One Algorithm, Two Goals: Dual Scoring for Parameter and Data Selection in LLM Fine-Tuning

Xinrui Chen[1] **Liu Yang**[2] **Ou Wu**[1*]
[1]Hangzhou Institute for Advanced Study,
University of Chinese Academy of Sciences, Hangzhou, China
[2]School of Artificial Intelligence, Tianjin University, Tianjin, China
chenxinrui25@mails.ucas.ac.cn yangliuyl@tju.edu.cn wuou@ucas.ac.cn

Abstract

In Large Language Model (LLM) fine-tuning, parameter and data selection are common strategies for reducing fine-tuning cost, yet they are typically driven by separate scoring mechanisms. When a parameter mask and data subset jointly determine restricted fine-tuning, this separation incurs redundant overhead and makes coordinated selection difficult. We cast parameter and data selection as two bilevel selection problems under a common validation objective and derive a shared local response-surrogate scoring rule. Under first- and second-order validation-improvement approximations, parameter importance and data utility emerge as column-wise and row-wise aggregations of a single gradient interaction matrix, yielding a closed-form row–column correspondence for co-extracting both signals. Building on this structure, we propose **DualSFT** (**Dual**-**S**election **F**ine-**T**uning), a one-shot dual-scoring algorithm that produces a parameter mask and data subset from shared gradient statistics. On 3B–9B LLMs, single-axis DualSFT variants strengthen target-task performance and stability–plasticity trade-offs within their comparison groups, while full DualSFT yields a more favorable joint-constrained trade-off than sequential hybrid baselines under matched budgets.

1 Introduction

Full fine-tuning adapts Large Language Models (LLMs) to downstream tasks [1, 2], but is computationally prohibitive at modern model and data scales [3–5]. This cost has motivated two complementary lines of research, targeting the parameters being updated and the data used for training. Parameter-efficient fine-tuning (PEFT) controls adaptation costs through approaches such as additive modules, low-rank updates, and parameter selection [3, 5–9]. Data-efficient fine-tuning (DEFT) optimizes data use through data selection, weighting, and scheduling [10–14].

Within PEFT and DEFT, parameter selection [15–22] and data selection [10, 23–26] both rank candidates under fixed budgets, but have largely developed independently. They target complementary resources: data selection reduces optimization workload, while parameter selection reduces trainable-state memory. Recent pipelines combine these methods to save resources [27–30], yet separate scoring stages can duplicate computation and evaluate the two selections using different criteria. The parameter mask and data subset jointly determine each restricted update during fine-tuning, motivating the question: ***can one algorithm serve both goals through a shared scoring rule?***

We formulate parameter and data selection as separate bilevel problems with a shared validation objective balancing target-task adaptation with preservation of prior capabilities, and approximate their training responses at a common checkpoint. This reveals a row–column duality in scoring:

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

under first- and second-order approximations of local validation improvement, parameter importance and data utility are column and row aggregations of the same gradient interaction matrix. ***This shared matrix connects the two scoring problems, allowing both sets of scores to be computed in closed form from shared gradient statistics.*** The scores admit a localized Shapley interpretation, attributing surrogate validation improvement to training examples and parameter-coordinate updates.

Building on this duality, we propose **DualSFT**, realizing ***one algorithm for two goals***: a lightweight warmup establishes a stable scoring checkpoint, one-shot dual scoring reads parameter importance and data utility from a shared projection vector, and restricted fine-tuning updates selected parameters on selected data. By scoring both axes before the final update, DualSFT coordinates them during scoring, avoiding independent scoring rules and repeated conditional passes of sequential hybrids.

DualSFT preserves pre-trained capabilities via corpus-free **CWSD** (**C**onfidence-**W**eighted **S**elf-**D**istillation), mitigating catastrophic forgetting during downstream fine-tuning [11, 31–36]. Experiments on 3B to 9B LLMs for code generation [37] and mathematical reasoning [38] show DualSFT reduces fine-tuning cost while improving target-task performance. Furthermore, DualSFT strengthens stability-plasticity trade-offs: one-sided variants achieve best mean trade-offs within their groups, and full DualSFT yields best mean joint-constrained trade-off against sequential pipelines under matched budgets. Ablations and sensitivity analyses support its scoring and preservation components.

Our main contributions are summarized as follows:

- We derive a shared response-surrogate scoring rule for parameter and data selection from separate bilevel formulations with a common validation objective. Within this tractable scoring regime, parameter importance and data utility are the column-wise and row-wise aggregations of one gradient interaction matrix, with a localized Shapley interpretation of the resulting closed-form scores.
- We propose DualSFT, a one-shot dual scoring algorithm that co-extracts both signals from shared gradient statistics, reducing redundant scoring and repeated conditional passes in sequential combinations while coordinating the two restricted axes at scoring time.
- We construct a practical preservation direction using CWSD without pre-training data.
- On 3B–9B LLMs, DualSFT reduces tunable parameters and training data, improving target performance. Single-axis variants strengthen intra-group performance and stability-plasticity trade-offs, while full DualSFT outperforms sequential hybrid baselines in joint-constrained trade-offs.

2 Related Work

2.1 Parameter and Data Efficient Fine-tuning

PEFT efficiently restricts updates via additive modules [6], low-rank updates [3, 7, 8], and parameter selection [5, 15, 16]. Recent works combine structured or dynamic sparsity with criteria like Fisher guidance, regularization, and importance tracking [9, 17–22, 33, 39, 40]. Although conceptually related, game-theoretic evaluations [41] and Shapley-based methods primarily target post-training compression, including rank allocation [42] and layer pruning [43].

DEFT improves efficiency via reweighting [11, 35], scheduling [12, 13], token filtering [44, 45], and sample selection [10, 23, 26, 46], reallocating effort, reordering signals, dropping low-value tokens, and choosing subsets. Sample selection evolved from static gradient/subspace matching [10, 25] to trajectory-aware, validation-guided scoring [24, 46–49], refined by Shapley attribution [50], submodular conflict resolution [26], and quality-diversity tradeoffs [51–53]. Influence-based attribution links examples to target losses [54] and informs layer-wise resource allocation [55]. DualSFT derives example and parameter-coordinate scores from a shared local update surrogate.

Recent cost-efficient fine-tuning methods increasingly couple efficiency and retention. They restrict adaptation via low-rank or subspace adapters [56–58], partial updates, optimizer filtering [21, 59, 60], and data/token selection [31, 48]. While preserving pre-trained capabilities downstream, they mostly operate on a single axis, lacking a shared scoring rule for parameter and data selection.

Recent cross-axis studies show that jointly considering parameter- and data-side resources can further reduce fine-tuning cost [27–30]. Yet they lack explicitly coordinated data-subset and parameter-mask scoring under a shared validation objective. DualSFT derives both scores from a shared validation-improvement surrogate, yielding a row–column scoring rule for coordination.

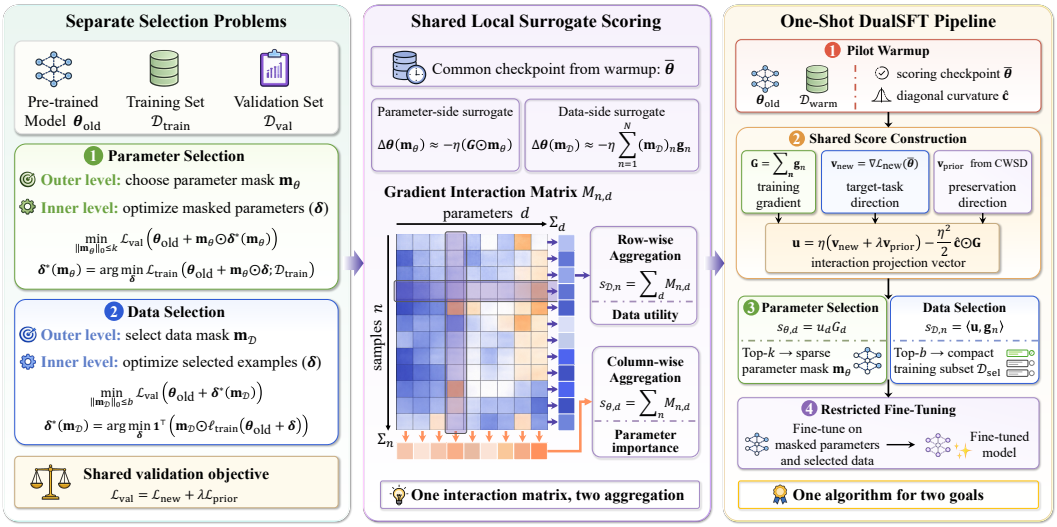


Figure 1: DualSFT overview. Separate parameter and data bilevel problems share a local surrogate: scores aggregate columns and rows of the same conceptual gradient interaction matrix. DualSFT uses one-shot streaming dual scoring with CWSD before restricted fine-tuning.

2.2 Bilevel Optimization

Bilevel optimization (BLO) is widely used for meta-learning, hyperparameter optimization, and architecture search [61–63]. It supports validation-guided data selection before and after LLMs [49, 64, 65], and efficient fine-tuning and compression [66–68]. Unlike prior BLO optimizing data, prompts, parameters, or compression targets separately, DualSFT links separate validation-guided bilevel formulations through a shared local surrogate, enabling coordinated data–parameter selection.

3 Bilevel Formulations and Surrogates for Parameter and Data Selection

Figure 1 summarizes our framework for parameter and data selection. We formulate them as separate budget-constrained bilevel problems sharing an outer validation objective that balances target-task adaptation and prior-capability preservation. Approximating their training responses at a common checkpoint yields tractable local surrogates, which underpin the shared scoring rule in Section 4.

3.1 Problem Setting

Consider a pre-trained LLM with $\theta_{\text{old}} \in \mathbb{R}^D$. Let $\delta \in \mathbb{R}^D$ denote the update applied to θ_{old}, and let $\theta = \theta_{\text{old}} + \delta$ be the adapted parameter vector. Let $\mathcal{D}_{\text{train}} = \{(x_n, y_n)\}_{n=1}^N$ be the training set and $\mathcal{D}_{\text{val}}$ a validation set. We index updatable coordinates by a binary mask $\mathbf{m}_\theta \in \{0,1\}^D$ with $\|\mathbf{m}_\theta\|_0 \le k$, so that $\delta_d = 0$ whenever $(\mathbf{m}_\theta)_d = 0$, and we index training examples by a binary mask $\mathbf{m}_\mathcal{D} \in \{0,1\}^N$ with $\|\mathbf{m}_\mathcal{D}\|_0 \le b$, inducing the selected subset $\mathcal{D}_{\text{sel}}(\mathbf{m}_\mathcal{D}) \triangleq \{(x_n, y_n) \in \mathcal{D}_{\text{train}} \mid (\mathbf{m}_\mathcal{D})_n = 1\}$. The two masks act on different objects but share the same outer criterion. To balance target-task adaptation and prior-capability preservation, we define the validation objective as $\mathcal{L}_{\text{val}}(\theta) \triangleq \mathcal{L}_{\text{new}}(\theta) + \lambda\mathcal{L}_{\text{prior}}(\theta)$, where $\mathcal{L}_{\text{new}}$ is the target-task validation loss, $\mathcal{L}_{\text{prior}}$ the prior-capability preservation loss, and $\lambda > 0$ controls the stability–plasticity trade-off.

3.2 Bilevel Optimization Problems for Parameter and Data Selection

Under the shared outer criterion $\mathcal{L}_{\text{val}}$, parameter selection and data selection can be formulated as two exact bilevel optimization problems. In both cases, the outer level selects a budget-constrained mask, while the inner level determines the training-induced parameter update under the corresponding restriction. Accordingly, the quality of a selection is determined by the induced update and its resulting validation performance. Here, $\ell(x_n, y_n; \theta)$ denotes the per-example training loss, $\boldsymbol{\ell}_{\text{train}}(\theta) \triangleq [\ell(x_1, y_1; \theta), \ldots, \ell(x_N, y_N; \theta)]^\top \in \mathbb{R}^N$ is the vector of per-sample training losses, $\mathbf{1} \in \mathbb{R}^N$ is the all-ones vector, and $\mathcal{L}_{\text{train}}(\theta; \mathcal{D}_{\text{train}}) \triangleq \mathbf{1}^\top \boldsymbol{\ell}_{\text{train}}(\theta)$ is the full training objective.

For parameter selection during LLM fine-tuning, the outer mask determines which parameters are allowed to update, and the inner problem computes the optimal update under this restriction:

$$\min_{\|\mathbf{m}_\theta\|_0 \le k} \mathcal{L}_{\text{val}}(\boldsymbol{\theta}_{\text{old}} + \mathbf{m}_\theta \odot \boldsymbol{\delta}^*(\mathbf{m}_\theta)) \quad \text{s.t.} \quad \boldsymbol{\delta}^*(\mathbf{m}_\theta) = \arg\min_{\boldsymbol{\delta}} \mathcal{L}_{\text{train}}(\boldsymbol{\theta}_{\text{old}} + \mathbf{m}_\theta \odot \boldsymbol{\delta}; \mathcal{D}_{\text{train}}). \quad (1)$$

For data selection during LLM fine-tuning, the outer mask determines which training examples are used for learning, and the inner problem computes the update induced by the selected subset:

$$\min_{\|\mathbf{m}_\mathcal{D}\|_0 \le b} \mathcal{L}_{\text{val}}(\boldsymbol{\theta}_{\text{old}} + \boldsymbol{\delta}^*(\mathbf{m}_\mathcal{D})) \quad \text{s.t.} \quad \boldsymbol{\delta}^*(\mathbf{m}_\mathcal{D}) = \arg\min_{\boldsymbol{\delta}} \mathbf{1}^\top \Big(\mathbf{m}_\mathcal{D} \odot \boldsymbol{\ell}_{\text{train}}(\boldsymbol{\theta}_{\text{old}} + \boldsymbol{\delta})\Big). \quad (2)$$

Here, $\boldsymbol{\delta}^*(\mathbf{m}_\theta)$ and $\boldsymbol{\delta}^*(\mathbf{m}_\mathcal{D})$ denote the optimal updates induced by the corresponding masks.

3.3 Local Surrogate Approximation of the Exact Bilevel Problems

Solving Eqs. (1) and (2) is intractable at LLM scale, as each candidate mask requires constrained fine-tuning. Following In-Run Data Shapley [50] and scalable proxy-based scoring methods [24, 46, 49, 69–71], we approximate both updates via one gradient step of the inner objective at a common checkpoint $\bar{\boldsymbol{\theta}} \in \mathbb{R}^D$, yielding local response surrogates sharing this reference point.

Let $\mathbf{g}_n \triangleq \nabla_{\boldsymbol{\theta}} \ell(x_n, y_n; \bar{\boldsymbol{\theta}})$ be the n-th example's gradient at $\bar{\boldsymbol{\theta}}$ and $\mathbf{G} \triangleq \sum_{n=1}^N \mathbf{g}_n$ the aggregated training gradient. The masked parameter-side step then descends along $\mathbf{G} \odot \mathbf{m}_\theta$, while the data-side step descends along $\sum_{n=1}^N (\mathbf{m}_\mathcal{D})_n \mathbf{g}_n$, giving the one-step inner-level updates:

$$\Delta\boldsymbol{\theta}(\mathbf{m}_\theta) \approx -\eta(\mathbf{G} \odot \mathbf{m}_\theta), \qquad \Delta\boldsymbol{\theta}(\mathbf{m}_\mathcal{D}) \approx -\eta \sum\nolimits_{n=1}^N (\mathbf{m}_\mathcal{D})_n \mathbf{g}_n, \quad (3)$$

with η the effective local step size. Appendix A.1 gives the derivation. The parameter-side surrogate masks the aggregate gradient, whereas the data-side surrogate sums selected-example gradients.

For notational convenience, let $[D] \triangleq \{1, \dots, D\}$, $[N] \triangleq \{1, \dots, N\}$, and identify each mask with its selected index set $S_\theta \triangleq \{d \in [D] \mid (\mathbf{m}_\theta)_d = 1\}$, $S_\mathcal{D} \triangleq \{n \in [N] \mid (\mathbf{m}_\mathcal{D})_n = 1\}$. Substituting Eq. (3) into the validation gain at $\bar{\boldsymbol{\theta}}$ yields the localized utilities:

$$U_\theta(S_\theta) \triangleq \mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}}) - \mathcal{L}_{\text{val}}\big(\bar{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}_\theta(S_\theta)\big), \qquad U_\mathcal{D}(S_\mathcal{D}) \triangleq \mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}}) - \mathcal{L}_{\text{val}}\big(\bar{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}_\mathcal{D}(S_\mathcal{D})\big), \quad (4)$$

where G_d denotes the d-th entry of $\mathbf{G}$, $\mathbf{e}_d \in \mathbb{R}^D$ the d-th standard basis vector, $\Delta\boldsymbol{\theta}_\theta(S_\theta) \triangleq -\eta \sum_{d \in S_\theta} G_d \mathbf{e}_d$, and $\Delta\boldsymbol{\theta}_\mathcal{D}(S_\mathcal{D}) \triangleq -\eta \sum_{n \in S_\mathcal{D}} \mathbf{g}_n$. These utilities measure validation-objective reduction at $\bar{\boldsymbol{\theta}}$ from local updates, providing a tractable basis for shared scoring analysis.

4 Shared Interaction Matrix and Dual Scores

Parameter and data scores arise as column and row aggregations of a shared interaction matrix built from a local validation-improvement expansion around $\bar{\boldsymbol{\theta}}$. The score identities and Shapley decompositions are exact within specified surrogate games; approximation and selection-stability bounds hold under stated assumptions. These scores guide one-shot selection for multi-step AdamW fine-tuning.

4.1 Shared Taylor Approximation, Interaction Matrix, and Closed-Form Scores

Let $\mathbf{v}_{\text{new}} \triangleq \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{new}}(\bar{\boldsymbol{\theta}})$, $\mathbf{v}_{\text{prior}} \triangleq \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{prior}}(\bar{\boldsymbol{\theta}})$, $\mathbf{v}_{\text{val}} \triangleq \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}}) = \mathbf{v}_{\text{new}} + \lambda \mathbf{v}_{\text{prior}}$, and $\mathbf{H} \triangleq \nabla^2_{\boldsymbol{\theta}} \mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}})$. Although the localized utilities in Eq. (4) act on different selection spaces, the surrogate updates in Eq. (3) place them around the same checkpoint $\bar{\boldsymbol{\theta}}$ under one outer objective, so both admit a common local second-order expansion. For either $a \in \{\theta, \mathcal{D}\}$ and its selected set S_a:

$$U_a(S_a) = -\mathbf{v}_{\text{val}}^\top \Delta\boldsymbol{\theta}_a(S_a) - \frac{1}{2} \Delta\boldsymbol{\theta}_a(S_a)^\top \mathbf{H} \Delta\boldsymbol{\theta}_a(S_a) + R_a(S_a), \quad (5)$$

where $R_a(S_a) = O(\|\Delta\boldsymbol{\theta}_a(S_a)\|^3)$ is the Taylor remainder. We use the first- or second-order truncation of Eq. (5); Appendix A.2 shows utility errors are $O(\eta^2)$ and $O(\eta^3)$, respectively.

Following tractable diagonal curvature approximations for scalable and stable LLM fine-tuning [72–74], we adopt the diagonal approximation $\mathbf{H} \approx \text{Diag}(c_1, \dots, c_D)$ for this instantiation. With G_d and $g_{n,d}$ denoting the d-th entries of $\mathbf{G}$ and $\mathbf{g}_n$, we define three interaction matrices:

$$M^{(1)}_{n,d} \triangleq \eta(\mathbf{v}_{\text{val}})_d g_{n,d}, \quad M^{(\text{diag})}_{n,d} \triangleq M^{(1)}_{n,d} - \frac{\eta^2}{2} c_d G_d g_{n,d}, \quad M^{(\text{full})}_{n,d} \triangleq M^{(1)}_{n,d} - \frac{\eta^2}{2} (\mathbf{HG})_d g_{n,d}, \quad (6)$$

and let $\mathbf{M}^{(\alpha)} \in \mathbb{R}^{N\times D}$ collect the entries $M^{(\alpha)}_{n,d}$. For $\alpha \in \{1, \text{diag}, \text{full}\}$, the corresponding scores are

$$s^{(\alpha)}_{\theta,d} \triangleq \sum_{n=1}^{N} M^{(\alpha)}_{n,d}, \qquad s^{(\alpha)}_{\mathcal{D},n} \triangleq \sum_{d=1}^{D} M^{(\alpha)}_{n,d}. \tag{7}$$

Within the local validation-improvement surrogate, the scoring relation takes an explicit form: parameter importance and data utility are column and row aggregations of one shared interaction matrix.

4.2 Shapley Interpretation of the Closed-Form Scores

The closed-form scores admit a Shapley interpretation. For each $\alpha \in \{1, \text{diag}, \text{full}\}$, localized utilities $U^{(\alpha)}_{\theta}$ and $U^{(\alpha)}_{\mathcal{D}}$ from Eq. (5) define cooperative games $\mathcal{G}^{(\alpha)}_{\theta} = ([D], U^{(\alpha)}_{\theta})$ and $\mathcal{G}^{(\alpha)}_{\mathcal{D}} = ([N], U^{(\alpha)}_{\mathcal{D}})$.

Theorem 1. *For any $\alpha \in \{1, \text{diag}, \text{full}\}$, the Shapley values of the two localized surrogate games are given by the closed-form scores in Eq. (7):*

$$\phi^{(\alpha)}_{\theta,d} = s^{(\alpha)}_{\theta,d} = \sum_{n=1}^{N} M^{(\alpha)}_{n,d}, \qquad \phi^{(\alpha)}_{\mathcal{D},n} = s^{(\alpha)}_{\mathcal{D},n} = \sum_{d=1}^{D} M^{(\alpha)}_{n,d}. \tag{8}$$

The Shapley view interprets closed-form scores as attributions: first-order localized utilities decompose additively; in diagonal and full second-order cases, permutation averaging redistributes pairwise interactions into row and column aggregations of Eq. (7). Within surrogate games, the interaction matrix yields practical scores and Shapley-consistent attributions. Appendix A.3 details derivations.

4.3 Coordination Gap under Isolated Scoring

The shared interaction matrix exposes the cost of isolated scoring under joint restricted updates. Let $\mathbf{G} = \sum_{n=1}^{N} \mathbf{g}_n$ and $\mathbf{G}_{S_{\mathcal{D}}} = \sum_{n\in S_{\mathcal{D}}} \mathbf{g}_n$. Under both restrictions, $\Delta\boldsymbol{\theta}(S_{\mathcal{D}}, \mathbf{m}_\theta) = -\eta(\mathbf{m}_\theta \odot \mathbf{G}_{S_{\mathcal{D}}})$, inducing mask-aware first-order scores $\bar{s}_{\mathcal{D},n}(\mathbf{m}_\theta) = \eta\langle \mathbf{v}_{\text{val}} \odot \mathbf{m}_\theta, \mathbf{g}_n\rangle$ and $\bar{s}_{\theta,d}(S_{\mathcal{D}}) = \eta(\mathbf{v}_{\text{val}})_d(\mathbf{G}_{S_{\mathcal{D}}})_d$.

Proposition 1. *Let $s^{\text{iso}}_{\mathcal{D},n} = \eta\langle \mathbf{v}_{val}, \mathbf{g}_n\rangle$ and $s^{\text{iso}}_{\theta,d} = \eta(\mathbf{v}_{val})_d G_d$ be the isolated first-order data and parameter scores. Let $\bar{S}_\theta = \{d \in [D] : (\mathbf{m}_\theta)_d = 0\}$ and $\bar{S}_{\mathcal{D}} = [N] \setminus S_{\mathcal{D}}$. Then:*

$$s^{\text{iso}}_{\mathcal{D},n} - \bar{s}_{\mathcal{D},n}(\mathbf{m}_\theta) = \eta\textstyle\sum_{d\in\bar{S}_\theta}(\mathbf{v}_{val})_d g_{n,d}, \quad s^{\text{iso}}_{\theta,d} - \bar{s}_{\theta,d}(S_{\mathcal{D}}) = \eta(\mathbf{v}_{val})_d\textstyle\sum_{n\in\bar{S}_{\mathcal{D}}} g_{n,d}. \tag{9}$$

Eq. (9) identifies contributions removed by the opposite restriction: frozen coordinates contribute to data scores, and excluded samples contribute to parameter scores. DualSFT coordinates the scoring criterion by extracting both marginal scores from shared gradient statistics under the same validation objective and checkpoint, followed by separate Top-b/Top-k selection. Dependencies on the final opposite mask remain unmodeled in these one-shot scores. Section 6.4 evaluates the resulting utility loss and the cost–benefit of conditional re-selection. Appendix A.4 proves Proposition 1.

5 The DualSFT Algorithm

5.1 Pilot Warmup and Global Dual Scoring

DualSFT instantiates the local scoring framework in Sections 3 and 4 using warmup and shared gradient statistics. Warmup on $\mathcal{D}_{\text{warm}}$ produces $\bar{\boldsymbol{\theta}}$ to stabilize scoring and reduce gradient noise [10, 24, 25]. It provides a diagonal curvature proxy [75–78]: after t_w AdamW updates, we bias-correct the second-moment buffer: $\hat{\mathbf{r}}_{t_w} = \mathbf{r}_{t_w}/(1-\beta_2^{t_w})$ and set $\hat{\mathbf{c}} = \sqrt{\hat{\mathbf{r}}_{t_w}} + \epsilon_{\text{adam}}$ elementwise. This RMS-gradient scale is fixed during scoring; Appendix B explains its surrogate role.

At $\bar{\boldsymbol{\theta}}$, we accumulate $\mathbf{G} = \sum_{n=1}^{N_{\text{sc}}} \mathbf{g}_n$ over $\mathcal{D}_{\text{pool}} = \mathcal{D}_{\text{train}} \setminus \mathcal{D}_{\text{warm}}$, where $N_{\text{sc}} = |\mathcal{D}_{\text{pool}}|$. Setting $\eta_{\text{sc}} = \eta_{\text{ft}}/N_{\text{sc}}$ makes $\eta_{\text{sc}}\mathbf{G}$ a mean-gradient step with learning-rate scale η_{ft}. Combining the target-validation gradient $\mathbf{v}_{\text{new}}$ with the CWSD preservation gradient $\mathbf{v}_{\text{prior}}$ gives the shared scoring projection:

$$\mathbf{u} \triangleq \eta_{\text{sc}}\mathbf{v}_{\text{val}} - \frac{\eta_{\text{sc}}^2}{2}\hat{\mathbf{c}} \odot \mathbf{G} = \eta_{\text{sc}}(\mathbf{v}_{\text{new}} + \lambda\mathbf{v}_{\text{prior}}) - \frac{\eta_{\text{sc}}^2}{2}\hat{\mathbf{c}} \odot \mathbf{G}. \tag{10}$$

This projection combines the first-order validation gradient and optimizer-aware diagonal correction. Aggregating interaction scores column-wise and row-wise yields parameter and data scores:

$$M^{(\text{diag})}_{n,d} = u_d g_{n,d}, \qquad s_{\theta,d} = \textstyle\sum_{n=1}^{N_{\text{sc}}} M^{(\text{diag})}_{n,d} = u_d G_d, \qquad s_{\mathcal{D},n} = \textstyle\sum_{d=1}^{D} M^{(\text{diag})}_{n,d} = \langle \mathbf{u}, \mathbf{g}_n\rangle. \tag{11}$$

We select the largest signed scores using Top-k for parameters and Top-b for data. Appendix A.5 bounds deviations from the local full-Hessian scores in terms of η_{sc}^2 and curvature-proxy error, and establishes selection stability when boundary gaps exceed the perturbation bounds. After accumulating $\mathbf{G}$ and constructing $\mathbf{u}$, a second scoring-pool pass computes data scores with Ghost Dot Product [50, 69], avoiding per-sample-gradient materialization. For an affine weight matrix, its contribution is $\sum_t \boldsymbol{\epsilon}_{n,t}^\top \mathbf{U}_{\text{layer}} \mathbf{a}_{n,t}$, where t indexes token positions, $\mathbf{a}_{n,t}$ is the input activation, $\boldsymbol{\epsilon}_{n,t}$ is the pre-activation gradient of the example loss, and $\mathbf{U}_{\text{layer}}$ reshapes the corresponding segment of $\mathbf{u}$.

CWSD supplies $\mathbf{v}_{\text{prior}}$ without pretraining data, using the frozen pre-trained model as teacher and the warmup model as student to measure predictive changes introduced during pilot warmup. Let $\mathcal{X}_{\text{anchor}}$ contain inputs from a held-out anchor set with $\mathcal{D}_{\text{anchor}} \cap \mathcal{D}_{\text{train}} = \emptyset$. We compute:

$$\mathbf{v}_{\text{prior}} \approx \nabla_{\bar{\theta}} \frac{1}{|\mathcal{X}_{\text{anchor}}|} \sum_{x \in \mathcal{X}_{\text{anchor}}} \omega(x) \tau^2 \mathrm{KL}\Big(P_\tau(\cdot \mid \boldsymbol{\theta}_{\text{old}}, x) \,\Big\|\, P_\tau(\cdot \mid \bar{\boldsymbol{\theta}}, x) \Big), \tag{12}$$

where τ is the distillation temperature and $\omega(x) = 1 - H_{\boldsymbol{\theta}_{\text{old}}}(x)/\log|\mathcal{V}|$ emphasizes teacher-confident inputs; $H_{\boldsymbol{\theta}_{\text{old}}}(x)$ is the mean per-token predictive entropy. Teacher predictions and confidence weights are fixed, and gradients flow through the student. Evaluating at the warmup checkpoint captures departures from the pre-trained teacher. We average token-wise forward KL over non-padding response tokens and then over anchor examples. For each weighted token loss, the student-logit gradient before averaging is $\omega(x)\tau(\mathbf{p} - \mathbf{q})$, with $\mathbf{p}$ and $\mathbf{q}$ denoting student and teacher distributions. Appendix C derives the resulting parameter gradient through the student Jacobian. CWSD is used only for selection; final restricted fine-tuning optimizes the target-task loss.

5.2 Restricted Fine-Tuning and Complexity

In Algorithm 1, pilot warmup yields $\bar{\boldsymbol{\theta}}$, $\hat{\mathbf{c}}$; one-shot dual scoring via $\mathbf{u}$ selects $\mathcal{D}_{\text{sel}}$, $\mathbf{m}_\theta$; restricted fine-tuning updates selected parameters on $\mathcal{D}_{\text{sel}}$. Restarting from $\boldsymbol{\theta}_{\text{old}}$ decouples scoring from adaptation, following warmup-then-select-from-scratch pipelines [10, 24, 25]. Final AdamW fine-tuning, $\textsc{MaskedFT}_{\eta_{\text{ft}}}(\boldsymbol{\theta}_{\text{old}}, \mathcal{D}_{\text{sel}}, \mathbf{m}_\theta)$, masks gradients, adaptive moments, and decoupled weight decay.

DualSFT avoids materializing the $O(N_{\text{sc}} \times D)$ interaction matrix, confining selection to lightweight warmup and two streaming passes. First accumulates $\mathbf{G}$, constructing $\mathbf{u}$; second computes sample scores, maintaining the Top-b subset. Parameter scoring uses full-model vectors ($\mathbf{G}$, $\mathbf{u}$, $\hat{\mathbf{c}}$), incurring $O(D)$ auxiliary memory, linear in model size, implementable blockwise or via sharding. For data scoring, Ghost Dot Product avoids materializing per-sample gradients $\mathbf{g}_n \in \mathbb{R}^D$. Restricted fine-tuning proceeds on $\mathcal{D}_{\text{sel}}$ masked by $\mathbf{m}_\theta$. Thus, DualSFT reduces scoring redundancy and avoids interaction-matrix materialization, with parameter scoring requiring linear auxiliary state.

Algorithm 1 DualSFT

Require: $\boldsymbol{\theta}_{\text{old}}, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{val}}, \mathcal{D}_{\text{warm}}, \mathcal{D}_{\text{anchor}}, \eta_{\text{ft}}, \lambda, \tau, b, k$

$(\bar{\boldsymbol{\theta}}, \hat{\mathbf{c}}) \leftarrow \textsc{Warmup}(\boldsymbol{\theta}_{\text{old}}, \mathcal{D}_{\text{warm}})$
$\mathcal{D}_{\text{pool}} \leftarrow \mathcal{D}_{\text{train}} \setminus \mathcal{D}_{\text{warm}}$
$\mathbf{G} \leftarrow \textsc{GradSum}(\bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{pool}})$
$N_{\text{sc}} \leftarrow |\mathcal{D}_{\text{pool}}|, \quad \eta_{\text{sc}} \leftarrow \eta_{\text{ft}}/N_{\text{sc}}$
$\mathbf{v}_{\text{new}} \leftarrow \nabla_{\bar{\boldsymbol{\theta}}} \mathcal{L}_{\text{new}}(\bar{\boldsymbol{\theta}})$
$\mathbf{v}_{\text{prior}} \leftarrow \text{CWSD}(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_{\text{old}}, \mathcal{D}_{\text{anchor}}, \tau)$
$\mathbf{u} \leftarrow \eta_{\text{sc}}(\mathbf{v}_{\text{new}} + \lambda \mathbf{v}_{\text{prior}}) - (\eta_{\text{sc}}^2/2)\hat{\mathbf{c}} \odot \mathbf{G}$
$\mathbf{s}_\theta \leftarrow \mathbf{u} \odot \mathbf{G}, \quad \mathbf{s}_{\mathcal{D}} \leftarrow \textsc{GhostDot}(\mathbf{u}, \bar{\boldsymbol{\theta}}, \mathcal{D}_{\text{pool}})$
$\mathbf{m}_\theta \leftarrow \textsc{TopK}_{\text{signed}}(\mathbf{s}_\theta, k)$
$\mathbf{m}_{\mathcal{D}} \leftarrow \textsc{TopB}_{\text{signed}}(\mathbf{s}_{\mathcal{D}}, b)$
$\mathcal{D}_{\text{sel}} \leftarrow \{(x_n, y_n) \in \mathcal{D}_{\text{pool}} : (\mathbf{m}_{\mathcal{D}})_n = 1\}$
$\boldsymbol{\theta}^\star \leftarrow \textsc{MaskedFT}_{\eta_{\text{ft}}}(\boldsymbol{\theta}_{\text{old}}, \mathcal{D}_{\text{sel}}, \mathbf{m}_\theta)$
return $\mathbf{m}_{\mathcal{D}}, \mathbf{m}_\theta, \boldsymbol{\theta}^\star$

6 Experiments

We evaluate DualSFT through matched-budget performance and efficiency comparisons, component ablations, sensitivity analyses, and diagnostics of score reliability and re-selection costs. Comparisons with forgetting-specific baselines further assess capability preservation.

6.1 Experimental Setup

Baselines. We compare against ❶ standard methods (Standard SFT, LoRA [3]); ❷ parameter-only sparse methods (S^2FT [17], SMT [19], LIFT [20], NanoAdam [21]); ❸ data-only methods (Random, LESS [10], FisherSFT [23], SPICE [26]); ❹ sequential hybrids (LESS + S^2FT, SPICE + NanoAdam); and ❺ forgetting-aware baselines (DavIR [79], FLOW [11], STM [31], FAPM [33], TALR [35]).

Datasets & Metrics. We evaluate on Magicoder [37] and MetaMathQA [38], each reserving 1,024 validation samples and 2,048 disjoint anchor samples for CWSD. Following prior studies [11, 32, 80],

Table 1: Fine-tuning results on Magicoder (%). Pre-trained, Standard SFT, and LoRA (gray) are excluded from group ranking. In each group, **best** and second-best results are highlighted.

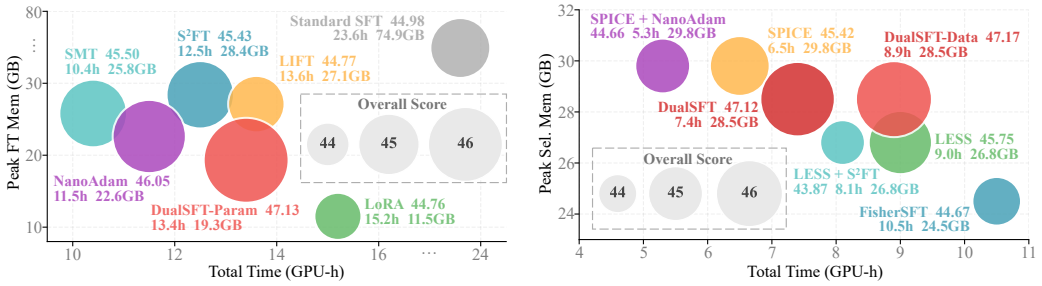
Model	Method	Stability ↑					Plasticity ↑	Overall ↑
		ARC-C	PIQA	MMLU	GSM8K	Avg	HE	Score
Llama-3.2-3B	Pre-trained	43.00$_{\pm 0.00}$	76.61$_{\pm 0.00}$	56.52$_{\pm 0.00}$	27.45$_{\pm 0.00}$	50.89$_{\pm 0.00}$	28.66$_{\pm 0.00}$	39.78$_{\pm 0.00}$
	Standard SFT	39.99$_{\pm 1.46}$	72.09$_{\pm 0.98}$	48.18$_{\pm 0.39}$	20.70$_{\pm 0.85}$	45.24$_{\pm 0.50}$	44.72$_{\pm 0.57}$	44.98$_{\pm 0.37}$
	LoRA ICLR'22	42.61$_{\pm 1.45}$	74.43$_{\pm 0.99}$	51.87$_{\pm 0.41}$	23.30$_{\pm 0.83}$	48.05$_{\pm 0.49}$	41.46$_{\pm 0.50}$	44.76$_{\pm 0.37}$
	S^{2}FT NIPS'24	41.81$_{\pm 1.45}$	**76.26**$_{\pm 0.98}$	52.10$_{\pm 0.40}$	22.57$_{\pm 0.72}$	48.18$_{\pm 0.49}$	42.68$_{\pm 0.86}$	45.43$_{\pm 0.44}$
	SMT ICLR'25	40.98$_{\pm 1.45}$	74.16$_{\pm 0.98}$	52.59$_{\pm 0.40}$	23.88$_{\pm 0.75}$	47.90$_{\pm 0.48}$	43.09$_{\pm 0.57}$	45.50$_{\pm 0.39}$
	LIFT ICML'25	41.21$_{\pm 1.45}$	72.51$_{\pm 0.99}$	52.14$_{\pm 0.40}$	23.22$_{\pm 0.73}$	47.27$_{\pm 0.48}$	42.28$_{\pm 0.57}$	44.77$_{\pm 0.39}$
	NanoAdam NIPS'25	**42.66**$_{\pm 1.44}$	74.68$_{\pm 0.98}$	52.71$_{\pm 0.40}$	24.39$_{\pm 0.73}$	48.61$_{\pm 0.48}$	43.50$_{\pm 0.76}$	46.05$_{\pm 0.50}$
	DualSFT-Param (Ours)	42.55$_{\pm 1.46}$	76.19$_{\pm 0.99}$	**54.01**$_{\pm 0.41}$	**26.28**$_{\pm 0.75}$	**49.76**$_{\pm 0.50}$	**44.51**$_{\pm 0.86}$	**47.13**$_{\pm 0.42}$
	Random	**43.00**$_{\pm 1.46}$	76.10$_{\pm 0.98}$	**55.31**$_{\pm 0.40}$	**26.99**$_{\pm 0.78}$	**50.35**$_{\pm 0.49}$	38.62$_{\pm 0.76}$	44.48$_{\pm 0.45}$
	LESS ICML'24	42.18$_{\pm 1.45}$	75.66$_{\pm 0.99}$	53.97$_{\pm 0.40}$	23.45$_{\pm 0.63}$	48.82$_{\pm 0.50}$	42.68$_{\pm 0.86}$	45.75$_{\pm 0.46}$
	FisherSFT ICML'25	41.55$_{\pm 1.44}$	74.17$_{\pm 0.99}$	52.83$_{\pm 0.40}$	22.92$_{\pm 0.62}$	47.87$_{\pm 0.47}$	41.46$_{\pm 0.86}$	44.67$_{\pm 0.43}$
	SPICE ICLR'26	40.53$_{\pm 1.45}$	76.10$_{\pm 0.98}$	53.13$_{\pm 0.40}$	24.51$_{\pm 0.73}$	48.57$_{\pm 0.44}$	42.28$_{\pm 0.57}$	45.42$_{\pm 0.39}$
	DualSFT-Data (Ours)	42.66$_{\pm 1.44}$	**76.44**$_{\pm 0.99}$	55.28$_{\pm 0.40}$	26.54$_{\pm 0.73}$	50.23$_{\pm 0.50}$	**44.11**$_{\pm 0.76}$	**47.17**$_{\pm 0.43}$
	LESS + S^{2}FT	42.58$_{\pm 1.44}$	**76.68**$_{\pm 0.96}$	53.55$_{\pm 0.39}$	22.87$_{\pm 0.71}$	48.92$_{\pm 0.48}$	38.82$_{\pm 0.57}$	43.87$_{\pm 0.40}$
	SPICE + NanoAdam	42.15$_{\pm 1.46}$	76.61$_{\pm 0.98}$	54.87$_{\pm 0.40}$	25.93$_{\pm 0.71}$	49.89$_{\pm 0.49}$	39.43$_{\pm 0.76}$	44.66$_{\pm 0.42}$
	DualSFT (Ours)	**42.95**$_{\pm 1.45}$	76.22$_{\pm 0.98}$	**56.09**$_{\pm 0.40}$	**26.89**$_{\pm 0.70}$	**50.54**$_{\pm 0.47}$	**43.70**$_{\pm 0.76}$	**47.12**$_{\pm 0.43}$
Gemma-3-4B-PT	Pre-trained	51.45$_{\pm 0.00}$	79.16$_{\pm 0.00}$	59.61$_{\pm 0.00}$	37.00$_{\pm 0.00}$	56.80$_{\pm 0.00}$	35.37$_{\pm 0.00}$	46.09$_{\pm 0.00}$
	Standard SFT	47.16$_{\pm 1.46}$	76.39$_{\pm 0.99}$	53.87$_{\pm 0.40}$	24.03$_{\pm 0.81}$	50.36$_{\pm 0.50}$	58.13$_{\pm 0.76}$	54.25$_{\pm 0.51}$
	LoRA ICLR'22	49.15$_{\pm 1.46}$	75.57$_{\pm 1.00}$	56.71$_{\pm 0.39}$	31.64$_{\pm 0.82}$	53.27$_{\pm 0.49}$	54.67$_{\pm 1.04}$	53.97$_{\pm 0.52}$
	S^{2}FT NIPS'24	49.74$_{\pm 1.46}$	78.29$_{\pm 0.96}$	55.75$_{\pm 0.40}$	31.08$_{\pm 0.60}$	53.72$_{\pm 0.47}$	55.28$_{\pm 0.76}$	54.50$_{\pm 0.50}$
	SMT ICLR'25	49.91$_{\pm 1.46}$	77.98$_{\pm 0.93}$	56.44$_{\pm 0.39}$	32.75$_{\pm 0.84}$	54.27$_{\pm 0.51}$	56.50$_{\pm 0.76}$	55.39$_{\pm 0.50}$
	LIFT ICML'25	50.03$_{\pm 1.46}$	77.17$_{\pm 0.95}$	54.56$_{\pm 0.40}$	30.60$_{\pm 0.63}$	53.09$_{\pm 0.48}$	55.69$_{\pm 0.57}$	54.39$_{\pm 0.39}$
	NanoAdam NIPS'25	50.77$_{\pm 1.45}$	78.20$_{\pm 0.94}$	**58.33**$_{\pm 0.39}$	33.13$_{\pm 0.84}$	55.11$_{\pm 0.49}$	56.30$_{\pm 1.04}$	55.70$_{\pm 0.51}$
	DualSFT-Param (Ours)	**51.14**$_{\pm 1.45}$	**78.55**$_{\pm 0.98}$	58.14$_{\pm 0.39}$	**36.34**$_{\pm 0.83}$	**56.04**$_{\pm 0.49}$	**57.52**$_{\pm 0.76}$	**56.78**$_{\pm 0.48}$
	Random	**51.25**$_{\pm 1.46}$	78.89$_{\pm 0.95}$	56.91$_{\pm 0.40}$	34.65$_{\pm 0.67}$	55.42$_{\pm 0.48}$	49.19$_{\pm 0.76}$	52.31$_{\pm 0.48}$
	LESS ICML'24	50.91$_{\pm 1.46}$	**78.93**$_{\pm 0.93}$	54.14$_{\pm 0.39}$	31.59$_{\pm 0.84}$	53.89$_{\pm 0.51}$	**57.11**$_{\pm 1.04}$	55.50$_{\pm 0.53}$
	FisherSFT ICML'25	50.26$_{\pm 1.45}$	74.50$_{\pm 0.94}$	53.71$_{\pm 0.38}$	32.17$_{\pm 0.77}$	52.66$_{\pm 0.48}$	53.46$_{\pm 0.76}$	53.06$_{\pm 0.45}$
	SPICE ICLR'26	50.48$_{\pm 1.44}$	78.16$_{\pm 0.95}$	55.12$_{\pm 0.40}$	32.63$_{\pm 0.72}$	54.10$_{\pm 0.48}$	56.91$_{\pm 0.76}$	55.50$_{\pm 0.46}$
	DualSFT-Data (Ours)	51.02$_{\pm 1.45}$	78.53$_{\pm 0.98}$	**58.33**$_{\pm 0.40}$	**36.54**$_{\pm 0.79}$	**56.11**$_{\pm 0.49}$	56.91$_{\pm 0.76}$	**56.51**$_{\pm 0.42}$
	LESS + S^{2}FT	51.02$_{\pm 1.46}$	78.65$_{\pm 0.93}$	55.68$_{\pm 0.39}$	31.13$_{\pm 0.78}$	54.12$_{\pm 0.49}$	53.05$_{\pm 0.86}$	53.59$_{\pm 0.48}$
	SPICE + NanoAdam	50.77$_{\pm 1.45}$	78.55$_{\pm 0.94}$	57.29$_{\pm 0.40}$	33.89$_{\pm 0.82}$	55.12$_{\pm 0.49}$	52.64$_{\pm 0.76}$	53.88$_{\pm 0.47}$
	DualSFT (Ours)	**51.14**$_{\pm 1.44}$	**79.02**$_{\pm 0.95}$	**59.16**$_{\pm 0.38}$	**35.61**$_{\pm 0.64}$	**56.23**$_{\pm 0.48}$	**56.71**$_{\pm 0.86}$	**56.47**$_{\pm 0.43}$
Qwen-3.5-9B-Base	Pre-trained	54.27$_{\pm 0.00}$	80.09$_{\pm 0.00}$	72.62$_{\pm 0.00}$	86.20$_{\pm 0.00}$	73.29$_{\pm 0.00}$	60.98$_{\pm 0.00}$	67.13$_{\pm 0.00}$
	Standard SFT	49.23$_{\pm 1.46}$	74.30$_{\pm 0.98}$	66.14$_{\pm 0.40}$	78.65$_{\pm 0.75}$	67.08$_{\pm 0.49}$	75.61$_{\pm 0.86}$	71.34$_{\pm 0.49}$
	LoRA ICLR'22	51.51$_{\pm 1.46}$	76.62$_{\pm 0.95}$	69.47$_{\pm 0.39}$	81.93$_{\pm 0.73}$	69.88$_{\pm 0.48}$	70.33$_{\pm 0.76}$	70.10$_{\pm 0.50}$
	S^{2}FT NIPS'24	51.73$_{\pm 1.45}$	78.96$_{\pm 0.93}$	**70.35**$_{\pm 0.37}$	82.06$_{\pm 0.45}$	70.78$_{\pm 0.45}$	72.56$_{\pm 0.86}$	71.67$_{\pm 0.45}$
	SMT ICLR'25	50.34$_{\pm 1.45}$	76.51$_{\pm 0.95}$	68.41$_{\pm 0.39}$	82.64$_{\pm 0.74}$	69.48$_{\pm 0.49}$	73.78$_{\pm 0.86}$	71.63$_{\pm 0.48}$
	LIFT ICML'25	51.22$_{\pm 1.44}$	77.06$_{\pm 0.96}$	68.75$_{\pm 0.40}$	82.13$_{\pm 0.72}$	69.79$_{\pm 0.49}$	71.75$_{\pm 0.76}$	70.77$_{\pm 0.43}$
	NanoAdam NIPS'25	52.87$_{\pm 1.45}$	**79.04**$_{\pm 0.95}$	70.22$_{\pm 0.39}$	**84.23**$_{\pm 0.80}$	71.59$_{\pm 0.49}$	73.37$_{\pm 0.76}$	72.48$_{\pm 0.46}$
	DualSFT-Param (Ours)	**53.53**$_{\pm 1.44}$	78.73$_{\pm 0.95}$	70.26$_{\pm 0.39}$	**84.23**$_{\pm 0.70}$	**71.69**$_{\pm 0.47}$	**75.20**$_{\pm 0.76}$	**73.44**$_{\pm 0.44}$
	Random	**54.10**$_{\pm 1.44}$	79.63$_{\pm 0.97}$	**71.55**$_{\pm 0.40}$	**85.47**$_{\pm 0.77}$	**72.69**$_{\pm 0.48}$	65.24$_{\pm 0.86}$	68.97$_{\pm 0.48}$
	LESS ICML'24	50.26$_{\pm 1.45}$	79.02$_{\pm 0.97}$	69.65$_{\pm 0.41}$	81.27$_{\pm 0.80}$	70.05$_{\pm 0.47}$	72.76$_{\pm 0.57}$	71.41$_{\pm 0.40}$
	FisherSFT ICML'25	50.94$_{\pm 1.46}$	76.15$_{\pm 0.93}$	68.49$_{\pm 0.37}$	81.35$_{\pm 0.71}$	69.23$_{\pm 0.50}$	71.54$_{\pm 0.76}$	70.39$_{\pm 0.43}$
	SPICE ICLR'26	50.85$_{\pm 1.44}$	76.15$_{\pm 0.98}$	69.22$_{\pm 0.40}$	83.02$_{\pm 0.63}$	69.81$_{\pm 0.48}$	72.15$_{\pm 0.76}$	70.98$_{\pm 0.43}$
	DualSFT-Data (Ours)	53.67$_{\pm 1.45}$	**80.03**$_{\pm 0.95}$	70.75$_{\pm 0.39}$	85.04$_{\pm 0.70}$	72.37$_{\pm 0.48}$	**74.19**$_{\pm 0.57}$	**73.28**$_{\pm 0.41}$
	LESS + S^{2}FT	50.97$_{\pm 1.45}$	79.54$_{\pm 0.96}$	69.14$_{\pm 0.40}$	81.27$_{\pm 0.75}$	70.23$_{\pm 0.48}$	67.89$_{\pm 0.76}$	69.06$_{\pm 0.44}$
	SPICE + NanoAdam	50.54$_{\pm 1.44}$	79.34$_{\pm 0.95}$	70.09$_{\pm 0.39}$	80.74$_{\pm 0.71}$	70.18$_{\pm 0.48}$	69.31$_{\pm 0.76}$	69.74$_{\pm 0.42}$
	DualSFT (Ours)	**54.01**$_{\pm 1.43}$	**79.63**$_{\pm 0.94}$	**70.33**$_{\pm 0.38}$	**84.61**$_{\pm 0.68}$	**72.15**$_{\pm 0.47}$	**73.78**$_{\pm 0.50}$	**72.96**$_{\pm 0.39}$

we assess both plasticity and stability. For code, plasticity is measured by HumanEval [81] and stability by the four out-of-domain benchmarks PIQA [82], ARC-C [83], MMLU [84], and GSM8K [85]; for math, GSM8K and HumanEval swap roles. We report Plasticity, Stability (four-benchmark mean), and Overall (their arithmetic mean); see Appendix D.1.1 for partitioning.

Implementation. Base models are Llama-3.2-3B [86], Gemma-3-4B-PT [87], and Qwen-3.5-9B-Base [88]. Training uses LlamaFactory [89] with AdamW and cosine decay (Magicoder: 3 epochs, batch 64; MetaMathQA: 2 epochs, batch 128). DualSFT uses $\lambda = 0.8$, $\tau = 1.0$, a 1-epoch warmup with fraction $r_{\text{warm}} = 5\%$, and budgets $b = 10\%$ for data and $k = 5\%$ for parameters. For scoring, $\eta_{\text{sc}} = \eta_{\text{ft}}/N_{\text{sc}}$, with η_{ft} reported in Appendix D.1.2; budget-constrained baselines use matched budgets within their comparison groups. We report 3-seed means and stds. Appendix D.1.2 details hyperparameters, separating adaptation budgets from scoring-only validation/anchors.

6.2 Main Results

Matched-Budget Target-Task Performance. Table 1 shows that DualSFT improves target-task performance within each variant's matched-budget comparison group. On Llama-3.2-3B with Magi-



(a) Parameter selection efficiency frontier. (b) Data selection efficiency frontier.

Figure 2: Efficiency frontiers comparing against various baselines on Magicoder with Llama-3.2-3B.

coder, single-axis variants improve over their counterparts: DualSFT-Param raises HumanEval over NanoAdam from 43.50% to 44.51%, and DualSFT-Data raises HumanEval over LESS from 42.68% to 44.11%. In the joint-constrained regime, where both axes are restricted, full DualSFT reaches 43.70% HumanEval, 4.27 points above SPICE+NanoAdam under the same joint budget.

Retention and Overall Performance. Single-axis DualSFT-Param/Data lead their groups in Overall; full DualSFT is compared to joint-budget hybrids. It keeps Stability Avg. near pre-trained references, maintains strong HumanEval, and leads joint-budget methods in Stability Avg. and Overall, outperforming the strongest sequential hybrid by 2.46, 2.59, and 3.22 points on Llama/Gemma/Qwen. Relative to the pre-trained baseline, DualSFT lowers Stability Avg. by 0.35/0.57/1.14 points, versus 5.65/6.44/6.21 for Standard SFT, indicating less forgetting. Appendix D.2 corroborates this advantage. Appendix D.3 details MetaMathQA trends. Appendix D.4 evaluates Alpaca-GPT4, where DualSFT-Param/Data and full DualSFT improve instruction-tuning Overall over matched-budget baselines by 1.72/1.69/2.13 points, supporting applicability beyond code and mathematics. Sections 4.3 and 6.4 examine coordination gaps and one-shot selection.

Efficiency and Trade-offs. Fig. 2 compares Overall against time and peak memory. For parameters, DualSFT-Param achieves the highest Overall with the lowest peak fine-tuning memory (19.3 GB) among parameter-selection methods, despite not being the fastest. On the data and joint side, DualSFT-Data attains the best Overall, and full DualSFT maintains 47.12% with 28.5 GB peak selection memory and 7.4 GPU-hours. Under the same joint budget, full DualSFT improves Overall over sequential hybrids, trading modest selection cost for stronger stability-plasticity outcomes.

6.3 Ablation & Sensitivity Analysis

We examine component contributions, the source of parameter-selection gains relative to Fisher-based scoring, and sensitivity to hyperparameters and the validation set.

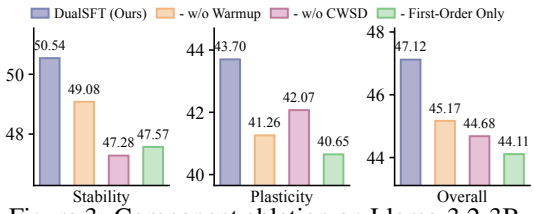


Figure 3: Component ablation on Llama-3.2-3B.

Component Ablation. Fig. 3 shows that removing any component lowers both Stability and Plasticity under joint data–parameter restrictions. Removing warmup reduces Plasticity from 43.70% to 41.26%, with Overall decreasing by 1.95 points to 45.17%. Omitting CWSD causes the largest Stability drop (50.54% → 47.28%) and lowers Plasticity to 42.07%, supporting preservation-guided selection on both axes. First-order-only scoring reduces Stability and Plasticity by 2.97 and 3.05 points, respectively, producing the largest Overall drop (47.12% → 44.11%) and highlighting the diagonal correction's contribution to both retention and adaptation. Appendix D.5 further compares preservation directions and anchor sizes.

Comparison with Fisher-Based Parameter Selection. Table 2 compares parameter-selection methods using full training data and a 5% parameter budget. DualSFT-Param exceeds FISH Mask [90] by 1.79 Overall points. With checkpoint, scoring pool, learning rate, and training protocol matched, replacing magnitude-only MB-Fisher-Full with DualSFT-Param-FO improves Overall by 1.04 points; adding the diagonal correction yields 0.64. Results support the contribution of validation-and-preservation-guided first-order scoring beyond Fisher-style saliency, with a complementary benefit from the diagonal correction. Appendix D.10 defines the controls and details the protocol.

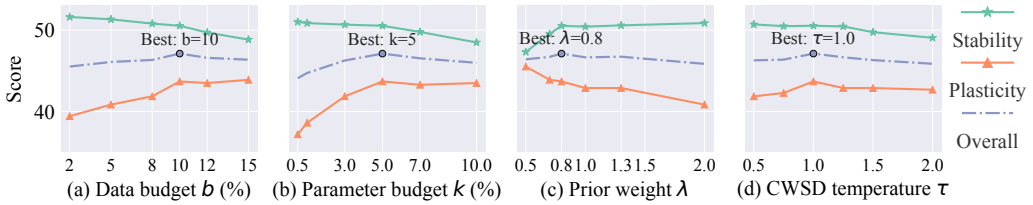


Figure 4: Sensitivity to key hyperparameters on Llama-3.2-3B (Magicoder).

Table 2: Parameter-Only Fisher Controls.

Method	Stability ↑	Plasticity ↑	Overall ↑
FISH Mask	$47.79_{\pm 0.44}$	$42.89_{\pm 0.76}$	$45.34_{\pm 0.60}$
EF-W	$48.14_{\pm 0.43}$	$43.09_{\pm 0.29}$	$45.61_{\pm 0.35}$
EF-W+LR	$47.64_{\pm 0.26}$	$42.68_{\pm 0.50}$	$45.16_{\pm 0.33}$
MB-Fisher-1K	$47.72_{\pm 0.25}$	$42.68_{\pm 0.50}$	$45.20_{\pm 0.33}$
MB-Fisher-Full	$47.61_{\pm 0.23}$	$43.29_{\pm 0.50}$	$45.45_{\pm 0.26}$
DualSFT-Param-FO	$49.07_{\pm 0.39}$	$43.90_{\pm 0.50}$	$46.49_{\pm 0.38}$
DualSFT-Param	$\mathbf{49.76}_{\pm 0.50}$	$\mathbf{44.51}_{\pm 0.86}$	$\mathbf{47.13}_{\pm 0.42}$

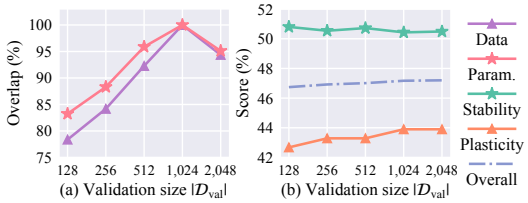


Figure 5: Validation-size sensitivity: overlap (left) and performance (right).

Hyperparameter Sensitivity and Robustness. Fig. 4 shows that Overall peaks at $b = 10\%$ and $k = 5\%$, reflecting the stability–plasticity trade-off as adaptation capacity grows. Performance remains near-optimal across $\lambda = 0.7$–1.3, with the best settings at $\lambda = 0.8$ and $\tau = 1.0$. Fig. 5 examines validation-set size: across 512–2,048 examples, data and parameter overlaps exceed 90% relative to the 1,024-example reference, while Overall ranges from 47.01 to 47.20. Doubling the reference size improves Overall by only 0.03 points. Appendix D.11 details the protocol and additional composition and sampling tests, which yield Overall spreads of 0.19 and 0.21 points at size 1,024. These results support this validation size as a reasonable cost–stability operating point under the tested conditions.

6.4 Diagnostic Analysis

Unless specified, all diagnostics in this section are conducted on Llama-3.2-3B with Magicoder.

Same-Surrogate Local-Regret Diagnostic. Table 3 examines whether one-shot scores lose aware utility once the other-side restriction is imposed under the shared surrogate $\mathbf{u}$. Jacc., Ratio, and Reg. denote the aware-unaware overlap, the aware-utility ratio, and the relative drop. We compare $\langle \mathbf{u} \odot \mathbf{m}_\theta, \mathbf{g}_n \rangle$ vs. $\langle \mathbf{u}, \mathbf{g}_n \rangle$ for data, and $u_d G_{S_\mathcal{D},d}$ vs. $u_d G_d$ for parameters. At warmup the regret is bounded (5.2% on data, 3.6% on parameters) with Jaccard 0.62–0.68; the smaller init regret is consistent with noisier early gradients. Together with the fixed-side re-selection results below, this is consistent with one-shot dual scoring acting as low-cost ex-ante coordination. Appendix D.6 gives definitions and protocol.

Table 3: Same-surrogate local-regret diagnostic.

Side	Ckpt.	Jacc.	Ratio	Reg. (%)
Data ($b = 10\%$)	Init	0.71	0.972	2.8
	Warmup	0.62	0.948	5.2
Param ($k = 5\%$)	Init	0.76	0.981	1.9
	Warmup	0.68	0.964	3.6

Table 4: Conditional re-selection.

Method	Stab.	Plas.	Overall	Re. Sel.	Re. E2E
DualSFT	50.54	43.70	47.12	1.00×	1.00×
D. → P.	50.48	43.90	47.19	1.06×	1.05×
P. → D.	50.61	43.90	47.26	1.45×	1.41×

Table 5: Selected data statistics.

Data statistic	Sel.	Rand.	Top-loss	Pool
Val-grad alignment	0.42	0.18	0.31	0.15
Avg. training loss	1.45	1.39	2.83	1.38
Diversity score	0.82	0.84	0.71	0.86

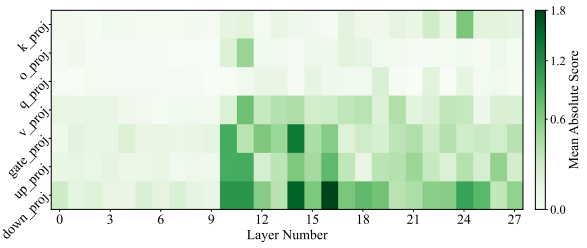


Figure 6: Layer-wise parameter selection on Llama-3.2-3B (Magicoder). Cells show mean absolute block scores. Appendix D.7 reports exact aggregates.

Conditional Re-selection Analysis. Table 4 compares DualSFT with fixed-side re-selection (D./P.: data/parameter; Rel. Sel./Rel. E2E: relative selection/E2E costs). Re-selection improves Overall by at

Table 6: Score fidelity diagnostic.

Setting	Data		Param	
	ρ	Pair	ρ	Pair
Init + First-Order	0.28	58.6	0.21	56.3
Init + Diag-Second-Order	0.37	64.1	0.30	60.7
Warmup + First-Order	0.56	73.1	0.49	69.1
Warmup + Diag-Second-Order	**0.82**	**79.8**	**0.79**	**75.3**

Table 7: Diagonal-vs-full rank agreement.

Checkpoint	Data			Param		
	ρ	Pair	Top-5%	ρ	Pair	Top-5%
Init	0.76	83.1	78.2	0.82	87.0	83.7
Warmup	**0.86**	**89.8**	**86.5**	**0.90**	**92.6**	**89.4**

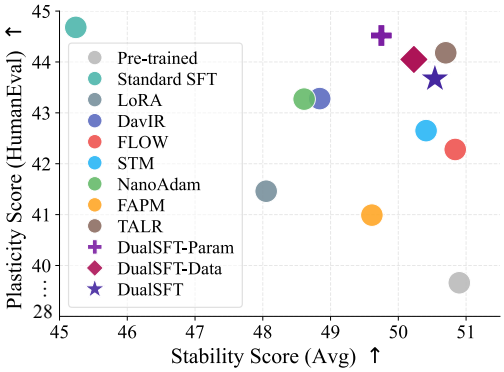


Figure 7: Forgetting-aware baseline comparison on Magicoder using Llama-3.2-3B.

most 0.14, while increasing selection and E2E costs by 6%–45% and 5%–41%. Results support one-shot selection as a cost-effective choice under evaluated budgets. The cost asymmetry follows from granularity: D. $\rightarrow$ P. aggregates selected-example gradients, whereas P. $\rightarrow$ D. rescores candidates under a fixed mask. Appendix D.8 reports the protocol and cost breakdowns.

Analysis of Selected Data and Parameters. Table 5 shows DualSFT-selected data have higher validation-gradient alignment than random, top-loss, and pool baselines, maintaining pool-like loss and diversity, proving selection avoids mere high loss or redundancy. Fig. 6 reveals concentrated parameter scores: top five layers contain 42.4% of selected parameters (2.38× enrichment), whereas early layers (L0 to L9) contain 8.2% (0.23×). Module shares closely match the pool (0.91× attention, 1.03× FFN), except over-selected FFN down_proj (1.44×).

Score Fidelity and Diagonal Approximation. We compare predicted and realized rankings of data subsets and budget-matched parameter masks under the same validation objective and selection budgets. Table 6 reports Init/Warmup × First/Diag-Second-Order variants averaged over $K \in \{50, 200, 500\}$ AdamW steps, using Spearman ρ and pairwise agreement; warmup and diagonal second-order scoring improve fidelity on both axes. Appendix D.9 reports declining score fidelity over longer optimization horizons. Table 7 compares diagonal scores with slice-level full-Hessian rankings, supporting the optimizer-aware diagonal proxy as a local ranking approximation.

6.5 Comparison with Forgetting-Specific Baselines

We compare DualSFT with some forgetting-specific baselines, including DavIR [79], FLOW [11], STM [31], NanoAdam [21], FAPM [33], and TALR [35]. We follow the main evaluation protocol on Magicoder with three seeds, reporting means and standard deviations over these three runs. Baselines retain their respective adaptation settings; DualSFT-Param/Data use 5% parameter/10% data budgets, respectively, and full DualSFT applies both. Fig. 7 shows that, on Llama-3.2-3B, DualSFT-Param leads the compared forgetting-aware methods in HumanEval (44.51%), while full DualSFT achieves the highest Stability Avg. among its variants (50.54%). Across backbones, the best DualSFT variant trails TALR in mean Overall by only 0.16–0.46 points (Table 20), indicating competitive stability–plasticity performance under restricted adaptation budgets.

7 Conclusion

DualSFT casts parameter and data selection as two views of a shared validation-improvement surrogate, co-extracting both signals. Under first-order and diagonal/full second-order approximations, they become row-wise and column-wise aggregations of a gradient interaction matrix, with a Shapley interpretation in the localized surrogate game. This turns two historically separate procedures into an efficient one-shot dual-scoring algorithm, reducing trainable parameters and training data in restricted fine-tuning. Experiments on 3B–9B LLMs show favorable stability–plasticity–efficiency trade-offs under matched budgets. DualSFT suggests treating parameters and data as coupled adaptation resources in future work, including longer horizons and post-training.

Acknowledgments and Disclosures

This work was supported by the National Natural Science Foundation of China under Grant Nos. 42550187 and 62476191, with computing resources provided by our research group. The authors declare no competing interests.

References

[1] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, R. Sauvestre, T. Remez *et al.*, "Code llama: Open foundation models for code," *arXiv preprint arXiv:2308.12950*, 2023.

[2] Z. Azerbayev, H. Schoelkopf, K. Paster, M. D. Santos, S. M. McAleer, A. Q. Jiang, J. Deng, S. Biderman, and S. Welleck, "Llemma: An open language model for mathematics," in *The Twelfth International Conference on Learning Representations*, 2024.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[5] L. Xu, H. Xie, S. J. Qin, X. Tao, and F. L. Wang, "Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[6] X. Yang, J. Y. Huang, W. Zhou, and M. Chen, "Parameter-efficient tuning with special token adaptation," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 865–872.

[7] J. Zhao, Z. Zhang, B. Chen, Z. Wang, A. Anandkumar, and Y. Tian, "Galore: Memory-efficient LLM training by gradient low-rank projection," in *Forty-first International Conference on Machine Learning*, 2024.

[8] S. yang Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "DoRA: Weight-decomposed low-rank adaptation," in *Forty-first International Conference on Machine Learning*, 2024.

[9] V. Lingam, A. Tejaswi, A. Vavre, A. Shetty, G. K. Gudur, J. Ghosh, A. Dimakis, E. Choi, A. Bojchevski, and S. Sanghavi, "Svft: Parameter-efficient fine-tuning with singular vectors," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 41 425–41 446.

[10] M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen, "LESS: Selecting influential data for targeted instruction tuning," in *Forty-first International Conference on Machine Learning*, 2024.

[11] S. Sanyal, H. Prairie, R. Das, A. Kavis, and S. Sanghavi, "Upweighting easy samples in fine-tuning mitigates forgetting," in *Forty-second International Conference on Machine Learning*, 2025.

[12] H. Xu, W. Shen, Y. Wei, A. Wang, G. Runfan, T. Wang, Y. Li, M. Li, and W. Jia, "Skrull: Towards efficient long context fine-tuning through dynamic data scheduling," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[13] Z. Jiang, J. Han, T. Li, X. Wang, S. Jiang, X. Meng, J. Wei, J. Liang, and Y. Xiao, "Difficulty is not enough: Curriculum learning for llms fine-tuning must consider utility," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 37, 2026, pp. 31 365–31 373.

[14] I. Agarwal, K. Killamsetty, L. Popa, and M. Danilevsky, "DELIFT: Data efficient language model instruction fine-tuning," in *The Thirteenth International Conference on Learning Representations*, 2025.

[15] A. Panigrahi, N. Saunshi, H. Zhao, and S. Arora, "Task-specific skill localization in fine-tuned language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 23–29 Jul 2023, pp. 27 011–27 033.

[16] W. Song, Z. Li, L. Zhang, H. Zhao, and B. Du, "Sparse is enough in fine-tuning pre-trained large language models," in *Forty-first International Conference on Machine Learning*, 2024.

[17] X. Yang, J. Leng, G. Guo, J. Zhao, R. Nakada, L. Zhang, H. Yao, and B. Chen, "S^2FT: Efficient, scalable and generalizable LLM fine-tuning by structured sparsity," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[18] R. Pan, X. Liu, S. Diao, R. Pi, J. Zhang, C. Han, and T. Zhang, "LISA: Layerwise importance sampling for memory-efficient large language model fine-tuning," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[19] H. He, J. B. Li, X. Jiang, and H. Miller, "SMT: Fine-tuning large language models with sparse matrices," in *The Thirteenth International Conference on Learning Representations*, 2025.

[20] Z. Liu, T. Pang, O. Balabanov, C. Yang, T. Huang, L. Yin, Y. Yang, and S. Liu, "LIFT the veil for the truth: Principal weights emerge after rank reduction for reasoning-focused supervised fine-tuning," in *Forty-second International Conference on Machine Learning*, 2025.

[21] C. Zhou, T. Jacobs, A. Gadhikar, and R. Burkholz, "Pay attention to small weights," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[22] K. Yao, Z. Song, K. Wu, M. Zhong, D. Cheng, Z. Tan, Y. Ji, and P. Gao, "Gast: Gradient-aligned sparse tuning of large language models with data-layer selection," in *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2026, pp. 4401–4416.

[23] R. Deb, K. K. Thekumparampil, K. Kalantari, G. Hiranandani, S. Sabach, and B. Kveton, "FisherSFT: Data-efficient supervised fine-tuning of language models using information gain," in *Forty-second International Conference on Machine Learning*, 2025.

[24] X. Liang, L. Yang, J. Wang, Y. Lu, R. Wu, H. Chen, and J. HAO, "Boosting multi-domain fine-tuning of large language models through evolving interactions between samples," in *Forty-second International Conference on Machine Learning*, 2025.

[25] G. Min, T. Huang, K. Wan, and C. Chen, "Gist: Targeted data selection for instruction tuning via coupled optimization geometry," *arXiv preprint arXiv:2602.18584*, 2026.

[26] P. Chang, J. Zhang, B. Chen, C. Wang, C. Guo, Y. Zhang, Y. Gao, J. Xiang, Y. Gao, C. Sun, Y. Chen, and D. Kong, "SPICE: Submodular penalized information–conflict selection for efficient large language model training," in *The Fourteenth International Conference on Learning Representations*, 2026.

[27] B. Zhao, H. Tu, C. Wei, J. Mei, and C. Xie, "Tuning layernorm in attention: Towards efficient multi-modal LLM finetuning," in *The Twelfth International Conference on Learning Representations*, 2024.

[28] S. Khaki, X. Li, J. Guo, L. Zhu, K. N. Plataniotis, A. Yazdanbakhsh, K. Keutzer, S. Han, and Z. Liu, "SparseloRA: Accelerating LLM fine-tuning with contextual sparsity," in *Forty-second International Conference on Machine Learning*, 2025.

[29] S. Luo, H. Yang, Y. Xin, M. Yi, G. Wu, G. Zhai, and X. Liu, "Tr-pts: Task-relevant parameter and token selection for efficient tuning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4360–4369.

[30] D. McCoy, Y. Wu, and Z. Butzin-Dozier, "AI progress should be measured by capability-per-resource, not scale alone: A framework for gradient-guided resource allocation in LLMs," in *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025.

[31] C.-C. Wu, Z. R. Tam, C.-Y. Lin, Y.-N. Chen, S.-H. Sun, and H. yi Lee, "Mitigating forgetting in LLM fine-tuning via low-perplexity token learning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[32] Y. Wang, A. Bai, N. Peng, and C.-J. Hsieh, "On the loss of context awareness in general instruction fine-tuning," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[33] W. Huang, A. Cheng, and Y. Wang, "Mitigating catastrophic forgetting in large language models with forgetting-aware pruning," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 21 853–21 867.

[34] J. Harmon, A. Hochlehnert, M. Bethge, and A. Prabhu, "Mapping post-training forgetting in language models at scale," in *The Fourteenth International Conference on Learning Representations*, 2026.

[35] J. Lin, Z. Wang, K. Qian, T. Wang, A. Srinivasan, H. Zeng, R. Jiao, X. Zhou, J. Gesi, D. Wang, Y. Guo, K. Zhong, W. Zhang, S. Sanghavi, C. Chen, H. Yun, and L. Li, "SFT doesn't always hurt general capabilities: Revisiting domain-specific fine-tuning in LLMs," in *The Fourteenth International Conference on Learning Representations*, 2026.

[36] J. Zhang, O. Wu, and Y. Du, "Nads: Navigator-guided data selection for mitigating catastrophic forgetting in fine-tuning," 2026.

[37] Y. Wei, Z. Wang, J. Liu, Y. Ding, and L. ZHANG, "Magicoder: Empowering code generation with OSS-instruct," in *Forty-first International Conference on Machine Learning*, 2024.

[38] L. Yu, W. Jiang, H. Shi, J. YU, Z. Liu, Y. Zhang, J. Kwok, Z. Li, A. Weller, and W. Liu, "Metamath: Bootstrap your own mathematical questions for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[39] D. Miao, Y. Liu, J. Wang, C. Sun, Y. Zhang, D. Yan, S. Dong, Q. Zhang, and Y. Wu, "Taso: Task-aligned sparse optimization for parameter-efficient model adaptation," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 22 746–22 758.

[40] X. Wang, L. Gao, H. Wang, Y. Zhang, and J. Zhao, "Floe: Fisher-based layer selection for efficient sparse adaptation of low-rank experts," *arXiv preprint arXiv:2506.00495*, 2025.

[41] X. Chu, X. Jiang, R. Qiu, J. Gao, and J. Zhao, "MODEL SHAPLEY: Find your ideal parameter player via one gradient backpropagation," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[42] C. Zhao, Q. Yao, X. Song, and W. Zhu, "ShaploRA: Allocation of low-rank adaption on large language models via shapley value inspired importance estimation," in *The Third Conference on Parsimony and Learning (Proceedings Track)*, 2026.

[43] X. Ding, P. Tong, R. Duan, Y. Zhang, R. Sun, and Y. Zhu, "Pruning as a cooperative game: Surrogate-assisted layer contribution estimation for large language models," in *The Fourteenth International Conference on Learning Representations*, 2026.

[44] J. Pang, N. Di, Z. Zhu, J. Wei, H. Cheng, C. Qian, and Y. Liu, "Token cleaning: Fine-grained data selection for LLM supervised fine-tuning," in *Forty-second International Conference on Machine Learning*, 2025.

[45] X. Qin, X. Wang, N. Liao, C. Zhang, X. Zhang, M. Feng, J. Wang, and J. Yan, "sstoken: Self-modulated and semantic-aware token selection for LLM fine-tuning," in *The Fourteenth International Conference on Learning Representations*, 2026.

[46] A. Jain, A. Montanari, and E. Sasoglu, "Train on validation (tov): Fast data selection with applications to fine-tuning," in *The Fourteenth International Conference on Learning Representations*, 2026.

[47] Z. Yu, S. Das, and C. Xiong, "MATES: Model-aware data selection for efficient pretraining with data influence models," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[48] Z. Liu, Y. Liu, S. Wang, Z. Song, J. Wang, J. Liu, Q. Liu, G. Chen, and Y. Wang, "Learn more, forget less: A gradient-aware data selection approach for llm," *Signal Processing*, p. 110611, 2026.

[49] H. Shen, P.-Y. Chen, P. Das, and T. Chen, "SEAL: Safety-enhanced aligned LLM fine-tuning via bilevel data selection," in *The Thirteenth International Conference on Learning Representations*, 2025.

[50] J. T. Wang, P. Mittal, D. Song, and R. Jia, "Data shapley in one training run," in *The Thirteenth International Conference on Learning Representations*, 2025.

[51] Y. Yan, M. Zhong, Q. Zhu, X. Gu, J. Chen, and H. Li, "CoIDO: Efficient data selection for visual instruction tuning via coupled importance-diversity optimization," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[52] Z. Ling, D. Chen, L. Yao, Q. Shen, Y. Li, and Y. Shen, "Diversity as a reward: Fine-tuning LLMs on a mixture of domain-undetermined data," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[53] A. M. Das, G. Bhatt, S. Verma, Y. Wang, V. V. Muppirala, and J. Bilmes, "Matched data, better models: Target aligned data filtering with sparse autoencoders," in *The Fourteenth International Conference on Learning Representations*, 2026.

[54] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1885–1894.

[55] H. Askari, S. Gupta, F. Wang, A. Chhabra, and M. Chen, "LayerIF: Estimating layer quality for large language models using influence functions," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[56] D. Biderman, J. Portes, J. J. G. Ortiz, M. Paul, P. Greengard, C. Jennings, D. King, S. Havens, V. Chiley, J. Frankle, C. Blakeney, and J. P. Cunningham, "LoRA learns less and forgets less," *Transactions on Machine Learning Research*, 2024.

[57] Y. Yang, X. Li, Z. Zhou, S. L. Song, J. Wu, L. Nie, and B. Ghanem, "CorDA: Context-oriented decomposition adaptation of large language models for task-aware parameter-efficient fine-tuning," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[58] M. Luo, F. Kuang, Y. Wang, Z. Liu, and T. He, "Sc-lora: Balancing efficient fine-tuning and knowledge preservation via subspace-constrained lora," *arXiv preprint arXiv:2505.23724*, 2025.

[59] Y. Chen, S. Wang, Y. Zhang, Z. Lin, H. Zhang, W. Sun, T. Ding, and R. Sun, "MoFO: Momentum-filtered optimizer for mitigating forgetting in LLM fine-tuning," *Transactions on Machine Learning Research*, 2025, j2C Certification.

[60] R. Wang, P. Ping, Z. Guo, X. Zhang, Q. Shi, L. Zhou, and T. Ji, "Loki: Low-damage knowledge implanting of large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 39, 2026, pp. 33 620–33 628.

[61] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[62] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning," in *International conference on machine learning*. PMLR, 2018, pp. 1568–1577.

[63] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *International Conference on Learning Representations*, 2019.

[64] K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, and R. Iyer, "Glister: Generalization based data subset selection for efficient and robust learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 9, 2021, pp. 8110–8118.

[65] Y. Yu, K. Han, H. Zhou, Y. Tang, K. Huang, Y. Wang, and D. Tao, "LLM data selection and utilization via dynamic bi-level optimization," in *Forty-second International Conference on Machine Learning*, 2025.

[66] Y. Liao, Z. Li, B. Wu, Z. Cheng, S. Zhao, S. Chen, and W. ZHANG, "Compress large language models via collaboration between learning and matrix approximation," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[67] R. Shirkavand, P. Yu, Q. He, and H. Huang, "Bilevel ZOFO: Efficient LLM fine-tuning and meta-training," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[68] L. Jiang, Q. Xiao, L. Chen, and T. Chen, "Beyond value functions: Single-loop bilevel optimization under flatness conditions," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[69] J. T. Wang, T. Wu, D. Song, P. Mittal, and R. Jia, "GREATS: Online selection of high-quality data for LLM training in every iteration," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[70] S. Chen, Y. Qi, M. Ai, Y. Sun, R. Qiu, J. Zou, and J. He, "Influence-preserving proxies for gradient-based data selection in LLM finetuning," in *The Fourteenth International Conference on Learning Representations*, 2026.

[71] M. Nikdan, V. Cohen-Addad, D. Alistarh, and V. Mirrokni, "Efficient data selection at scale via influence distillation," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[72] Y. Zhao, S. Dang, H. Ye, G. Dai, Y. Qian, and I. Tsang, "Second-order fine-tuning without pain for LLMs: A hessian informed zeroth-order optimizer," in *The Thirteenth International Conference on Learning Representations*, 2025.

[73] H. Zhao, S. Ding, C. Fang, and Z. Lin, "PaZO: Preconditioned accelerated zeroth-order optimization for fine-tuning LLMs," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[74] J. Sliwa, F. Schneider, P. Hennig, and J. M. Hernández-Lobato, "Mitigating forgetting in low rank adaptation," in *Second Workshop on Test-Time Adaptation: Putting Updates to the Test! at ICML 2025*, 2025.

[75] Y. Zhang, C. Chen, Z. Li, T. Ding, C. Wu, D. P. Kingma, Y. Ye, Z.-Q. Luo, and R. Sun, "Adam-mini: Use fewer learning rates to gain more," in *The Thirteenth International Conference on Learning Representations*, 2025.

[76] D. S. Kalra, J. Kirchenbauer, M. Barkeshli, and T. Goldstein, "When can you get away with low memory adam?" in *High-dimensional Learning Dynamics 2025*, 2025.

[77] T. Robert, M. Safaryan, I.-V. Modoranu, and D. Alistarh, "LDAdam: Adaptive optimization from low-dimensional gradient statistics," in *The Thirteenth International Conference on Learning Representations*, 2025.

[78] N. Vyas, D. Morwani, R. Zhao, I. Shapira, D. Brandfonbrener, L. Janson, and S. M. Kakade, "SOAP: Improving and stabilizing shampoo using adam for language modeling," in *The Thirteenth International Conference on Learning Representations*, 2025.

[79] H. Zhou, T. Liu, Q. Ma, Y. Zhang, J. Yuan, P. Liu, Y. You, and H. Yang, "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)*, 2025, pp. 9220–9237.

[80] X. Yuan, C. Shen, S. Yan, K. Liu, X. Zhang, S. Fan, L. Xie, W. Wang, R. Guan, Y. Wang, and J. Ye, "Differential fine-tuning large language models towards better diverse reasoning abilities," in *The Fourteenth International Conference on Learning Representations*, 2026.

[81] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[82] Y. Bisk, R. Zellers, R. Le Bras, J. Gao, and Y. Choi, "Piqa: Reasoning about physical commonsense in natural language," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 05, 2020, pp. 7432–7439.

[83] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv preprint arXiv:1803.05457*, 2018.

[84] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *International Conference on Learning Representations*, 2021.

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

[86] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[87] Gemma Team, "Gemma 3 technical report," 2025.

[88] Qwen Team, "Qwen3.5: Towards native multimodal agents," February 2026.

[89] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, and Z. Luo, "Llamafactory: Unified efficient fine-tuning of 100+ language models," in *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 3: system demonstrations)*, 2024, pp. 400–410.

[90] Y.-L. Sung, V. Nair, and C. Raffel, "Training neural networks with fixed sparse masks," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 24 193–24 205.

[91] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff *et al.*, "A framework for few-shot language model evaluation," *Zenodo*, 2021.

[92] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with gpt-4," 2023.

[93] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou, "Instruction-following evaluation for large language models," 2023.

A Detailed Theoretical Derivations and Shapley Proofs

This appendix provides four technical complements to the main text: the derivation of the shared one-step surrogate from exact bilevel problems, local truncation guarantees for the Taylor surrogate, detailed proofs of Theorem 1, and stability analysis of the practical diagonal scores used by DualSFT.

A.1 From Exact Inner Objectives to One-Step Surrogate Updates

We derive one-step surrogate updates in Eq. (3) from exact bilevel formulations in Eqs. (1)–(2). Following Section 3, we linearize the inner response at common checkpoint $\bar{\boldsymbol{\theta}}$ and use local gradients

$$\mathbf{g}_n \triangleq \nabla_{\boldsymbol{\theta}} \ell(x_n, y_n; \bar{\boldsymbol{\theta}}), \qquad \mathbf{G} \triangleq \sum_{n=1}^{N} \mathbf{g}_n. \tag{13}$$

For the parameter-side problem in Eq. (1), a single gradient step on the inner objective under the coordinate mask $\mathbf{m}_\theta$ gives the local update

$$\Delta\boldsymbol{\theta}(\mathbf{m}_\theta) \approx -\eta\, \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{train}}(\bar{\boldsymbol{\theta}}; \mathcal{D}_{\text{train}}) \odot \mathbf{m}_\theta = -\eta(\mathbf{G} \odot \mathbf{m}_\theta). \tag{14}$$

For the data-side problem in Eq. (2), the selected subset induces the masked empirical objective $\mathbf{1}^\top\big(\mathbf{m}_{\mathcal{D}} \odot \boldsymbol{\ell}_{\text{train}}(\boldsymbol{\theta})\big)$. A single gradient step at $\bar{\boldsymbol{\theta}}$ gives

$$\Delta\boldsymbol{\theta}(\mathbf{m}_{\mathcal{D}}) \approx -\eta\, \nabla_{\boldsymbol{\theta}} \big[\mathbf{1}^\top\big(\mathbf{m}_{\mathcal{D}} \odot \boldsymbol{\ell}_{\text{train}}(\bar{\boldsymbol{\theta}})\big)\big] = -\eta \sum_{n=1}^{N} (\mathbf{m}_{\mathcal{D}})_n \mathbf{g}_n. \tag{15}$$

Identifying each binary mask with its selected index set gives the localized functions in Eq. (4). These one-step surrogates underlie the Taylor analysis in Eq. (5) and shared-score derivation in Section 4.

A.2 Local Truncation Error

We now justify the first- and second-order truncations of Eq. (5). For either the parameter-side or data-side utility, define

$$U_{\text{step}}(\mathbf{S}) \triangleq \mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}}) - \mathcal{L}_{\text{val}}\big(\bar{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}(\mathbf{S})\big), \qquad \Delta\boldsymbol{\theta}(\mathbf{S}) = -\eta\, \mathbf{g}(\mathbf{S}). \tag{16}$$

Proposition 2 (Local truncation error)**.** *Assume $\mathcal{L}_{val}$ is β-smooth in a neighborhood of $\bar{\boldsymbol{\theta}}$. Then the first-order truncation*

$$U_{1\text{st}}(\mathbf{S}) = -\mathbf{v}_{val}^\top \Delta\boldsymbol{\theta}(\mathbf{S}) \tag{17}$$

satisfies

$$\big|U_{\text{step}}(\mathbf{S}) - U_{1\text{st}}(\mathbf{S})\big| \le \frac{\beta}{2}\eta^2 \|\mathbf{g}(\mathbf{S})\|_2^2. \tag{18}$$

If, in addition, $\nabla^2 \mathcal{L}_{val}$ is ρ-Lipschitz on a neighborhood containing the segment between $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\theta}} + \Delta\boldsymbol{\theta}(\mathbf{S})$, then the second-order truncation

$$U_{2\text{nd}}(\mathbf{S}) = -\mathbf{v}_{val}^\top \Delta\boldsymbol{\theta}(\mathbf{S}) - \frac{1}{2}\Delta\boldsymbol{\theta}(\mathbf{S})^\top \mathbf{H} \Delta\boldsymbol{\theta}(\mathbf{S}) \tag{19}$$

satisfies

$$\big|U_{\text{step}}(\mathbf{S}) - U_{2\text{nd}}(\mathbf{S})\big| \le \frac{\rho}{6}\eta^3 \|\mathbf{g}(\mathbf{S})\|_2^3. \tag{20}$$

Proof. By β-smoothness of $\mathcal{L}_{\text{val}}$, for any Δ in a neighborhood of $\bar{\boldsymbol{\theta}}$,

$$\big|\mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}} + \Delta) - \mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}}) - \nabla\mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}})^\top \Delta\big| \le \frac{\beta}{2}\|\Delta\|_2^2. \tag{21}$$

Substituting $\Delta = \Delta\boldsymbol{\theta}(\mathbf{S})$ gives

$$\big|U_{\text{step}}(\mathbf{S}) - U_{1\text{st}}(\mathbf{S})\big| \le \frac{\beta}{2}\eta^2 \|\mathbf{g}(\mathbf{S})\|_2^2. \tag{22}$$

If $\nabla^2\mathcal{L}_{\text{val}}$ is ρ-Lipschitz, then

$$\left|\mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}} + \Delta) - \mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}}) - \nabla\mathcal{L}_{\text{val}}(\bar{\boldsymbol{\theta}})^\top \Delta - \frac{1}{2}\Delta^\top \mathbf{H} \Delta\right| \le \frac{\rho}{6}\|\Delta\|_2^3. \tag{23}$$

Substituting $\Delta = \Delta\boldsymbol{\theta}(\mathbf{S})$ gives

$$\big|U_{\text{step}}(\mathbf{S}) - U_{2\text{nd}}(\mathbf{S})\big| \le \frac{\rho}{6}\eta^3 \|\mathbf{g}(\mathbf{S})\|_2^3. \tag{24}$$

□

A.3 Analytical Shapley Values under Approximation Hierarchies

We now prove Theorem 1. For each $\alpha \in \{1, \text{diag}, \text{full}\}$, let

$$\mathcal{G}_\theta^{(\alpha)} = ([D], U_\theta^{(\alpha)}), \qquad \mathcal{G}_\mathcal{D}^{(\alpha)} = ([N], U_\mathcal{D}^{(\alpha)}) \tag{25}$$

be the localized surrogate games defined in Section 4, and write

$$\phi_{\theta,d}^{(\alpha)} \triangleq \phi_d(\mathcal{G}_\theta^{(\alpha)}), \qquad \phi_{\mathcal{D},n}^{(\alpha)} \triangleq \phi_n(\mathcal{G}_\mathcal{D}^{(\alpha)}). \tag{26}$$

For a finite player set $\mathcal{P}$, utility function $v : 2^\mathcal{P} \to \mathbb{R}$, and player $i \in \mathcal{P}$, the Shapley value is

$$\phi_i(v) = \sum_{S \subseteq \mathcal{P}\setminus\{i\}} \frac{|S|!(|\mathcal{P}| - |S| - 1)!}{|\mathcal{P}|!} \left[v(S \cup \{i\}) - v(S)\right]. \tag{27}$$

Equivalently, if π is a random permutation of $\mathcal{P}$ and P_i^π denotes players preceding i in π,

$$\phi_i(v) = \mathbb{E}_\pi\left[v(P_i^\pi \cup \{i\}) - v(P_i^\pi)\right]. \tag{28}$$

Lemma 1 (Half-Averaging under Permutations)**.** *Fix $i \in \mathcal{P}$. For any coefficients $\{a_j\}_{j\in\mathcal{P}\setminus\{i\}}$,*

$$\mathbb{E}_\pi\left[\sum_{j \in P_i^\pi} a_j\right] = \frac{1}{2}\sum_{j\neq i} a_j. \tag{29}$$

The same identity holds componentwise for vector-valued a_j.

Proof. For each fixed $j \neq i$, under a uniformly random permutation, the probability that j appears before i is $1/2$. Applying linearity of expectation gives the result. □

Proof of Theorem 1. Let $M_{n,d}^{(1)}$, $M_{n,d}^{(\text{diag})}$, and $M_{n,d}^{(\text{full})}$ be defined as in Eq. (6).

First-order case. Using Eq. (5), the first-order parameter-side utility is

$$U_\theta^{(1)}(S) = -\mathbf{v}_{\text{val}}^\top \Delta\boldsymbol{\theta}_\theta(S) = \sum_{d\in S}\sum_{n=1}^{N} M_{n,d}^{(1)}. \tag{30}$$

For any $S \subseteq [D] \setminus \{d\}$,

$$U_\theta^{(1)}(S \cup \{d\}) - U_\theta^{(1)}(S) = \sum_{n=1}^{N} M_{n,d}^{(1)} = \phi_{\theta,d}^{(1)} = s_{\theta,d}^{(1)}. \tag{31}$$

$$U_\mathcal{D}^{(1)}(S) = -\mathbf{v}_{\text{val}}^\top \Delta\boldsymbol{\theta}_\mathcal{D}(S) = \sum_{n\in S}\sum_{d=1}^{D} M_{n,d}^{(1)}. \tag{32}$$

$$U_\mathcal{D}^{(1)}(S \cup \{n\}) - U_\mathcal{D}^{(1)}(S) = \sum_{d=1}^{D} M_{n,d}^{(1)} = \phi_{\mathcal{D},n}^{(1)} = s_{\mathcal{D},n}^{(1)}. \tag{33}$$

Diagonal-curvature case. For the parameter-side game,

$$U_\theta^{(\text{diag})}(S) = \sum_{d\in S}\sum_{n=1}^{N} M_{n,d}^{(1)} - \frac{\eta^2}{2}\sum_{d\in S} c_d G_d^2, \tag{34}$$

$$U_\theta^{(\text{diag})}(S \cup \{d\}) - U_\theta^{(\text{diag})}(S) = \sum_{n=1}^{N} M_{n,d}^{(\text{diag})} = \phi_{\theta,d}^{(\text{diag})} = s_{\theta,d}^{(\text{diag})}. \tag{35}$$

For the data-side game,

$$U_\mathcal{D}^{(\text{diag})}(S) = \sum_{n\in S}\sum_{d=1}^{D} M_{n,d}^{(1)} - \frac{\eta^2}{2}\sum_{d=1}^{D} c_d \left(\sum_{m\in S} g_{m,d}\right)^2. \tag{36}$$

$$U_{\mathcal{D}}^{(\mathrm{diag})}(S\cup\{n\}) - U_{\mathcal{D}}^{(\mathrm{diag})}(S) = \sum_{d=1}^{D} M_{n,d}^{(1)} - \frac{\eta^2}{2}\sum_{d=1}^{D} c_d\left[g_{n,d}^2 + 2g_{n,d}\sum_{m\in S} g_{m,d}\right]. \tag{37}$$

Applying Eq. (28) and Lemma 1 coordinatewise replaces $\sum_{m\in S} g_{m,d}$ by $\frac{1}{2}\sum_{m\neq n} g_{m,d}$ under Shapley averaging, yielding

$$\phi_{\mathcal{D},n}^{(\mathrm{diag})} = \sum_{d=1}^{D}\left(M_{n,d}^{(1)} - \frac{\eta^2}{2}c_d G_d g_{n,d}\right) = \sum_{d=1}^{D} M_{n,d}^{(\mathrm{diag})} = s_{\mathcal{D},n}^{(\mathrm{diag})}. \tag{38}$$

Full-Hessian case. For the parameter-side game,

$$U_{\theta}^{(\mathrm{full})}(S) = \sum_{d\in S}\sum_{n=1}^{N} M_{n,d}^{(1)} - \frac{\eta^2}{2}\sum_{i\in S}\sum_{j\in S} G_i H_{ij} G_j. \tag{39}$$

For any $S \subseteq [D]\setminus\{d\}$, by symmetry of $\mathbf{H}$,

$$U_{\theta}^{(\mathrm{full})}(S\cup\{d\}) - U_{\theta}^{(\mathrm{full})}(S) = \sum_{n=1}^{N} M_{n,d}^{(1)} - \frac{\eta^2}{2}\left[G_d H_{dd} G_d + 2G_d\sum_{l\in S} H_{dl} G_l\right]. \tag{40}$$

Applying Eq. (28) and Lemma 1 with $a_l = H_{dl}G_l$ gives

$$\phi_{\theta,d}^{(\mathrm{full})} = \sum_{n=1}^{N} M_{n,d}^{(1)} - \frac{\eta^2}{2}(\mathbf{HG})_d G_d = \sum_{n=1}^{N} M_{n,d}^{(\mathrm{full})} = s_{\theta,d}^{(\mathrm{full})}. \tag{41}$$

For the data-side game,

$$U_{\mathcal{D}}^{(\mathrm{full})}(S) = \sum_{n\in S}\sum_{d=1}^{D} M_{n,d}^{(1)} - \frac{\eta^2}{2}\left(\sum_{m\in S}\mathbf{g}_m\right)^{\top}\mathbf{H}\left(\sum_{m\in S}\mathbf{g}_m\right). \tag{42}$$

$$U_{\mathcal{D}}^{(\mathrm{full})}(S\cup\{n\}) - U_{\mathcal{D}}^{(\mathrm{full})}(S) = \sum_{d=1}^{D} M_{n,d}^{(1)} - \frac{\eta^2}{2}\left[\mathbf{g}_n^{\top}\mathbf{H}\mathbf{g}_n + 2\mathbf{g}_n^{\top}\mathbf{H}\sum_{m\in S}\mathbf{g}_m\right]. \tag{43}$$

Applying Eq. (28) and Lemma 1 in vector form yields

$$\phi_{\mathcal{D},n}^{(\mathrm{full})} = \sum_{d=1}^{D}\left(M_{n,d}^{(1)} - \frac{\eta^2}{2}(\mathbf{HG})_d g_{n,d}\right) = \sum_{d=1}^{D} M_{n,d}^{(\mathrm{full})} = s_{\mathcal{D},n}^{(\mathrm{full})}. \tag{44}$$

This completes the proof. □

A.4 Proof of Proposition 1

Under the joint restrictions, the first-order update is $\Delta\boldsymbol{\theta}(S_{\mathcal{D}}, \mathbf{m}_\theta) = -\eta(\mathbf{m}_\theta \odot \mathbf{G}_{S_{\mathcal{D}}})$, where $\mathbf{G}_{S_{\mathcal{D}}} = \sum_{n\in S_{\mathcal{D}}}\mathbf{g}_n$ and $S_\theta = \{d\in[D] : (\mathbf{m}_\theta)_d = 1\}$. The corresponding first-order utility decomposes as

$$-\mathbf{v}_{\mathrm{val}}^{\top}\Delta\boldsymbol{\theta}(S_{\mathcal{D}}, \mathbf{m}_\theta) = \eta\sum\nolimits_{n\in S_{\mathcal{D}}}\langle \mathbf{v}_{\mathrm{val}}\odot\mathbf{m}_\theta, \mathbf{g}_n\rangle = \eta\sum\nolimits_{d\in S_\theta}(\mathbf{v}_{\mathrm{val}})_d G_{S_{\mathcal{D}},d}. \tag{45}$$

Thus, the mask-aware marginal scores compatible with the joint restricted update are $\bar{s}_{\mathcal{D},n}(\mathbf{m}_\theta) = \eta\langle \mathbf{v}_{\mathrm{val}}\odot\mathbf{m}_\theta, \mathbf{g}_n\rangle$ and $\bar{s}_{\theta,d}(S_{\mathcal{D}}) = \eta(\mathbf{v}_{\mathrm{val}})_d G_{S_{\mathcal{D}},d}$.

The isolated first-order scores are $s_{\mathcal{D},n}^{\mathrm{iso}} = \eta\langle \mathbf{v}_{\mathrm{val}}, \mathbf{g}_n\rangle$ and $s_{\theta,d}^{\mathrm{iso}} = \eta(\mathbf{v}_{\mathrm{val}})_d G_d$, where $\mathbf{G} = \sum_{n=1}^{N}\mathbf{g}_n$. Subtracting the mask-aware data score from the isolated data score gives

$$s_{\mathcal{D},n}^{\mathrm{iso}} - \bar{s}_{\mathcal{D},n}(\mathbf{m}_\theta) = \eta\langle \mathbf{v}_{\mathrm{val}}\odot(\mathbf{1}_D - \mathbf{m}_\theta), \mathbf{g}_n\rangle = \eta\sum\nolimits_{d:(\mathbf{m}_\theta)_d=0}(\mathbf{v}_{\mathrm{val}})_d g_{n,d}. \tag{46}$$

Similarly, subtracting the mask-aware parameter score from the isolated parameter score gives

$$s_{\theta,d}^{\mathrm{iso}} - \bar{s}_{\theta,d}(S_{\mathcal{D}}) = \eta(\mathbf{v}_{\mathrm{val}})_d(G_d - G_{S_{\mathcal{D}},d}) = \eta(\mathbf{v}_{\mathrm{val}})_d\sum\nolimits_{n\in[N]\setminus S_{\mathcal{D}}} g_{n,d}. \tag{47}$$

These two identities prove Proposition 1. They show that the data-side gap is supported on frozen coordinates, whereas the parameter-side gap is supported on samples excluded by the data selection. This characterizes the isolated-scoring gap, but it does not mean that DualSFT performs fully conditional re-scoring. DualSFT instead extracts both marginal scores once from the same local surrogate, leaving fixed-side re-selection as an optional refinement analyzed in Section 6.4.

A.5 Practical Diagonal Scores and Selection Stability

We analyze scores in Eqs. (10)–(11). Let $\widehat{D} \triangleq \mathrm{Diag}(\hat{\mathbf{c}})$, and write $\widehat{s}_{\theta,d}$ and $\widehat{s}_{\mathcal{D},n}$ for scores in Eq. (11). For comparison, write $s^{(\mathrm{full})}_{\theta,d}$ and $s^{(\mathrm{full})}_{\mathcal{D},n}$ for the local full-Hessian counterparts obtained from Eq. (7) with the step size instantiated as η_{sc}.

Proposition 3 (Diagonal-Proxy Score Perturbation)**.** *For parameter $d \in [D]$ and sample $n \in [N]$,*

$$\widehat{s}_{\theta,d} - s^{(\mathrm{full})}_{\theta,d} = \frac{\eta_{\mathrm{sc}}^2}{2}\big((\mathbf{H} - \widehat{D})\mathbf{G}\big)_d G_d, \qquad \widehat{s}_{\mathcal{D},n} - s^{(\mathrm{full})}_{\mathcal{D},n} = \frac{\eta_{\mathrm{sc}}^2}{2}\mathbf{g}_n^\top(\mathbf{H} - \widehat{D})\mathbf{G}. \tag{48}$$

Consequently,

$$\|\widehat{\mathbf{s}}_\theta - \mathbf{s}^{(\mathrm{full})}_\theta\|_\infty \le \frac{\eta_{\mathrm{sc}}^2}{2}\|\mathbf{H} - \widehat{D}\|_{\mathrm{op}}\|\mathbf{G}\|_2^2, \qquad \|\widehat{\mathbf{s}}_{\mathcal{D}} - \mathbf{s}^{(\mathrm{full})}_{\mathcal{D}}\|_\infty \le \frac{\eta_{\mathrm{sc}}^2}{2}B_g\|\mathbf{H} - \widehat{D}\|_{\mathrm{op}}\|\mathbf{G}\|_2 \tag{49}$$

if $\|\mathbf{g}_n\|_2 \le B_g$ for all n.

Proof. Subtracting the definitions of the diagonal scores and the full-Hessian scores gives the two identities directly. The ℓ_∞ bounds follow from Cauchy–Schwarz and the operator-norm inequality. □

Corollary 1 (Top-k parameter selection stability)**.** *Let $\widehat{s}_{\theta,(1)} \ge \cdots \ge \widehat{s}_{\theta,(D)}$ denote the sorted practical parameter scores, and define*

$$\Delta^\theta_k \triangleq \widehat{s}_{\theta,(k)} - \widehat{s}_{\theta,(k+1)}. \tag{50}$$

If

$$\Delta^\theta_k > \eta_{\mathrm{sc}}^2\|\mathbf{H} - \widehat{D}\|_{\mathrm{op}}\|\mathbf{G}\|_2^2, \tag{51}$$

then the Top-k parameter set induced by $\widehat{\mathbf{s}}_\theta$ is identical to that induced by $\mathbf{s}^{(\mathrm{full})}_\theta$.

Proof. By Proposition 3,

$$\|\widehat{\mathbf{s}}_\theta - \mathbf{s}^{(\mathrm{full})}_\theta\|_\infty \le \varepsilon_\theta \triangleq \frac{\eta_{\mathrm{sc}}^2}{2}\|\mathbf{H} - \widehat{D}\|_{\mathrm{op}}\|\mathbf{G}\|_2^2. \tag{52}$$

$\Delta^\theta_k > 2\varepsilon_\theta$ implies Top-k stability by the standard boundary-gap argument. □

Corollary 2 (Top-b data selection stability)**.** *Let $\widehat{s}_{\mathcal{D},(1)} \ge \cdots \ge \widehat{s}_{\mathcal{D},(N)}$ denote the sorted practical data scores, and define*

$$\Delta^{\mathcal{D}}_b \triangleq \widehat{s}_{\mathcal{D},(b)} - \widehat{s}_{\mathcal{D},(b+1)}. \tag{53}$$

If

$$\Delta^{\mathcal{D}}_b > \eta_{\mathrm{sc}}^2 B_g\,\|\mathbf{H} - \widehat{D}\|_{\mathrm{op}}\,\|\mathbf{G}\|_2, \tag{54}$$

then the Top-b data set induced by $\widehat{\mathbf{s}}_{\mathcal{D}}$ is identical to that induced by $\mathbf{s}^{(\mathrm{full})}_{\mathcal{D}}$.

Proof. By Proposition 3,

$$\|\widehat{\mathbf{s}}_{\mathcal{D}} - \mathbf{s}^{(\mathrm{full})}_{\mathcal{D}}\|_\infty \le \varepsilon_{\mathcal{D}} \triangleq \frac{\eta_{\mathrm{sc}}^2}{2}B_g\,\|\mathbf{H} - \widehat{D}\|_{\mathrm{op}}\,\|\mathbf{G}\|_2. \tag{55}$$

$\Delta^{\mathcal{D}}_b > 2\varepsilon_{\mathcal{D}}$ implies Top-b stability by the same argument. □

Remark 1 (Scope of the comparison)**.** *This subsection compares practical diagonal scores with local full-Hessian counterparts. It does not claim to recover exact bilevel optimum but shows diagonal approximation is controlled and, under a boundary-gap condition, preserves Top-k/Top-b selection.*

B Optimizer Compatibility: SGD Surrogate and Practical Curvature Proxy

Sections 3 to 4 analyze a localized single-step surrogate, for which Theorem 1 gives closed-form Shapley decompositions of its surrogate games. In practice, DualSFT retains this surrogate utility and instantiates its diagonal second-order coefficient with an optimizer-derived proxy. Let $\mathbf{r}_{t_w}$ denote AdamW's second-moment buffer after t_w warmup updates, with decay β_2 and stabilizer ϵ_{adam}. Define $\hat{\mathbf{c}} = \sqrt{\mathbf{r}_{t_w}/(1-\beta_2^{t_w})} + \epsilon_{\text{adam}}$ and keep it fixed during one-shot scoring. This is a positive damped RMS-gradient scale, not an exact Hessian diagonal. It aligns with recent analyses of Adam-family preconditioning and memory-reduced adaptive optimizers [75–78]. Accordingly, the exact row-column Shapley decomposition remains that of the SGD-defined localized surrogate, while the practical second-order score should be interpreted as its optimizer-aware instantiation.

Accordingly, any implementation-level mismatch introduced by the practical diagonal proxy is absorbed into the curvature/proxy utility term analyzed in Appendix A.5, so that the resulting score perturbation and selection stability can be controlled within the same error-propagation framework.

C Detailed Derivation of Confidence-Weighted Self-Distillation

Section 5 introduces CWSD in Eq. (12) as a tractable surrogate for prior-preserving direction $\mathbf{v}_{\text{prior}}$ without the pre-training corpus. This appendix proves CWSD is non-degenerate at warmup checkpoint $\bar{\boldsymbol{\theta}}$, and its gradient reduces to a vector-Jacobian product.

C.1 Objective and Evaluation Point

Let $\mathcal{D}_{\text{anchor}}$ be the held-out anchor set with $\mathcal{D}_{\text{anchor}} \cap \mathcal{D}_{\text{train}} = \emptyset$, and let $\mathcal{X}_{\text{anchor}}$ denote its input set.

$$\mathcal{L}_{\text{CWSD}}(\bar{\boldsymbol{\theta}}) = \frac{1}{|\mathcal{X}_{\text{anchor}}|} \sum_{x \in \mathcal{X}_{\text{anchor}}} \omega(x)\tau^2 \text{KL}\Big(P_\tau(\cdot \mid \boldsymbol{\theta}_{\text{old}}, x) \,\Big\|\, P_\tau(\cdot \mid \bar{\boldsymbol{\theta}}, x)\Big). \tag{56}$$

The surrogate preservation direction is

$$\mathbf{v}_{\text{prior}} \triangleq \nabla_{\bar{\boldsymbol{\theta}}} \mathcal{L}_{\text{CWSD}}(\bar{\boldsymbol{\theta}}). \tag{57}$$

CWSD is evaluated at the warmup checkpoint $\bar{\boldsymbol{\theta}}$ rather than $\boldsymbol{\theta}_{\text{old}}$. If evaluated at $\boldsymbol{\theta}_{\text{old}}$, teacher and student distributions would coincide, and the KL term and its gradient would both be zero. Using $\bar{\boldsymbol{\theta}}$ avoids this degeneracy and yields a non-trivial surrogate preservation direction for one-shot scoring.

C.2 Gradient Derivation

For one anchor example x, let

$$\mathbf{q} = P_\tau(\cdot \mid \boldsymbol{\theta}_{\text{old}}, x), \qquad \mathbf{p} = P_\tau(\cdot \mid \bar{\boldsymbol{\theta}}, x), \tag{58}$$

and let $\mathbf{z}$ denote the student logits at $\bar{\boldsymbol{\theta}}$. The per-example CWSD loss is

$$\ell_{\text{CWSD}}(x) = \omega(x)\tau^2 \text{KL}(\mathbf{q} \,\|\, \mathbf{p}) = \omega(x)\tau^2 \sum\nolimits_i q_i(\log q_i - \log p_i). \tag{59}$$

Since $\mathbf{q}$ is produced by the frozen reference model $\boldsymbol{\theta}_{\text{old}}$, gradients flow only through $\mathbf{p}$.

Differentiating with respect to the student logit z_j gives

$$\frac{\partial \ell_{\text{CWSD}}(x)}{\partial z_j} = -\omega(x)\tau^2 \sum\nolimits_i q_i \frac{\partial \log p_i}{\partial z_j}. \tag{60}$$

Using

$$\frac{\partial \log p_i}{\partial z_j} = \frac{1}{\tau}(\delta_{ij} - p_j), \tag{61}$$

we obtain

$$\frac{\partial \ell_{\text{CWSD}}(x)}{\partial z_j} = \omega(x)\tau(p_j - q_j). \tag{62}$$

Therefore, in column-vector form,

$$\nabla_{\mathbf{z}} \ell_{\mathrm{CWSD}}(x) = \omega(x)\tau(\mathbf{p} - \mathbf{q}) \in \mathbb{R}^{|\mathcal{V}|}. \tag{63}$$

Let $\mathbf{J}_z(x; \bar{\boldsymbol{\theta}}) \triangleq \frac{\partial \mathbf{z}}{\partial \bar{\boldsymbol{\theta}}} \in \mathbb{R}^{|\mathcal{V}| \times D}$. Applying the chain rule from logits to parameters gives

$$\nabla_{\bar{\boldsymbol{\theta}}} \ell_{\mathrm{CWSD}}(x) = \mathbf{J}_z(x; \bar{\boldsymbol{\theta}})^\top \omega(x)\tau(\mathbf{p} - \mathbf{q}) \in \mathbb{R}^D. \tag{64}$$

Summing over anchor inputs yields

$$\mathbf{v}_{\mathrm{prior}} = \nabla_{\bar{\boldsymbol{\theta}}} \mathcal{L}_{\mathrm{CWSD}}(\bar{\boldsymbol{\theta}}) = \frac{1}{|\mathcal{X}_{\mathrm{anchor}}|} \sum\nolimits_{x \in \mathcal{X}_{\mathrm{anchor}}} \mathbf{J}_z(x; \bar{\boldsymbol{\theta}})^\top \omega(x)\tau(\mathbf{p} - \mathbf{q}). \tag{65}$$

Equation (63) shows CWSD induces the logit-level gradient $\omega(x)\tau(\mathbf{p} - \mathbf{q})$, and Eq. (65) shows $\mathbf{v}_{\mathrm{prior}}$ is obtained by backpropagating this logit gradient through the student Jacobian. In practice, this gradient is obtained exactly with one backward pass at $\bar{\boldsymbol{\theta}}$, without the pre-training corpus. CWSD is used once at the warmup checkpoint to define the one-shot scoring geometry; it is not used as an online regularizer during final restricted fine-tuning.

D Additional Experimental Details and Results

D.1 Detailed Experimental Settings

D.1.1 Datasets and Evaluation Details

Datasets. We evaluate across two domains: Code Generation using Magicoder [37] and Mathematical Reasoning using MetaMathQA [38]. We denote each full corpus as $\mathcal{D}_{\mathrm{full}}$ and use it as the common budget base so that data-side budgets are directly comparable across all methods.

For each corpus, we reserve a 1,024-sample validation set $\mathcal{D}_{\mathrm{val}}$ that provides the meta-validation signal for scoring, and final performance is evaluated only on the benchmark suites described below, then draw a disjoint 2,048-sample anchor set $\mathcal{D}_{\mathrm{anchor}}$ for the prior-preserving direction $\mathbf{v}_{\mathrm{prior}}$ via CWSD. The remainder forms the candidate training pool $\mathcal{D}_{\mathrm{train}} = \mathcal{D}_{\mathrm{full}} \setminus (\mathcal{D}_{\mathrm{val}} \cup \mathcal{D}_{\mathrm{anchor}})$, further partitioned during scoring as follows:

- **Warmup Set ($\mathcal{D}_{\mathrm{warm}}$):** A random $5\% \cdot |\mathcal{D}_{\mathrm{full}}|$ subset of $\mathcal{D}_{\mathrm{train}}$, used for a 1-epoch pilot warmup to establish the scoring checkpoint $\bar{\boldsymbol{\theta}}$ and diagonal curvature proxy $\hat{\mathbf{c}}$. After warmup, $\mathcal{D}_{\mathrm{warm}}$ is excluded from the scoring pool to avoid bias from already-updated samples.
- **Scoring Pool ($\mathcal{D}_{\mathrm{pool}} = \mathcal{D}_{\mathrm{train}} \setminus \mathcal{D}_{\mathrm{warm}}$):** The remainder of the candidate pool, over which one-shot shared-surrogate scoring is performed.
- **Selected Subset ($\mathcal{D}_{\mathrm{sel}}$):** Top-$b$ samples drawn from $\mathcal{D}_{\mathrm{pool}}$, with $b = 10\% \cdot |\mathcal{D}_{\mathrm{full}}|$.

CWSD estimates a preservation direction, so we fix the anchor count at 2,048, where the direction and selections stabilize, well below corpus scale (Appendix D.5.1). The 3,072 validation/anchor samples (~2.8% of Magicoder; ~0.8% of MetaMathQA) serve only scoring and preservation-direction estimation, excluding final fine-tuning. Final adaptation subset sizes match within budget groups.

Evaluation Protocol. We use `lm-evaluation-harness` [91]: HumanEval [81], 0-shot pass@1 with greedy decoding (temperature 0, maximum 512 tokens); PIQA [82] and ARC-Challenge [83], 0-shot accuracy via option log-likelihoods; MMLU [84], 5-shot accuracy averaged over 57 subjects; GSM8K [85], 5-shot chain-of-thought with final-answer numeric exact match. Stability averages four out-of-domain scores; Plasticity is HumanEval for code or GSM8K for math; Overall averages both. Unless otherwise stated, results are mean ± standard deviation over seeds $\{42, 123, 2026\}$.

D.1.2 Implementation Details

For reproducibility, we distinguish shared optimization settings from method-specific configurations. All methods use the same dataset split, evaluation protocol, and reporting convention, run on four NVIDIA A100 (80GB) GPUs, and report the mean and standard deviation over three random seeds. Standard SFT, LoRA, and the supervised fine-tuning stage of data-selection baselines are implemented with LlamaFactory, whereas parameter-sparse baselines, including S^2FT, SMT, LIFT, and

NanoAdam, follow their original codebases. Optimization hyper-parameters are selected separately for each method, base model, and dataset, and are reported in Tables 8 and 9.

For sparse baselines, we also report the structural hyper-parameters that define their update spaces and sparsity patterns; these settings are summarized in Table 10. Specifically, LoRA is defined by rank, scaling, dropout, and target modules; S^2FT by structured selection ratios over attention and FFN subspaces; SMT by matrix/block sparsity and warm-start configuration; LIFT by mask type, no-gradient module setting, rank-reduction configuration, and projection update interval; and NanoAdam by density budget, mask interval, dynamic-density schedule, and mask criterion.

For DualSFT, the reported results use a 1-epoch warmup with $r_{\text{warm}} = 5\%$, balance factor $\lambda = 0.8$, CWSD temperature $\tau = 1.0$, data budget $b = 10\%$, and parameter budget $k = 5\%$. The final fine-tuning learning rate η_{ft} is listed in Tables 8–9; scoring uses $\eta_{\text{sc}} = \eta_{\text{ft}}/N_{\text{sc}}$, where N_{sc} is the number of examples accumulated in $\mathbf{G}$.

Table 8: Fine-Tuning Hyper-parameters for Magicoder.

Hyper-parameter	Std. SFT	LoRA	Data-Only	S^2FT	SMT	LIFT	NanoAdam	DualSFT (Ours)
LR (Llama-3.2-3B)	2e-5	2e-4	2e-5	5e-5	4e-5	4e-5	3e-5	2e-5
LR (Gemma-3-4B)	2e-5	2e-4	2e-5	5e-5	4e-5	4e-5	3e-5	2e-5
LR (Qwen-3.5-9B)	1e-5	2e-4	1e-5	2e-5	2e-5	1.5e-5	2e-5	1e-5
Shared settings: AdamW, cosine decay, global batch size 64, max sequence length 4096, 3 epochs, LR warmup ratio 0.03, weight decay 0.0.								

Table 9: Fine-Tuning Hyper-parameters for MetaMathQA.

Hyper-parameter	Std. SFT	LoRA	Data-Only	S^2FT	SMT	LIFT	NanoAdam	DualSFT (Ours)
LR (Llama-3.2-3B)	1e-5	2e-4	1e-5	5e-5	4e-5	3e-5	3e-5	1e-5
LR (Gemma-3-4B)	1e-5	2e-4	1e-5	5e-5	4e-5	3e-5	3e-5	1e-5
Shared settings: AdamW, cosine decay, global batch size 128, max sequence length 1024, 2 epochs, LR warmup ratio 0.03, weight decay 0.0.								

Table 10: Method-specific structural hyper-parameters.

Method	Reported method-specific hyper-parameters
LoRA	rank $r = 64$, scaling $\alpha = 128$, dropout = 0.05, target modules = all linear layers
S^2FT	structured selection ratios: $d_ratio = 0.03$, $u_ratio = o_ratio = v_ratio = 0$
SMT	warm-start steps = 100, attention block ratio = 0.25, MLP block ratio = 0.10
LIFT	mask type = `principle`, no-gradient modules = `gateupdown`, LoRA rank = 64, filter rank = 64, update interval = 200
NanoAdam	mask interval = 600, dynamic density = `False`, density interval = 1000, mask criterion = `weights`
DualSFT	warmup epochs = 1, warmup fraction $r_{\text{warm}} = 5\%$ of $\lvert\mathcal{D}_{\text{full}}\rvert$, balance factor $\lambda = 0.8$, CWSD temperature $\tau = 1.0$, data budget $b = 10\%$, parameter budget $k = 5\%$

D.1.3 Prompt Templates

To ensure complete reproducibility, we provide the exact instruction templates used for fine-tuning and evaluation. We adhere to the standard Alpaca-style formats.

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:

Table 11: Additional 2×2 sequential-combination baselines on Magicoder (%). Best results are **bold**, and second-best results are underlined. All methods use the same final data and parameter budgets.

Model	Method	Stability ↑					Plasticity ↑	Overall ↑
		ARC-C	PIQA	MMLU	GSM8K	Avg	HE	Score
Llama-3.2	LESS + S^2FT	42.58$_{\pm 1.44}$	**76.68**$_{\pm 0.96}$	53.55$_{\pm 0.39}$	22.87$_{\pm 0.71}$	48.92$_{\pm 0.48}$	38.82$_{\pm 0.57}$	43.87$_{\pm 0.40}$
	LESS + NanoAdam	42.75$_{\pm 1.46}$	76.31$_{\pm 0.97}$	54.06$_{\pm 0.41}$	24.84$_{\pm 0.76}$	49.49$_{\pm 0.50}$	39.84$_{\pm 0.76}$	44.66$_{\pm 0.43}$
	SPICE + S^2FT	41.41$_{\pm 1.45}$	76.55$_{\pm 0.97}$	53.96$_{\pm 0.40}$	23.07$_{\pm 0.69}$	48.75$_{\pm 0.48}$	39.23$_{\pm 0.57}$	43.99$_{\pm 0.42}$
	SPICE + NanoAdam	42.15$_{\pm 1.46}$	76.61$_{\pm 0.98}$	54.87$_{\pm 0.40}$	25.93$_{\pm 0.71}$	49.89$_{\pm 0.49}$	39.43$_{\pm 0.76}$	44.66$_{\pm 0.42}$
	DualSFT (Ours)	**42.95**$_{\pm 1.45}$	76.22$_{\pm 0.98}$	**56.09**$_{\pm 0.40}$	**26.89**$_{\pm 0.70}$	**50.54**$_{\pm 0.47}$	**43.70**$_{\pm 0.76}$	**47.12**$_{\pm 0.43}$
Gemma-3	LESS + S^2FT	51.02$_{\pm 1.46}$	78.65$_{\pm 0.93}$	55.68$_{\pm 0.39}$	31.13$_{\pm 0.78}$	54.12$_{\pm 0.49}$	53.05$_{\pm 0.86}$	53.59$_{\pm 0.48}$
	LESS + NanoAdam	50.34$_{\pm 1.47}$	78.20$_{\pm 0.96}$	56.18$_{\pm 0.40}$	32.04$_{\pm 0.81}$	54.19$_{\pm 0.49}$	53.05$_{\pm 0.86}$	53.62$_{\pm 0.47}$
	SPICE + S^2FT	50.46$_{\pm 1.46}$	78.38$_{\pm 0.95}$	56.04$_{\pm 0.40}$	32.27$_{\pm 0.84}$	54.29$_{\pm 0.48}$	52.44$_{\pm 0.86}$	53.36$_{\pm 0.48}$
	SPICE + NanoAdam	50.77$_{\pm 1.45}$	78.55$_{\pm 0.94}$	57.29$_{\pm 0.40}$	33.89$_{\pm 0.82}$	55.12$_{\pm 0.49}$	52.64$_{\pm 0.76}$	53.88$_{\pm 0.47}$
	DualSFT (Ours)	**51.14**$_{\pm 1.44}$	**79.02**$_{\pm 0.95}$	**59.16**$_{\pm 0.38}$	**35.61**$_{\pm 0.64}$	**56.23**$_{\pm 0.48}$	**56.71**$_{\pm 0.86}$	**56.47**$_{\pm 0.43}$
Qwen-3.5	LESS + S^2FT	50.97$_{\pm 1.45}$	79.54$_{\pm 0.96}$	69.14$_{\pm 0.40}$	81.27$_{\pm 0.75}$	70.23$_{\pm 0.48}$	67.89$_{\pm 0.76}$	69.06$_{\pm 0.44}$
	LESS + NanoAdam	51.42$_{\pm 1.46}$	79.16$_{\pm 0.95}$	69.78$_{\pm 0.39}$	80.62$_{\pm 0.74}$	70.25$_{\pm 0.50}$	68.90$_{\pm 0.86}$	69.57$_{\pm 0.42}$
	SPICE + S^2FT	50.82$_{\pm 1.44}$	79.27$_{\pm 0.96}$	69.61$_{\pm 0.40}$	80.41$_{\pm 0.71}$	70.03$_{\pm 0.47}$	67.48$_{\pm 0.76}$	68.75$_{\pm 0.43}$
	SPICE + NanoAdam	50.54$_{\pm 1.44}$	79.34$_{\pm 0.95}$	70.09$_{\pm 0.39}$	80.74$_{\pm 0.71}$	70.18$_{\pm 0.48}$	69.31$_{\pm 0.76}$	69.74$_{\pm 0.42}$
	DualSFT (Ours)	**54.01**$_{\pm 1.43}$	**79.63**$_{\pm 0.94}$	**70.33**$_{\pm 0.38}$	**84.61**$_{\pm 0.68}$	**72.15**$_{\pm 0.47}$	**73.78**$_{\pm 0.50}$	**72.96**$_{\pm 0.39}$

MetaMathQA Instruction Template

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

### Instruction:
{instruction}

### Response:
```

D.2 Additional Sequential-Combination Baselines

To examine pipeline-style hybrids, we evaluate a 2×2 grid combining two data selectors, LESS and SPICE, with two parameter selectors, S^2FT and NanoAdam. All combinations use the same final data budget (10%) and parameter budget (5%) as DualSFT. This comparison tests whether sequentially composing strong single-axis selectors can close the gap to coordinated dual scoring.

Table 11 complements the main comparison by covering all pairings between the selected data-side and parameter-side baselines. Across all three backbones, SPICE + NanoAdam remains the strongest sequential hybrid in Overall, while DualSFT consistently achieves higher Stability Avg., Plasticity, and Overall under the same final data and parameter budgets.

D.3 Additional Results on MetaMathQA

Table 12 reports additional MetaMathQA results, which match the Magicoder trends. For Llama-3.2-3B and Gemma-3-4B-PT, DualSFT-Param/Data lead mean Overall within single-axis groups, while full DualSFT leads joint-budget alternatives. DualSFT scores 57.95 on Llama-3.2-3B, surpassing SPICE + NanoAdam by 1.63 points. On Gemma-3-4B-PT, it reaches 61.47, improving over the strongest sequential composition by 2.19 points. These results support the advantage of shared dual scoring over isolated or sequential selection under evaluated budgets. We omit Qwen results because its strong GSM8K prior makes downstream mathematical fine-tuning less informative.

D.4 General Instruction Tuning

We evaluate Llama-3.2-3B Base on Alpaca-GPT4 [92] with training seed 42, retaining the main experiments' matched-budget settings for parameter-only, data-only, and joint selection. General Avg. averages ARC-Challenge, PIQA, MMLU, GSM8K, and HumanEval. IFEval [93] measures prompt- and instruction-level accuracy under strict and loose matching; IFEval Avg. averages these four scores. Overall averages General Avg. and IFEval Avg. Table 13 shows that all three DualSFT variants improve both General Avg. and IFEval Avg. over their matched-budget baselines. DualSFT-Data and full DualSFT improve all four IFEval metrics, while DualSFT-Param improves three. These single-seed results support the applicability of shared scoring to general instruction tuning.

Table 12: Fine-tuning results on MetaMathQA (%). Pre-trained, Standard SFT, and LoRA (gray) are excluded from group ranking. In each group, **best** and second-best results are highlighted.

Model	Method	Stability ↑					Plasticity ↑	Overall ↑
		ARC-C	PIQA	MMLU	HE	Avg	GSM8K	Score
Llama-3.2-3B	Pre-trained	$43.00_{\pm 0.00}$	$76.61_{\pm 0.00}$	$56.52_{\pm 0.00}$	$28.66_{\pm 0.00}$	$51.20_{\pm 0.00}$	$27.45_{\pm 0.00}$	$39.32_{\pm 0.00}$
	Standard SFT	$36.60_{\pm 1.41}$	$74.48_{\pm 1.02}$	$48.51_{\pm 0.41}$	$15.04_{\pm 0.57}$	$43.66_{\pm 0.44}$	$65.96_{\pm 0.81}$	$54.81_{\pm 0.48}$
	LoRA ICLR'22	$40.61_{\pm 1.44}$	$74.21_{\pm 0.98}$	$52.11_{\pm 0.39}$	$25.00_{\pm 0.50}$	$47.98_{\pm 0.48}$	$61.08_{\pm 0.75}$	$54.53_{\pm 0.45}$
	S^2FT NIPS'24	$40.78_{\pm 1.45}$	**$75.95_{\pm 0.99}$**	$52.88_{\pm 0.40}$	$25.41_{\pm 0.57}$	$48.76_{\pm 0.48}$	$64.95_{\pm 0.76}$	$56.85_{\pm 0.45}$
	SMT ICLR'25	$41.21_{\pm 1.45}$	$72.32_{\pm 0.98}$	$52.74_{\pm 0.40}$	$26.22_{\pm 0.50}$	$48.12_{\pm 0.47}$	$65.05_{\pm 0.75}$	$56.59_{\pm 0.44}$
	LIFT ICML'25	$40.93_{\pm 1.44}$	$73.58_{\pm 0.99}$	$52.61_{\pm 0.40}$	$25.00_{\pm 0.50}$	$48.03_{\pm 0.47}$	$64.72_{\pm 0.74}$	$56.37_{\pm 0.44}$
	NanoAdam NIPS'25	**$42.55_{\pm 1.44}$**	$74.50_{\pm 0.98}$	$53.74_{\pm 0.40}$	$26.63_{\pm 0.76}$	$49.35_{\pm 0.48}$	$65.18_{\pm 0.78}$	$57.26_{\pm 0.46}$
	DualSFT-Param (Ours)	$42.43_{\pm 1.45}$	$75.83_{\pm 0.99}$	**$55.96_{\pm 0.40}$**	**$28.05_{\pm 0.50}$**	**$50.57_{\pm 0.47}$**	**$65.83_{\pm 0.76}$**	**$58.20_{\pm 0.45}$**
	Random	$42.18_{\pm 1.45}$	**$76.22_{\pm 0.99}$**	$55.92_{\pm 0.40}$	**$28.66_{\pm 0.86}$**	$50.75_{\pm 0.47}$	$59.67_{\pm 0.77}$	$55.21_{\pm 0.46}$
	LESS ICML'24	$41.52_{\pm 1.45}$	$74.66_{\pm 0.99}$	$53.58_{\pm 0.40}$	$26.42_{\pm 0.76}$	$49.05_{\pm 0.50}$	$63.56_{\pm 0.79}$	$56.30_{\pm 0.46}$
	FisherSFT ICML'25	$40.70_{\pm 1.44}$	$74.25_{\pm 0.99}$	$55.12_{\pm 0.40}$	$27.03_{\pm 0.76}$	$49.27_{\pm 0.48}$	$61.69_{\pm 0.76}$	$55.48_{\pm 0.45}$
	SPICE ICLR'26	$41.27_{\pm 1.45}$	$74.88_{\pm 0.98}$	$55.42_{\pm 0.40}$	$27.24_{\pm 0.57}$	$49.70_{\pm 0.48}$	$62.37_{\pm 0.75}$	$56.04_{\pm 0.44}$
	DualSFT-Data (Ours)	**$42.89_{\pm 1.44}$**	$76.13_{\pm 0.99}$	**$56.03_{\pm 0.40}$**	**$28.66_{\pm 0.86}$**	**$50.93_{\pm 0.47}$**	**$65.13_{\pm 0.77}$**	**$58.03_{\pm 0.44}$**
	LESS + S^2FT	$41.75_{\pm 1.44}$	$75.35_{\pm 0.97}$	$53.67_{\pm 0.39}$	$25.61_{\pm 0.50}$	$49.10_{\pm 0.48}$	$61.54_{\pm 0.74}$	$55.32_{\pm 0.44}$
	SPICE + NanoAdam	$42.46_{\pm 1.45}$	$76.02_{\pm 0.98}$	$54.83_{\pm 0.40}$	$26.63_{\pm 0.57}$	$49.99_{\pm 0.50}$	$62.65_{\pm 0.75}$	$56.32_{\pm 0.44}$
	DualSFT (Ours)	**$42.89_{\pm 1.45}$**	**$76.42_{\pm 0.98}$**	**$56.34_{\pm 0.40}$**	**$29.88_{\pm 0.86}$**	**$51.38_{\pm 0.48}$**	**$64.52_{\pm 0.75}$**	**$57.95_{\pm 0.42}$**
Gemma-3-4B-PT	Pre-trained	$51.45_{\pm 0.00}$	$79.16_{\pm 0.00}$	$59.61_{\pm 0.00}$	$35.37_{\pm 0.00}$	$56.40_{\pm 0.00}$	$37.00_{\pm 0.00}$	$46.70_{\pm 0.00}$
	Standard SFT	$43.37_{\pm 1.44}$	$74.48_{\pm 0.97}$	$53.00_{\pm 0.40}$	$29.47_{\pm 0.76}$	$50.08_{\pm 0.47}$	$67.78_{\pm 0.81}$	$58.93_{\pm 0.48}$
	LoRA ICLR'22	$48.92_{\pm 1.44}$	$77.95_{\pm 0.95}$	$55.80_{\pm 0.40}$	$30.49_{\pm 0.86}$	$53.29_{\pm 0.47}$	$62.47_{\pm 0.79}$	$57.88_{\pm 0.46}$
	S^2FT NIPS'24	$49.52_{\pm 1.46}$	$77.68_{\pm 0.96}$	$55.73_{\pm 0.40}$	$31.30_{\pm 0.76}$	$53.56_{\pm 0.49}$	$64.34_{\pm 0.80}$	$58.95_{\pm 0.47}$
	SMT ICLR'25	$49.83_{\pm 1.46}$	$77.29_{\pm 0.93}$	$56.17_{\pm 0.39}$	$31.91_{\pm 0.76}$	$53.80_{\pm 0.48}$	$65.50_{\pm 0.84}$	$59.65_{\pm 0.49}$
	LIFT ICML'25	$50.14_{\pm 1.46}$	$76.59_{\pm 0.95}$	$55.21_{\pm 0.40}$	$30.28_{\pm 0.57}$	$53.06_{\pm 0.49}$	$64.72_{\pm 0.78}$	$58.89_{\pm 0.46}$
	NanoAdam NIPS'25	$50.68_{\pm 1.45}$	$78.16_{\pm 0.94}$	$58.25_{\pm 0.39}$	$33.54_{\pm 0.86}$	$55.16_{\pm 0.49}$	$65.88_{\pm 0.81}$	$60.52_{\pm 0.48}$
	DualSFT-Param (Ours)	**$51.05_{\pm 1.45}$**	**$78.87_{\pm 0.98}$**	**$59.79_{\pm 0.39}$**	**$34.96_{\pm 0.76}$**	**$56.17_{\pm 0.50}$**	**$67.15_{\pm 0.80}$**	**$61.66_{\pm 0.47}$**
	Random	$51.25_{\pm 1.45}$	**$79.04_{\pm 0.99}$**	$59.37_{\pm 0.40}$	**$35.16_{\pm 0.76}$**	**$56.20_{\pm 0.48}$**	$61.59_{\pm 0.77}$	$58.90_{\pm 0.44}$
	LESS ICML'24	$48.66_{\pm 1.45}$	$77.53_{\pm 0.99}$	$56.07_{\pm 0.40}$	$30.89_{\pm 0.76}$	$53.29_{\pm 0.49}$	$65.28_{\pm 0.79}$	$59.28_{\pm 0.49}$
	FisherSFT ICML'25	$49.80_{\pm 1.45}$	$77.60_{\pm 0.94}$	$56.49_{\pm 0.38}$	$32.52_{\pm 0.76}$	$54.10_{\pm 0.48}$	$63.81_{\pm 0.77}$	$58.96_{\pm 0.45}$
	SPICE ICLR'26	$50.34_{\pm 1.44}$	$78.42_{\pm 0.95}$	$57.52_{\pm 0.40}$	$33.94_{\pm 0.76}$	$55.06_{\pm 0.49}$	$64.72_{\pm 0.79}$	$59.89_{\pm 0.46}$
	DualSFT-Data (Ours)	**$51.31_{\pm 1.45}$**	$78.62_{\pm 0.98}$	**$59.44_{\pm 0.40}$**	**$35.16_{\pm 0.76}$**	$56.13_{\pm 0.49}$	**$66.94_{\pm 0.80}$**	**$61.54_{\pm 0.48}$**
	LESS + S^2FT	$50.91_{\pm 1.46}$	$78.55_{\pm 0.93}$	$56.48_{\pm 0.39}$	$31.50_{\pm 0.76}$	$54.36_{\pm 0.50}$	$63.86_{\pm 0.78}$	$59.11_{\pm 0.45}$
	SPICE + NanoAdam	$49.74_{\pm 1.45}$	$78.65_{\pm 0.94}$	$58.21_{\pm 0.40}$	$33.74_{\pm 0.76}$	$55.09_{\pm 0.48}$	$63.48_{\pm 0.80}$	$59.28_{\pm 0.46}$
	DualSFT (Ours)	**$51.39_{\pm 1.44}$**	**$79.04_{\pm 0.95}$**	**$59.61_{\pm 0.38}$**	**$35.37_{\pm 0.86}$**	**$56.35_{\pm 0.49}$**	**$66.59_{\pm 0.79}$**	**$61.47_{\pm 0.47}$**

Table 13: General instruction-tuning results on Alpaca-GPT4 with Llama-3.2-3B Base.

Method	General Avg.	Prompt Strict	Inst. Strict	Prompt Loose	Inst. Loose	IFEval Avg.	Overall
Pre-trained	46.45	22.00	34.77	24.58	38.01	29.84	38.15
Standard SFT	39.85	31.24	40.89	32.72	44.96	37.45	38.65
NanoAdam	44.75	28.47	40.41	30.13	44.24	35.81	40.28
DualSFT-Param	46.94	30.68	40.29	32.90	44.36	37.06	42.00
LESS	45.26	28.84	39.33	30.50	42.45	35.28	40.27
DualSFT-Data	47.12	30.50	40.17	32.53	44.00	36.80	41.96
SPICE + NanoAdam	45.54	28.10	38.85	29.76	39.45	34.04	39.79
DualSFT	47.32	30.31	40.05	31.79	43.88	36.51	41.92

D.5 Preservation Direction and Anchor-size Ablations

The main ablation in Fig. 3 shows that removing CWSD substantially reduces stability. We further compare CWSD with simpler preservation-direction alternatives under the same DualSFT joint setting. All variants use the same data budget, parameter budget, warmup checkpoint, and final fine-tuning hyperparameters; only the construction of $\mathbf{v}_{\text{prior}}$ is changed. Table 14 indicates that CWSD is more effective than parameter-space anchoring alternatives. L2-to-$\boldsymbol{\theta}_{\text{old}}$ improves stability over removing preservation entirely, while diagonal Fisher anchoring gives a stronger stability signal but remains below CWSD. CWSD provides the best balance because it constructs a behavior-level preservation direction from high-confidence old-model predictions at the scoring checkpoint.

Table 14: Preservation-direction ablation on Llama-3.2-3B with Magicoder in joint DualSFT.

Preservation direction	Stability	Plasticity	Overall
No preservation	47.28	42.07	44.68
L2-to-$\boldsymbol{\theta}_{\text{old}}$	48.84	42.89	45.86
Diagonal Fisher anchoring	49.93	43.50	46.71
CWSD	**50.54**	**43.70**	**47.12**

D.5.1 Anchor-size sensitivity

CWSD uses $\mathcal{D}_{\text{anchor}}$ to estimate a preservation direction at the warmup checkpoint, rather than to provide additional fine-tuning examples. We therefore test whether increasing the anchor set changes

Table 15: Anchor-size sensitivity for CWSD on Llama-3.2-3B with Magicoder.

$\lvert\mathcal{D}_{\text{anchor}}\rvert$	256	512	1024	2048	4096
Cosine w.r.t. 4096-anchor $\mathbf{v}_{\text{prior}}$	0.82	0.89	0.94	0.98	1.00
Top-b data overlap	0.71	0.78	0.85	0.92	1.00
Top-k parameter overlap	0.74	0.81	0.87	0.94	1.00
Stability	49.42	49.86	50.21	50.54	50.57
Plasticity	43.29	43.50	43.50	43.70	43.70
Overall	46.36	46.68	46.85	47.12	47.13

the estimated direction and the induced selections. Using Llama-3.2-3B on Magicoder, we vary $|\mathcal{D}_{\text{anchor}}| \in \{256, 512, 1024, 2048, 4096\}$ and compare each setting with the 4096-anchor reference. We report the cosine similarity of $\mathbf{v}_{\text{prior}}$, Top-$b$ data overlap, Top-k parameter overlap, and final performance after restricted fine-tuning. Anchor samples are never used in final restricted fine-tuning.

Table 15 shows that both the CWSD direction and the induced selections stabilize as the anchor size increases. Moving from 2048 to 4096 anchors gives only marginal changes in direction similarity, selected data, selected parameters, and final performance. This supports using a moderate fixed anchor size as a scoring-only resource, rather than scaling the anchor set with the full corpus size.

D.6 Protocol for Same-Surrogate Local-Regret Diagnostic

We compute the local-regret diagnostic at the init and warmup checkpoints without additional training. For the data side, we fix the DualSFT parameter mask $\mathbf{m}_\theta$ and define the aware and unaware scores as $s^{\text{aware}}_{\mathcal{D},n} = \langle \mathbf{u} \odot \mathbf{m}_\theta, \mathbf{g}_n \rangle$ and $s^{\text{unaware}}_{\mathcal{D},n} = \langle \mathbf{u}, \mathbf{g}_n \rangle$. Let $S^{\text{aware}}_{\mathcal{D}}$ and $S^{\text{unaware}}_{\mathcal{D}}$ be their Top-b subsets. We evaluate both under the aware utility $U^{\text{aware}}_{\mathcal{D}}(S) = \sum_{n \in S} s^{\text{aware}}_{\mathcal{D},n}$. The data-side utility ratio and regret are

$$\text{Ratio}_{\mathcal{D}} = \frac{U^{\text{aware}}_{\mathcal{D}}(S^{\text{unaware}}_{\mathcal{D}})}{U^{\text{aware}}_{\mathcal{D}}(S^{\text{aware}}_{\mathcal{D}}) + \epsilon}, \qquad \text{Regret}_{\mathcal{D}} = 100(1 - \text{Ratio}_{\mathcal{D}}).$$

For the parameter side, we fix the DualSFT data subset $S_{\mathcal{D}}$ and define $s^{\text{aware}}_{\theta,d} = u_d G_{S_{\mathcal{D}},d}$ and $s^{\text{unaware}}_{\theta,d} = u_d G_d$. Let m^{aware}_θ and $m^{\text{unaware}}_\theta$ be their Top-k masks. We use $U^{\text{aware}}_\theta(m) = \sum_{d:m_d=1} s^{\text{aware}}_{\theta,d}$ and compute the same utility ratio and regret. Top-q Jaccard reports the overlap between aware and unaware selections at $q = b$ for data and $q = k$ for parameters.

D.7 Statistic Definitions

We define statistics used in Table 5, Fig. 6, and Table 17. For data, Random denotes a same-size random subset, Top-loss selects the same number of examples with the largest training losses at $\bar{\boldsymbol{\theta}}$, and Pool denotes the full candidate pool. For a selected subset $\mathcal{D}_{\text{sel}}$, validation-gradient alignment is

$$\text{Align}(\mathcal{D}_{\text{sel}}) = \frac{1}{|\mathcal{D}_{\text{sel}}|} \sum_{n \in \mathcal{D}_{\text{sel}}} \frac{\langle \mathbf{g}_n, \mathbf{v}_{\text{val}} \rangle}{\|\mathbf{g}_n\|_2 \, \|\mathbf{v}_{\text{val}}\|_2}, \tag{66}$$

where $\mathbf{g}_n$ and $\mathbf{v}_{\text{val}}$ are computed at checkpoint $\bar{\boldsymbol{\theta}}$. The average training loss is the mean per-example loss over the corresponding subset. For diversity, let $\phi(x)$ denote the mean-pooled sentence embedding extracted from the pre-trained backbone. We compute the mean pairwise cosine distance:

$$\text{Div}(\mathcal{D}_{\text{sel}}) = \frac{1}{\binom{|\mathcal{D}_{\text{sel}}|}{2}} \sum_{i<j} \left(1 - \cos\big(\phi(x_i), \phi(x_j)\big)\right). \tag{67}$$

This score is used only as a redundancy diagnostic; DualSFT does not optimize it directly.

For parameter statistics, all shares are computed within the trainable candidate pool. Let S_θ denote the selected parameter mask and $\mathcal{P}$ the trainable parameter pool. For a layer or module group $\mathcal{G}$, we define enrichment as a foreground-to-background share ratio:

$$\text{Enrichment}(\mathcal{G}) = \frac{|\mathcal{G} \cap S_\theta| / |S_\theta|}{|\mathcal{G} \cap \mathcal{P}| / |\mathcal{P}|}. \tag{68}$$

Enrichment above 1 indicates over-selection relative to the pool. Top-5 layers have the largest selected-parameter shares. The early-layer group (L0–L9) is predefined only for reporting and remains in the trainable candidate pool. The down_proj statistic compares its share among selected FFN parameters with its FFN-pool share.

Table 16: Cost breakdown of fixed-side re-selection on Llama-3.2-3B (Magicoder).

Stage	DualSFT	Data → Param	Param → Data
Shared one-shot scoring	6.5 h	6.5 h	6.5 h
Fixed-side re-scoring	–	0.4 h	2.9 h
Restricted fine-tuning	0.9 h	0.9 h	1.0 h
Selection total	6.5 h	6.9 h (1.06×)	9.4 h (1.45×)
End-to-end total	7.4 h	7.8 h (1.05×)	10.4 h (1.41×)

Table 17: Selected parameter statistics on Llama-3.2-3B (Magicoder).

Param statistic	Sel.	Pool	Enrich.
Top-5 selected layers	42.4%	17.9%	2.38×
Early layers (L0–L9)	8.2%	35.7%	0.23×
Attention share	22.6%	25.0%	0.91×
FFN share	77.4%	75.0%	1.03×
down_proj within FFN	47.9%	33.3%	1.44×

For the layer-wise heatmap in Fig. 6, let $\mathcal{B}_{\ell,m}$ denote the trainable parameter coordinates in layer ℓ and module m. We compute the block-level mean absolute score as

$$H_{\ell,m} = \frac{1}{|\mathcal{B}_{\ell,m}|} \sum_{d \in \mathcal{B}_{\ell,m}} |s_{\theta,d}|, \tag{69}$$

where $s_{\theta,d}$ is the DualSFT parameter score in Eq. (11). This heatmap visualizes the spatial distribution of parameter scores, while aggregate shares in Table 17 provide exact numerical summaries.

D.8 Fixed-side Re-selection Details

We provide definitions and diagnostics for the fixed-side re-selection variants in Table 4. Let $(S_{\mathcal{D}}^{(0)}, S_{\theta}^{(0)})$ denote the one-shot DualSFT selection at checkpoint $\bar{\boldsymbol{\theta}}$. Data → Param fixes $S_{\mathcal{D}}^{(0)}$ and re-selects a parameter mask $S_{\theta}^{(1)}$ using scores recomputed on the selected subset. Param → Data fixes $S_{\theta}^{(0)}$ and re-selects a data subset $S_{\mathcal{D}}^{(1)}$ under the fixed mask. Checkpoint, budgets, and final fine-tuning hyperparameters remain unchanged. The larger cost of Param → Data comes from its per-candidate scoring pass, whereas Data → Param only recomputes an aggregate gradient on the selected subset.

Cost breakdown. Table 16 reports GPU-hour costs. Selection total includes warmup, first-round scoring, and fixed-side re-scoring; end-to-end total includes restricted fine-tuning under selected data and parameter budgets. Results show that small fixed-side re-selection gains incur non-negligible cost, especially for Param → Data, where data-side re-scoring dominates added overhead.

D.9 Score Fidelity Protocol and Extended Horizons

D.9.1 Local Surrogate Drift and Horizon Decay

DualSFT computes one-shot scores at the warmup checkpoint $\bar{\boldsymbol{\theta}}$, whereas realized validation gains are measured after restricted optimization. This mismatch creates a natural horizon dependence: as the optimization trajectory moves away from $\bar{\boldsymbol{\theta}}$, the gradients defining the local surrogate may drift.

To make this dependence explicit, consider the first-order data score at a generic checkpoint $\boldsymbol{\theta}$:

$$s_{\mathcal{D},n}(\boldsymbol{\theta}) = \eta \, \langle \nabla \mathcal{L}_{\text{val}}(\boldsymbol{\theta}), \nabla \ell_n(\boldsymbol{\theta}) \rangle \, . \tag{70}$$

Assume that, along the restricted optimization trajectory, $\|\nabla \mathcal{L}_{\text{val}}(\boldsymbol{\theta})\| \le B_v$, $\|\nabla \ell_n(\boldsymbol{\theta})\| \le B_g$, and the two gradients are Lipschitz with constants L_v and L_g. Then

$$\left| s_{\mathcal{D},n}(\boldsymbol{\theta}_t) - s_{\mathcal{D},n}(\bar{\boldsymbol{\theta}}) \right| \le \eta \left(B_v L_g + B_g L_v \right) \left\| \boldsymbol{\theta}_t - \bar{\boldsymbol{\theta}} \right\| . \tag{71}$$

One-shot ranking remains stable when Eq. (71) is small relative to score margins, but fidelity decays as the trajectory moves farther from $\bar{\boldsymbol{\theta}}$. The same argument applies to parameter scores by replacing per-sample gradients with coordinate-wise training-gradient terms, motivating extended-horizon diagnostics and leaving dynamic score updates for longer adaptation horizons.

D.9.2 Protocol

We provide extended diagnostics for Table 6. The main table averages over $K \in \{50, 200, 500\}$ restricted AdamW steps, while this section reports $K \in \{10, 50, 200, 500, 1000\}$ and the full Init/Warmup × First/Diag grid. For data selection, we sample 64 diagnostic subsets with 256 examples, scoring them by $\hat{U}_{\mathcal{D}}(S) = \sum_{n \in S} s_{\mathcal{D},n}$. For parameter selection, we sample 64 budget-matched masks using block-stratified layer-module sampling, scoring them by $\hat{U}_{\theta}(S) = \sum_{d \in S} s_{\theta,d}$. From the

Table 18: Extended-horizon score fidelity on Llama-3.2-3B with Magicoder.

Variant	K	Data ρ	Param ρ	Data Pair	Param Pair
Init + First-Order	10	0.40	0.33	65.3	62.0
	50	0.34	0.27	61.8	58.9
	200	0.28	0.21	58.4	56.1
	500	0.23	0.16	55.7	53.8
	1000	0.17	0.12	53.6	52.0
Init + Diag-Second-Order	10	0.48	0.41	70.4	66.2
	50	0.43	0.36	67.4	63.6
	200	0.37	0.30	64.0	60.6
	500	0.31	0.25	60.8	58.0
	1000	0.25	0.20	57.6	55.3
Warmup + First-Order	10	0.68	0.61	80.1	76.0
	50	0.62	0.55	76.8	72.4
	200	0.56	0.49	73.0	69.0
	500	0.49	0.42	69.5	65.8
	1000	0.40	0.35	65.5	62.2
Warmup + Diag-Second-Order	10	**0.90**	**0.87**	**85.1**	**81.8**
	50	**0.87**	**0.84**	**82.8**	**79.5**
	200	**0.83**	**0.80**	**79.8**	**75.2**
	500	**0.76**	**0.72**	**76.8**	**71.3**
	1000	**0.66**	**0.61**	**70.2**	**65.9**

Table 19: Short-horizon score fidelity on Llama-3.2-3B, averaged over $K \in \{10, 50, 200\}$.

Variant	Data ρ	Param ρ	Data Pair	Param Pair
Init + First-Order	0.33	0.26	61.0	58.2
Init + Diag-Second-Order	0.43	0.35	66.5	62.7
Warmup + First-Order	0.62	0.54	75.0	70.2
Warmup + Diag-Second-Order	**0.80**	**0.76**	**80.6**	**76.0**

corresponding checkpoint, we run K restricted AdamW steps and compare predicted rankings with realized validation gains $\Delta_K(S) = \mathcal{L}_{\text{val}}(\bar{\theta}) - \mathcal{L}_{\text{val}}(\theta_K(S))$ using Spearman correlation and pairwise agreement. $K = 1000$ serves as a stress test for the local surrogate.

Per-K results. Table 18 shows the expected horizon decay: fidelity decreases as optimization moves away from $\bar{\theta}$. Warmup improves over initialization, diagonal second-order scoring improves over first-order scoring, and Warmup + Diag-Second-Order remains the most predictive variant.

Additional domain. Table 19 repeats the short-horizon analysis on MetaMathQA. The same ordering holds, with Warmup + Diag-Second-Order giving the strongest agreement with realized gains.

These diagnostics show that warmup and diagonal second-order scoring improve ranking fidelity, with fidelity declining over longer optimization horizons.

D.10 Fisher-Based Parameter Selection Controls

Empirical Fisher (EF) averages squared per-example gradients; MB-Fisher averages squared mini-batch gradients. FISH Mask [90] computes EF before warmup, while EF-W and EF-W+LR compute it after warmup. MB-Fisher-1K and MB-Fisher-Full also score after warmup, using the shared subset and full scoring pool, respectively. Fisher controls retain the highest-scoring coordinates in fixed masks. DualSFT-Param-FO removes the diagonal correction from DualSFT-Param.

FISH Mask and EF-W use a development-selected learning rate of 3×10^{-5}. EF-W+LR and all remaining variants use DualSFT's 2×10^{-5}, providing matched-learning-rate comparisons.

D.11 Sensitivity to the Validation Set Driving Selection

For the validation-size analysis in Fig. 5, we construct a dedicated 2,048-example reservoir disjoint from warmup, anchor, candidate scoring, and final-training data. Apart from this controlled split, settings follow the main experiments, with training seed 42 and data/parameter budgets of 10%/5%. Validation subsets of sizes $\{128, 256, 512, 1024, 2048\}$ are nested within one fixed reservoir permu-

Table 20: Comparison with forgetting-specific methods on Magicoder (%). Pre-trained, Standard SFT, and LoRA are excluded from ranking. **Best** , second-best , and third-best are highlighted.

Model	Method	Stability ↑					Plasticity ↑	Overall ↑
		ARC-C	PIQA	MMLU	GSM8K	Avg	HE	Score
Llama-3.2-3B	Pre-trained	$43.00_{\pm 0.00}$	$76.61_{\pm 0.00}$	$56.52_{\pm 0.00}$	$27.45_{\pm 0.00}$	$50.89_{\pm 0.00}$	$28.66_{\pm 0.00}$	$39.78_{\pm 0.00}$
	Standard SFT	$39.99_{\pm 1.46}$	$72.09_{\pm 0.98}$	$48.18_{\pm 0.39}$	$20.70_{\pm 0.85}$	$45.24_{\pm 0.50}$	$44.72_{\pm 0.57}$	$44.98_{\pm 0.37}$
	LoRA ICLR'22	$42.61_{\pm 1.45}$	$74.43_{\pm 0.99}$	$51.87_{\pm 0.41}$	$23.30_{\pm 0.83}$	$48.05_{\pm 0.49}$	$41.46_{\pm 0.50}$	$44.76_{\pm 0.37}$
	DavIR ACL'25	$42.72_{\pm 1.45}$	$74.92_{\pm 0.99}$	$53.09_{\pm 0.40}$	$24.64_{\pm 0.75}$	$48.84_{\pm 0.49}$	$43.29_{\pm 0.86}$	$46.07_{\pm 0.47}$
	FLOW ICML'25	**$42.95_{\pm 1.44}$**	$76.22_{\pm 0.98}$	**$56.91_{\pm 0.40}$**	$27.27_{\pm 0.72}$	**$50.84_{\pm 0.49}$**	$42.28_{\pm 0.76}$	$46.56_{\pm 0.47}$
	STM NIPS'25	$42.38_{\pm 1.46}$	$76.12_{\pm 0.99}$	$56.04_{\pm 0.41}$	$27.09_{\pm 0.78}$	$50.41_{\pm 0.49}$	$42.68_{\pm 0.50}$	$46.54_{\pm 0.41}$
	NanoAdam NeurIPS'25	$42.66_{\pm 1.44}$	$74.68_{\pm 0.98}$	$52.71_{\pm 0.40}$	$24.39_{\pm 0.73}$	$48.61_{\pm 0.48}$	$43.50_{\pm 0.76}$	$46.05_{\pm 0.50}$
	FAPM EMNLP'25	$42.55_{\pm 1.45}$	$76.01_{\pm 0.98}$	$53.86_{\pm 0.40}$	$26.00_{\pm 0.74}$	$49.60_{\pm 0.49}$	$41.06_{\pm 1.04}$	$45.33_{\pm 0.52}$
	TALR ICLR'26	$42.49_{\pm 1.45}$	$76.33_{\pm 0.99}$	$56.41_{\pm 0.40}$	**$27.55_{\pm 0.76}$**	$50.69_{\pm 0.50}$	$44.11_{\pm 0.76}$	**$47.40_{\pm 0.46}$**
	DualSFT-Param (Ours)	$42.55_{\pm 1.46}$	$76.19_{\pm 0.99}$	$54.01_{\pm 0.41}$	$26.28_{\pm 0.75}$	$49.76_{\pm 0.50}$	**$44.51_{\pm 0.86}$**	$47.13_{\pm 0.42}$
	DualSFT-Data (Ours)	$42.66_{\pm 1.44}$	**$76.44_{\pm 0.99}$**	$55.28_{\pm 0.40}$	$26.54_{\pm 0.73}$	$50.23_{\pm 0.50}$	$44.11_{\pm 0.76}$	$47.17_{\pm 0.43}$
	DualSFT (Ours)	**$42.95_{\pm 1.45}$**	$76.22_{\pm 0.98}$	$56.09_{\pm 0.40}$	$26.89_{\pm 0.70}$	$50.54_{\pm 0.47}$	$43.70_{\pm 0.76}$	$47.12_{\pm 0.43}$
Gemma-3-4B-PT	Pre-trained	$51.45_{\pm 0.00}$	$79.16_{\pm 0.00}$	$59.61_{\pm 0.00}$	$37.00_{\pm 0.00}$	$56.80_{\pm 0.00}$	$35.37_{\pm 0.00}$	$46.09_{\pm 0.00}$
	Standard SFT	$47.16_{\pm 1.46}$	$76.39_{\pm 0.99}$	$53.87_{\pm 0.40}$	$24.03_{\pm 0.81}$	$50.36_{\pm 0.50}$	$58.13_{\pm 0.76}$	$54.25_{\pm 0.51}$
	LoRA ICLR'22	$49.15_{\pm 1.46}$	$75.57_{\pm 1.00}$	$56.71_{\pm 0.39}$	$31.64_{\pm 0.82}$	$53.27_{\pm 0.49}$	$54.67_{\pm 1.04}$	$53.97_{\pm 0.52}$
	DavIR ACL'25	$50.26_{\pm 1.45}$	$78.04_{\pm 0.96}$	$58.01_{\pm 0.40}$	$33.71_{\pm 0.75}$	$55.00_{\pm 0.48}$	$56.30_{\pm 0.76}$	$55.65_{\pm 0.47}$
	FLOW ICML'25	$50.91_{\pm 1.44}$	$78.82_{\pm 0.94}$	$58.92_{\pm 0.39}$	**$36.82_{\pm 0.72}$**	**$56.37_{\pm 0.49}$**	$55.69_{\pm 0.76}$	$56.03_{\pm 0.47}$
	STM NIPS'25	$50.63_{\pm 1.46}$	$78.47_{\pm 0.98}$	$58.58_{\pm 0.41}$	$36.09_{\pm 0.78}$	$55.94_{\pm 0.50}$	$55.89_{\pm 0.76}$	$55.92_{\pm 0.45}$
	NanoAdam NIPS'25	$50.77_{\pm 1.45}$	$78.20_{\pm 0.94}$	$58.33_{\pm 0.39}$	$33.13_{\pm 0.84}$	$55.11_{\pm 0.49}$	$56.30_{\pm 1.04}$	$55.70_{\pm 0.51}$
	FAPM EMNLP'25	$50.60_{\pm 1.44}$	$78.49_{\pm 0.95}$	$58.60_{\pm 0.40}$	$35.91_{\pm 0.75}$	$55.90_{\pm 0.49}$	$52.24_{\pm 1.04}$	$54.07_{\pm 0.53}$
	TALR ICLR'26	$50.97_{\pm 1.45}$	$78.84_{\pm 0.97}$	$58.94_{\pm 0.38}$	$36.69_{\pm 0.76}$	$56.36_{\pm 0.49}$	**$57.52_{\pm 0.76}$**	**$56.94_{\pm 0.49}$**
	DualSFT-Param (Ours)	**$51.14_{\pm 1.45}$**	$78.55_{\pm 0.98}$	$58.14_{\pm 0.39}$	$36.34_{\pm 0.83}$	$56.04_{\pm 0.49}$	**$57.52_{\pm 0.76}$**	$56.78_{\pm 0.48}$
	DualSFT-Data (Ours)	$51.02_{\pm 1.45}$	$78.53_{\pm 0.98}$	$58.33_{\pm 0.40}$	$36.54_{\pm 0.79}$	$56.11_{\pm 0.49}$	$56.91_{\pm 0.76}$	$56.51_{\pm 0.42}$
	DualSFT (Ours)	**$51.14_{\pm 1.44}$**	**$79.02_{\pm 0.95}$**	**$59.16_{\pm 0.38}$**	$35.61_{\pm 0.64}$	$56.23_{\pm 0.48}$	$56.71_{\pm 0.86}$	$56.47_{\pm 0.43}$
Qwen-3.5-9B-Base	Pre-trained	$54.27_{\pm 0.00}$	$80.09_{\pm 0.00}$	$72.62_{\pm 0.00}$	$86.20_{\pm 0.00}$	$73.29_{\pm 0.00}$	$60.98_{\pm 0.00}$	$67.13_{\pm 0.00}$
	Standard SFT	$49.23_{\pm 1.46}$	$74.30_{\pm 0.98}$	$66.14_{\pm 0.40}$	$78.65_{\pm 0.75}$	$67.08_{\pm 0.49}$	$75.61_{\pm 0.86}$	$71.34_{\pm 0.49}$
	LoRA ICLR'22	$51.51_{\pm 1.46}$	$76.62_{\pm 0.95}$	$69.47_{\pm 0.39}$	$81.93_{\pm 0.73}$	$69.88_{\pm 0.48}$	$70.33_{\pm 0.76}$	$70.10_{\pm 0.50}$
	DavIR ACL'25	$52.39_{\pm 1.45}$	$78.42_{\pm 0.95}$	$70.02_{\pm 0.39}$	$83.60_{\pm 0.72}$	$71.11_{\pm 0.48}$	$73.37_{\pm 0.76}$	$72.24_{\pm 0.47}$
	FLOW ICML'25	$53.21_{\pm 1.44}$	$79.56_{\pm 0.94}$	$70.62_{\pm 0.38}$	$85.29_{\pm 0.70}$	$72.17_{\pm 0.47}$	$73.17_{\pm 0.86}$	$72.67_{\pm 0.48}$
	STM NIPS'25	$52.79_{\pm 1.46}$	$79.18_{\pm 0.96}$	$70.38_{\pm 0.40}$	$84.56_{\pm 0.74}$	$71.73_{\pm 0.50}$	$73.17_{\pm 0.86}$	$72.45_{\pm 0.46}$
	NanoAdam NIPS'25	$52.87_{\pm 1.45}$	$79.04_{\pm 0.95}$	$70.22_{\pm 0.39}$	$84.23_{\pm 0.80}$	$71.59_{\pm 0.49}$	$73.37_{\pm 0.76}$	$72.48_{\pm 0.46}$
	FAPM EMNLP'25	$52.87_{\pm 1.44}$	$79.49_{\pm 0.94}$	$70.18_{\pm 0.39}$	$83.88_{\pm 0.71}$	$71.60_{\pm 0.48}$	$70.93_{\pm 1.04}$	$71.27_{\pm 0.48}$
	TALR ICLR'26	$53.44_{\pm 1.45}$	$79.71_{\pm 0.95}$	$70.72_{\pm 0.38}$	**$85.72_{\pm 0.72}$**	**$72.40_{\pm 0.49}$**	**$75.41_{\pm 0.76}$**	**$73.90_{\pm 0.49}$**
	DualSFT-Param (Ours)	$53.53_{\pm 1.44}$	$78.73_{\pm 0.95}$	$70.26_{\pm 0.39}$	$84.23_{\pm 0.70}$	$71.69_{\pm 0.47}$	$75.20_{\pm 0.76}$	$73.44_{\pm 0.44}$
	DualSFT-Data (Ours)	$53.67_{\pm 1.45}$	**$80.03_{\pm 0.95}$**	**$70.75_{\pm 0.39}$**	$85.04_{\pm 0.70}$	$72.37_{\pm 0.48}$	$74.19_{\pm 0.57}$	$73.28_{\pm 0.41}$
	DualSFT (Ours)	**$54.01_{\pm 1.43}$**	$79.63_{\pm 0.94}$	$70.33_{\pm 0.38}$	$84.61_{\pm 0.68}$	$72.15_{\pm 0.47}$	$73.78_{\pm 0.50}$	$72.96_{\pm 0.39}$

Table 21: Validation-composition and sampling sensitivity (%) at 1,024 examples. Overlaps reference Natural with validation seed 0. Sampling variants preserve Natural cluster counts.

Setting	Data overlap	Param. overlap	Stability	Plasticity	Overall
Natural (seed 0)	100.00	100.00	50.43	43.90	47.17
Balanced	93.18	95.62	50.52	43.90	47.21
Skewed	86.74	90.81	50.74	43.29	47.02
Natural, seed 1	92.76	95.41	50.69	43.29	46.99
Natural, seed 2	91.83	94.72	50.49	43.90	47.20

tation, using 1,024 examples as reference. For equal-size selected supports S_a and S_a^{ref}, we define

$$\mathrm{Overlap}_a = 100\,|S_a \cap S_a^{\mathrm{ref}}|/|S_a^{\mathrm{ref}}|, \qquad a \in \{\mathcal{D}, \theta\}. \tag{72}$$

This overlap is normalized by the selection budget, rather than the union used by Jaccard similarity.

At validation size 1,024, we apply K-means ($K = 4$, seed 0) to mean-pooled, ℓ_2-normalized prompt embeddings from frozen Llama-3.2-3B. Reservoir cluster sizes are $(612, 544, 481, 411)$; Natural, Balanced, and Skewed draws use counts $(297, 280, 235, 212)$, $(256, 256, 256, 256)$, and $(512, 194, 171, 147)$, respectively. Sampling variants preserve Natural cluster counts and independently resample within clusters using validation seeds 1 and 2, while training seed remains 42. Table 21 reports the complete results against the shared Natural/seed-0 reference.

Limitations

DualSFT relies on a local validation-improvement surrogate for tractable data–parameter selection. Separate Top-b/Top-k selection carries no guarantee of joint optimality over data subsets and parameter masks. Its diagonal second-order score uses an optimizer-aware curvature proxy instead of the full Hessian, and score fidelity may weaken as optimization moves far from the warmup checkpoint.

Our experiments cover 3B–9B LLMs; extensions to larger models, longer adaptive horizons, and preference-based post-training remain important future directions. For very large models, parameter-side scoring may also require block-wise or sharded implementations because it maintains linear full-vector state.

Broader Impacts

DualSFT reduces the resource cost of LLM fine-tuning by selecting both data and parameter subsets, which can lower compute requirements and make model adaptation more accessible. At the same time, more efficient fine-tuning may also lower the barrier to harmful, unsafe, or low-quality downstream adaptation. Our experiments use public datasets and standard benchmarks, and do not collect user data or involve human-subject studies. Deployments in sensitive domains should include task-specific safety evaluation, data-quality review, and misuse-risk assessment.

Reproducibility

Our code is publicly available for reproducing the main experiments. All datasets and base models used in this work are publicly available.

Licenses and assets

We use publicly available datasets, benchmarks, and base models, following their respective licenses and terms of use. The proposed method does not introduce a new dataset. We do not release new model checkpoints; the released assets consist of code, configurations, and scripts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims match the proposed method, theory, and empirical results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses local-surrogate scope, proxy approximations, and future extensions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.

- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions and complete derivations are provided in the main text and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Datasets, models, metrics, splits, budgets, and training settings are specified.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.

- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and reproduction instructions are provided with public data and models.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper reports training, evaluation, baselines, and method-specific hyperparameters.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Main results report mean and standard deviation over three random seeds.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Hardware, runtime, memory, and selection costs are reported.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The work uses public assets and does not involve private or human-subject data.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses efficient adaptation benefits and potential misuse risks.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new high-risk models or scraped datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing datasets, models, benchmarks, and tools are cited with license information.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code is documented with reproduction instructions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects or personally identifiable information.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Public LLMs are used as fine-tuning backbones and frozen references.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.