92 $\pm$ 0.32	27.63 $\pm$ 0.34	42.35 $\pm$ 0.18	66.91 $\pm$ 0.32	54.63 $\pm$ 0.20
Gemma3-4B	Pre-trained	57.41 $\pm$ 0.00	34.62 $\pm$ 0.00	60.24 $\pm$ 0.00	35.77 $\pm$ 0.00	47.01 $\pm$ 0.00	36.73 $\pm$ 0.00	41.87 $\pm$ 0.00
	Standard FT	51.42 $\pm$ 0.43	30.20 $\pm$ 0.26	53.01 $\pm$ 0.37	29.62 $\pm$ 0.43	41.06 $\pm$ 0.22	67.84 $\pm$ 0.29	54.45 $\pm$ 0.20
	LoRA ICLR 22	54.34 $\pm$ 0.37	32.12 $\pm$ 0.35	55.84 $\pm$ 0.34	30.36 $\pm$ 0.34	43.17 $\pm$ 0.19	63.04 $\pm$ 0.28	53.11 $\pm$ 0.21
	FLOW ICML 25	54.92 $\pm$ 0.36	32.64 $\pm$ 0.31	58.27 $\pm$ 0.35	33.24 $\pm$ 0.35	44.77 $\pm$ 0.20	64.78 $\pm$ 0.35	54.78 $\pm$ 0.22
	DFT ICLR 26	53.49 $\pm$ 0.39	32.20 $\pm$ 0.34	56.15 $\pm$ 0.33	32.51 $\pm$ 0.31	43.59 $\pm$ 0.18	67.91 $\pm$ 0.32	55.75 $\pm$ 0.19
	TALR ICLR 26	55.03 $\pm$ 0.36	32.85 $\pm$ 0.35	58.16 $\pm$ 0.34	33.28 $\pm$ 0.39	44.83 $\pm$ 0.21	64.87 $\pm$ 0.27	54.85 $\pm$ 0.20
	ADAPT ICLR 26	53.51 $\pm$ 0.41	31.86 $\pm$ 0.27	55.76 $\pm$ 0.36	31.71 $\pm$ 0.34	43.21 $\pm$ 0.19	67.96 $\pm$ 0.32	55.59 $\pm$ 0.21
	UFT (Ours)	56.04 $\pm$ 0.33	33.69 $\pm$ 0.32	59.07 $\pm$ 0.31	34.46 $\pm$ 0.33	45.82 $\pm$ 0.17	69.45 $\pm$ 0.29	57.63 $\pm$ 0.19
Qwen3.5-9B	Pre-trained	59.67 $\pm$ 0.00	34.17 $\pm$ 0.00	72.39 $\pm$ 0.00	61.05 $\pm$ 0.00	56.82 $\pm$ 0.00	86.23 $\pm$ 0.00	71.53 $\pm$ 0.00
	Standard FT	54.39 $\pm$ 0.40	30.78 $\pm$ 0.29	67.52 $\pm$ 0.34	57.14 $\pm$ 0.34	52.46 $\pm$ 0.20	90.78 $\pm$ 0.36	71.62 $\pm$ 0.22
	LoRA ICLR 22	56.64 $\pm$ 0.34	32.80 $\pm$ 0.33	69.75 $\pm$ 0.31	58.20 $\pm$ 0.39	54.35 $\pm$ 0.18	88.64 $\pm$ 0.33	71.50 $\pm$ 0.19
	FLOW ICML 25	57.13 $\pm$ 0.32	33.25 $\pm$ 0.28	70.42 $\pm$ 0.30	59.28 $\pm$ 0.43	55.02 $\pm$ 0.17	88.27 $\pm$ 0.32	71.65 $\pm$ 0.20
	DFT ICLR 26	55.49 $\pm$ 0.36	32.93 $\pm$ 0.31	70.06 $\pm$ 0.31	58.64 $\pm$ 0.48	54.28 $\pm$ 0.21	90.93 $\pm$ 0.29	72.61 $\pm$ 0.18
	TALR ICLR 26	57.62 $\pm$ 0.31	33.14 $\pm$ 0.33	70.74 $\pm$ 0.29	58.87 $\pm$ 0.45	55.09 $\pm$ 0.19	89.07 $\pm$ 0.35	72.08 $\pm$ 0.22
	ADAPT ICLR 26	55.27 $\pm$ 0.38	32.19 $\pm$ 0.32	69.95 $\pm$ 0.32	58.34 $\pm$ 0.32	53.94 $\pm$ 0.20	90.81 $\pm$ 0.41	72.38 $\pm$ 0.23
	UFT (Ours)	58.10 $\pm$ 0.32	33.68 $\pm$ 0.29	71.43 $\pm$ 0.28	60.24 $\pm$ 0.36	55.86 $\pm$ 0.18	91.37 $\pm$ 0.34	73.62 $\pm$ 0.21

Table 8: Performance comparison on the MetaMathQA dataset (%). Pre-trained and Standard FT are excluded from the ranking. The best results are highlighted in red, and the second-best results are highlighted in blue.

epochs. During UFT training, we evaluate checkpoints on the held-out validation split and select the checkpoint with the lowest target-task validation negative log-likelihood as M_{uft} . The selected teacher is then used to compute UFT reliability, source-relative divergence, and sample utilities for downstream student fine-tuning.

D Experimental Results and Analysis

D.1 Results on MetaMathQA Benchmark

In this subsection, we present the experimental results on the MetaMathQA benchmark. We report results on LLaMA3.2-3B, Gemma3-4B, and Qwen3.5-9B to evaluate whether UFT can improve mathematical reasoning while preserving general capabilities across different backbone models.

Table 8 reports MetaMathQA results across three models. On LLaMA3.2-3B, Standard FT raises GSM8K from 27.89% to 65.94% but reduces the general capability average from 43.04% to 35.55%. UFT achieves the best target, general, and overall averages among fine-tuned methods, reaching 66.91%, 42.35%, and 54.63%, and surpassing the

strongest baseline overall average of 53.29%. On Gemma3-4B, Standard FT improves GSM8K from 36.73% to 67.84% but lowers the general average from 47.01% to 41.06%, while UFT further improves GSM8K to 69.45% and retains 45.82% general capability, reducing the general capability drop from 5.95 to 1.19 points and achieving the best overall average of 57.63%. On Qwen3.5-9B, the pre-trained model already reaches 86.23% on GSM8K, making additional adaptation harder. UFT improves GSM8K to 91.37% while retaining 55.86% general capability, outperforming Standard FT and achieving the best overall average of 73.62%. These results show that UFT does not merely increase target-task fitting. Its value discovery favors samples with stronger target signals and lower preservation cost, while dual-teacher training helps convert this selection into stronger adaptation with less general capability degradation.

D.2 Pareto Analysis of Target and General Capability

Fig. 8 provides a Pareto-style visualization of the adaptation-retention trade-off on Magicode and

Target vs. General Capability Pareto Frontier

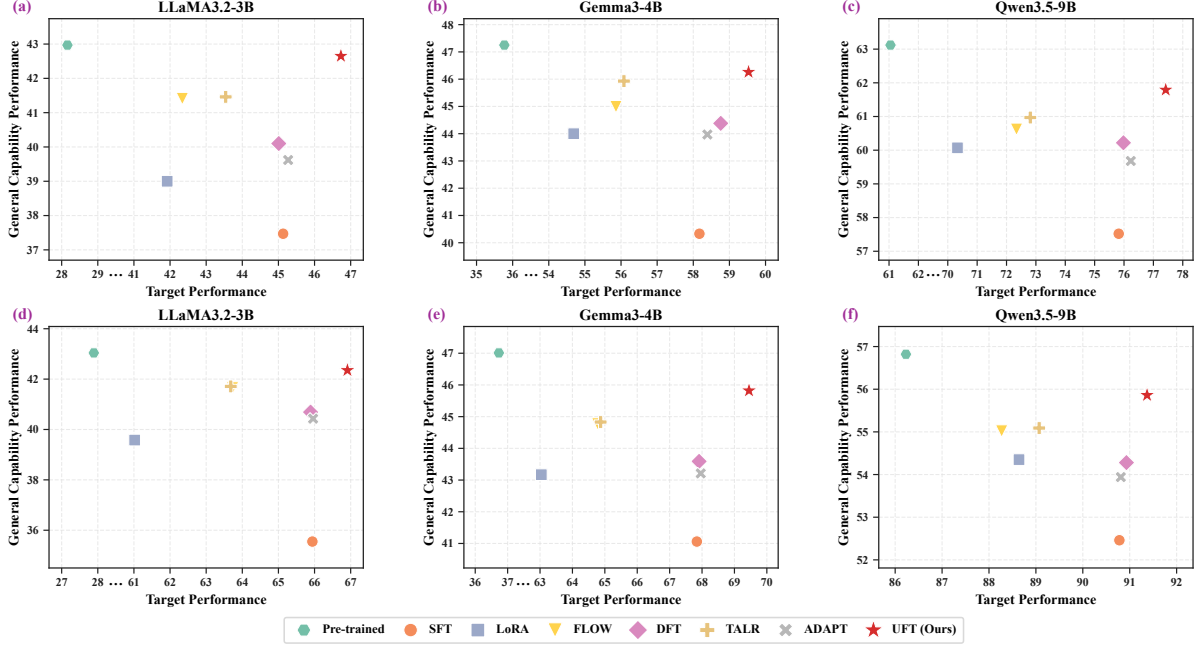


Figure 8: Target-general capability Pareto frontier across two scenarios and three backbone models. (a)–(c) show Magicoder results with HumanEval as the target benchmark, while (d)–(f) show MetaMathQA results with GSM8K as the target benchmark. Each point denotes one fine-tuning method, and UFT consistently lies near the upper-right region, indicating stronger target adaptation with better evaluated general-capability retention.

MetaMathQA. The horizontal axis shows target performance, measured by HumanEval for Magicoder and GSM8K for MetaMathQA, and the vertical axis shows average general capability. Across both scenarios and all three backbones, UFT consistently lies in the upper-right region or near the empirical Pareto frontier, indicating stronger target-task adaptation without a large sacrifice in general capability. Compared with methods that either preserve general capability but under-adapt or improve target performance with larger general-capability degradation, UFT achieves a more favorable adaptation-retention balance.

D.3 Statistical Significance and Dynamic Utility Stability

We provide two complementary analyses to assess the reliability of UFT. The first examines whether its performance gains are consistent across random seeds, while the second evaluates whether the estimated utilities and resulting training weights remain stable as the student model evolves during downstream optimization.

Statistical Significance. We conduct two sided paired t tests on the Overall scores for LLaMA3.2-3B and Gemma3-4B. Runs using the same random seed are paired across methods. Each score differ-

ence is defined as the Overall score of UFT minus that of the corresponding baseline. Holm correction is applied to the six comparisons within each backbone. Table 9 reports the 95% confidence intervals of the paired differences and the adjusted p values. Every confidence interval lies above zero and every adjusted p value is below 0.05. These results show that the advantage of UFT exceeds random seed variation and is not driven by any particular seed.

Dynamic Utility Stability. UFT computes static utilities before downstream student training. Proposition 1 establishes the stability of the resulting weights when the ideal dynamic utilities exhibit bounded drift. We empirically examine this condition using a fixed diagnostic subset under the LLaMA3.2 and Magicoder setting. At student initialization and after every epoch, we recompute the dynamic utility v_i at checkpoint θ_k . Utility drift is measured relative to initialization. We report its mean, P95, and maximum value together with the Spearman correlation and top 20% overlap between the original and dynamic utility rankings. We also compare the empirical total variation distance between the fixed and dynamic weight distributions with the theoretical bound in Proposition 1.

As shown in Table 10, utility drift increases grad-

Backbone	Comparison	95% CI of Overall Difference	Holm Adjusted p
LLaMA3.2-3B	UFT vs. Standard FT	[2.60, 4.18]	1.23×10^{-2}
	UFT vs. LoRA	[3.47, 4.99]	1.03×10^{-2}
	UFT vs. FLOW	[2.21, 3.41]	1.23×10^{-2}
	UFT vs. DFT	[1.36, 2.90]	1.68×10^{-2}
	UFT vs. TALR	[1.48, 2.90]	1.68×10^{-2}
	UFT vs. ADAPT	[1.41, 3.07]	1.68×10^{-2}
Gemma3-4B	UFT vs. Standard FT	[2.84, 4.46]	1.60×10^{-2}
	UFT vs. LoRA	[2.69, 4.41]	1.60×10^{-2}
	UFT vs. FLOW	[1.76, 3.18]	1.77×10^{-2}
	UFT vs. DFT	[0.57, 2.09]	3.33×10^{-2}
	UFT vs. TALR	[1.03, 2.75]	3.30×10^{-2}
	UFT vs. ADAPT	[0.75, 2.69]	3.33×10^{-2}

Table 9: Statistical significance of UFT on Magicoder.

Student	Utility Drift ↓			Ranking Stability ↑		Weight Distance ↓	
	Mean	P95	Max $\hat{\Delta}_k$	Spearman	Top 20% Overlap	Empirical TV	Proposition 1 Bound
Initialization	0.0000	0.0000	0.0000	1.000	100.0%	0.000	0.000
Epoch 1	0.0016	0.0038	0.0062	0.978	92.4%	0.015	0.060
Epoch 2	0.0032	0.0076	0.0124	0.951	86.3%	0.031	0.117
Epoch 3	0.0051	0.0121	0.0192	0.926	80.8%	0.052	0.175

Table 10: Dynamic utility stability during student training on Magicoder with LLaMA3.2-3B.

ually but remains small throughout student training. The original and dynamic utility rankings remain highly consistent. The empirical total variation distance also remains below the theoretical bound at every checkpoint. These results support the bounded drift condition in Proposition 1 and show that the fixed UFT utilities preserve stable sample rankings and training weights during downstream optimization.

D.4 Efficient UFT Approximation with LoRA

The unconstrained fine-tuned teacher in UFT is designed to expose target-task adaptation preferences without preservation regularization. In principle, this teacher can be constructed by full-parameter fine-tuning under the same target-task recipe as the student training stage, so that its predictions provide a recipe-aligned signal for value discovery. However, UFT does not require the teacher to reproduce the exact full-parameter trajectory. The teacher is used to compute continuous sample utilities and provide target-side distillation guid-

ance. Therefore, a LoRA-based unconstrained teacher can serve as an efficient approximation if it preserves the induced utility distribution, sample weights, and teacher predictive behavior.

Specifically, we train a LoRA-based unconstrained teacher by adapting the source model M_{ref} on the target-task training split without preservation regularization. The resulting teacher $M_{\text{uft}}^{\text{LoRA}}$ approximates the full-parameter unconstrained teacher $M_{\text{uft}}^{\text{Full}}$ within a low-rank update subspace. For each target-training example x_i , we compute the UFT utility using both teachers, denoted by u_i^{LoRA} and u_i^{Full} . We then transform these utilities into soft training weights:

$$\begin{aligned}
 w_i^{\text{LoRA}} &= \frac{\exp(u_i^{\text{LoRA}}/\lambda)}{\sum_{j=1}^N \exp(u_j^{\text{LoRA}}/\lambda)} \\
 w_i^{\text{Full}} &= \frac{\exp(u_i^{\text{Full}}/\lambda)}{\sum_{j=1}^N \exp(u_j^{\text{Full}}/\lambda)}.
 \end{aligned} \tag{70}$$

Since UFT uses target examples with continuous weights, the primary approximation criterion is

UFT Teacher	Utility Pearson $\uparrow$	Weight TV $\downarrow$	Teacher KL $\downarrow$	Rel. Time $\downarrow$	Rel. Mem. $\downarrow$
LoRA teacher	0.86	0.14	0.16	0.75 $\times$	0.16 $\times$
Full teacher	1.00	0.00	0.00	1.00 $\times$	1.00 $\times$

Table 11: Comparison between full-parameter and LoRA-based unconstrained fine-tuned teachers in UFT. Utility Pearson measures the agreement of raw sample utilities, Weight TV measures the distance between induced soft training-weight distributions, and Teacher KL measures the discrepancy between target-side teacher predictions.

Method	General Capability $\uparrow$					Target $\uparrow$	Overall $\uparrow$
	HASW	OBQA	MMLU	GSM8K	Avg.	HE	Avg.
UFT + LoRA teacher	55.07	30.71	55.87	27.58	42.31	46.05	44.18
UFT + Full teacher	55.29	30.96	56.22	28.12	42.65	46.73	44.69

Table 12: Comparison between LoRA-based and full-parameter unconstrained fine-tuned teachers under the same UFT training objective on Magicoder with LLaMA3.2-3B.

whether the LoRA-based teacher induces a similar soft weighting distribution to the full-parameter teacher. We therefore measure the agreement between the two induced weight vectors using total variation distance:

$$\text{TV}(\mathbf{w}^{\text{LoRA}}, \mathbf{w}^{\text{Full}}) = \frac{1}{2} \sum_{i=1}^N |w_i^{\text{LoRA}} - w_i^{\text{Full}}|. \quad (71)$$

This metric directly evaluates whether the two teachers assign similar training mass to target examples. As an auxiliary diagnostic before softmax normalization, we also report the Pearson correlation between $\{u_i^{\text{LoRA}}\}_{i=1}^N$ and $\{u_i^{\text{Full}}\}_{i=1}^N$, which measures whether the two teachers produce a similar continuous valuation pattern.

Because the unconstrained teacher is also used for target-side distillation, we further measure the predictive discrepancy between $M_{\text{uft}}^{\text{LoRA}}$ and $M_{\text{uft}}^{\text{Full}}$ on target responses:

$$\text{KL}_{\text{teach}} = \frac{1}{\sum_{i=1}^N T_i} \sum_{i=1}^N \sum_{t=1}^{T_i} D_{\text{KL}} \left(P_{\text{uft}}^{\text{Full}}(\cdot | x_i, y_i, < t) \parallel P_{\text{uft}}^{\text{LoRA}}(\cdot | x_i, y_i, < t) \right). \quad (72)$$

This metric directly tests whether the LoRA teacher provides a similar target-side distillation signal.

Table 11 first compares the LoRA-based and full-parameter unconstrained teachers in terms of utility estimation, induced soft weighting, target-side predictive behavior, and construction cost. The LoRA-based teacher achieves a high utility correlation with the full-parameter teacher, induces

a close soft weighting distribution, and yields a small target-side teacher discrepancy, while reducing teacher-construction cost.

We further compare the downstream student performance obtained with the two teacher instantiations under the same utility-weighted training objective and dual-teacher distillation setup. As shown in Table 12, the LoRA-based teacher achieves performance close to the full-parameter teacher on both target-task adaptation and evaluated general-capability retention. These results suggest that UFT mainly relies on a stable continuous value distribution and compatible teacher predictions, rather than exact reproduction of the full-parameter adaptation trajectory. Thus, LoRA provides a practical approximation when teacher construction cost is a concern, while the full-parameter teacher remains the most recipe-matched instantiation.

D.5 Teacher Construction Cost

Before value discovery, UFT constructs an unconstrained fine-tuned teacher on the target-task data. Given the constructed teacher, value discovery estimates which target-task samples provide useful adaptation signals while remaining compatible with retention requirements. Specifically, it uses UFT predictions and source-relative divergence to compute continuous sample utilities. In the full-parameter setting on Magicoder with LLaMA3.2-3B, teacher construction **takes about 4 hours on 4 A100 GPUs, corresponding to 0.67 GPU-days**.

To reduce this cost, Appendix D.4 examines a

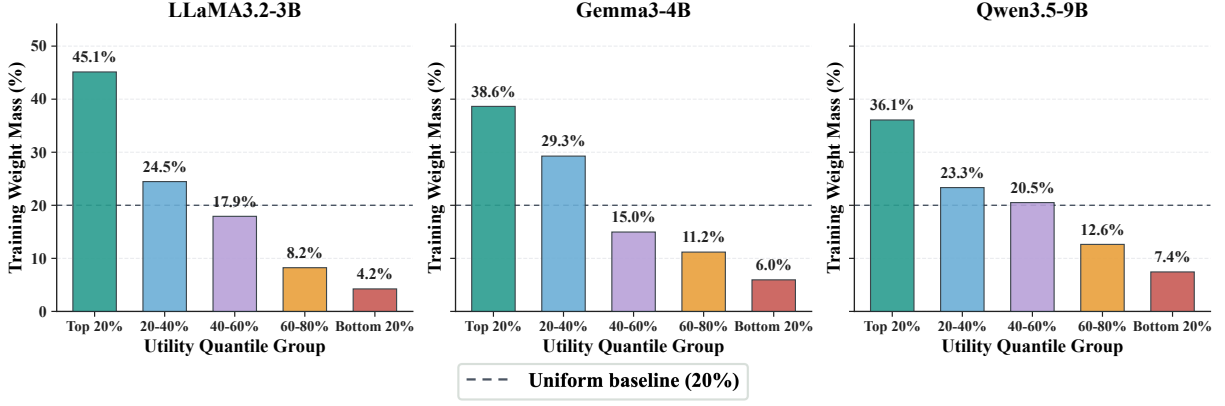


Figure 9: Utility-stratified training-weight distribution under UFT on Magicoder. For each backbone, target-training examples are sorted by UFT utility and divided into five equal-sized quantile groups. Bars show the fraction of total continuous training weight assigned to each group, and the dashed line marks the uniform 20% allocation baseline.

LoRA-based approximation to the unconstrained teacher. Under the same setting, constructing the LoRA-based teacher **takes about 3 hours on 4 A100 GPUs, corresponding to 0.5 GPU-days**. Although the LoRA-based teacher provides a slightly weaker valuation and distillation signal than the full-parameter teacher, it induces a similar soft weighting distribution and leads to only a small drop in final performance. Therefore, UFT incurs additional cost for teacher-based value discovery compared with lightweight baselines, but this cost remains moderate and should be evaluated together with its gains in the adaptation–retention trade-off.

D.6 Utility-Stratified Training-Weight Distribution.

To examine how UFT utilities affect optimization, we analyze the continuous training-weight distribution induced by UFT. For each backbone on Magicoder, we sort all target-training examples by UFT utility and divide them into five equal-sized quantile groups, each containing 20% of the examples. We compute the total training mass assigned to each group as

$$\text{WeightMass}(\mathcal{G}) = \sum_{i \in \mathcal{G}} w_i^*, \quad (73)$$

where w_i^* is sample i ’s normalized UFT weight.

As shown in Fig. 9, UFT assigns larger continuous training mass to higher-utility groups across all three backbones. On LLaMA3.2-3B, Gemma3-4B, and Qwen3.5-9B, the top 20% utility group receives 45.13%, 38.64%, and 36.08% of the total training mass, respectively, all above the uniform 20% baseline, while the bottom 20% group receives 4.24%, 5.95%, and 7.44%. This pattern shows that

learned utilities directly shape sample emphasis during fine-tuning. Since all target-training examples remain in the objective, lower-utility samples still contribute with smaller continuous weights. Together with the shuffled and inverse weighting comparisons in Table 3, this result shows that both the magnitude and ordering of UFT utilities contribute to the final adaptation–retention trade-off.

D.7 Results on Instruction Tuning

We further evaluate all methods by fine-tuning LLaMA3.2-3B on Alpaca-GPT4 (Peng et al., 2023) to assess broad instruction-tuning performance. Alpaca-GPT4 contains diverse instruction-following tasks, covering question answering, reasoning, writing, and formatting. For general capability, we report results on ARC-Challenge and ARC-Easy (Clark et al., 2018), PIQA (Bisk et al., 2020), Winogrande (Sakaguchi et al., 2020), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), TruthfulQA (Lin et al., 2022), MMLU (Hendrycks et al., 2021), GSM8K (Cobbe et al., 2021), and HumanEval (Chen et al., 2021), and compute General Avg. as their arithmetic mean. Following recent instruction-tuning studies (Li et al., 2025a; Han and Huang, 2026; Mirza et al., 2025), we use IFEval (Zhou et al., 2023) and report prompt-level strict accuracy (PS), instruction-level strict accuracy (IS), prompt-level loose accuracy (PL), instruction-level loose accuracy (IL), and their arithmetic mean as IFEval Avg. Overall performance is computed as the average of General Avg. and IFEval Avg. Following the main experimental setup, we fine-tune on Alpaca-GPT4 for 3 epochs with a global batch size of 128 and a

Tasks	Pre-trained	Standard FT	LoRA	FLOW	DFT	TALR	ADAPT	UFT
General Capability $\uparrow$								
ARC-Challenge	42.92 $\pm$ 0.00	44.54 $\pm$ 0.36	44.97 $\pm$ 0.32	43.86 $\pm$ 0.30	45.12 $\pm$ 0.33	45.47 $\pm$ 0.31	45.83 $\pm$ 0.35	46.26 $\pm$ 0.29
ARC-Easy	74.58 $\pm$ 0.00	73.23 $\pm$ 0.34	75.19 $\pm$ 0.30	74.21 $\pm$ 0.32	75.94 $\pm$ 0.28	74.91 $\pm$ 0.35	75.06 $\pm$ 0.30	76.43 $\pm$ 0.28
PIQA	76.28 $\pm$ 0.00	75.14 $\pm$ 0.32	78.09 $\pm$ 0.29	76.07 $\pm$ 0.30	77.68 $\pm$ 0.28	76.34 $\pm$ 0.30	77.02 $\pm$ 0.31	77.86 $\pm$ 0.27
HellaSwag	55.83 $\pm$ 0.00	57.59 $\pm$ 0.33	57.32 $\pm$ 0.30	56.14 $\pm$ 0.29	57.37 $\pm$ 0.31	58.71 $\pm$ 0.30	58.96 $\pm$ 0.32	59.43 $\pm$ 0.28
OpenBookQA	31.23 $\pm$ 0.00	33.40 $\pm$ 0.36	32.60 $\pm$ 0.32	31.86 $\pm$ 0.34	33.78 $\pm$ 0.29	32.63 $\pm$ 0.35	33.42 $\pm$ 0.31	34.16 $\pm$ 0.30
TruthfulQA	39.13 $\pm$ 0.00	50.15 $\pm$ 0.39	48.80 $\pm$ 0.35	47.36 $\pm$ 0.33	49.22 $\pm$ 0.36	49.57 $\pm$ 0.34	50.86 $\pm$ 0.38	51.92 $\pm$ 0.32
Winogrande	69.22 $\pm$ 0.00	65.98 $\pm$ 0.42	66.72 $\pm$ 0.38	68.41 $\pm$ 0.36	67.03 $\pm$ 0.37	67.65 $\pm$ 0.36	68.77 $\pm$ 0.40	68.64 $\pm$ 0.34
MMLU	56.92 $\pm$ 0.00	53.67 $\pm$ 0.29	55.97 $\pm$ 0.26	55.86 $\pm$ 0.27	54.18 $\pm$ 0.26	55.12 $\pm$ 0.28	55.84 $\pm$ 0.29	56.08 $\pm$ 0.25
GSM8K	27.89 $\pm$ 0.00	23.81 $\pm$ 0.45	26.10 $\pm$ 0.40	26.29 $\pm$ 0.41	24.67 $\pm$ 0.38	26.09 $\pm$ 0.42	24.86 $\pm$ 0.44	27.01 $\pm$ 0.36
HumanEval	28.16 $\pm$ 0.00	25.83 $\pm$ 0.42	27.08 $\pm$ 0.38	27.14 $\pm$ 0.39	26.51 $\pm$ 0.36	27.29 $\pm$ 0.37	26.76 $\pm$ 0.40	27.71 $\pm$ 0.34
Avg.	50.22 $\pm$ 0.00	50.33 $\pm$ 0.16	51.28 $\pm$ 0.14	50.72 $\pm$ 0.15	51.15 $\pm$ 0.14	51.38 $\pm$ 0.15	51.74 $\pm$ 0.16	52.55 $\pm$ 0.13
IFEval (Target) $\uparrow$								
Prompt (S)	21.81 $\pm$ 0.00	29.72 $\pm$ 0.46	27.36 $\pm$ 0.42	28.29 $\pm$ 0.44	32.74 $\pm$ 0.41	30.68 $\pm$ 0.39	33.21 $\pm$ 0.37	34.24 $\pm$ 0.32
Inst (S)	34.77 $\pm$ 0.00	40.05 $\pm$ 0.57	38.82 $\pm$ 0.50	39.25 $\pm$ 0.52	41.93 $\pm$ 0.49	41.05 $\pm$ 0.55	42.58 $\pm$ 0.43	43.29 $\pm$ 0.40
Prompt (L)	24.40 $\pm$ 0.00	31.42 $\pm$ 0.42	29.18 $\pm$ 0.53	29.88 $\pm$ 0.61	33.36 $\pm$ 0.38	32.45 $\pm$ 0.57	34.14 $\pm$ 0.34	35.03 $\pm$ 0.31
Inst (L)	38.13 $\pm$ 0.00	42.57 $\pm$ 0.46	40.53 $\pm$ 0.59	41.48 $\pm$ 0.54	43.88 $\pm$ 0.44	42.77 $\pm$ 0.48	44.32 $\pm$ 0.39	44.21 $\pm$ 0.36
Avg.	29.78 $\pm$ 0.00	35.94 $\pm$ 0.25	33.97 $\pm$ 0.23	34.73 $\pm$ 0.22	37.98 $\pm$ 0.20	36.74 $\pm$ 0.24	38.56 $\pm$ 0.21	39.19 $\pm$ 0.19
Overall $\uparrow$								
Avg.	40.00 $\pm$ 0.00	43.14 $\pm$ 0.18	42.63 $\pm$ 0.17	42.73 $\pm$ 0.16	44.57 $\pm$ 0.15	44.06 $\pm$ 0.18	45.15 $\pm$ 0.16	45.87 $\pm$ 0.14

Table 13: Performance comparison on the Alpaca-GPT4 dataset (%). Pre-trained and Standard FT are excluded from the ranking. The **best results** are highlighted in red, and the **second-best** results are highlighted in blue.

maximum sequence length of 4096. For UFT, the learning rate selected on Alpaca-GPT4 is 4×10^{-5} .

Table 13 reports the Alpaca-GPT4 results. Standard fine-tuning improves IFEval Avg. from 29.78 to 35.94, but this gain comes with drops on retention-sensitive benchmarks such as MMLU, GSM8K, and HumanEval. LoRA and FLOW better recover these general capabilities, but their lower IFEval Avg. scores indicate limited adaptation to instruction-following requirements. DFT, TALR, and ADAPT further improve IFEval Avg. over Standard FT, with ADAPT achieving the strongest baseline Overall score of 45.15. UFT achieves the highest General Avg., IFEval Avg., and Overall scores, reaching 52.55, 39.19, and 45.87, respectively. Although ADAPT slightly outperforms UFT on instruction-level loose accuracy, UFT attains the best average instruction-following performance across the four IFEval metrics. These results show that UFT remains effective in broad instruction tuning, improving instruction following while main-

taining strong average general capability.

D.8 Comparison with DavIR

DavIR (Zhou et al., 2025) is originally designed for supervised fine-tuning data selection. It scores each sample by the normalized loss reduction between a base model and a fine-tuned reference model:

$$S_{\text{DavIR}}(x_i, y_i) = \frac{\mathcal{L}_{\text{base}}(x_i, y_i) - \mathcal{L}_{\text{ref}}(x_i, y_i)}{\mathcal{L}_{\text{base}}(x_i, y_i)}. \quad (74)$$

Here, $\mathcal{L}_{\text{base}}(x_i, y_i)$ and $\mathcal{L}_{\text{ref}}(x_i, y_i)$ denote the base-model and fine-tuned-reference losses on sample (x_i, y_i) , respectively. A larger $S_{\text{DavIR}}(x_i, y_i)$ indicates a larger relative loss reduction after fine-tuning. The original DavIR protocol ranks samples by this score, selects the top- k subset, and re-trains the base model on the selected data.

To compare DavIR with UFT under the same full-data training setting, we construct a DavIR-style reweighting baseline, denoted as DavIR-W. DavIR-W keeps the original DavIR score in

Method	General Capability $\uparrow$					Target $\uparrow$	Overall $\uparrow$
	HASW	OBQA	MMLU	GSM8K	Avg.	HE	Avg.
<i>LLaMA3.2-3B</i>							
DavIR-W	54.10 $\pm$ 0.42	29.92 $\pm$ 0.36	53.65 $\pm$ 0.35	24.65 $\pm$ 0.38	40.58 $\pm$ 0.24	45.35 $\pm$ 0.43	42.97 $\pm$ 0.26
UFT	55.29 $\pm$ 0.31	30.96 $\pm$ 0.26	56.22 $\pm$ 0.28	28.12 $\pm$ 0.31	42.65 $\pm$ 0.22	46.73 $\pm$ 0.40	44.69 $\pm$ 0.23
<i>Gemma3-4B</i>							
DavIR-W	54.89 $\pm$ 0.39	31.85 $\pm$ 0.34	57.58 $\pm$ 0.40	32.06 $\pm$ 0.41	44.10 $\pm$ 0.23	58.31 $\pm$ 0.44	51.21 $\pm$ 0.27
UFT	56.93 $\pm$ 0.30	32.56 $\pm$ 0.29	59.28 $\pm$ 0.28	36.27 $\pm$ 0.46	46.26 $\pm$ 0.24	59.53 $\pm$ 0.34	52.90 $\pm$ 0.21

Table 14: Comparison with DavIR-style reweighting on the Magicoder dataset (%).

Method	General Capability $\uparrow$					Target $\uparrow$	Overall $\uparrow$
	HASW	OBQA	MMLU	HE	Avg.	GSM8K	Avg.
<i>LLaMA3.2-3B</i>							
DavIR-W	52.94 $\pm$ 0.43	29.52 $\pm$ 0.37	54.48 $\pm$ 0.34	25.97 $\pm$ 0.42	40.73 $\pm$ 0.24	66.23 $\pm$ 0.36	53.48 $\pm$ 0.25
UFT	55.37 $\pm$ 0.34	30.46 $\pm$ 0.33	55.92 $\pm$ 0.32	27.63 $\pm$ 0.34	42.35 $\pm$ 0.18	66.91 $\pm$ 0.32	54.63 $\pm$ 0.20
<i>Gemma3-4B</i>							
DavIR-W	55.23 $\pm$ 0.38	32.15 $\pm$ 0.33	57.49 $\pm$ 0.36	32.64 $\pm$ 0.40	44.38 $\pm$ 0.22	68.15 $\pm$ 0.35	56.27 $\pm$ 0.24
UFT	56.04 $\pm$ 0.33	33.69 $\pm$ 0.32	59.07 $\pm$ 0.31	34.46 $\pm$ 0.33	45.82 $\pm$ 0.17	69.45 $\pm$ 0.29	57.63 $\pm$ 0.19

Table 15: Comparison with DavIR-style reweighting on the MetaMathQA dataset (%).

Eq. (74) unchanged and converts it into continuous sample weights:

$$w_i^{\text{DavIR}} = \frac{\exp(S_{\text{DavIR}}(x_i, y_i)/\lambda)}{\sum_{j=1}^N \exp(S_{\text{DavIR}}(x_j, y_j)/\lambda)}. \quad (75)$$

Here, λ is the same temperature parameter as in UFT. After obtaining w_i^{DavIR} , we fine-tune the model under the same configuration as UFT. Tables 14 and 15 compare results on Magicoder and MetaMathQA. On Magicoder, UFT exceeds DavIR-W in Overall score by 1.72 on LLaMA3.2-3B and 1.69 on Gemma3-4B; on MetaMathQA, the gains are 1.15 and 1.36, respectively. UFT also consistently improves the evaluated general-capability average and target-task performance across both datasets and backbones. These results show that preservation-aware value discovery provides a stronger reweighting signal than loss-reduction-based learnability alone.

D.9 Compute-Matched Comparison

UFT introduces an additional teacher-construction stage before downstream student training. To examine whether the resulting improvements can be attributed to this additional computation, we conduct a compute-matched comparison on Magicoder with

LLaMA3.2-3B. Starting from the same pre-trained checkpoint, we train each baseline for up to six epochs, consisting of the original three epochs and three additional epochs that approximately match the cost of constructing the UFT teacher. We reconstruct the cosine learning-rate schedule over the full six-epoch horizon and apply early stopping using the same held-out target validation criterion as in the main experiments. For each method, we report the checkpoint with the lowest validation negative log-likelihood. All other training and evaluation settings remain unchanged.

As shown in Table 16, UFT continues to achieve the strongest balance between target task performance and general capability retention under the matched compute budget. It obtains the highest General Avg. of 42.65%, the highest HumanEval score of 46.73%, and the highest Overall Avg. of 44.69% among all fine tuned methods. FLOW and TALR preserve relatively strong general capability but achieve lower HumanEval performance. Standard FT, DFT, and ADAPT obtain stronger HumanEval results but retain less general capability. The additional training epochs therefore do not close the gap between these baselines and UFT. These results show that the gains of UFT are not

Method	General Capability $\uparrow$					Target $\uparrow$	Overall $\uparrow$
	HASW	OBQA	MMLU	GSM8K	Avg.	HE	Avg.
Pre-trained	55.83 $\pm$ 0.00	31.23 $\pm$ 0.00	56.92 $\pm$ 0.00	27.89 $\pm$ 0.00	42.97 $\pm$ 0.00	28.16 $\pm$ 0.00	35.57 $\pm$ 0.00
Standard FT	51.55 $\pm$ 0.52	28.10 $\pm$ 0.40	47.85 $\pm$ 0.46	19.70 $\pm$ 0.38	36.80 $\pm$ 0.31	46.35 $\pm$ 0.52	41.58 $\pm$ 0.32
LoRA	53.25 $\pm$ 0.41	29.40 $\pm$ 0.35	51.90 $\pm$ 0.37	20.85 $\pm$ 0.42	38.85 $\pm$ 0.28	42.57 $\pm$ 0.43	40.71 $\pm$ 0.27
FLOW	54.58 $\pm$ 0.36	30.09 $\pm$ 0.31	55.74 $\pm$ 0.33	24.91 $\pm$ 0.27	41.33 $\pm$ 0.23	42.79 $\pm$ 0.37	42.06 $\pm$ 0.22
DFT	53.15 $\pm$ 0.40	29.62 $\pm$ 0.24	51.78 $\pm$ 0.36	24.65 $\pm$ 0.29	39.80 $\pm$ 0.24	45.96 $\pm$ 0.41	42.88 $\pm$ 0.26
TALR	54.80 $\pm$ 0.36	30.22 $\pm$ 0.35	53.92 $\pm$ 0.32	26.50 $\pm$ 0.37	41.36 $\pm$ 0.25	44.19 $\pm$ 0.39	42.78 $\pm$ 0.24
ADAPT	53.60 $\pm$ 0.44	29.55 $\pm$ 0.32	52.68 $\pm$ 0.39	21.65 $\pm$ 0.34	39.37 $\pm$ 0.26	46.22 $\pm$ 0.45	42.80 $\pm$ 0.28
UFT (Ours)	55.29 $\pm$ 0.31	30.96 $\pm$ 0.26	56.22 $\pm$ 0.28	28.12 $\pm$ 0.31	42.65 $\pm$ 0.22	46.73 $\pm$ 0.40	44.69 $\pm$ 0.23

Table 16: Compute-matched comparison on the Magicoder dataset with LLaMA3.2-3B (%).

merely due to additional compute. They also depend on how UFT uses the available computation for value discovery and downstream training.