SAFE-DRIFT: Data Selection for Supervised Fine-tuning with Controllable Off-Target Drifts

Yeo Jin Jung*
Department of Statistics
University of Chicago
yeojinjung@uchicago.edu

Yating Liu*
Department of Statistics
University of Chicago
yatingliu@uchicago.edu

Lalchand Pandia
Department of Computer Science
University of Chicago
lcpandia@uchicago.edu

Claire Donnat
Department of Statistics
University of Chicago
cdonnat@uchicago.edu

Abstract

We propose SAFE-DRIFT, a data-selection framework for supervised fine-tuning that improves a target capability while controlling unwanted drift in the model's other behaviors. We formalize this as a constrained optimization problem that maximizes alignment with the target gradient, subject to a budget on off-target drift which we quantify using the Fisher information of a reference distribution. Under this objective, the optimal, closed-form update is a damped natural-gradient step that downweights reference-sensitive directions. To estimate these high-dimensional gradients and work within the right geometry, we further derive a scalable algorithm that approximates the Fisher information and the target gradient in a shared low-dimensional subspace and analyze how entanglement of the target and reference task can affect the update. We evaluate SAFE-DRIFT against state-of-the-art data selection methods across diverse domains including coding and medical question answering, yielding competitive target performance while reducing degradation on specified off-target behaviors.

1 Introduction

Supervised fine-tuning (SFT) is an effective technique to adapt large language models (LLMs) to specialized tasks. In many of such tasks, a critical bottleneck is the scarcity of high-quality examples that precisely define the desired behavior. For instance, in medicine, clinician-validated cases may serve as gold-standard target examples; larger biomedical corpora may provide a much wider but noisier training pool [1, 2]. Similarly, in law, a small number of expert-annotated examples may define the target task, while larger collections of legal documents may offer a broader but less aligned training pool [3–5]. In all of these examples, estimates of the target distribution are high-value and expensive to label, whereas the available training pool is broader, cheaper, and only partially aligned with the task of interest. This pool of candidate examples, which we refer to as the candidate pool, can provide a useful training signal, but training on all of it is not necessarily safe or effective. Carefully selected subsets have indeed been shown to outperform training on the full candidate pool in various settings [6–9] (see Table 2 for additional examples).

At the same time, fine-tuning does not solely affect the capability we intend to improve [10]. Growing evidence shows that seemingly benign finetuning can weaken safety alignment and degrade behavior outside the target task [11, 12]. Beyond safety, fine-tuning can also lead to the forgetting of more general capabilities [13]. We refer to these unintended changes as *off-target drifts*.

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

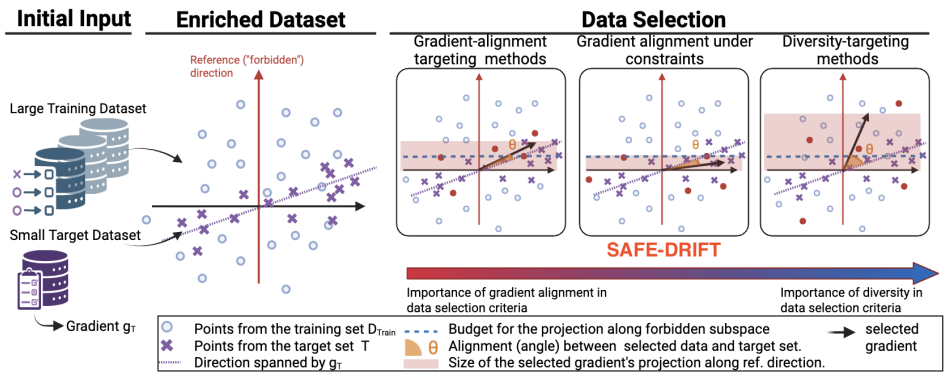


Figure 1: *Trade-offs in data selection: gradient alignment selects candidate gradients aligned with the task but can cause substantial off-target drift (e.g., movement along a "forbidden dimension", in red). Diversity-based selection can mitigate this but does not reliably prevent off-target drift.* SAFE-DRIFT *combines both approaches by constraining updates along the forbidden dimension.*

These observations suggest that fine-tuning is not just a matter of collecting more data, but selecting the right ones: given a large and noisy candidate pool, we must choose examples that improve the target task without inducing unnecessary off-target drift. The key question is therefore: ***which examples from the candidate pool should we train on in order to improve the target task while safeguarding other model behaviors?***

Existing data selection methods broadly fall into two (and often mutually exclusive) families. The first selects candidate examples that appear most valuable for the target task, either through explicit quality proxies [14–17] or through direct influence estimation [6, 18]. The second prioritizes diversity or distributional coverage of selected candidates to improve generalization [19, 20]. We refer the reader to [21] for a comprehensive review. None of these methods explicitly constrain the off-target drift induced by the selected subset, nor do they account for the possible entanglement between target-improving and reference-sensitive parameter directions [22–25].

If the gradients that yield the largest improvement in the target task overlap with directions associated with worse performance on a reference task, then aggressively selecting them can induce substantial off-target drift. Diversity-based selection, on the other hand, spreads the update across many gradient directions, but may not deliver the best target performance. Optimal data selection must therefore navigate potential tensions between target improvement and reference degradation. Here, we propose an algorithm that explicitly models data selection as a constrained optimization problem to limit unwanted movement along reference-sensitive directions, as illustrated in Figure 1.

Contributions. We summarize our contributions as follows:

- **A drift-constrained formulation of data selection.** We introduce SAFE-DRIFT, which selects data that maximizes target gradient alignment subject to an explicit budget on off-target drift, quantified by the Fisher information of a user-specified reference distribution. (Section 2).
- **A connection between target-aligned and diversity-based selection.** We show that the optimal update under a reference-drift budget is a damped natural-gradient direction. This clarifies when target-aligned selection cannot safeguard against off-target drift, when diversity can reduce the drift, and how both arise as limiting or special cases of a single constrained update problem (Section 2.1).
- **A scalable low-rank implementation.** We build a shared low-rank basis that merges leading reference-sensitive directions with residual task-relevant directions (Section 2.3) and work with gradients projected onto the derived low-rank basis, making the selection algorithm practical. We also propose diagnostics of target–reference entanglement using the low-rank projections and analyze how it affects the off-target drift control.
- **Experiments demonstrating controllable drift.** Across four settings—code generation, medical question answering, emotional support dialogues, and instruction-following experiments—SAFE-DRIFT achieves competitive target gains while reducing off-target drift on the specified reference behaviors. We provide further robustness and ablation results on the off-target drift budget and different design choices (Section 3).

2 Method

Consider a base language model with parameters $\theta_0 \in \mathbb{R}^d$ that defines a conditional distribution over output tokens $y \in \mathcal{Y}$ given input tokens $x \in \mathcal{X}$. Let $z = (x, y)$ denote an input–output pair, and let $\ell(z; \theta)$ denote the training loss (typically the negative log-likelihood of the output sequence, or of a single answer token in a multiple-choice setting). We assume that we are given a large candidate training pool $\mathcal{D}_{\text{train}} = \{z_i\}_{i=1}^{N_C}$ and a small target set $\mathcal{D}_{\text{tar}} = \{z_j\}_{j=1}^{N_T}$ with $N_T \ll N_C$ that encodes the capability we wish to improve. We are also given a reference set $\mathcal{D}_{\text{ref}} = \{x_i\}_{i=1}^{N_R}$ of prompts from a reference distribution $\mathcal{R}$ over inputs, on which the model's output distribution should be preserved. Let the target risk at parameter $\theta \in \mathbb{R}^d$ be defined as:

$$L_{\mathcal{T}}(\theta) := \mathbb{E}_{z\sim\mathcal{T}}\big[\ell(z;\theta)\big] = \mathbb{E}_{x\sim\mathcal{T}_X}\mathbb{E}_{y\sim q_{\mathcal{T}}(\cdot|x)}\big[\ell((x,y);\theta)\big], \tag{1}$$

where $\mathcal{T}_X$ is the marginal of the target distribution $\mathcal{T}$ on inputs and $q_{\mathcal{T}}(\cdot \mid x)$ denotes the gold-standard conditional distribution. Our goal is to select a subset $\mathcal{S} \subset \mathcal{D}_{\text{train}}$ of size S such that fine-tuning on $\mathcal{S}$ minimizes the target risk $L_{\mathcal{T}}(\theta)$ while preserving the model's output distribution on $\mathcal{R}$.

We proceed in two stages. We first characterize the optimal local update, which minimizes $L_{\mathcal{T}}(\theta)$ while preserving behavior under $\mathcal{R}$. We then identify a subset $\mathcal{S} \subset \mathcal{D}_{\text{train}}$ of size S whose aggregate effect most closely matches the optimal update. The model is fine-tuned on $\mathcal{S}$ in one pass, without any online reweighting or iterative refinement. The first subsection characterizes the oracle update, assuming access to the full target and reference populations. We then derive the subset-selection algorithm and its low-rank implementation in the noisy empirical setting using the finite datasets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{tar}}$, and $\mathcal{D}_{\text{ref}}$.

2.1 Optimal parameter update under a drift budget

We begin with a population-level oracle problem in which both the population gradient of the target loss $g_{\mathcal{T}} = \mathbb{E}[\nabla_\theta L_{\mathcal{T}}(\theta_0)]$ and the reference distribution $\mathcal{R}$ are known exactly. Here $\theta_0 \in \mathbb{R}^d$ denotes the base model parameters, and the fine-tuned model is parametrized by $\theta = \theta_0 + \Delta\theta$ for some update $\Delta\theta \in \mathbb{R}^d$. The oracle question we ask is: *what is the optimal parameter update $\Delta\theta$ that maximizes the target performance while safeguarding against off-target drift?*

Off-target drift on reference distribution. For the fine-tuned model $\theta_0 + \Delta\theta$, define the expected drift on the reference distribution as the Kullback-Leibler divergence with the base model,

$$D_{\mathcal{R}}(\Delta\theta) := \mathbb{E}_{x\sim\mathcal{R}}\left[D_{\text{KL}}\big(p_{\theta_0}(\cdot \mid x) \,\|\, p_{\theta_0+\Delta\theta}(\cdot \mid x)\big)\right]. \tag{2}$$

The quantity $D_{\mathcal{R}}(\Delta\theta)$ captures the off-target effects of fine-tuning: how much the parameter update $\Delta\theta$ alters model behavior on $\mathcal{R}$, outside the target task.We note here that $D_{\mathcal{R}}(\Delta\theta)$ is defined globally for any $\Delta\theta$. However, because our results consider small fine-tuning updates, it is natural to study the local behavior of $D_{\mathcal{R}}(\Delta\theta)$ around $\Delta\theta = 0$. We next show that, to second order, off-target drift is a quadratic form induced by the Fisher information of the model on the reference distribution.

Definition 1 (Reference Fisher)**.** Let $s^R_{\theta_0}(x, y) := \nabla_\theta \log p_{\theta_0}(y \mid x)$. The *reference Fisher information* at θ_0 is:

$$\mathbf{F}_{\mathcal{R}} := \mathbb{E}_{x\sim\mathcal{R}}\mathbb{E}_{y\sim p_{\theta_0}(\cdot|x)}\big[s^R_{\theta_0}(x,y)s^R_{\theta_0}(x,y)^\top\big]. \tag{3}$$

Proposition 1 (Local reference behavior)**.** *Assume a common parameter-independent support and standard smoothness and integrability conditions so that differentiation may be exchanged with expectation and with the normalization integral near θ_0. Then, as $\Delta\theta \to 0$,*

$$D_{\mathcal{R}}(\Delta\theta) = \frac{1}{2}\Delta\theta^\top \mathbf{F}_{\mathcal{R}}\Delta\theta + o\big(\|\Delta\theta\|_2^2\big). \tag{4}$$

The smoothness conditions and proof are provided in Appendix B.1. Proposition 1 shows that, for small fine-tuning updates, the cost $D_{\mathcal{R}}(\Delta\theta)$ is estimated as a quadratic form with the curvature given by $\mathbf{F}_{\mathcal{R}}$, rather than the classic Euclidean norm $\|\Delta\theta\|_2^2$: directions with large curvature under $\mathbf{F}_{\mathcal{R}}$ induce larger second-order changes in the model's output distribution on $x \sim \mathcal{R}$, whereas directions lying near the null space of $\mathbf{F}_{\mathcal{R}}$ have small quadratic off-target drift.

Target gain as gradient alignment. Recall the population gradient loss of the target at the base model θ_0: $g_{\mathcal{T}} := \nabla_\theta L_{\mathcal{T}}(\theta_0)$. For a small update $\Delta\theta$, a first order Taylor approximation at θ_0 gives,

$$L_{\mathcal{T}}(\theta_0 + \Delta\theta) - L_{\mathcal{T}}(\theta_0) = g_{\mathcal{T}}^\top \Delta\theta + o(\|\Delta\theta\|_2). \tag{5}$$

Under a Euclidean norm bound and $g_{\mathcal{T}} \neq 0$, the update providing the steepest decrease in target loss is proportional to $g_{\mathcal{T}}$. However, this strategy might induce updates in the space spanned by $\mathbf{F}_{\mathcal{R}}$, creating a potential trade-off between maximizing target accuracy, and minimizing off-target drift.

The oracle update problem. We construct a bridge of data selection by formulating it as a constrained optimization problem of $\Delta\theta$: we maximize the target-side gain $-g_{\mathcal{T}}^\top \Delta\theta$, while allowing only a small budget of quadratic off-target drift $\frac{1}{2}\Delta\theta^\top \mathbf{F}_{\mathcal{R}} \Delta\theta$ on the reference distribution. This leads to the following oracle problem: *among all small parameter updates, choose the one that maximizes target gain while spending at most ρ, a fixed drift budget:*

$$\Delta\theta^\star \in \arg\max_{\Delta\theta \in \mathbb{R}^d} \left\{ -g_{\mathcal{T}}^\top \Delta\theta \right\} \quad \text{subject to} \quad \frac{1}{2}\Delta\theta^\top \mathbf{F}_{\mathcal{R}} \Delta\theta \leq \rho, \qquad \|\Delta\theta\|_2 \leq \epsilon. \tag{6}$$

There are two central parameters: $\rho > 0$, which sets the budget on quadratic off-target drift, and $\epsilon > 0$ which constrains the magnitude of the update.

The constraint ϵ serves two purposes: it ensures that $\Delta\theta$ is small enough for the local approximations in 4 and 5 to hold, and it regularizes the update in directions where $g_{\mathcal{T}}$ lies in the null space of $\mathbf{F}_{\mathcal{R}}$. In practice, a fixed optimizer and its learning rate η can determine ϵ. For stochastic gradient descent, for instance, we have $\|\Delta\theta\|_2 \leq \eta/\sqrt{S}\|\mathbf{G}\|_{\text{op}}$ where S is the selection budget and $\|\mathbf{G}\|_{\text{op}}$ is the spectral norm of the candidate gradient matrix, both quantities that are fixed characteristics of the problem.

With ϵ fixed, the sole quantity in 6 that must be specified is ρ, the maximum off-target drift we are willing to tolerate. To make this quantity interpretable, let $B_x \subseteq \mathcal{Y}$ be a set of undesirable outputs for prompt x, for instance, a set of toxic or stereotyped responses. By Pinsker's and Jensen's inequalities,

$$\mathbb{E}_{x\sim\mathcal{R}}[(p_\theta(B_x \mid x) - p_{\theta_0}(B_x \mid x))_+] \leq \sqrt{D_{\mathcal{R}}(\Delta\theta)/2}.$$

For a local oracle update, Appendix B.1 gives $D_{\mathcal{R}}(\Delta\theta) \leq \rho + M\epsilon^3/6$, where M is the expected third-derivative envelope and ϵ lies within the stated neighborhood. Therefore, for any undesirable reference output $B_x \subseteq \mathcal{Y}$, the expected increase in the probability of B_x is bounded by $\sqrt{(\rho + M\epsilon^3/6)/2}$. In practice, we treat ρ as a hyperparameter that requires tuning to efficiently safeguard against deleterious off-target drifts; we ablate over a range of ρ's in Section 3. We now provide the optimal oracle solution of 6.

Proposition 2 (Oracle Update $\Delta\theta^*$)**.** *Assume the target gradient vector satisfies $g_{\mathcal{T}} \neq 0$. When both the drift and the norm constraint multipliers are strictly positive, the solution to* (6) *is,*

$$\Delta\theta^\star = -\epsilon \frac{(\mathbf{F}_{\mathcal{R}} + \alpha\mathbf{I})^{-1} g_{\mathcal{T}}}{\|(\mathbf{F}_{\mathcal{R}} + \alpha\mathbf{I})^{-1} g_{\mathcal{T}}\|_2},$$

where $\alpha > 0$ is chosen to satisfy $\frac{1}{2}(\Delta\theta^\star)^\top \mathbf{F}_{\mathcal{R}} \Delta\theta^\star = \rho$.

The remaining boundary and singular Fisher cases are characterized in Appendix B.2; a tight constraint can have zero multiplier and drift-only solution may include null space components. In the strictly positive-multiplier regime, where both constraints are active, the maximum-gain direction is proportional to the damped natural gradient of the target objective under the reference distribution.

A complete characterization of all cases, along with proofs and an explicit formula for α as a function of (ϵ, ρ), is provided in Appendices B.2 and B.3.

Remark 1 (Connection to existing data selection methods)**.** Proposition 2 places existing data selection methods within a common framework and helps explain the conditions under which each paradigm is preferred. When $\rho \to \infty$, the drift constraint is never active and the optimal update reduces to $\Delta\theta^\star \propto -g_{\mathcal{T}}$, recovering pure gradient alignment with the target—the principle underlying LESS [6]. This connection also reveals when target-aligned selection can be insufficient: if g_T has large components along the leading eigenvectors of $\mathbf{F}_{\mathcal{R}}$, which we define as *entanglement*, then selecting based on gradient alignment with $g_{\mathcal{T}}$ necessarily induces large off-target drift. In this regime, diversity-based selection can be preferable as the parameter update is not necessarily concentrated along g_T and may avoid reference-sensitive directions.

2.2 Algorithm: Safe-Drift Data Selection

Let $w \in \{0,1\}^{N_C}$ denote the selection vector, where $w_i = 1$ indicates that candidate example i is selected for fine-tuning. Let $\mathbf{G} \in \mathbb{R}^{d \times N_C}$ be the matrix of candidate loss gradients at θ_0. We approximate the subset-induced parameter update by one step with fixed gradients and preconditioner: $\Delta\theta(w) = -\eta \mathbf{P}\mathbf{G}w / |\sum_i w_i|$, where $\mathbf{P} \in \mathbb{R}^{d \times d}$ denotes the preconditioner induced by the optimizer and $\eta > 0$ is the learning rate. For example, $\mathbf{P} = \mathbf{I}_d$ recovers stochastic gradient descent, while adaptive methods such as Adam [26] use a diagonal second-order approximation for local curvature. We defer a further discussion of the choice of $\mathbf{P}$ to Appendix B.4. In our setting, the goal is to approximate the optimal safe update $\Delta\theta^\star$ in Proposition 2 that maximizes the target gain while satisfying the constraint on off-target drift. Accordingly, we aim to construct a subset of exactly S examples by,

$$w^* = \arg \min_{w \in \{0,1\}^{N_C},\, \mathbf{1}^\top w = S} \|\Delta\theta(w) - \Delta\theta^\star\|^2_{\mathbf{F}_\mathcal{R} + \alpha\mathbf{I}}, \tag{7}$$

where $\|x\|^2_{\mathbf{F}_\mathcal{R}+\alpha\mathbf{I}} := x^\top(\mathbf{F}_\mathcal{R} + \alpha\mathbf{I})x$ for a vector $x \in \mathbb{R}^d$ with $\alpha > 0$.

To approximately solve (7) in practice, we use a greedy forward selector, following the common use of greedy matching procedures in sparse approximation [27] and gradient-based subset selection [28]. We detail the full pipeline in Algorithm 1 below and Algorithm 3 in Appendix C.2.

Remark 2. We carry out the reconstruction within the geometry defined by the reference set, characterized as the norm $\|\cdot\|^2_{\mathbf{F}_\mathcal{R}+\alpha\mathbf{I}}$ in (7), which enables us to assign higher penalties to reconstructions that have larger components along costly off-target directions. In Appendix D.7.1, we show that this approach yields stronger safeguarding mechanisms than reconstructions based on the standard Euclidean norm.

2.3 Low-rank estimation of Fisher information and gradients

We turn to the problem of estimating the gradients of the candidates and target, as well as the reference Fisher information required to solve problems (6) and (7). In a very high-dimensional parameter space, plug-in estimation in the full parameter space is computationally unstable, memory-intensive and noisy. In particular, the empirical reference Fisher is rank-deficient whenever the number of reference samples is smaller than the number of trainable parameters. Luckily, prior works have shown that language-model updates often lie in a low-dimensional subspace or admit a low-rank update structure [29–31]. Our goal is therefore to construct a low-rank projection, $\widehat{\mathbf{U}}_K \in \mathbb{R}^{d \times K}$ with $K \ll d$, that captures the reference-sensitive directions while also retaining the dominant task-relevant gradient structure.

Raw estimation. Define the estimated gradient by $\widehat{g}_i^{\text{raw}} := \nabla_\theta \ell(z_i; \theta_0)$. The target gradient is then estimated by the sample mean $\widehat{g}_{\text{tar}}^{\text{raw}} := \frac{1}{N_T} \sum_{z_i \in \mathcal{D}_{\text{tar}}} \nabla_\theta \ell(z_i; \theta_0)$. Let $\mathcal{D}_{\text{ref}} = \{x_j\}_{j=1}^{N_R}$ be the held-out set of reference inputs, and let the score vector for the reference data be denoted:

$$s_j^R := \nabla_\theta \log p_{\theta_0}(y_j \mid x_j) \in \mathbb{R}^d, \qquad y_j \sim p_{\theta_0}(\cdot \mid x_j).$$

Let $\mathbf{S}_R \in \mathbb{R}^{N_R \times d}$ denote the matrix of stacked score vectors.

Estimating the low-rank projection $\widehat{\mathbf{U}}_K$. We construct $\widehat{\mathbf{U}}_K$ in two stages.

Step 1: low-rank approximation of $\mathbf{F}_\mathcal{R}$. We estimate the leading eigenspace of $\mathbf{F}_\mathcal{R}$, and retain the top K_R strictly positive eigenpairs, $(\widehat{\mathbf{U}}_{K_R}, \widehat{\mathbf{\Lambda}}_{K_R})$, from the raw empirical reference Fisher $\widehat{\mathbf{F}}_R^{\text{raw}} = \frac{1}{N_R}\mathbf{S}_R^\top \mathbf{S}_R$[1].

Step 2: low-rank approximation of target and candidate gradients on the residual subspace. Since the target and candidate gradients may contain informative structure outside this reference eigenspace, we project them onto its orthogonal complement and extract a residual task basis from the projected gradients. Let $\widetilde{g}_i$ denote the residualized gradients:

$$\widetilde{g}_i := (\mathbf{I}_d - \widehat{\mathbf{U}}_{K_R}\widehat{\mathbf{U}}_{K_R}^\top)\widehat{g}_i^{\text{raw}}, \qquad \forall i \in [N_C + N_T]. \tag{8}$$

[1] In implementation, we never form the ambient $d \times d$ matrix explicitly. Instead, we work with the smaller $N_R \times N_R$ Gram matrix whose nonzero eigenvalues coincide with those of $\widehat{\mathbf{F}}_R^{\text{raw}}$ (see Appendix C.)

Stacking the residualized gradients gives $\widetilde{\mathbf{G}}_{TC} \in \mathbb{R}^{d\times(N_T+N_C)}$. We then estimate an additional residual basis by retaining the top K_T left singular vectors corresponding to top positive singular values of $\widetilde{\mathbf{G}}_{TC}$, yielding, $\widehat{\mathbf{U}}_{K_T} \in \mathbb{R}^{d\times K_T}$.

Stacking the two bases yields the final working basis $\widehat{\mathbf{U}}_K$, which by construction spans both the reference-sensitive and residual task-relevant directions:

$$\widehat{\mathbf{U}}_K := \begin{bmatrix}\widehat{\mathbf{U}}_{K_R} & \widehat{\mathbf{U}}_{K_T}\end{bmatrix} \in \mathbb{R}^{d\times K}, \qquad K := K_R + K_T, \qquad \widehat{\mathbf{U}}_K^\top \widehat{\mathbf{U}}_K = \mathbf{I}_K. \tag{9}$$

We evaluate the effect of the approximation of $\mathbf{F}_\mathcal{R}$ and the gradients under this low rank projection in Appendices D.1 and D.7.3. As expected, the suitability of the approximation depends on the spectrum of the eigenvalues of $\mathbf{F}_\mathcal{R}$ and the residual target gradient, and the projection error of $g_\mathcal{T}$. The selection of K is discussed in Appendix C.1 and D.7.3.

Entanglement diagnostics. Proposition 2 shows that how strongly the drift budget ρ affects the update depends on the target–reference *entanglement*. With the low-rank estimation described above, entanglement can be characterized as the overlap between the target-relevant eigenspace and the reference-sensitive eigenspace. We first illustrate the effect of entanglement in a simple synthetic experiment where we use a linear-Gaussian model with known exact per-example gradients and exact $\mathbf{F}_\mathcal{R} = \sigma^{-2}\mathbf{\Sigma}_\mathcal{R}$ (Appendix D.1), where $\mathbf{\Sigma}_\mathcal{R}$ has one expensive direction u_1 that costs $25\times$ more drift than u_2. The two regimes differ only in the target: in the *entangled* regime $-g_\mathcal{T}$ lies mostly along u_1, and in the *disentangled* regime $-g_\mathcal{T} = u_2$. In the entangled regime (Figure 2a), other baseline selectors incur drift cost, whereas SAFE-DRIFT achieves the lowest mean plotted drift at the cost of some target alignment. In the disentangled regime (Figure 2b) the budget does not bind and SAFE-DRIFT matches the target-aligned baseline LESS [6].

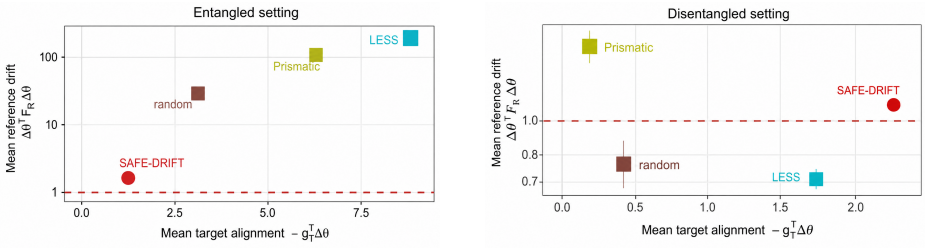


Figure 2: *Trade-off between target gradient alignment and off-target drift for an "entangled regime" and a "disentangled regime" in the synthetic linear experiment (Appendix D.1). The results are averaged over 100 runs. The red line denotes the budget ρ.* SAFE-DRIFT *is the only method able to control the drift.*

On real data, we measure entanglement using the working basis $\widehat{\mathbf{U}}_K$: we report the proportion of candidates with positive damped target alignment in both reference and residual blocks, denoted as $\pi_{Q_{\text{ent}}}$, while we use $\pi_{Q_{\text{clean}}}$ to indicate the proportion of candidates with positive residual alignment and non-positive reference-block alignment. We provide their definitions and discuss their interpretation in Appendix C.3.

2.4 Safe update $\widehat{\Delta}\theta^\star_K$ and selection $\hat{w}$ in practice.

The target and candidate coordinates in the common basis can be calculated by,

$$\widehat{b}_i^{(K)} := \widehat{\mathbf{U}}_K^\top \widehat{g}_i^{\text{raw}} \text{ for all } z_i \in \mathcal{D}_{\text{train}} \bigcup \mathcal{D}_{\text{tar}}, \qquad \widehat{b}_{\text{tar}}^{(K)} := \frac{1}{N_T} \sum_{i:z_i\in\mathcal{D}_{\text{tar}}} \widehat{b}_i^{(K)}. \tag{10}$$

Given $\alpha > 0$ and the scaling factor $\widehat{\mu}_K \geq 1$ that satisfy the reference and norm budgets in (6), which we detail in Appendix C.1, the practical safe update $\Delta\theta^*$ using the low rank basis is,

$$\widehat{\Delta}\theta^\star_K(\alpha) = -\frac{1}{\widehat{\mu}_K}\left(\text{diag}(\widehat{\mathbf{\Lambda}}_{K_R}, \mathbf{0}_{K_T}) + \alpha\mathbf{I}_K\right)^{-1} \widehat{\mathbf{U}}_K^\top \widehat{g}_{\text{tar}}^{\text{raw}} \in \mathbb{R}^K. \tag{11}$$

Define the reduced preconditioner $\mathbf{P}_K := \widehat{\mathbf{U}}_K^\top \mathbf{P} \widehat{\mathbf{U}}_K \in \mathbb{R}^{K \times K}$, where $\mathbf{P}$ is the fixed preconditioner. Similarly, for the selection vector $w \in \{0,1\}^{N_C}$ with $\mathbf{1}^\top w = S$, the projected-gradient surrogate update is,

$$\widehat{\Delta}\theta_K(w) = -\frac{\eta}{S} \mathbf{P}_K \mathbf{B}_K w \in \mathbb{R}^K, \qquad \mathbf{B}_K := \begin{bmatrix} \widehat{b}_1^{(K)} & \cdots & \widehat{b}_{N_C}^{(K)} \end{bmatrix}. \tag{12}$$

We can now plug in $\widehat{\Delta}\theta_K^\star(\alpha)$ and $\widehat{\Delta}\theta_K(w)$ in (7) to solve for the final selection $\hat{w}$. In practice, we turn the norm objective to a score for each candidate and select the top S points using the forward-greedy scores as detailed in Algorithm 1 and Appendix C.1.

Algorithm 1 SAFE-DRIFT data selection in the low-rank space

Input: Damping parameter $\alpha > 0$, target update $\widehat{\Delta}\theta_K^\star(\alpha)$, preconditioner $\mathbf{P}_K$, candidate gradients $\mathbf{B}_K$, learning rate η, subset budget $1 \le S \le N_C$, shortlist size $S \le m \le N_C$.

1: Form the reduced reference Fisher $\mathbf{M}_{\alpha,K} = \mathrm{diag}(\widehat{\mathbf{\Lambda}}_{K_R}, \mathbf{0}_{K_T}) + \alpha \mathbf{I}_K$.

2: Whiten the reduced objective via $\mathbf{M}_{\alpha,K}^{1/2}$, and set

$$\widetilde{\mathbf{\Delta}}_K \leftarrow -\frac{\eta}{S} \mathbf{M}_{\alpha,K}^{1/2} \mathbf{P}_K \mathbf{B}_K, \qquad \widetilde{\Delta}\theta_K^\star \leftarrow \mathbf{M}_{\alpha,K}^{1/2} \widehat{\Delta}\theta_K^\star(\alpha).$$

3: Compute singleton scores for each sample in $\mathcal{D}_{train}$

$$s_i = \langle \widetilde{\Delta}_K(i), \widetilde{\Delta}\theta_K^\star \rangle - \frac{1}{2} \|\widetilde{\Delta}_K(i)\|_2^2, \qquad i = 1, \ldots, N_C.$$

4: Choose the top S candidates using the forward greedy selector (Algorithm 3 in Appendix C.2).

Output: Selection $\widehat{w}$, selected subset $\widehat{\mathcal{S}} := \{i \in [N_C] : \widehat{w}_i = 1\}$, and the update $\widehat{\Delta}\theta_K(\widehat{w})$.

3 Experiments

We compare SAFE-DRIFT against four established data selection baselines. These include (1) random sampling; (2) LESS [6], a target-alignment method which leverages LoRA-based low-dimensional gradient similarity to select data; (3) DSIR [17], a target-alignment method based on hashed n-gram features; (4) Prismatic Synthesis [19], a diversity-based method that clusters candidate gradients and samples from underrepresented clusters.[2] None of these explicitly controls off-target drift. We additionally compare against two methods that preserve reference behavior: (a) replay, the standard method for catastrophic forgetting [32], and KL-regularized fine-tuning, which penalizes divergence from the base model on reference data.

All experiments use open-weight OLMo-2-7B models (OLMo-2-1124-7B or OLMo-2-1124-7B-Instruct) [33] as the base model with parameter θ_0; the robustness study of Section 3.1 additionally evaluates two model families across five scales. The full experimental protocol is summarized in Table 1. We have additional results about off-target drifts across the settings in Table 4.

Measuring off-target drift. There are three ways we measure off-target drifts. We write $D_{\mathrm{F}} = \Delta\theta^\top \hat{\mathbf{F}}_R^{(K_R)} \Delta\theta$ for the *predicted* drift, the reference Fisher cost the selection procedure optimizes; D_{KL} for the *realized* drift, the token-weighted KL divergence between the base and fine-tuned model on held-out reference data; and D_{task} for the *behavioral* drift, the change in accuracy on the reference benchmarks. SAFE-DRIFT controls D_{F} by construction, and as the drift budget changes, D_{F} tracks the realized drift D_{KL} closely (Spearman correlation 0.967 and 0.670 in Settings 2 and 3; Table 4). Unless stated otherwise, results report D_{task} and D_{KL}; we invoke D_{F} only where the claim concerns the mechanism.

3.1 Main Results

The settings in Table 1 fall into two regimes, depending on whether fine-tuning on the candidate pool degrades or improves model behavior on the reference task. Settings 1, 2, and 4 fall in the first,

[2] Although introduced for synthetic data generation, we adapt it for data selection, motivated by its use of gradient diversity to capture meaningful variation across data instances.

Setting	Candidate (N_C)	Target (N_T)	Reference (N_R)	Base Model	LoRA rank	LoRA alpha	Rank $K_R + K_T$	$(\rho,\ \epsilon)$	Entanglement $(\pi_{Q_{\text{clean}}},\ \pi_{Q_{\text{ent}}})$
1	CodeAlpaca (20K) [34]	HumanEval (164) [35]	StereoSet (4.2K) [36]	Base	16	16	$512 + 128$	$(1 \times 10^{-5},\ 1)$	(0.16, 0.31)
2	MedMCQA (2K) [37]	MedQA (256) [38]	BBQ (253) [39] TruthfulQA (256) [40]	Base-Instruct	16	32	$64 + 16$	(1, 1)	(0.22, 0.77)
3	Empathetic Dialogues (5K) [41]	ESConv (100) [42]	GSM8K (256) [43]	Base-Instruct	8	16	$64 + 16$	(0.01, 1)	(0.36, 0.08)
4	FLAN (100K) [44]	IFEval (42) [45]	MMLU (285) [46]	Base	32	128	$128 + 32$	$(2 \times 10^{-5},\ 2)$	(0.22, 0.26)

Table 1: *Summary of experimental settings. N_C: candidate pool size. N_T: target set size (used for $\widehat{g}_{\text{tar}}$, validation, and test). N_R: reference set size (used for Fisher estimation and off-target drift evaluation). Sizes here describe the source pools; estimator denominators in Section 2.3 use the corresponding estimation splits. Setting-specific evaluation sets are detailed in Appendix D.2. The definitions of $\pi_{Q_{\text{clean}}}$ and $\pi_{Q_{\text{ent}}}$ are defined and illustrated in Appendix C.3. All runs use learning rate $\eta = 2 \times 10^{-4}$.*

Setting	Metric	Random	DSIR	LESS	Prismatic	SAFE-DRIFT
1 (15%)	Δ HumanEval pass@1	$\mathbf{0.136 \pm 0.005}$	0.122 ± 0.015	0.116 ± 0.003	0.118 ± 0.016	$\underline{0.132 \pm 0.004}$
	Δ HumanEval pass@10	$\mathbf{0.207 \pm 0.023}$	0.150 ± 0.010	0.162 ± 0.017	0.175 ± 0.026	$\underline{0.183 \pm 0.023}$
	Δ StereoSet score	$\underline{0.537 \pm 0.519}$	1.162 ± 0.455	0.999 ± 0.706	1.012 ± 0.745	$\mathbf{0.148 \pm 0.662}$
2 (25%)	Δ MedQA acc.	$\underline{0.033 \pm 0.004}$	$\mathbf{0.039 \pm 0.016}$	0.021 ± 0.003	0.027 ± 0.013	0.015 ± 0.006
	Δ TruthfulQA score	-0.069 ± 0.023	-0.083 ± 0.019	$\underline{-0.062 \pm 0.005}$	-0.080 ± 0.022	$\mathbf{-0.038 \pm 0.010}$
	Δ BBQ acc.	-0.026 ± 0.036	-0.004 ± 0.011	$\mathbf{0.014 \pm 0.042}$	-0.071 ± 0.000	$\underline{0.007 \pm 0.033}$
3 (10%)	Δ ESConv	$\underline{0.047 \pm 0.080}$	0.039 ± 0.007	$\mathbf{0.053 \pm 0.007}$	0.037 ± 0.008	0.039 ± 0.006
	Δ GSM8K acc.	0.064 ± 0.007	0.065 ± 0.004	$\underline{0.072 \pm 0.008}$	0.069 ± 0.015	$\mathbf{0.084 \pm 0.004}$
	Δ MMLU acc.	0.030 ± 0.007	0.025 ± 0.008	$\underline{0.039 \pm 0.004}$	0.029 ± 0.006	$\mathbf{0.044 \pm 0.010}$
4 (5%)	Δ IFEval prompt acc.	$\underline{3.530 \pm 2.435}$	-0.407 ± 0.665	2.172 ± 1.545	3.259 ± 1.090	$\mathbf{4.141 \pm 2.145}$
	Δ IFEval inst. acc.	$\underline{3.728 \pm 1.886}$	0.877 ± 1.236	2.632 ± 1.162	3.158 ± 1.058	$\mathbf{4.649 \pm 1.911}$
	Δ MMLU acc.	-0.833 ± 1.434	-0.667 ± 0.850	$\underline{0.500 \pm 0.816}$	-0.500 ± 2.828	$\mathbf{0.500 \pm 0.408}$

Table 2: *Target and reference performance changes across all settings. All values are fine-tuned minus base; higher is better for task scores, whereas smaller absolute StereoSet change means closer preservation. Reference-score changes therefore have the opposite sign to degradation D_{task} for higher-is-better metrics. For each setting, target metrics (shaded) are separated by a horizontal rule from reference metrics (details in Appendix D). Bold denotes the best result per row; underline the second best. Scores are means ± standard deviation over three random seeds.*

where target gains come at the cost of off-target drift. Setting 3 falls in the second: its entanglement score $\pi_{Q_{\text{ent}}} = 0.08$ (Table 1) is far below that of the other settings, so target-improving candidates lie largely outside the reference-sensitive directions and the target can be improved at zero off-target cost. Table 2 reports the target performance and the off-target behavioral drift D_{task} across all settings.

Case 1: fine-tuning degrades reference behavior (Settings 1, 2, and 4). In Setting 1, fine-tuning on CodeAlpaca improves Python programming ability but increases stereotype bias: all baselines increase the StereoSet score relative to the base model, where a smaller change indicates better preservation of the base model's stereotype bias. SAFE-DRIFT achieves competitive pass@1 and pass@10 rates on HumanEval while producing the smallest ΔStereoSet score among all methods.

Setting 2 illustrates the trade-off more sharply. Although fine-tuning improves MedQA performance, reference metrics deteriorate for most baselines: TruthfulQA drops for all baselines, and BBQ drops for Random and Prismatic. SAFE-DRIFT sacrifices some MedQA accuracy in exchange for substantially smaller losses on off-target drift. The high entanglement score $\pi_{Q_{\text{ent}}} = 0.77$ indicates that most target-improving candidates are entangled with reference-sensitive directions, explaining why controlling drift necessarily constrains target performance in this setting.

Setting 4 presents a complementary picture. With $\pi_{Q_{\text{ent}}} = 0.26$ and $\pi_{Q_{\text{clean}}} = 0.22$, a larger share of candidates carry target signal in directions decoupled from the reference geometry. SAFE-DRIFT achieves the best performance on both target metrics and improves MMLU accuracy, while all baselines except LESS degrade it, confirming that clean task-relevant signal allows the method to improve the target without incurring significant off-target drift.

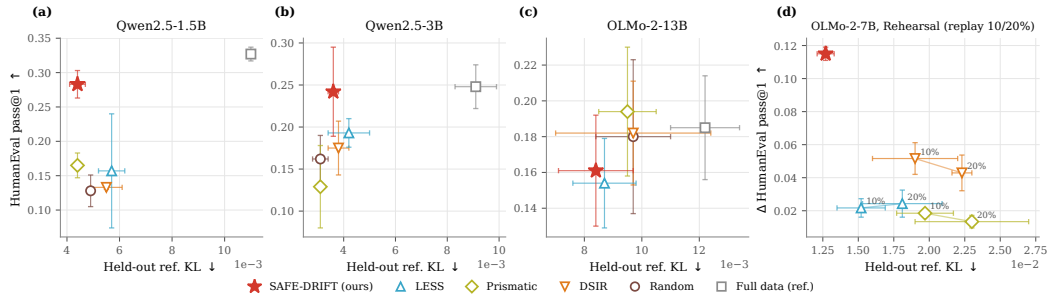


Figure 3: *Realized drift D_{KL} vs. target gain (Setting 1).* ***(a–c)*** *Ablation across Qwen2.5 and OLMo-2 base models.* ***(d)*** *Comparison against rehearsal baselines on OLMo-2-7B: each baseline is fine-tuned with 10% or 20% of replayed reference data, whereas* SAFE-DRIFT *uses no replay. We plot the mean and standard deviation across three seeds; upper left is better on the Pareto plots.*

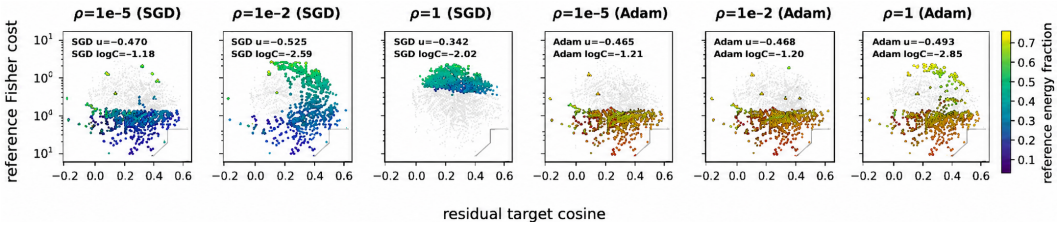


Figure 4: *Effect of the drift budget ρ on target accuracy and off-target drift under SGD and Adam preconditioners. Colored markers denote the examples selected by* SAFE-DRIFT *in Setting 2.*

Case 2: fine-tuning improves reference behavior (Setting 3). Setting 3 has by far the lowest entanglement of the four, $\pi_{Q_{\mathrm{ent}}} = 0.08$ (Table 1): almost no candidate carries its target signal through reference-sensitive directions, so SAFE-DRIFT can place greater emphasis on target alignment. The reference metrics improve—emotional support dialogue draws on broad world knowledge and social reasoning, the same general capabilities probed by MMLU and GSM8K, so gains transfer rather than trade-off. SAFE-DRIFT matches DSIR and Prismatic on the target task while achieving the largest gains on both GSM8K and MMLU. Accordingly, selection is driven almost entirely by target alignment and the selected set includes points whose residual Fisher cost is not small.

3.2 Robustness and ablation results

SAFE-DRIFT outperforms behavior-preservation methods without training on reference data. We compare against rehearsal, which mixes held-out reference examples into training at 10% and 20% of the dataset, to the baselines in Setting 1 (Figure 3d). To the best our knowledge, SAFE-DRIFT is the first to formulate drift control from a reference Fisher perspective and to use it to select data under an explicit budget. SAFE-DRIFT attains the largest target gain compared to any rehearsal-based baselines, with the lowest realized drift, and without ever training on the reference set. We also augment LESS and Prismatic with the reference KL penalty in the SFT loss. The penalty reduces drift as intended, but generally at the cost of target accuracy, so the regularized baselines and SAFE-DRIFT occupy different points on the same trade-off. Full numbers are in Appendix D.4.

The drift budget ρ trades target gain against drift, and its cost depends on the gradient update method and entanglement. We sweep ρ over $\{10^{-5}, 10^{-2}, 10^{0}\}$ under the SGD and Adam preconditioners in Setting 2 (Figure 4). Under the SGD preconditioner ($\mathbf{P} = \mathbf{I}$), increasing ρ moves the selected set from mostly low-cost candidates toward a high-cost, high-entanglement region, consistent with $\pi_{Q_{\mathrm{ent}}} = 0.77$; under Adam, the selected set is nearly unchanged between $\rho = 10^{-5}$ and 10^{-2} and enters that region only at $\rho = 1$, so the preconditioner keeps the effective update in a low-cost region for longer. Across a larger sweep (Table 4), larger ρ consistently improves target accuracy (Spearman 0.753 and 0.672 in Settings 2 and 3), but its association with reference drift reverses sign: 0.483 in the entangled Setting 2 against -0.735 in Setting 3, where improving the target also improves reference behavior. Throughout the sweep, the predicted drift D_{F}, which

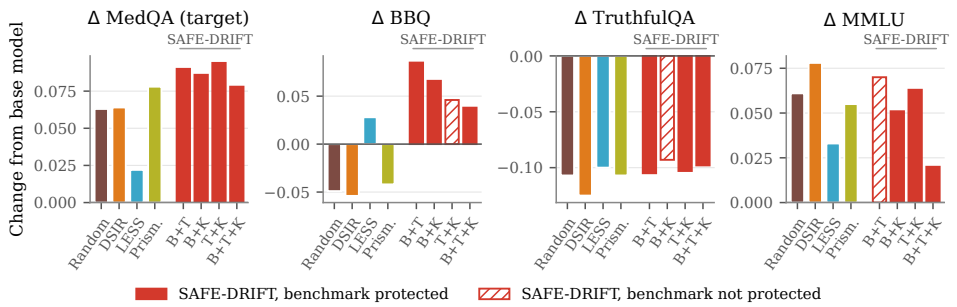


Figure 5: *Composite reference distributions (Setting 2). Each panel reports the change from the base model on the target (MedQA) and on each reference benchmark (*B = *BBQ,* T = *TruthfulQA,* K = *MMLU). Results of* SAFE-DRIFT *are shown for each reference composition. Full results, including held-out KL, are in Table 6.*

SAFE-DRIFT controls, tracks the realized drift D_{KL} (Spearman 0.967 and 0.670 in Settings 2 and 3; Table 4). In practice, we selecting ρ on a validation drift set (Appendix D.5).

The drift control generalizes across models, mixtures of references, and full parameter fine-tuning. We repeat Setting 1 across Qwen2.5 at 1.5B/3B/7B and OLMo-2 at 7B/13B, at a 15% budget with three seeds (Figure 3a–c; full table in Appendix D.6.1). SAFE-DRIFT attains the lowest predicted drift D_{F} in every configuration, reducing D_{KL} by 31.1–60% relative to full data at competitive target gain. Its realized drift D_{KL} is lowest at Qwen2.5-1.5B and OLMo-2-13B and its target gain is largest on the smaller Qwen models. When the reference combines several datasets, SAFE-DRIFT matches or exceeds the best baseline on the target and keeps BBQ above the base model, even when BBQ is not protected; on TruthfulQA and MMLU, it performs comparably to the baselines (Figure 5). This suggests that a reference Fisher estimated from a subset of benchmarks can capture directions shared with the others (Appendix D.6.2). Also, running SAFE-DRIFT with full SFT instead of LoRA trades some target accuracy for better reference preservation (Appendix D.6.3).

Design choices and computational cost. Using reference-Fisher-induced norm in Equation (7) makes the data selection more responsive to the budget ρ; Euclidean norm objective selects more reference-sensitive gradients regardless of ρ (Appendix D.7.1). The best preconditioner varies by setting, and we recommend matching the downstream optimizer (Appendix D.7.2). Results are stable across ranks (K_R, K_T), with the largest basis giving the smallest drift (Appendix D.7.3). Finally, SAFE-DRIFT shares its dominant cost—gradient construction over the candidate pool—with LESS and Prismatic, and its reported selection FLOPs match those of other gradient-based selectors at the displayed precision (Appendix D.8).

4 Conclusion

Data selection methods are typically evaluated on target task alone, without accounting for their impact on model behavior outside the target domain. In this paper, we explicitly formalize this trade-off between target improvement and off-target drift. Our method, SAFE-DRIFT, enables controllable off-target drift during targeted data selection. Through systematic experiments across various settings and domains, we demonstrate the effectiveness of our method. However, our method relies on two assumptions. First, it requires a well-approximated reference distribution. If the reference set does not fully cover all the behaviors we want to preserve, such as reasoning ability or domain knowledge, the observed reference distribution may underestimate the true off-target drift. We believe a future direction is to study how reference distribution should be constructed, expanded, and adapted rather than treated as fixed inputs, for example, following related work on broad-coverage evaluation, red-teaming evaluation, and trustworthiness [47–50]. Second, our analysis is based on a one-step or local update approximation which is a reasonable assumption for supervised fine-tuning, but it might not hold for fine-tuning based on reinforcement learning, where parameter updates can grow large as the policy diverges from initialization. We encourage future work to handle the trade-off in such settings where drift accumulates across updates.

Acknowledgments

C.D. would like to acknowledge support for this project from the National Science Foundation (NSF grant IIS-2238616). This work was completed in part with resources provided by the University of Chicago's Research Computing Center.

References

[1] Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*, 2023.

[2] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[3] Pierre Colombo, Telmo Pires, Malik Boudiaf, Rui Melo, Dominic Culver, Etienne Malaboeuf, Gabriel Hautreux, Johanne Charpentier, and Michael Desa. Saullm-54b & saullm-141b: Scaling up domain adaptation for the legal domain. *Advances in Neural Information Processing Systems*, 37:129672–129695, 2024.

[4] Joel Niklaus, Lucia Zheng, Arya D McCarthy, Christopher Hahn, Brian M Rosen, Peter Henderson, Daniel E Ho, Garrett Honke, Percy Liang, and Christopher Manning. Flawn-t5: An empirical examination of effective instruction-tuning data mixtures for legal reasoning. *arXiv preprint arXiv:2404.02127*, 2024.

[5] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168, 2021.

[6] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.

[7] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less, 2024. URL `https://arxiv.org/abs/2405.09673`.

[8] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=KBMOKmX2he`.

[9] Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL `https://arxiv.org/abs/2502.03387`.

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

[11] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*, 2023.

[12] Jan Betley, Niels Warncke, Anna Sztyber-Betley, Daniel Tan, Xuchan Bao, Martín Soto, Megha Srivastava, Nathan Labenz, and Owain Evans. Training large language models on narrow tasks can lead to broad misalignment. *Nature*, 649(8097):584–589, 2026.

[13] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[14] Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. Instruction mining: Instruction data selection for tuning large language models. *arXiv preprint arXiv:2307.06290*, 2023.

[15] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4003–4012, 2020.

[16] Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*, 2024.

[17] Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling, 2023. URL `https://arxiv.org/abs/2302.03169`.

[18] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *International Conference on Machine Learning*, pages 9525–9587. PMLR, 2022.

[19] Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in llm reasoning. *arXiv preprint arXiv:2505.20161*, 2025.

[20] Xu Yang, Chen Liu, and Ying Wei. Data selection matters: Towards robust instruction tuning of large multimodal models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[21] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*, 2024.

[22] Kiho Park, Todd Nief, Yo Joong Choe, and Victor Veitch. The information geometry of softmax: Probing and steering. *arXiv preprint arXiv:2602.15293*, 2026.

[23] Vincent Siu, Nicholas Crispino, David Park, Nathan W Henry, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. Steeringsafety: A systematic safety evaluation framework of representation steering in llms. *arXiv preprint arXiv:2509.13450*, 2025.

[24] Jinman Wu, Yi Xie, Shen Lin, Shiqian Zhao, and Xiaofeng Chen. Knowing without acting: The disentangled geometry of safety mechanisms in large language models. *arXiv preprint arXiv:2603.05773*, 2026.

[25] Austin C Kozlowski, Callin Dai, and Andrei Boutyline. Semantic structure in large language model embeddings. *arXiv preprint arXiv:2508.10003*, 2025.

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Stéphane G Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on signal processing*, 41(12):3397–3415, 1993.

[28] Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pages 5464–5474. PMLR, 2021.

[29] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 7319–7328, 2021.

[30] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[31] Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*, 2024.

[32] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2025. URL https://arxiv.org/abs/2308.08747.

[33] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

[34] Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. https://github.com/sahil280114/codealpaca, 2023.

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

[36] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models, 2020.

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

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

[39] Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. BBQ: A hand-built bias benchmark for question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.165. URL https://aclanthology.org/2022.findings-acl.165/.

[40] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022. URL https://arxiv.org/abs/2109.07958.

[41] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Annual Meeting of the Association for Computational Linguistics*, 2018. URL https://api.semanticscholar.org/CorpusID:195069365.

[42] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *ACL*, 2021.

[43] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[44] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *Advances in Neural Information Processing Systems*, 36:74764–74786, 2023.

[45] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL `https://arxiv.org/abs/2311.07911`.

[46] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[47] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[48] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *Advances in Neural Information Processing Systems*, 36: 31232–31339, 2023.

[49] Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[50] Hwan Chang and Hwanhee Lee. Which retain set matters for llm unlearning? a case study on entity unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5966–5982, 2025.

[51] Mayee F Chen, Nicholas Roberts, Kush Bhatia, Jue WANG, Ce Zhang, Frederic Sala, and Christopher Re. Skill-it! a data-driven skills framework for understanding and training language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL `https://openreview.net/forum?id=Ioizw01NLf`.

[52] Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale, 2023. URL `https://arxiv.org/abs/2309.04564`.

[53] Hamish Ivison, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Data-efficient finetuning using cross-task nearest neighbors. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9036–9061, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.576. URL `https://aclanthology.org/2023.findings-acl.576/`.

[54] Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Baker Grosse, and Eric P. Xing. What is your data worth to GPT? LLM-scale data valuation with influence functions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=zPKeJAEo27`.

[55] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach, 2018. URL `https://arxiv.org/abs/1708.00489`.

[56] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models, 2020. URL https://arxiv.org/abs/1906.01827.

[57] Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens, 2022. URL https://arxiv.org/abs/2205.11320.

[58] Gantavya Bhatt, Yifang Chen, Arnav Das, Jifan Zhang, Sang Truong, Stephen Mussmann, Yinglun Zhu, Jeff Bilmes, Simon Du, Kevin Jamieson, Jordan Ash, and Robert Nowak. An experimental design framework for label-efficient supervised finetuning of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6549–6560, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.390. URL https://aclanthology.org/2024.findings-acl.390/.

[59] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

[60] Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10658–10671, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.715. URL https://aclanthology.org/2023.findings-emnlp.715/.

[61] Hippolyt Ritter, Aleksandar Botev, and David Barber. Online structured laplace approximations for overcoming catastrophic forgetting, 2018. URL https://arxiv.org/abs/1805.07810.

[62] Ashwinee Panda, Berivan Isik, Xiangyu Qi, Sanmi Koyejo, Tsachy Weissman, and Prateek Mittal. Lottery ticket adaptation: Mitigating destructive interference in llms, 2024. URL https://arxiv.org/abs/2406.16797.

[63] Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32, 2019.

[64] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, 2017. arXiv:1703.04730.

[65] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

[66] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language model with self generated instructions, 2022.

[67] Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. Olmes: A standard for language model evaluations, 2025. URL https://arxiv.org/abs/2406.08446.

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: Our abstract accurately reflects all the findings presented in the paper

   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: Limitations are discussed in the conclusion section.

   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: We provide the detailed proof in the corresponding section as well as extended proof in the 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: The appendix sections contain the necessary hyperparameters and implementation details to reproduce our work. Furthermore, we provide code to reproduce the results and figures.

   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: We provide as a supplemental file the code necessary to reproduce all results.

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 results and appendix provides all necessary experimental details to fully understand and interpret our results.

   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: We use error bars for all the reported results.

   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: We provide information on the compute resources used in the appendix.

   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: Our research fully conforms to the code of ethics.

   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: This paper provides new tools for safeguarding large language models against unwanted behavioral drift during fine-tuning, helping preserve important capabilities and reduce unintended misalignment.

    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: There are no risks associated with our paper or code.

    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: We properly attribute all relevant software and datasets used.

    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: The new asset introduced in this paper is the code implementation of our SAFE-DRIFT algorithm used in the experiments. We plan to release the code with accompanying documentation, including instructions for setup, preprocessing, running experiments, and reproducing the main results.

    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: There were no human subjects involved in our 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: There were no human subjects involved in our research.

    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: [N/A]

    Justification: The paper does not involve the use of large language models (LLMs) as an important, original, or non-standard component of the core methods.

    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.

A Related Work & Notations

A.1 Related Works

Data Selection Data selection aims to find a subset of training examples that induces performance similar to training on the full dataset. In particular, [21] divide data selection methods in two main classes: Distribution Matching methods, and Distribution Diversification methods.

Distribution matching methods aim to find samples that closely match the target distribution. In this setting, a crucial component is to assess the similarity of a given data point with the target data. A plethora of methods have introduced non-model specific notions of useful data [51] or n-gram features [17] to perform data selection for pre-training. Because these methods lack the finesse of detecting "high-quality" data points by solely relying on surface attributes, the general trend in the literature – particularly in the context of fine-tuning or instruction tuning — seems to have turned towards using model-specific metrics based on scores like perplexity [52], model embedding [53]. Recent methods have for instance relied on formulating the data selection problem as a multi-armed bandit, where the utility function is an aggregated reward function of the cosine similarity and magnitude similarity between gradients of batches from the auxiliary and target datasets. Another line of recent work relies on model gradients [6, 54] for data selection, and calculate the pointwise data influence between pair of training and validation instances.

Distribution diversification methods, by contrast, seek to emphasize diversity of instances. These algorithms aim to efficiently approximate various geometric extent measures over the training set using a small subset. These include coreset methods [55, 56] and submodular algorithms [57, 58].

Catastrophic Forgetting Catastrophic forgetting [59, 32] happens when adaptation to a new distribution generally results in a largely reduced ability to capture the old ones. Existing approaches to remedy this issue can be typically divided into two categories: data-aware and data-oblivious. Data-aware methods [60, 61] assume access to earlier data distributions to some extent. In contrast, data-oblivious methods operate solely with the pre-trained model, without revisiting any prior data. Data-oblivious methods assume access to the pretrained model without revisiting any prior data. Previous works [62, 7] selectively update a subset of parameters in a pretrained model to limit interference with previously knowledge.

A.2 Notations.

Throughout the paper, we use bold uppercase letters, such as $\mathbf{A}$, to denote matrices. For a matrix $\mathbf{A} \in \mathbb{R}^{d\times n}$, we write $\mathbf{A}(i)$ for its i-th column and $\mathbf{A}_{ij}$ for its (i,j)-th entry. For a vector $a \in \mathbb{R}^d$, we write a_i for its i-th coordinate. Let $\mathrm{diag}(\mathbf{A}, \mathbf{B})$ denote the block-diagonal matrix $\begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{B} \end{bmatrix}$ given matrices $\mathbf{A}$ and $\mathbf{B}$. Let $\mathbf{0}_K \in \mathbb{R}^{K\times K}$ denote the zero matrix, $\mathbf{I}_K \in \mathbb{R}^{K\times K}$ denote the identity matrix for an integer $K \geq 0$, with empty matrices when $K = 0$.

B Technical Appendices

B.1 Proof of Proposition 1

Proof. Assume there exists $r_0 > 0$ such that the following hold in the neighborhood $\mathcal{N}_{r_0}(\theta_0) := \{\theta \in \mathbb{R}^d : \|\theta - \theta_0\|_2 < r_0\}$:

(i) For $\mathcal{R}$-almost every x, the distributions $p_\theta(\cdot \mid x)$ have a common support $\mathcal{Y}_x$ throughout $\mathcal{N}_{r_0}(\theta_0)$ and are strictly positive on that support. For every $y \in \mathcal{Y}_x$, the map $\theta \mapsto \log p_\theta(y \mid x)$ is three times continuously differentiable there.

(ii) Differentiating $\int_{\mathcal{Y}_x} p_\theta(y \mid x)\, dy = 1$ twice under the integral is valid; for example, the first two derivatives of p_θ admit integrable dominating envelopes on this neighborhood. The first two log derivatives at θ_0 and the squared score have finite joint expectations under $x \sim \mathcal{R}$, $y \sim p_{\theta_0}(\cdot \mid x)$.

(iii) There exists an integrable envelope $G(x, y)$ such that

$$\sup_{\theta \in \mathcal{N}_{r_0}(\theta_0)} \left\| \nabla_\theta^3 \log p_\theta(y \mid x) \right\| \le G(x, y), \quad \mathbb{E}_{x \sim \mathcal{R}} \mathbb{E}_{y \sim p_{\theta_0}(\cdot|x)}[G(x, y)] < \infty.$$

For a fixed x, define

$$d_x(\theta) = D_{\mathrm{KL}}(p_{\theta_0}(\cdot \mid x) \,\|\, p_\theta(\cdot \mid x)) \,.$$

Then,

$$D_{\mathcal{R}}(\Delta\theta) = \mathbb{E}_{x \sim \mathcal{R}}[\, d_x(\theta_0 + \Delta\theta)].$$

From the definition of KL divergence,

$$d_x(\theta) = \mathbb{E}_{y \sim p_{\theta_0}(\cdot|x)}[\log p_{\theta_0}(y \mid x) - \log p_\theta(y \mid x)]$$

and

$$\nabla_\theta d_x(\theta) = \mathbb{E}_{p_{\theta_0}(\cdot|x)}[-\nabla_\theta \log p_\theta(\cdot|x)], \qquad \nabla_\theta^2 d_x(\theta) = \mathbb{E}_{p_{\theta_0}(\cdot|x)}[-\nabla_\theta^2 \log p_\theta(\cdot|x)].$$

At $\theta = \theta_0$,

$$\begin{aligned}
\nabla_\theta d_x(\theta)|_{\theta=\theta_0} &= \mathbb{E}_{p_{\theta_0}(\cdot|x)}[-\nabla_\theta \log p_{\theta_0}(\cdot|x)] \\
&= -\int_{\mathcal{Y}_x} p_{\theta_0}(y|x) \frac{\nabla_\theta p_{\theta_0}(y|x)}{p_{\theta_0}(y|x)} \, dy \\
&= -\nabla_\theta \int_{\mathcal{Y}_x} p_{\theta_0}(y|x) \, dy \\
&= \mathbf{0}_d
\end{aligned}$$

and

$$\nabla_\theta^2 d_x(\theta)|_{\theta=\theta_0} = \mathbb{E}_{p_{\theta_0}(\cdot|x)}[-\nabla_\theta^2 \log p_{\theta_0}(\cdot|x)].$$

By a second-order Taylor expansion around $\theta = \theta_0$,

$$\begin{aligned}
d_x(\theta) &= d_x(\theta_0) + \nabla_\theta d_x(\theta)|_{\theta=\theta_0}^\top (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^\top \nabla_\theta^2 d_x(\theta)|_{\theta=\theta_0}(\theta - \theta_0) + R_x(\Delta\theta) \\
&= \frac{1}{2}\Delta\theta^\top \mathbb{E}_{p_{\theta_0}(\cdot|x)}[-\nabla_\theta^2 \log p_{\theta_0}(\cdot|x)]\Delta\theta + R_x(\Delta\theta),
\end{aligned}$$

where $R_x(\Delta\theta)$ is Taylor remainder. The common-support and differentiation assumptions give the Fisher identity [63],

$$\mathbf{F}_{\mathcal{R}} = \mathbb{E}_{x \sim \mathcal{R}} \mathbb{E}_{y \sim p_{\theta_0}(\cdot|x)} \big[-\nabla_\theta^2 \log p_{\theta_0}(\cdot|x) \big],$$

which leads to,

$$D_{\mathcal{R}}(\Delta\theta) = \frac{1}{2}\Delta\theta^\top \mathbf{F}_{\mathcal{R}} \Delta\theta + \mathbb{E}_{x \sim \mathcal{R}}[R_x(\Delta\theta)].$$

For $\|\Delta\theta\|_2 < r_0$, Taylor's theorem applied to each log density, followed by expectation over y, gives

$$\begin{aligned}
|R_x(\Delta\theta)| &\le \frac{\|\Delta\theta\|_2^3}{6} \mathbb{E}_{y \sim p_{\theta_0}(\cdot|x)} \left[\sup_{\theta \in \mathcal{N}_{r_0}(\theta_0)} \left\| \nabla_\theta^3 \log p_\theta(y \mid x) \right\| \right] \\
&\le \frac{\|\Delta\theta\|_2^3}{6} \mathbb{E}_{y \sim p_{\theta_0}(\cdot|x)}[G(x, y)].
\end{aligned}$$

Taking expectation over $x \sim \mathcal{R}$ gives

$$\left| D_{\mathcal{R}}(\Delta\theta) - \frac{1}{2}\Delta\theta^\top \mathbf{F}_{\mathcal{R}} \Delta\theta \right| \le \frac{\|\Delta\theta\|_2^3}{6} \, \mathbb{E}_{x \sim \mathcal{R}} \mathbb{E}_{y \sim p_{\theta_0}(\cdot|x)}[G(x, y)].$$

Since the envelope is integrable,

$$D_{\mathcal{R}}(\Delta\theta) = \frac{1}{2}\Delta\theta^\top \mathbf{F}_{\mathcal{R}} \Delta\theta + o(\|\Delta\theta\|_2^2).$$

This proves the claim. □

B.2 Proof of Proposition 2

Theorem 1 (Oracle safe update (full))**.** *Let $\mathbf{F}_{\mathcal{R}} \succeq 0$ and let $g_{\mathcal{T}} \neq 0$. For $\rho > 0$ and $\epsilon > 0$, consider*

$$\Delta^\star \in \arg\min_{\Delta \in \mathbb{R}^d} g_{\mathcal{T}}^\top \Delta \quad \textit{subject to} \quad \frac{1}{2}\Delta^\top \mathbf{F}_{\mathcal{R}} \Delta \leq \rho, \qquad \|\Delta\|_2 \leq \epsilon.$$

The optimal solutions are characterized by the following cases, checked in order. Here a constraint is tight *when it holds with equality; tightness alone does not imply a strictly positive KKT multiplier.*

1. ***Feasible norm-only optimum.*** *If*

$$\frac{\epsilon^2}{2\|g_{\mathcal{T}}\|_2^2} g_{\mathcal{T}}^\top \mathbf{F}_{\mathcal{R}} g_{\mathcal{T}} \leq \rho,$$

   then a zero drift multiplier is valid, and the unique solution is

$$\Delta^\star = -\epsilon \frac{g_{\mathcal{T}}}{\|g_{\mathcal{T}}\|_2}.$$

2. ***Feasible drift-only optimum.*** *Suppose $g_{\mathcal{T}} \in \mathrm{Range}(\mathbf{F}_{\mathcal{R}})$ and define*

$$\Delta_R = -\sqrt{\frac{2\rho}{g_{\mathcal{T}}^\top \mathbf{F}_{\mathcal{R}}^\dagger g_{\mathcal{T}}}}\, \mathbf{F}_{\mathcal{R}}^\dagger g_{\mathcal{T}},$$

   where $\mathbf{F}_{\mathcal{R}}^\dagger$ denotes the Moore–Penrose pseudoinverse. If

$$\|\Delta_R\|_2 \leq \epsilon,$$

   then a zero norm multiplier is valid, and the complete optimal set is

$$\left\{\Delta_R + v : \; v \in \mathrm{Null}(\mathbf{F}_{\mathcal{R}}), \quad \|v\|_2^2 \leq \epsilon^2 - \|\Delta_R\|_2^2\right\}.$$

   In particular, Δ_R is the minimum-norm optimum. The norm constraint can still be tight for some of these optima.

3. ***Both multipliers strictly positive.*** *In all remaining cases, both constraints are tight and both KKT multipliers are strictly positive. There exists a unique damping parameter $\alpha > 0$ such that*

$$\frac{\epsilon^2}{2} \frac{g_{\mathcal{T}}^\top (\mathbf{F}_{\mathcal{R}} + \alpha I)^{-1} \mathbf{F}_{\mathcal{R}} (\mathbf{F}_{\mathcal{R}} + \alpha I)^{-1} g_{\mathcal{T}}}{g_{\mathcal{T}}^\top (\mathbf{F}_{\mathcal{R}} + \alpha I)^{-2} g_{\mathcal{T}}} = \rho,$$

   and the solution is

$$\Delta^\star = -\epsilon \frac{(\mathbf{F}_{\mathcal{R}} + \alpha I)^{-1} g_{\mathcal{T}}}{\|(\mathbf{F}_{\mathcal{R}} + \alpha I)^{-1} g_{\mathcal{T}}\|_2}.$$

In case 3, the unique solution has the KKT representation

$$\Delta^\star = -c\,(\mathbf{F}_{\mathcal{R}} + \alpha I)^{-1} g_{\mathcal{T}}$$

for some $c > 0$ and $\alpha > 0$. The endpoint $\alpha \to \infty$ yields the normalized Euclidean direction in case 1. In case 2, the pseudoinverse gives the minimum-norm optimum with its drift-determined scale; arbitrary feasible nullspace components must be included to recover every optimum.

Proof. Rewrite (6) as a minimization:

$$\min_{\Delta\theta \in \mathbb{R}^d} \quad g_{\mathcal{T}}^\top \Delta\theta \qquad \text{subject to} \qquad \frac{1}{2}\Delta\theta^\top \mathbf{F}_{\mathcal{R}} \Delta\theta \leq \rho, \qquad \|\Delta\theta\|_2^2 \leq \epsilon^2. \tag{13}$$

The objective is linear and both constraints are convex since $\mathbf{F}_{\mathcal{R}} \succeq 0$ and $\|\cdot\|_2$ is convex. Note here that $\mathbf{F}_{\mathcal{R}} \succeq 0$ always holds since for any $v \in \mathbb{R}^d$, $v^\top \mathbf{F}_{\mathcal{R}} v = \mathbb{E}\big[(v^\top s_{\theta_0}(x,y))^2\big] \geq 0$. Taking $\Delta\theta = 0$ gives $0 < \rho$ and $0 < \epsilon^2$, which is strictly feasible for any $\rho, \epsilon > 0$. The feasible set is compact and nonempty, so an optimum exists. Slater's condition and strong duality hold, and the KKT conditions are necessary and sufficient for a global optimum.

Form the Lagrangian with multipliers $\sigma \geq 0$ and $\mu \geq 0$:

$$\mathcal{L}(\Delta\theta, \sigma, \mu) = g_{\mathcal{T}}^{\top}\Delta\theta + \sigma\left(\frac{1}{2}\Delta\theta^{\top}\mathbf{F}_{\mathcal{R}}\Delta\theta - \rho\right) + \mu\big(\|\Delta\theta\|_2^2 - \epsilon^2\big)\,.$$

Stationarity. Setting $\nabla_{\Delta\theta}\mathcal{L} = 0$:

$$g_{\mathcal{T}} + \sigma\mathbf{F}_{\mathcal{R}}\Delta\theta + 2\mu\Delta\theta = 0 \implies (\sigma\mathbf{F}_{\mathcal{R}} + 2\mu\mathbf{I})\Delta\theta = -g_{\mathcal{T}}. \tag{14}$$

For any $\mu > 0$, the matrix $\sigma\mathbf{F}_{\mathcal{R}} + 2\mu\mathbf{I}$ has an inverse regardless of whether $\mathbf{F}_{\mathcal{R}}$ is singular, and:

$$\Delta\theta^{\star} = -(\sigma\mathbf{F}_{\mathcal{R}} + 2\mu\mathbf{I})^{-1}g_{\mathcal{T}}. \tag{15}$$

Complementary slackness.

$$\sigma\left(\frac{1}{2}(\Delta\theta^{\star})^{\top}\mathbf{F}_{\mathcal{R}}\Delta\theta^{\star} - \rho\right) = 0, \tag{16}$$

$$\mu\big(\|\Delta\theta^{\star}\|_2^2 - \epsilon^2\big) = 0. \tag{17}$$

We consider the three non-trivial cases.

Case 1: $\sigma = 0$**,** $\mu = 0$**.** Stationarity gives $g_{\mathcal{T}} = 0$, which is trivial. Assume $g_{\mathcal{T}} \neq 0$, so this case does not arise.

Case 2: $\sigma > 0$**,** $\mu = 0$ **(zero norm multiplier).** Stationarity requires $g_{\mathcal{T}} \in \mathrm{Range}(\mathbf{F}_{\mathcal{R}})$ and gives $\Delta\theta^{\star} = -\sigma^{-1}\mathbf{F}_{\mathcal{R}}^{\dagger}g_{\mathcal{T}} + v$ with $v \in \mathrm{Null}(\mathbf{F}_{\mathcal{R}})$. A zero norm multiplier allows either a slack or a tight norm constraint. By (16), the quadratic drift constraint is tight:

$$\frac{1}{2}(\Delta\theta^{\star})^{\top}\mathbf{F}_{\mathcal{R}}\Delta\theta^{\star} = \rho.$$

Substituting this stationarity solution and using $\mathbf{F}_{\mathcal{R}}v = 0$:

$$\frac{1}{2\sigma^2}g_{\mathcal{T}}^{\top}\mathbf{F}_{\mathcal{R}}^{\dagger}\mathbf{F}_{\mathcal{R}}\mathbf{F}_{\mathcal{R}}^{\dagger}g_{\mathcal{T}} = \frac{1}{2\sigma^2}g_{\mathcal{T}}^{\top}\mathbf{F}_{\mathcal{R}}^{\dagger}g_{\mathcal{T}} = \rho,$$

giving $\sigma^{\star} = \sqrt{g_{\mathcal{T}}^{\top}\mathbf{F}_{\mathcal{R}}^{\dagger}g_{\mathcal{T}}/(2\rho)}$, and every solution in this case is:

$$\Delta\theta^{\star} = -\sqrt{\frac{2\rho}{g_{\mathcal{T}}^{\top}\mathbf{F}_{\mathcal{R}}^{\dagger}g_{\mathcal{T}}}}\,\mathbf{F}_{\mathcal{R}}^{\dagger}g_{\mathcal{T}} + v, \qquad v \in \mathrm{Null}(\mathbf{F}_{\mathcal{R}}), \quad \|v\|_2^2 \leq \epsilon^2 - \|\Delta_R\|_2^2. \tag{18}$$

where $\mathbf{F}_{\mathcal{R}}^{\dagger}$ is the Moore–Penrose pseudoinverse. Let the decomposition of $g_{\mathcal{T}}$

$$g_{\mathcal{T}} = g_{\mathrm{range}} + g_{\mathrm{null}}, \qquad g_{\mathrm{null}} \in \mathrm{Null}(\mathbf{F}_{\mathcal{R}}),$$

The range and nullspace components in (18) are orthogonal, so this set is nonempty exactly when $\|\Delta_R\|_2 \leq \epsilon$. Every such point satisfies KKT and is optimal; equality in the drift-only Cauchy–Schwarz bound forces this range component, so the set is complete. If $g_{\mathrm{null}} \neq 0$, the problem without the norm constraint is unbounded below along $-g_{\mathrm{null}}$. The original norm-constrained problem remains bounded, and projecting stationarity onto the nullspace shows that its norm multiplier must satisfy $\mu > 0$.

Case 3: $\sigma = 0$**,** $\mu > 0$ **(norm constraint active).** By the slackness condition (17),

$$\|\Delta\theta^{*}\|_2^2 = \frac{1}{4\mu^2}\|g_{\mathcal{T}}\|_2^2 = \epsilon^2$$

leading to $\mu^{*} = \|g_{\mathcal{T}}\|_2/2\epsilon$. Thus,

$$\Delta\theta^{*} = -\frac{\epsilon}{\|g_{\mathcal{T}}\|_2}g_{\mathcal{T}}$$

This point is feasible exactly under the condition in case 1 of the theorem and is the unique minimizer over the Euclidean ball. Its drift constraint may be slack or tight, while the drift multiplier can be zero.

Case 4: $\sigma > 0$**,** $\mu > 0$ **(both multipliers positive).** If neither of the preceding feasible cases applies, both multipliers must be positive. In this case the solution takes the form (15), which we write compactly:

$$\Delta\theta^\star = -\frac{1}{\sigma^*}\left(\mathbf{F}_\mathcal{R} + \alpha I\right)^{-1} g_\mathcal{T},$$

where $\alpha = 2\mu^\star/\sigma^\star > 0$. Both constraints are tight by complementary slackness, and the Lagrangian has positive-definite Hessian $\sigma^\star \mathbf{F}_\mathcal{R} + 2\mu^\star I$, so this optimum is unique. The update is a damped natural gradient with damping determined by the ratio of the two multipliers.

Using the active norm constraint ($\|\Delta\theta^*\|_2 = \epsilon$, we let $\sigma^* = \frac{\|(\mathbf{F}_\mathcal{R}+\alpha\mathbf{I})^{-1} g_\mathcal{T}\|_2}{\epsilon}$), and $\Delta\theta^\star$ becomes

$$\Delta\theta^\star = -\epsilon \frac{(\mathbf{F}_\mathcal{R} + \alpha\mathbf{I})^{-1} g_\mathcal{T}}{\|(\mathbf{F}_\mathcal{R} + \alpha\mathbf{I})^{-1} g_\mathcal{T}\|_2}, \tag{19}$$

Then, α is chosen to satisfy (16)

$$\frac{1}{2(\sigma^*)^2} g_\mathcal{T}^\top (\mathbf{F}_\mathcal{R} + \alpha I)^{-1} \mathbf{F}_\mathcal{R} (\mathbf{F}_\mathcal{R} + \alpha I)^{-1} g_\mathcal{T} = \rho \tag{20}$$

The existence of positive multipliers gives a positive root of this scalar equation. In this remaining case the target gradient must have components on at least two distinct eigenvalues: for a single supported eigenvalue, either the norm-only optimum is feasible or the drift-only optimum is feasible. Proposition 3 below therefore gives uniqueness of α.

□

B.3 Characterization of α

Let $(\lambda_i, v_i)_{i=1}^d$ be a complete orthonormal eigensystem of $\mathbf{F}_\mathcal{R}$, including zero eigenvalues, and write $g_\mathcal{T} = \sum_i c_i v_i$. Proposition 2 gives the oracle direction when both multipliers are positive. When both multipliers are positive, α is a positive root of the scalar equation $\phi(\alpha) = 2\rho/\epsilon^2$, where the sums include all eigenvectors, including those with $\lambda_i = 0$:

$$\phi(\alpha) := \frac{\sum_i \frac{\lambda_i c_i^2}{(\lambda_i+\alpha)^2}}{\sum_i \frac{c_i^2}{(\lambda_i+\alpha)^2}}.$$

Proposition 3. *Let ϕ be defined as above with $g_\mathcal{T} \neq 0$. Then ϕ is nondecreasing on $(0, \infty)$. If there exist $i \neq j$ such that $c_i c_j \neq 0$ and $\lambda_i \neq \lambda_j$, then $\phi'(\alpha) > 0$ for all $\alpha > 0$; under this condition, any positive solution is unique. Otherwise, ϕ is constant at the single eigenvalue supported by $g_\mathcal{T}$, so a matching level has every $\alpha > 0$ as a solution.*

Proof.

$$N(\alpha) := \sum_i \frac{\lambda_i c_i^2}{(\lambda_i+\alpha)^2}, \qquad D(\alpha) := \sum_i \frac{c_i^2}{(\lambda_i+\alpha)^2}, \qquad \phi(\alpha) = \frac{N(\alpha)}{D(\alpha)}.$$

Since $D(\alpha) > 0$ for all $\alpha > 0$, the quotient is well defined. We then differentiate both $N(a)$ and $D(a)$.

$$\phi'(\alpha) = \frac{N'(\alpha)D(\alpha) - N(\alpha)D'(\alpha)}{D(\alpha)^2}.$$

Moreover,

$$N'(\alpha) = -2\sum_i \frac{\lambda_i c_i^2}{(\lambda_i+\alpha)^3}, \qquad D'(\alpha) = -2\sum_i \frac{c_i^2}{(\lambda_i+\alpha)^3},$$

so

$$N'(\alpha)D(\alpha) - N(\alpha)D'(\alpha) = 2\sum_{i,j} \frac{\lambda_i c_i^2 c_j^2 (\lambda_i - \lambda_j)}{(\lambda_i+\alpha)^3(\lambda_j+\alpha)^3}.$$

Pairing the (i, j) and (j, i) terms gives

$$N'(\alpha)D(\alpha) - N(\alpha)D'(\alpha) = \sum_{i<j} \frac{2c_i^2 c_j^2(\lambda_i - \lambda_j)^2}{(\lambda_i + \alpha)^3(\lambda_j + \alpha)^3} \;\geq\; 0.$$

Hence $\phi'(\alpha) \geq 0$ for all $\alpha > 0$. If there exist $i \neq j$ such that $c_i c_j \neq 0$ and $\lambda_i \neq \lambda_j$, then the corresponding pairwise term is strictly positive for every $\alpha > 0$, so $\phi'(\alpha) > 0$. □

The endpoint values are

$$\phi_\infty = \frac{g_\mathcal{T}^\top \mathbf{F}_\mathcal{R} g_\mathcal{T}}{\|g_\mathcal{T}\|_2^2}, \qquad \phi_0 = \lim_{\alpha \downarrow 0} \phi(\alpha) = \begin{cases} 0, & g_{\text{null}} \neq 0, \\ \dfrac{g_\mathcal{T}^\top \mathbf{F}_\mathcal{R}^\dagger g_\mathcal{T}}{g_\mathcal{T}^\top (\mathbf{F}_\mathcal{R}^\dagger)^2 g_\mathcal{T}}, & g_{\text{null}} = 0. \end{cases}$$

Under the strict-monotonicity condition, a finite positive root exists exactly when $\phi_0 < 2\rho/\epsilon^2 < \phi_\infty$. A level at or above ϕ_∞ uses the norm-only optimum; a level at or below ϕ_0 uses the drift-only optimum when $g_{\text{null}} = 0$, with its own scale and feasible nullspace components. When $g_{\text{null}} \neq 0$, the positive budget excludes the lower endpoint $\phi_0 = 0$. These endpoint cases are checked before a one-dimensional monotone solve. See Algorithm 2 for the reduced-coordinate implementation.

B.4 Choices of the Preconditioner P

Influence functions describe the derivative of a re-optimized solution under infinitesimal data perturbations. Explicit fine-tuning steps instead depend on the optimizer and its state. We compare their local update models through

$$\Delta\theta(w) = -\eta \mathbf{P}\mathbf{G}w,$$

where $\mathbf{P} \in \mathbb{R}^{d \times d}$ is a fixed *preconditioner*, $w \in \mathbb{R}^{N_C}_{\geq 0}$ satisfies $\sum_i w_i = 1$, and $\eta > 0$ is a step size. For a size-S subset, this appendix uses normalized weights $w_i = \mathbf{1}\{i \in \mathcal{S}\}/S$, equivalent to the binary indicator divided by S in the main text. The *linearized target gain* is

$$\text{Gain}(w) := -g_\mathcal{T}^\top \Delta\theta(w) = \eta\, g_\mathcal{T}^\top \mathbf{P}\mathbf{G}w = \eta \sum_{i=1}^{N_C} w_i\, g_\mathcal{T}^\top \mathbf{P} g_i.$$

For a target risk with locally bounded Hessian, the actual risk decrease equals this gain plus $O(\|\Delta\theta(w)\|_2^2)$. Different choices of $\mathbf{P}$ recover exact one-step models or useful local proxies, as specified below; they do not identify a full multi-step training trajectory.

B.4.1 Special Cases of the Preconditioner

SGD: $\mathbf{P} = \mathbf{I}$. A single stochastic gradient step on example z_i gives:

$$\Delta\theta(e_i) = -\eta\, g_i,$$

corresponding to $\mathbf{P} = \mathbf{I}$. The gain is:

$$\text{Gain}_{\text{SGD}}(w) = \eta\, g_\mathcal{T}^\top \mathbf{G}w = \eta \sum_i w_i\, g_\mathcal{T}^\top g_i,$$

which is Euclidean gradient alignment. It does not generally equal influence alignment, which additionally depends on the training Hessian.

Influence / Newton: $\mathbf{P} = \mathbf{H}^{-1}$. Let $R(\theta) = N_C^{-1} \sum_i \ell(z_i; \theta)$ and $\mathbf{H} = \nabla^2 R(\theta_0)$. Assume twice-continuously differentiable losses, $\nabla R(\theta_0) = 0$, and $\mathbf{H} \succ 0$. For the local minimizer $\theta_{a,w}$ of $R(\theta) + a \sum_i w_i \ell(z_i; \theta)$ that continues from θ_0, implicit differentiation gives [64]

$$\left.\frac{d\theta_{a,w}}{da}\right|_{a=0} = -\mathbf{H}^{-1}\mathbf{G}w, \qquad \theta_{a,w} - \theta_0 = -a\mathbf{H}^{-1}\mathbf{G}w + o(a).$$

Thus $\mathbf{P} = \mathbf{H}^{-1}$ matches the infinitesimal influence direction, with linearized gain

$$\Delta\theta(w) = -\eta \mathbf{H}^{-1}\mathbf{G}w, \qquad \text{Gain}_{\text{inf}}(w) = \eta \sum_i w_i\, g_\mathcal{T}^\top \mathbf{H}^{-1} g_i.$$

Finite re-optimization generally differs from this first-order update. Away from a stationary point, $\mathbf{H}^{-1}$ may still define a Newton-style curvature model when it exists, but the influence derivative above requires the stated stationary-solution assumptions.

Adam: $\mathbf{P} = \mathbf{P}_{\mathrm{Adam}}$. Let $g_t := \nabla_\theta \ell(z_t; \theta^t)$ denote the stochastic gradient used at optimization step t. The Adam optimizer maintains uncorrected moment accumulators m^t and v^t, together with their bias-corrected versions $\widehat{m}^t$ and $\widehat{v}^t$. With steps indexed from $t = 0$, its coordinate-wise update is [26]:

$$\theta^{t+1} - \theta^t = -\eta_t \frac{\widehat{m}^{t+1}}{\sqrt{\widehat{v}^{t+1}} + \epsilon_{\mathrm{Adam}}},$$
$$m^{t+1} = \beta_1 m^t + (1-\beta_1) g_t, \qquad \widehat{m}^{t+1} = \frac{m^{t+1}}{1-\beta_1^{t+1}},$$
$$v^{t+1} = \beta_2 v^t + (1-\beta_2)(g_t \odot g_t), \qquad \widehat{v}^{t+1} = \frac{v^{t+1}}{1-\beta_2^{t+1}}.$$

Here $\beta_1, \beta_2 \in [0, 1)$, $\epsilon_{\mathrm{Adam}} > 0$, and $m^0 = v^0 = 0$. Division and square roots are elementwise.

To obtain a fixed preconditioned-gradient proxy for selection, we freeze the coordinate scales and omit first-moment history. Let $s_j = \mathbb{E}[g_j^2] < \infty$ be a finite second moment under a fixed warm-up gradient distribution, shared across candidates, and recall $g_i = \nabla_\theta \ell(z_i; \theta_0)$. The proxy update is

$$(\Delta\theta)_j = -\eta \frac{(g_i)_j}{\sqrt{s_j} + \epsilon_{\mathrm{Adam}}}.$$

This defines the frozen diagonal preconditioner

$$\mathbf{P}_{\mathrm{Adam}} = \mathrm{diag}\left(\frac{1}{\sqrt{s_1} + \epsilon_{\mathrm{Adam}}}, \ldots, \frac{1}{\sqrt{s_d} + \epsilon_{\mathrm{Adam}}}\right) \in \mathbb{R}^{d \times d}.$$

This is a second-moment scaling proxy, not a general approximation to $\mathbf{H}^{-1}$: large second moments reduce the coordinate scale, and cross-coordinate interactions are omitted. Second moments need not equal gradient variances. Because actual Adam also updates its moments and uses first-moment history, a fixed linear map of the current candidate gradients does not reproduce the complete Adam trajectory.

The gain under Adam preconditioning is:

$$\begin{aligned}
\mathrm{Gain}_{\mathrm{Adam}}(w) &= \eta\, g_{\mathcal{T}}^\top \mathbf{P}_{\mathrm{Adam}} \mathbf{G} w \\
&= \eta \sum_{i=1}^{N_C} w_i\, g_{\mathcal{T}}^\top \mathbf{P}_{\mathrm{Adam}} g_i \\
&= \eta \sum_{i=1}^{N_C} w_i \sum_{j=1}^{d} \frac{(g_{\mathcal{T}})_j (g_i)_j}{\sqrt{s_j} + \epsilon_{\mathrm{Adam}}}.
\end{aligned}$$

This is the coordinate-wise inner product between the target gradient and each training gradient, normalized by the Adam scaling factor.

The frozen scales can be estimated during a warm-up phase, following the use of optimizer information for selection in LESS [6]. For example, averaging the bias-corrected second moments over T warm-up steps gives

$$\mathbf{P}_{\mathrm{Adam}} = \mathrm{diag}\left(\frac{1}{\sqrt{\bar{v}_j} + \epsilon_{\mathrm{Adam}}}\right), \qquad \bar{v}_j = \frac{1}{T}\sum_{t=1}^{T} \widehat{v}_j^t,$$

making the score efficient to compute.

Natural gradient: $\mathbf{P} = \mathbf{F}^{-1}$. For a chosen prompt distribution $\mathcal{Q}$, assume the common-support, smoothness, and integrability conditions of Appendix B.1 with $\mathcal{R}$ replaced by $\mathcal{Q}$, and let $\mathbf{F} = \mathbb{E}_{x\sim\mathcal{Q}} \mathbb{E}_{y\sim p_{\theta_0}(\cdot|x)}[s_{\theta_0}(x,y) s_{\theta_0}(x,y)^\top]$. If $\mathbf{F} \succ 0$, choosing $\mathbf{P} = \mathbf{F}^{-1}$ gives the natural-gradient update $\Delta\theta = -\eta \mathbf{F}^{-1} g_i$, whose local geometry is the quadratic KL metric under these regularity conditions. For singular $\mathbf{F}$, a damped inverse is well defined; a pseudoinverse instead requires separate treatment of nullspace directions, as in Theorem 1.

C Further details on SAFE-DRIFT data selector in Section 2.3

C.1 Low-Rank Working Basis

All downstream computation is carried out in a K-dimensional working basis that combines the retained reference Fisher eigenspace with a residualized target-side subspace.

Estimating $g_{\mathcal{T}}$ Given the target sample $\mathcal{D}_{\mathrm{tar}} = \{z_j\}_{j=1}^{N_T}$, we estimate the target gradient by the sample mean

$$\widehat{g}_{\mathrm{tar}}^{\mathrm{raw}} := \frac{1}{N_T} \sum_{j=1}^{N_T} \nabla_\theta \ell(z_j; \theta_0) \in \mathbb{R}^d. \tag{21}$$

This is the direct plug-in estimator of the oracle target gradient.

Reference basis Retain the top K_R strictly positive eigenpairs $(\widehat{\mathbf{V}}_{K_R}, \widehat{\mathbf{\Lambda}}_{K_R})$ of the dual Gram matrix $\frac{1}{N_R}\mathbf{S}_R\mathbf{S}_R^\top$, with K_R no larger than its rank. The parameter-space reference basis and truncated Fisher are

$$\widehat{\mathbf{U}}_{K_R} = \mathbf{S}_R^\top \widehat{\mathbf{V}}_{K_R} (N_R \widehat{\mathbf{\Lambda}}_{K_R})^{-1/2}, \qquad \widehat{\mathbf{F}}_R^{(K_R)} = \widehat{\mathbf{U}}_{K_R} \widehat{\mathbf{\Lambda}}_{K_R} \widehat{\mathbf{U}}_{K_R}^\top.$$

The recovered basis $\widehat{\mathbf{U}}_{K_R} \in \mathbb{R}^{d \times K_R}$ satisfies $\widehat{\mathbf{U}}_{K_R}^\top \widehat{\mathbf{U}}_{K_R} = \mathbf{I}_{K_R}$ and $\widehat{\mathbf{F}}_R^{(K_R)} \widehat{\mathbf{U}}_{K_R} = \widehat{\mathbf{U}}_{K_R} \widehat{\mathbf{\Lambda}}_{K_R}$.

Residualized task basis Collect the per-example gradients $\widehat{g}_i^{\mathrm{raw}}$ for all $i \in [N_T + N_C]$ into

$$\widehat{\mathbf{G}}_{TC} := \begin{bmatrix} \widehat{g}_1^{\mathrm{raw}} & \cdots & \widehat{g}_{N_C+N_T}^{\mathrm{raw}} \end{bmatrix} \in \mathbb{R}^{d \times (N_C+N_T)}.$$

A low-rank approximation of the reference Fisher alone is not sufficient to represent target-side structure outside $\mathrm{span}(\widehat{\mathbf{U}}_{K_R})$. We therefore build an additional task block by first residualizing against the reference subspace:

$$\widetilde{\mathbf{G}}_{TC} := (\mathbf{I}_d - \widehat{\mathbf{U}}_{K_R} \widehat{\mathbf{U}}_{K_R}^\top) \widehat{\mathbf{G}}_{TC}.$$

On the residualized matrix, we form the dual Gram matrix $\frac{1}{N_T+N_C} \widetilde{\mathbf{G}}_{TC}^\top \widetilde{\mathbf{G}}_{TC}$ and compute its top-K_T strictly positive eigenpairs $(\widehat{\mathbf{V}}_{K_T}, \widehat{\mathbf{\Omega}}_{K_T})$, with K_T no larger than the residual rank. We then define the added task basis by

$$\widehat{\mathbf{U}}_{K_T} := \widetilde{\mathbf{G}}_{TC} \, \widehat{\mathbf{V}}_{K_T} \big((N_C + N_T) \widehat{\mathbf{\Omega}}_{K_T} \big)^{-1/2} \in \mathbb{R}^{d \times K_T}.$$

By construction, $\widehat{\mathbf{U}}_{K_R}^\top \widehat{\mathbf{U}}_{K_T} = 0$ and $\widehat{\mathbf{U}}_{K_T}^\top \widehat{\mathbf{U}}_{K_T} = \mathbf{I}_{K_T}$. If either matrix has rank zero, its retained basis is an empty block and its retained rank is zero; no inverse of a zero eigenvalue is taken.

The motivation for this residualization is as follows. If we were to project all quantities onto $\widehat{U}_{K_R}$ alone, the quality of the approximation would depend entirely on how much of the target and candidate gradient signal lies within the reference eigenspace. When the target task and the reference distribution are largely disentangled in parameter space, the target and candidate gradients may be nearly orthogonal to $\widehat{\mathbf{U}}_{K_R}$. Projecting onto the reference basis alone would then collapse all gradient coordinates to near zero, making every candidate look equally uninformative. The residual basis retains dominant gradient directions outside that subspace. Truncation can nevertheless discard target-specific directions, especially when the candidate pool dominates the residual spectrum; a nonzero raw target gradient may have zero projection into the working basis.

Common working basis Stacking the two orthonormal bases gives (9). Define the reduced surrogate Fisher by

$$\widehat{\mathbf{F}}_{R,K} := \mathbf{U}_K^\top \widehat{\mathbf{F}}_R^{(K_R)} \mathbf{U}_K = \mathrm{diag}(\widehat{\mathbf{\Lambda}}_{K_R}, \mathbf{0}_{K_T}) \in \mathbb{R}^{K \times K}.$$

The reduced ridge-regularized matching metric is

$$\mathbf{M}_{\alpha,K} := \widehat{\mathbf{F}}_{R,K} + \alpha \mathbf{I}_K = \begin{bmatrix} \widehat{\mathbf{\Lambda}}_{K_R} + \alpha \mathbf{I}_{K_R} & 0 \\ 0 & \alpha \mathbf{I}_{K_T} \end{bmatrix}. \tag{22}$$

The zero residual block belongs to the truncated surrogate: orthogonality to the retained reference eigenspace does not imply zero cost under the full empirical or population Fisher. In fact, exact spectral truncation gives

$$\widehat{\mathbf{F}}_R^{\mathrm{raw}} = \widehat{\mathbf{F}}_R^{(K_R)} + E_R, \qquad E_R \succeq 0, \qquad \|E_R\|_{\mathrm{op}} = \widehat{\lambda}_{K_R+1},$$

where $\widehat{\lambda}_{K_R+1} = 0$ if all positive eigenvalues are retained. Consequently, for any ambient update δ satisfying $\|\delta\|_2 \le \epsilon$ and $\frac{1}{2}\delta^\top \widehat{\mathbf{F}}_R^{(K_R)}\delta \le \rho$,

$$\frac{1}{2}\delta^\top \widehat{\mathbf{F}}_R^{\mathrm{raw}}\delta \le \rho + \frac{\epsilon^2}{2}\widehat{\lambda}_{K_R+1}.$$

This is an empirical quadratic bound. Transferring it to the population Fisher requires additional estimation-error control, and transferring it to KL drift also requires control of the local Taylor remainder.

Choosing K_R and K_T One optional rank heuristic measures how much each eigendirection modifies $(\widehat{\mathbf{F}}_R^{\mathrm{raw}} + \alpha\mathbf{I})^{-1}$ relative to $\alpha^{-1}\mathbf{I}$ at a fixed positive α. Let the nonzero eigenvalues of the empirical reference Fisher $\widehat{\mathbf{F}}_R^{\mathrm{raw}}$ be $\widehat{\lambda}_1 \ge \widehat{\lambda}_2 \ge \cdots \ge \widehat{\lambda}_{r_R} > 0$, with corresponding orthonormal eigenvectors $\widehat{u}_1, \ldots, \widehat{u}_{r_R}$ when $r_R > 0$, where $0 \le r_R \le \min\{N_R, d\}$. Then,

$$(\widehat{\mathbf{F}}_R^{\mathrm{raw}} + \alpha\mathbf{I})^{-1} = \alpha^{-1}\mathbf{I} - \sum_{j=1}^{r_R} \frac{\widehat{\lambda}_j}{\alpha(\widehat{\lambda}_j + \alpha)}\widehat{u}_j\widehat{u}_j^\top.$$

The j-th direction contributes through the relative correction factor

$$r_j(\alpha) := \frac{\widehat{\lambda}_j}{\widehat{\lambda}_j + \alpha}. \tag{23}$$

When $\widehat{\lambda}_j \ll \alpha$, the correction to the isotropic inverse is small in that eigendirection. This observation alone does not guarantee that discarding it preserves a particular target direction or selection.

For $r_R > 0$ and a chosen tolerance $\delta_R \in (0, 1)$, the heuristic is

$$K_R^{\mathrm{auto}} := \min\left\{ K \in \{1, \ldots, r_R\} : \frac{\sum_{j=1}^{K} r_j(\alpha)}{\sum_{j=1}^{r_R} r_j(\alpha)} \ge 1 - \delta_R \right\}.$$

Similarly, when the residual rank $r_T > 0$, let the positive eigenvalues of $\frac{1}{N_T+N_C}\widetilde{G}_{TC}^\top\widetilde{G}_{TC}$ be $\widehat{\omega}_1 \ge \cdots \ge \widehat{\omega}_{r_T} > 0$. An optional residual-energy heuristic is

$$K_T^{\mathrm{auto}} := \min\left\{ K \in \{1, \ldots, r_T\} : \frac{\sum_{j=1}^{K} \widehat{\omega}_j}{\sum_{j=1}^{r_T} \widehat{\omega}_j} \ge 1 - \delta_T \right\},$$

for $\delta_T \in (0, 1)$. Set $K_R^{\mathrm{auto}} = 0$ when $r_R = 0$ and $K_T^{\mathrm{auto}} = 0$ when $r_T = 0$. These are aggregate spectral criteria, not uniform guarantees on target-gradient or drift preservation. The reference criterion depends on α: if it is used, choose a pilot α and fix the resulting basis before solving for damping, or explicitly repeat rank selection and damping as a heuristic. The fixed-basis monotonicity used in Algorithm 2 does not itself justify a joint rank-and-damping search.

Choosing α based on (ρ, ϵ). We replace the population Fisher and target gradient by their reduced empirical counterparts from Section 2.3. For a fixed working basis, the resulting Rayleigh quotient $\widehat{\phi}(\alpha)$ is nondecreasing, as in Proposition 3. All K target coordinates must be included, including the residual task coordinates whose surrogate Fisher eigenvalues are zero. Algorithm 2 detects an endpoint root or bisects when a root is strictly bracketed within the prescribed positive interval; its status distinguishes a root found to tolerance from an interval or iteration limit. Subsequent downscaling enforces the reduced surrogate budgets even when no root is returned. A zero projected target gives the zero continuous update and carries no target-alignment information in this basis.

Algorithm 2 Finite-interval choice of $\widehat{\alpha}$ from (ρ, ϵ)

Input: Target feature $b = \widehat{b}_{\mathrm{tar}}^{(K)}$, diagonal reduced Fisher $\widehat{\mathbf{F}}_{R,K} = \mathrm{diag}(d_1, \ldots, d_K) \succeq 0$ from (22), budgets $\rho, \epsilon > 0$, interval $0 < \alpha_{\min} < \alpha_{\max}$, tolerance $\tau > 0$, integer iteration limit $T \geq 1$.
Output: Damping parameter $\widehat{\alpha}$ and status.

if $b = 0$ **then**
  return $(\alpha_{\min}, \texttt{zero_target})$ ▷ continuous update is zero
end if
Set $d_j \leftarrow [\widehat{\mathbf{F}}_{R,K}]_{jj}$ and $c_j \leftarrow b_j$ for $j = 1, \ldots, K$; set $r \leftarrow 2\rho/\epsilon^2$.
Define

$$\widehat{\phi}(a) := \frac{\sum_{j=1}^{K} d_j c_j^2/(d_j + a)^2}{\sum_{j=1}^{K} c_j^2/(d_j + a)^2}, \qquad \widehat{\phi}_\infty := \frac{\sum_{j=1}^{K} d_j c_j^2}{\sum_{j=1}^{K} c_j^2}.$$

if $r = \widehat{\phi}(\alpha_{\max})$ **then**
  return $(\alpha_{\max}, \texttt{endpoint_root})$
else if $r = \widehat{\phi}(\alpha_{\min})$ **then**
  return $(\alpha_{\min}, \texttt{endpoint_root})$
else if $r \geq \widehat{\phi}_\infty$ **then**
  return $(\alpha_{\max}, \texttt{inactive_approx})$
else if $r < \widehat{\phi}(\alpha_{\min})$ **then**
  return $(\alpha_{\min}, \texttt{floor_limited})$
else if $r > \widehat{\phi}(\alpha_{\max})$ **then**
  return $(\alpha_{\max}, \texttt{cap_limited})$
end if
$\ell \leftarrow \alpha_{\min}$; $h \leftarrow \alpha_{\max}$ ▷ $\widehat{\phi}(\ell) < r < \widehat{\phi}(h)$
for $t = 1, \ldots, T$ **do**
  $a \leftarrow \exp((\log \ell + \log h)/2)$; $r_a \leftarrow \widehat{\phi}(a)$.
  if $|r_a - r| \leq \tau r$ **then**
    return $(a, \texttt{root_tolerance})$
  else if $r_a < r$ **then**
    $\ell \leftarrow a$
  else
    $h \leftarrow a$
  end if
end for
return $(\ell, \texttt{iteration_limit})$

Endpoint and termination guarantees. The pseudocode is stated in exact arithmetic. If b is supported on one eigenvalue $d_\star$, then $\widehat{\phi}(a) \equiv d_\star$: a matching level returns an endpoint root, but this root is not unique. If all supported eigenvalues are zero, then $\widehat{\phi} \equiv 0$ and the positive level r takes the `inactive_approx` branch. Otherwise, support on two distinct eigenvalues gives strict monotonicity, including when one is zero. The `floor_limited` status means either that a root lies below the floor or that no positive root exists. The `cap_limited` branch has $\widehat{\phi}(\alpha_{\max}) < r < \widehat{\phi}_\infty$ and therefore a unique root above the cap. The `inactive_approx` status means that the norm-only optimum of the reduced surrogate problem is feasible, including equality at $r = \widehat{\phi}_\infty$; it does not imply strictly slack drift.

For a strictly bracketed root $\alpha_\star$, every unsuccessful iteration preserves $\widehat{\phi}(\ell) < r < \widehat{\phi}(h)$ and halves $\log(h/\ell)$. Thus, after T unsuccessful iterations,

$$0 \leq \log(\alpha_\star/\ell) \leq 2^{-T} \log(\alpha_{\max}/\alpha_{\min}), \qquad 0 \leq \frac{r - \widehat{\phi}(\ell)}{r} \leq 2^{1-T} \log(\alpha_{\max}/\alpha_{\min}).$$

For the second bound, the positive numerator and denominator defining $\widehat{\phi}$ satisfy $0 \leq a\widehat{\phi}'(a)/\widehat{\phi}(a) \leq 2$: differentiating their logarithms gives twice the difference of two weighted means of $a/(d_j + a) \in [0, 1]$, and monotonicity gives the lower bound. Integrating this inequality and using $1 - e^{-x} \leq x$ proves the result. The `root_tolerance` status certifies only the stated relative quotient residual; an

arbitrary iteration limit need not meet that tolerance. These guarantees assume exact spectral data and arithmetic; a numerical implementation must also handle rounding and overflow.

C.1.1 Scaling value for $\Delta\theta^*$

For fixed $\alpha > 0$, use the reduced matrices $\widehat{\mathbf{F}}_{R,K}$ and $\mathbf{M}_{\alpha,K}$ from (22). Define the unscaled reduced direction and practical update by

$$h_{\alpha,K} := \mathbf{M}_{\alpha,K}^{-1} \widehat{b}_{\mathrm{tar}}^{(K)} \in \mathbb{R}^K, \qquad \widehat{\Delta}\theta_K^\star := -h_{\alpha,K}/\widehat{\mu}_K \in \mathbb{R}^K.$$

Their ambient lifts are $\mathbf{U}_K h_{\alpha,K}$ and $\mathbf{U}_K \widehat{\Delta}\theta_K^\star \in \mathbb{R}^d$, respectively. We use the smallest downscaling factor subject to $\widehat{\mu}_K \geq 1$ that satisfies the imposed surrogate reference and norm budgets:

$$\widehat{\mu}_K := \max\left\{1,\ \sqrt{\frac{h_{\alpha,K}^\top \widehat{\mathbf{F}}_{R,K} h_{\alpha,K}}{2\rho}},\ \frac{\|h_{\alpha,K}\|_2}{\epsilon}\right\}. \tag{24}$$

The second and third terms are included only when their respective budgets are imposed. When both budgets are imposed, direct substitution gives $\frac{1}{2}(\widehat{\Delta}\theta_K^\star)^\top \widehat{\mathbf{F}}_{R,K} \widehat{\Delta}\theta_K^\star \leq \rho$ and $\|\widehat{\Delta}\theta_K^\star\|_2 \leq \epsilon$. If either budget is omitted, only the retained budget is guaranteed. The corresponding bounds also hold for the ambient lift with the truncated Fisher, because $\mathbf{U}_K$ has orthonormal columns. If $b = 0$, then $h_{\alpha,K} = 0$, $\widehat{\mu}_K = 1$, and the update is zero. The budgets need not be saturated: the restriction $\widehat{\mu}_K \geq 1$ deliberately permits only shrinking the unscaled direction. In particular, a finite cap used in the inactive case is an approximation to a direction, not an oracle-optimal step length. Indeed, $\|\widehat{\Delta}\theta_K^\star\|_2 \leq \|b\|_2/\alpha$, so with fixed nonzero b the practical step tends to zero as the cap tends to infinity, whereas the norm-only oracle step has norm ϵ. Even at an exact finite root, both budgets are saturated only if $\|h_{\alpha,K}\|_2/\epsilon \geq 1$; otherwise $\widehat{\mu}_K = 1$ and both are slack. A root residual within τr likewise controls the ratio of the two unscaled costs, not optimality or budget saturation.

This practical construction should be contrasted with the oracle formulation $\Delta\theta^\star = -\epsilon \frac{(\mathbf{F}_\mathcal{R}+\alpha\mathbf{I})^{-1} g_\mathcal{T}}{\|(\mathbf{F}_\mathcal{R}+\alpha\mathbf{I})^{-1} g_\mathcal{T}\|_2}$ in the interior regime with positive multipliers for both constraints, where α is chosen so that the normalized update satisfies the reference-cost constraint $\frac{1}{2}(\Delta\theta^\star)^\top \mathbf{F}_\mathcal{R} \Delta\theta^\star = \rho$. The practical construction is a feasible surrogate heuristic; it does not inherit oracle optimality, a population drift budget, or a guarantee for subsequent fine-tuning.

C.2 Selection Algorithm

Recall the practical reduced surrogate matching target $\widehat{\Delta}\theta_K^\star(\alpha)$ and the subset-induced reduced update $\widehat{\Delta}\theta_K(w)$ from (11) and (12), respectively. Throughout this section, $\alpha > 0$ is finite. Define the whitened candidate-update matrix and target update by

$$\widetilde{\mathbf{\Delta}}_K := \mathbf{M}_{\alpha,K}^{1/2}\left(-\frac{\eta}{S}\mathbf{P}_K \mathbf{B}_K\right) \in \mathbb{R}^{K \times N_C},$$
$$\widetilde{\Delta}\theta_K^\star := \mathbf{M}_{\alpha,K}^{1/2} \widehat{\Delta}\theta_K^\star(\alpha).$$

Then the matching objective

$$w_K^{\mathrm{opt}} \in \arg\min_{\substack{w \in \{0,1\}^{N_C} \\ \mathbf{1}^\top w = S}} \frac{1}{2}\left\|\widehat{\Delta}\theta_K(w) - \widehat{\Delta}\theta_K^\star(\alpha)\right\|_{\mathbf{M}_{\alpha,K}}^2$$

can be rewritten as the whitened objective

$$w_K^{\mathrm{opt}} \in \arg\min_{\substack{w \in \{0,1\}^{N_C} \\ \mathbf{1}^\top w = S}} \frac{1}{2}\left\|\widetilde{\mathbf{\Delta}}_K w - \widetilde{\Delta}\theta_K^\star\right\|_2^2.$$

The forward greedy procedure below approximates this cardinality-constrained optimum; no exact solution or approximation factor is claimed. For a singleton contribution $w = e_i$, keeping the final normalization $1/S$ fixed even when $S > 1$, we have

$$\frac{1}{2}\left\|\widetilde{\Delta}_K(i) - \widetilde{\Delta}\theta_K^\star\right\|_2^2 = \frac{1}{2}\|\widetilde{\Delta}\theta_K^\star\|_2^2 - \left\langle \widetilde{\Delta}_K(i), \widetilde{\Delta}\theta_K^\star \right\rangle + \frac{1}{2}\|\widetilde{\Delta}_K(i)\|_2^2.$$

Since the first term is independent of i, ranking singleton subsets by smaller objective value is equivalent to ranking them by

$$s_i := \left\langle \widetilde{\Delta}_K(i), \widetilde{\Delta}\theta^\star_K \right\rangle - \frac{1}{2}\|\widetilde{\Delta}_K(i)\|_2^2.$$

Thus, s_i is exactly the signed decrease in the whitened quadratic objective from the empty subset to candidate i alone. A singleton is not a feasible final subset when $S > 1$. Compared with LESS [6], which ranks examples by alignment alone, our selector adds a quadratic term that penalizes costly components under the empirical reference geometry. We provide the complete selection algorithm in Algorithm 3.

Algorithm 3 Forward greedy approximation to the whitened reduced objective

Input: Whitened candidate-update matrix $\widetilde{\mathbf{\Delta}}_K = [\widetilde{\Delta}_K(1) \ \cdots \ \widetilde{\Delta}_K(N_C)]$, whitened target update $\widetilde{\Delta}\theta^\star_K$, integer subset budget $1 \le S \le N_C$, integer shortlist size $S \le m \le N_C$.
Output: Selected subset $\widehat{\mathcal{S}} \subseteq [N_C]$ with $|\widehat{\mathcal{S}}| = S$, and selection vector $\widehat{w}$.

For each $i \in [N_C]$, compute the singleton score

$$s_i \leftarrow \left\langle \widetilde{\Delta}_K(i), \widetilde{\Delta}\theta^\star_K \right\rangle - \frac{1}{2}\|\widetilde{\Delta}_K(i)\|_2^2.$$

Let $\mathcal{M}_m \subseteq [N_C]$ be the indices of the top-m scores $\{s_i\}_{i=1}^{N_C}$, using any fixed rule to break ties.
Initialize $\widehat{\mathcal{S}} \leftarrow \emptyset$ and $\tilde{s} \leftarrow 0 \in \mathbb{R}^K$.
for $t = 1, \ldots, S$ **do**
    Choose

$$j^\star \in \arg\min_{j \in \mathcal{M}_m \setminus \widehat{\mathcal{S}}} \left(\frac{1}{2}\|\widetilde{\Delta}_K(j)\|_2^2 + \left\langle \tilde{s}, \widetilde{\Delta}_K(j) \right\rangle - \left\langle \widetilde{\Delta}\theta^\star_K, \widetilde{\Delta}_K(j) \right\rangle \right).$$

    Update

$$\widehat{\mathcal{S}} \leftarrow \widehat{\mathcal{S}} \cup \{j^\star\}, \qquad \tilde{s} \leftarrow \tilde{s} + \widetilde{\Delta}_K(j^\star).$$

end for
Set $\widehat{w}_i \leftarrow \mathbf{1}\{i \in \widehat{\mathcal{S}}\}$ for all $i \in [N_C]$.
return $\widehat{\mathcal{S}}, \widehat{w}$.

The singleton scores construct the shortlist. Each greedy choice uses the current sum $\tilde{s}$ and hence includes interactions with already selected examples; ties may be resolved arbitrarily. For $m > S$, these interactions can change the final subset relative to static top-S ranking. When $m = S$, the final subset is necessarily the entire shortlist, irrespective of the greedy order. The procedure takes all S steps even when the best remaining addition increases the objective, since the cardinality constraint is exact. Both shortlisting and greedy optimization can introduce matching error.

What matching error implies for drift. For any population-oracle update $\delta^\star$ satisfying $\frac{1}{2}(\delta^\star)^\top \mathbf{F}_\mathcal{R} \delta^\star \le \rho$ and any candidate-induced ambient update $\delta_\mathcal{S}$, let $e := \delta_\mathcal{S} - \delta^\star$. The triangle inequality for the Fisher seminorm gives

$$\frac{1}{2}\delta_\mathcal{S}^\top \mathbf{F}_\mathcal{R} \delta_\mathcal{S} \le \left(\sqrt{\rho} + \frac{\|e\|_{\mathbf{F}_\mathcal{R}}}{\sqrt{2}} \right)^2, \qquad \|e\|_{\mathbf{F}_\mathcal{R}} \le \|e\|_{\mathbf{F}_\mathcal{R} + \alpha \mathbf{I}}.$$

Thus matching error can increase the quadratic cost beyond ρ; the matching objective does not impose a hard feasibility constraint on the selected subset. The same inequality holds within the empirical surrogate geometry, but a small reduced matching error alone does not control omitted directions, Fisher estimation error, or the difference between the surrogate and actual optimizer update. Under Proposition 1, actual KL drift has the additional local remainder $o(\|\delta_\mathcal{S}\|_2^2)$. These local statements do not guarantee the drift of a subsequent multi-step fine-tuning trajectory.

C.3 Entanglement diagnostics in the working subspace

C.3.1 Entanglement quadrant

A key insight from Proposition 2 is that the target gradient's orientation in the reference Fisher geometry affects the local gain–drift trade-off. As a descriptive diagnostic, we use the low-rank basis $\mathbf{U}_K = [\,\widehat{\mathbf{U}}_{K_R}\ \ \widehat{\mathbf{U}}_{K_T}\,]$ to describe where candidate alignment with the reduced target direction occurs. Decompose the candidate- and target-gradient coordinates in this basis as,

$$\widehat{b}_i^{(K)} = \mathbf{U}_K^\top \widehat{g}_i^{\mathrm{raw}} = \begin{bmatrix} \widehat{\mathbf{U}}_{K_R}^\top \widehat{g}_i^{\mathrm{raw}} \\ \widehat{\mathbf{U}}_{K_T}^\top \widehat{g}_i^{\mathrm{raw}} \end{bmatrix} = \begin{bmatrix} \widehat{b}_i^R \\ \widehat{b}_i^T \end{bmatrix}, \qquad \widehat{b}_{\mathrm{tar}}^{(K)} = \mathbf{U}_K^\top \widehat{g}_{\mathrm{tar}}^{\mathrm{raw}} = \begin{bmatrix} \widehat{\mathbf{U}}_{K_R}^\top \widehat{g}_{\mathrm{tar}}^{\mathrm{raw}} \\ \widehat{\mathbf{U}}_{K_T}^\top \widehat{g}_{\mathrm{tar}}^{\mathrm{raw}} \end{bmatrix} = \begin{bmatrix} \widehat{b}_{\mathrm{tar}}^R \\ \widehat{b}_{\mathrm{tar}}^T \end{bmatrix}.$$

Here, $\widehat{b}_i^R$ refers to the projection of the candidate gradients along the reference space, while $\widehat{b}_i^T$ refers to their projection in the residualized target space. We define two target-alignment scores under the reference subspace and the residual task subspace, respectively,

$$d_i^R = (\widehat{b}_i^R)^\top (\widehat{\mathbf{\Lambda}}_{K_R} + \alpha \mathbf{I}_{K_R})^{-1} \widehat{b}_{\mathrm{tar}}^R, \qquad d_i^T = \alpha^{-1} (\widehat{b}_i^T)^\top \widehat{b}_{\mathrm{tar}}^T.$$

Here, $d_i^R > 0$ indicates positive damped target alignment within the retained reference block, while $d_i^T > 0$ indicates positive alignment in the residual task block. The latter block is orthogonal to the retained reference subspace, but need not have zero cost under the full Fisher. An empty block contributes zero to its alignment score. Their sum is $(\widehat{b}_i^{(K)})^\top \mathbf{M}_{\alpha,K}^{-1} \widehat{b}_{\mathrm{tar}}^{(K)}$, which in general differs from the linear term in the whitened matching score. Indeed, substituting (11) gives

$$\left\langle \widetilde{\Delta}_K(i), \widetilde{\Delta}\theta_K^\star \right\rangle = \frac{\eta}{S\widehat{\mu}_K} (\widehat{b}_i^{(K)})^\top \mathbf{P}_K^\top \widehat{b}_{\mathrm{tar}}^{(K)},$$

because the matching metric cancels the inverse metric in the target update. We partition candidates by the signs of (d_i^R, d_i^T):

$$Q_{\mathrm{ref}} := \{\, z_i \in \mathcal{D}_{\mathrm{train}} : d_i^T \le 0,\ d_i^R > 0 \,\}, \qquad Q_{\mathrm{clean}} := \{\, z_i \in \mathcal{D}_{\mathrm{train}} : d_i^T > 0,\ d_i^R \le 0 \,\},$$
$$Q_{\mathrm{ent}} := \{\, z_i \in \mathcal{D}_{\mathrm{train}} : d_i^T > 0,\ d_i^R > 0 \,\}, \qquad Q_{\mathrm{mis}} := \{\, z_i \in \mathcal{D}_{\mathrm{train}} : d_i^T \le 0,\ d_i^R \le 0 \,\}.$$

We then measure the proportions of the candidate examples in each quadrant by,

$$\pi_Q := \frac{1}{N_C} \sum_{z_i \in \mathcal{D}_{\mathrm{train}}} \mathbf{1}\{z_i \in Q\}, \qquad Q \in \{Q_{\mathrm{clean}}, Q_{\mathrm{ent}}, Q_{\mathrm{ref}}, Q_{\mathrm{mis}}\}. \tag{25}$$

These four proportions summarize signs of blockwise alignment in the candidate pool. Q_{ent} has positive alignment in both blocks; Q_{clean} has positive residual alignment and non-positive reference alignment; Q_{ref} has the reverse pattern; and Q_{mis} has non-positive alignment in both blocks. The signs alone give neither the magnitudes of the two contributions nor a reference-cost bound. In particular, membership in Q_{clean} need not imply positive total target alignment: a negative reference contribution can dominate the positive residual contribution. Total gain for a preconditioned optimizer also depends on $\mathbf{P}$. We provide $\pi_{Q_{\mathrm{clean}}}$ and $\pi_{Q_{\mathrm{ent}}}$ of our experiments in Table 1.

The quadrant percentages in (25) are descriptive diagnostics, not selection rules or guarantees of target improvement or drift control. SAFE-DRIFT does not simply select all examples from Q_{clean}. These pool-level percentages describe the available blockwise alignment patterns; the selected subset is determined by the matching procedure.

C.3.2 Reference-entanglement diagnostics

The same decomposition also permits diagnostics of blockwise gradient energy, complementing the alignment signs above. These energy ratios describe the retained subspace, without establishing gain or drift for an update.

We use two complementary *reference-entanglement diagnostics* for a reduced gradient, when $K_R, K_T > 0$ and the corresponding denominators are positive. Here $\mathbf{\Lambda}_R := \widehat{\mathbf{\Lambda}}_{K_R}$ is the diagonal matrix of retained reference eigenvalues:

$$\rho_E(b) = \frac{\|b^R\|_2^2/K_R}{\|b^R\|_2^2/K_R + \|b^T\|_2^2/K_T}, \qquad \rho_{\mathrm{safe}}(b) = \frac{(b^R)^\top \mathbf{\Lambda}_R b^R/K_R}{(b^R)^\top \mathbf{\Lambda}_R b^R/K_R + \alpha\|b^T\|_2^2/K_T}. \tag{26}$$

The quantity $\rho_E(b)$ compares mean squared gradient coordinates in the two blocks; it is dimension-normalized rather than the fraction of total gradient energy. We retain the numerical definition of $\rho_{\mathrm{safe}}(b)$ as a hybrid diagnostic, weighting the reference block by $\mathbf{\Lambda}_R$ and the residual block by α. It is not the energy allocation under $\mathbf{M}_{\alpha,K}$, whose reference block is $\mathbf{\Lambda}_R + \alpha\mathbf{I}$, nor the allocation of the preconditioned update.

For equal, strictly positive squared magnitude in every coordinate, the baselines are

$$\rho_E = \frac{1}{2}, \qquad \rho_{\mathrm{safe}} = \frac{\bar{\lambda}_R}{\bar{\lambda}_R + \alpha}, \qquad \bar{\lambda}_R := \frac{\mathrm{tr}(\mathbf{\Lambda}_R)}{K_R}.$$

Thus K_R/K is not a general baseline for equal energy per coordinate in these diagnostics. When a block dimension or a denominator is zero, the corresponding ratio is undefined and should not be interpreted or plotted as an observed value.

We apply these diagnostics to target and candidate gradients together with the alignment signs from Section C.3.1. A large value indicates greater relative energy in the retained reference block under the specified weighting; it does not measure absolute update magnitude, establish target gain, or bound reference drift. Their role is to describe the structured working subspace used by the selector.

D Experiments

D.1 Synthetic linear experiments

To build better intuition on the performance of our method in simple settings, we begin with a family of synthetic linear models designed to isolate the two statistical ingredients highlighted by our theory: the estimation problem (estimating the target gradient $g_{\mathcal{T}}$ from few target examples, and estimating the reference geometry $\mathbf{F}_{\mathcal{R}}$ from limited reference data) and the optimization problem (maximizing first-order target gain under an off-target budget).

Linear-Gaussian setup. We first consider a linear Gaussian model with $\sigma^2 > 0$ and noise independent of the input in each data-generating distribution. For an input $x \in \mathbb{R}^d$, the base model is

$$p_\theta(y \mid x) = \mathcal{N}(x^\top\theta, \sigma^2), \qquad \ell((x,y);\theta) = \frac{1}{2\sigma^2}(y - x^\top\theta)^2.$$

Target examples are generated by

$$x \sim \mathcal{N}(0, \mathbf{\Sigma}_{\mathcal{T}}), \qquad y = x^\top\beta_{\mathcal{T}} + \xi, \qquad \xi \sim \mathcal{N}(0,\sigma^2),$$

while reference inputs are drawn from

$$x \sim \mathcal{N}(0, \mathbf{\Sigma}_{\mathcal{R}}).$$

We set the base point to $\theta_0 = 0$. In this model, the per-example gradient and target gradient at the base point are

$$g_i = \nabla_\theta \ell(z_i;\theta_0) = -\frac{1}{\sigma^2} y_i x_i, \qquad g_{\mathcal{T}} = \nabla_\theta L_{\mathcal{T}}(\theta_0) = -\frac{1}{\sigma^2}\mathbf{\Sigma}_{\mathcal{T}}\beta_{\mathcal{T}},$$

and the reference Fisher is

$$\mathbf{F}_{\mathcal{R}} = \mathbb{E}_{x\sim\mathcal{R}}\,\mathbb{E}_{y\sim p_{\theta_0}(\cdot|x)}\left[\nabla_\theta \log p_{\theta_0}(y \mid x)\nabla_\theta \log p_{\theta_0}(y \mid x)^\top\right] = \frac{1}{\sigma^2}\mathbf{\Sigma}_{\mathcal{R}}.$$

In this synthetic setting, the raw per-example gradient g_i is available in closed form; it plays the same role as $\widehat{g}_i^{\mathrm{raw}}$ in the empirical algorithm. Moreover, unlike the general nonlinear setting where the quadratic expansion is local, the reference drift is *exact* here:

$$D_{\mathcal{R}}(\Delta\theta) = \mathbb{E}_{x\sim\mathcal{R}}\left[D_{\mathrm{KL}}(p_{\theta_0}(\cdot \mid x) \,\|\, p_{\theta_0+\Delta\theta}(\cdot \mid x))\right] = \frac{1}{2\sigma^2}\Delta\theta^\top\mathbf{\Sigma}_{\mathcal{R}}\Delta\theta = \frac{1}{2}\Delta\theta^\top\mathbf{F}_{\mathcal{R}}\Delta\theta.$$

The exact target improvement also includes curvature:

$$L_{\mathcal{T}}(0) - L_{\mathcal{T}}(\Delta\theta) = -g_{\mathcal{T}}^\top\Delta\theta - \frac{1}{2\sigma^2}\Delta\theta^\top\mathbf{\Sigma}_{\mathcal{T}}\Delta\theta.$$

Thus first-order alignment is not identical to realized target improvement, even though reference drift is exactly quadratic here.

$$v_\alpha = \frac{u_1}{25+\alpha} + \frac{u_2}{1+\alpha}$$

(a) $\mathbf{U} = \mathbf{I}_2$: u_1 is 45° from $-g_\mathcal{T}$.

$$v_\alpha = \frac{(\sqrt{3}+1)u_1}{2(25+\alpha)} + \frac{(\sqrt{3}-1)u_2}{2(1+\alpha)}$$

(b) $\mathbf{U}$ rotated by $\pi/6$: u_1 is 15° from $-g_\mathcal{T}$.

Figure 6: Analytic damped directions $v_\alpha/\|v_\alpha\|_2$, $\alpha \geq 0$, for the two geometries in (27). The reference spectrum is $(25, 1)$, so u_2 has positive curvature. Normalization describes direction only; the feasible step length depends on both budgets. In particular, $\rho = 0$ permits only the zero update in this positive-definite example.

Updates and evaluation. Given a selected subset $\mathcal{S} \subset \mathcal{D}_{\text{train}}$ of size S, we form the update

$$\Delta\theta(\mathcal{S}) = -\frac{\eta}{S}\mathbf{P}\sum_{i\in\mathcal{S}} g_i,$$

where $\mathbf{P}$ is the preconditioner associated with the update rule. In particular, we use $\mathbf{P} = \mathbf{I}_d$ for stochastic gradient descent. We evaluate each selected update using

$$\text{Gain}_{\text{held}}(\mathcal{S}) = L_\mathcal{T}(\theta_0) - L_\mathcal{T}(\theta_0 + \Delta\theta(\mathcal{S})),$$

$$\text{Drift}_{\text{held}}(\mathcal{S}) = D_\mathcal{R}(\Delta\theta(\mathcal{S})),$$

and, when the population quantities are known and both updates are nonzero, the directional error relative to the oracle update in the positive-multiplier regime

$$\Delta\theta^\star \propto -(\mathbf{F}_\mathcal{R} + \alpha\mathbf{I}_d)^{-1} g_\mathcal{T}, \qquad \text{DirErr}(\mathcal{S}) = 1 - \cos\big(\Delta\theta(\mathcal{S}), \Delta\theta^\star\big).$$

Experiment 1: Low-dimensional geometry ($d = 2$). In the first experiment, we ask how the reference geometry reshapes $\Delta\theta^\star$ as the damping α varies, and also how this reshaping depends on the orientation of the expensive direction u_1 relative to the target direction $-g_\mathcal{T}$. We choose

$$\mathbf{\Sigma}_\mathcal{R} = \mathbf{U}\text{diag}(25, 1)\mathbf{U}^\top, \qquad \mathbf{\Sigma}_\mathcal{T} = \mathbf{I}_2, \qquad \beta_\mathcal{T} = (1, 1)^\top, \qquad \sigma^2 = 1, \tag{27}$$

with $\mathbf{U} = (u_1 \quad u_2)$ a rotation, so $-g_\mathcal{T} = \beta_\mathcal{T}$ points at 45° and u_1 costs $25\times$ more drift than u_2. We compare (a) $\mathbf{U} = \mathbf{I}_2$, where u_1 is 45° from $-g_\mathcal{T}$, and (b) $\mathbf{U}$ a rotation by $\pi/6$, where u_1 is only 15° away (used in Experiments 2–3).

In both cases reference weighting turns the direction toward u_2, and increasing damping rotates it back toward $-g_\mathcal{T}$. The rotation depends on both the initial angle and α; neither orientation has a uniformly larger rotation for all damping values.

Experiment 2: Finite-pool subset selection in $d = 2$. To visualize which points each method would preferentially select, we consider the setting of Experiment 1 (Equation 27) and introduce a finite candidate pool, since Experiment 1 contains no data to select from. We draw it from a three-component Gaussian mixture with different input-variance orientations:

$$x_i \sim \frac{1}{3}\mathcal{N}(0, \mathbf{\Sigma}_1) + \frac{1}{3}\mathcal{N}(0, \mathbf{\Sigma}_2) + \frac{1}{3}\mathcal{N}(0, \mathbf{\Sigma}_3), \qquad y_i = x_i^\top \beta_\mathcal{T} + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, \sigma^2), \tag{28}$$

with

$$\mathbf{\Sigma}_1 = \mathbf{U}\text{diag}(9, 0.25)\mathbf{U}^\top, \qquad \mathbf{\Sigma}_2 = \mathbf{U}\text{diag}(0.25, 9)\mathbf{U}^\top, \qquad \mathbf{\Sigma}_3 = \mathbf{I}_2.$$

The first two components concentrate input variance along high- and low-curvature reference directions, respectively, and the third supplies isotropic inputs. The induced gradients also depend on labels and noise and need not be isotropic; component membership alone does not guarantee usefulness or a particular update cost. The raw data—and the corresponding descent directions—are shown in Figure 7. For each method, we visualize the candidate descent directions, highlight the selected subset, and plot the resulting update direction (Figure 8).

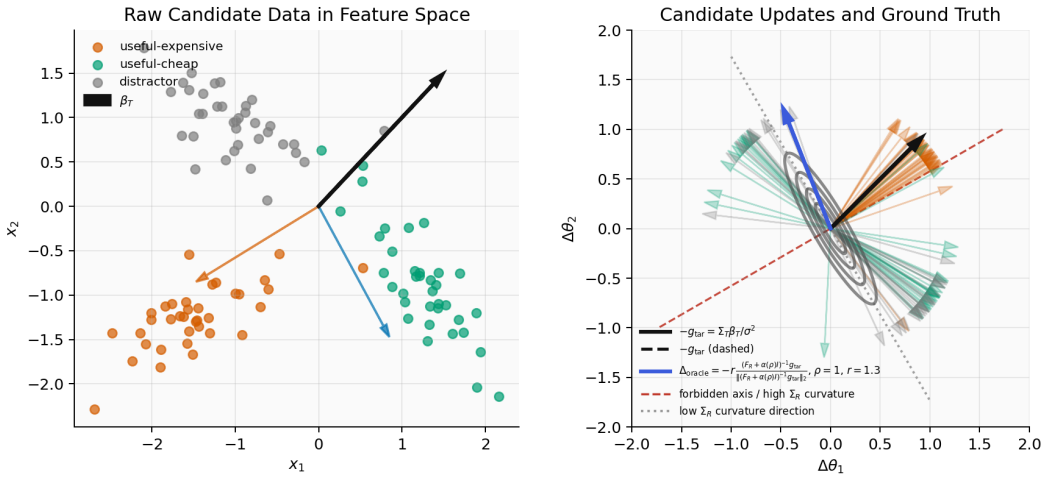


Figure 7: The candidate pool in the setting of Experiment 2 (Equation (28)). *Left:* the inputs x_i. *Right:* the candidate descent directions, proportional to $-g_i$, displayed in update coordinates. Here $-g_{\mathcal{T}} = \sigma^{-2}\mathbf{\Sigma}_{\mathcal{T}}\beta_{\mathcal{T}} = \beta_{\mathcal{T}}$, so the black arrow is the target direction $-g_{\mathcal{T}}$.

Experiment 3: Trade-offs in Alignment/ off-target drift in $d = 2$. Experiments 1 and 2 hold the target direction fixed. We now compare two target specifications at fixed reference geometry to illustrate how the target–drift trade-off depends on *entanglement*: the extent to which $-g_{\mathcal{T}}$ overlaps the expensive directions of $\mathbf{F}_{\mathcal{R}}$. Results are reported in Figure 2.

We compare the two instances:

- **entangled:** $\beta_{\mathcal{T}} = (1, 1)^\top$, which lies 15° from u_1 (the rotated case of Experiment 1);
- **disentangled:** $\beta_{\mathcal{T}} = u_2 = (-\frac{1}{2}, \frac{\sqrt{3}}{2})^\top$, exactly orthogonal to u_1.

The two target gradients also differ in norm ($\sqrt{2}$ versus 1), so this comparison does not isolate orientation alone. For each method we measure, for its selected update $\Delta\theta$:

$$\text{alignment: } -g_{\mathcal{T}}^\top \Delta\theta, \qquad \text{drift: } \tfrac{1}{2}\Delta\theta^\top \mathbf{F}_{\mathcal{R}} \Delta\theta,$$

The plotted drift axis in Figure 2 is labeled $\Delta\theta^\top \mathbf{F}_{\mathcal{R}} \Delta\theta$, twice the theoretical quadratic cost above; the displayed red line is a nominal reference level and is not used here to certify feasibility. Results are averaged over 100 random seeds.

In the entangled panel, SAFE-DRIFT reduces mean plotted drift relative to the other selectors while giving up some target alignment. In the disentangled panel, the selectors attain different alignment–drift trade-offs; SAFE-DRIFT is not the minimum-drift method. A target direction lying in a lower-curvature eigenspace does not by itself make the budget inactive, and approximate subset matching does not certify every selected update as feasible.

Experiment 4: Estimation error and spectral decay in $d = 200$. Experiments 1–3 assume the population quantities are known. We now reintroduce the estimation problem: $g_{\mathcal{T}}$ and $\mathbf{F}_{\mathcal{R}}$ must be recovered from finite samples and represented in a shared low-rank basis. We use a synthetic high-dimensional teacher–student model in which the reference and target geometries can have different spectral structure. In all runs, the ambient dimension is $d = 200$, the true reference rank is $r = 40$, the candidate pool contains $N_C = 800$ examples, the target gradient is estimated from $N_T = 32$ samples, the reference Fisher is estimated from $N_R = 300$ samples, and the subset budget is $S = 40$. We generate the reference Fisher as $\mathbf{F}_{\mathcal{R}} = \sigma^{-2}\mathbf{U}_R \mathrm{diag}(\lambda^R)\mathbf{U}_R^\top$ and the target covariance as $\mathbf{\Sigma}_{\mathcal{T}} = \mathbf{U}_T \mathrm{diag}(\lambda^T)\mathbf{U}_T^\top$, where $\mathbf{U}_R$ and $\mathbf{U}_T$ are independent random orthonormal bases. We sweep both the reference and target spectra over geometric, linear, and power-law decay, and in a denser variant we sweep power-law exponents $(q_R, q_T) \in \{0.5, 0.75, 1.0, 1.5, 2.0\}^2$. For each spectrum pair, we compare SAFE-DRIFT-full, SAFE-DRIFT-low-rank, SAFE-DRIFT-reference-only, SAFE-DRIFT-task-only, SAFE-DRIFT-random-K, SAFE-DRIFT-diagonal, LESS, random, and

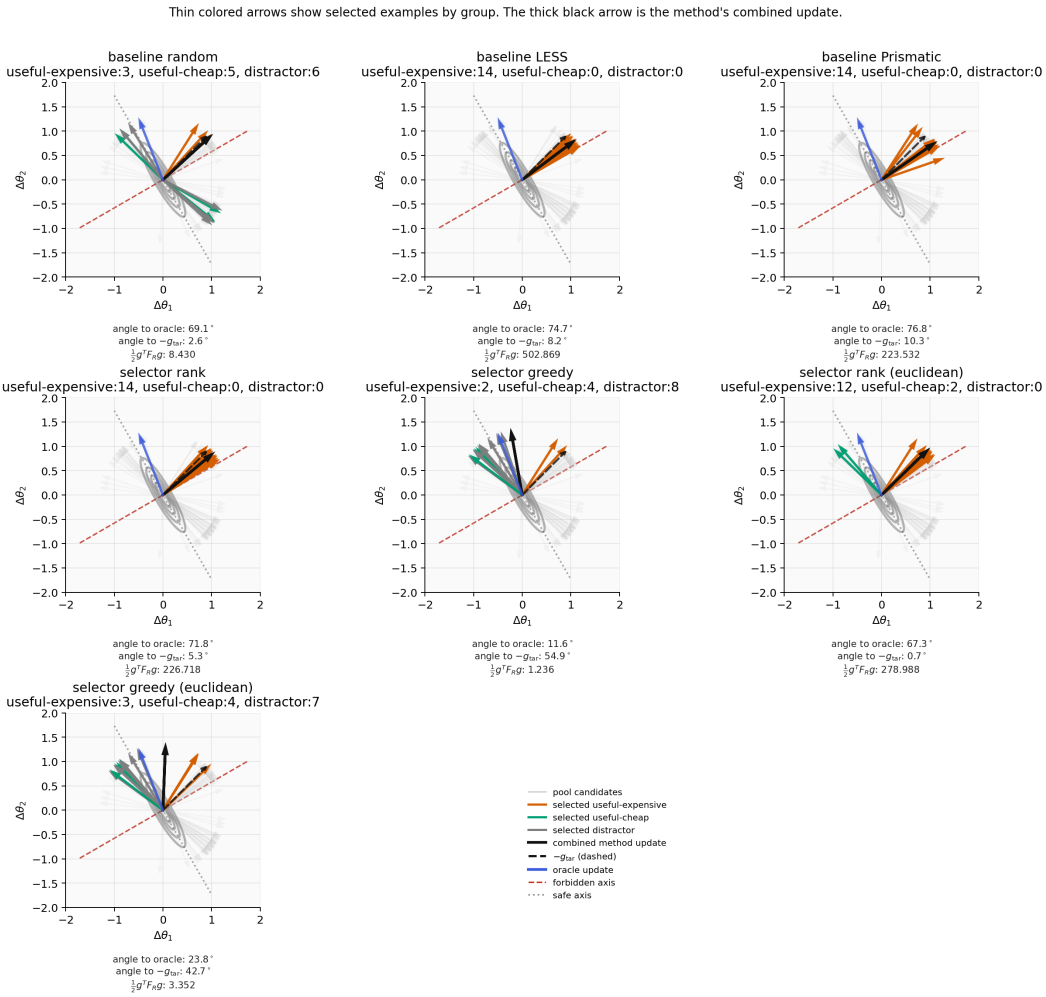


Figure 8: Selection methods on the pool of Figure 7. For each method: candidate descent directions proportional to $-g_i$, the selected subset (highlighted), and the resulting update direction $\Delta\theta(\mathcal{S})$. The graphic's "forbidden" and "safe" labels denote the high- and low-curvature axes; neither is prohibited or has zero reference cost under the spectrum $(25, 1)$.

Prismatic over requested working ranks $K_R + K_T \in \{10, 20, 40, 60, 80, 120\}$. All SAFE-DRIFT methods are run under a fixed reference-drift budget $\rho = 1$, and the realized subset update is always the SGD-style average $\Delta\theta(\mathcal{S}) = -(\eta/S)\sum_{i\in\mathcal{S}} g_i$. We report the exact realized reference drift $\frac{1}{2}\Delta\theta^\top \mathbf{F}_\mathcal{R}\Delta\theta$, the exact target gain, and the absolute regret relative to SAFE-DRIFT full, averaged over 10 random seeds and 3 target regimes (for a total of 30 instances per spectrum pair), with standard-error bars shown in all plots.

D.2 Additional details on experiments

Unless a setting specifies a separate evaluation set, experiments use the following configuration. The candidate pool $\mathcal{D}_{\text{train}}$ is split into a 3% warmup subset used to estimate the Adam preconditioner $\mathbf{P}$ and a remaining 97% from which training points are selected. The target set $\mathcal{D}_{\text{tar}}$ is divided into three parts: one for computing the target gradient $\widehat{g}_T$, one for validation, and one as a held-out test set. The reference set $\mathcal{D}_{\text{ref}}$ is split evenly between estimating $\widehat{\mathbf{F}}_R$ and evaluating reference drift. Each selected subset is then used for fine-tuning with the LoRA training recipe in Table 1 using dropout 0.1, target modules `{q,k,v,o}`, Adam update with learning rate 2×10^{-4}, 1 epoch, effective batch size 32, and sequence length 4096. Gradients for selection and model updates are computed over LoRA adapters [30] applied to the attention modules.

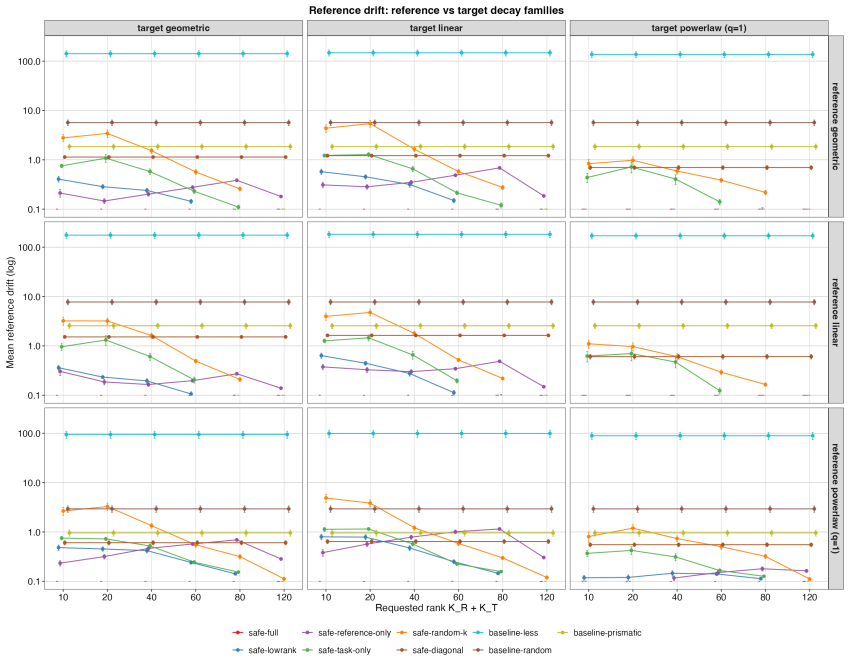


Figure 9: Off-target drift for the baselines and different variants of SAFE-DRIFT (colors), per requested rank (x-axis) as a function of the spectra of $\mathbf{\Sigma}_{\mathcal{T}}$ and $\mathbf{F}_{\mathcal{R}}$.

Throughout this section, we write *reference Fisher cost* to denote the predicted drift D_{F} and *held-out KL* to denote the realized drift D_{KL}.

Compute Resources For model fine-tuning, we use Huggingface Accelerate [65] and utilize Zero3 optimizer in DeepSpeed on two A100 GPUs.

D.2.1 Baseline selectors

We compare SAFE-DRIFT against five baselines. Full fine-tunes on the entire candidate pool without selection, and Random uniformly subsamples the pool at the target budget.

All gradient-based methods—LESS, Prismatic Synthesis, and SAFE-DRIFT—operate on the same set of low-rank projected gradients computed in the shared working basis $\mathbf{U}_K$, ensuring that differences in selection performance reflect the selection criterion rather than gradient featurization. LESS [6] scores each candidate by the cosine similarity between its projected gradient and the projected target gradient, using the low-rank Adam-style preconditioner as the inner product metric. We adapt Prismatic Synthesis [19] to promote diversity by clustering candidate gradients via k-means and sampling inversely proportional to cluster size, with the number of clusters set to 10% of the pool and 20 Lloyd iterations.

DSIR [17] operates on hashed n-gram features rather than gradients, scoring candidates by the dot product between their hashed n-gram count vectors and the log importance ratio $\log p - \log q$, estimated from a 30% subsample of the pool with $10{,}000$ hash buckets.

D.2.2 Setting 1: Code generation

The goal is to improve Python programming ability while leaving the base model's behavior on stereotypes unchanged. The candidate pool is CodeAlpaca [34], a 20K dataset of general-purpose programming instructions generated via self-instruct [66]. The target set is HumanEval [35], a benchmark of 164 hand-written Python programming problems; we split these into 109 test, 27 validation, and 28 examples for computing the target gradient $\widehat{g}_T$. The reference distribution is defined by the intrasentence split of StereoSet [36], which presents each prompt with a stereotypical, an anti-stereotypical, and an unrelated completion. We retain the stereotypical and anti-stereotypical

completions in equal proportion to construct a balanced reference set. This results in 4.21K samples which we split into half for reference Fisher estimation and for evaluating off-target drift. Since computing gradients over the full candidate pool is memory-intensive, the task-side basis $\widehat{\mathbf{U}}_{K_T}$ is estimated from $4{,}000$ randomly subsampled candidates. We ablate over reference ranks $K_R \in \{128, 256, 512\}$, with the residual ranks detailed in Appendix D.7.3, at selection budget 15%; results are reported in Figure 22 and 23.

Evaluation metrics. For the target task, we evaluate code generation quality via pass@k rates following the protocol of [35]: for each problem, k candidate solutions are sampled and the problem is considered solved if at least one passes all unit tests. We report pass@1 and pass@10 (mean and standard deviation) for all selectors in Table 2. For the reference distribution, we use StereoSet score, defined as the proportion of prompts for which the model assigns higher average token probability to the stereotypical completion than to the anti-stereotypical one [36]. A score of 50 means that the stereotypical completion receives higher average token probability on half the prompts; this aggregate statistic alone does not establish that a model is unbiased. The base model scores 64.9 on this metric. We additionally record the probability margin, the difference in average token probability between the stereotypical and anti-stereotypical completions. All reference metrics are reported as changes from the base model.

D.2.3 Setting 2: Improving medical knowledge

The goal of this experiment is to improve medical question-answering ability of a model while leaving the base model's broader non-medical behavior unchanged. The candidate pool is MedMCQA [37], a large collection of multiple-choice medical exam questions. For the target set, we subsample 256 samples from MedQA [38], which consists of medical licensing-style questions. The reference set combines BBQ [39] and TruthfulQA [40], covering social bias and factual reliability respectively, forming a non-medical alignment distribution.

Evaluation metrics. We measure fine-tuning performance on the target task using MedQA multiple-choice accuracy, reporting development accuracy for model selection and held-out test accuracy for final evaluation. To assess off-target drift, we evaluate the fine-tuned model on several non-medical benchmarks drawn from the reference evaluation suite. On BBQ [39], we report multiple-choice accuracy as a measure of robustness to social bias. On TruthfulQA [40], we report token-level F1 to capture factual reliability under free-form generation. On IFEval [45], we report prompt accuracy and instruction accuracy, which measure whether the model satisfies explicit formatting and content constraints.

Additional experiments To further show the importance of finetuning in this context, we compare the performance of random selection across various pool sizes. The results are shown in Figure 10. This figure highlights, in particular, that increasing the pool size does not automatically lead to better performance on the target task (here, MedQA), thereby making a case for the use of data selection in this context.

D.2.4 Setting 3: Improving emotional support

We seek to improve emotional-support dialogue performance while leaving the base model's behavior on mathematical reasoning unchanged. The candidate pool is EmpatheticDialogues [41], a dataset of emotion-grounded open-domain conversations. The target set is ESConv [42], a benchmark of emotional-support conversations. To define the protected reference distribution, we use GSM8K [43], a dataset of grade-school math word problems that probes multi-step mathematical reasoning.

Evaluation metrics. We evaluate target-task performance on ESConv using ROUGE-L F1, which is better suited than exact match for this open-ended generation setting. To measure drift away from the reference capability, we report accuracy on GSM8K and additionally on MMLU (zero-shot) as a broader out-of-domain measure of reasoning and knowledge preservation.

Additional experiments and diagnostics. As for setting 2, we compare the performance of random selection across various pool sizes. The results are shown in Figure 11. Again, we observe that increasing the pool size does not automatically lead to better performance on the target task.

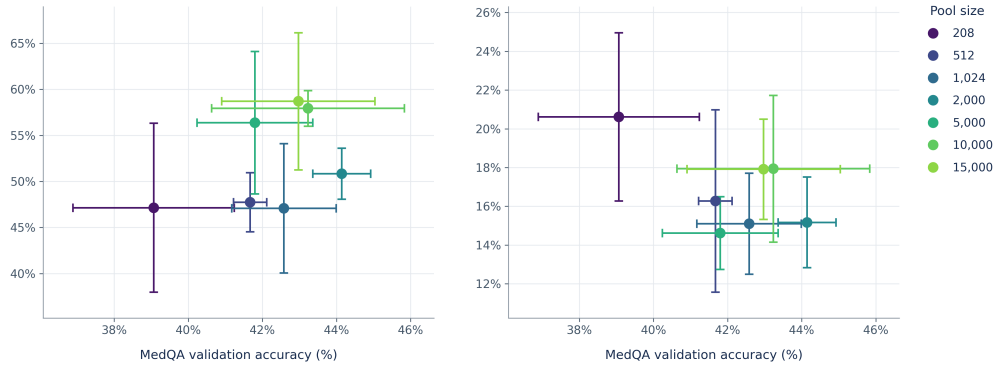


Figure 10: Pool-size effects on MedQA validation accuracy versus BBQ validation accuracy (left) and TruthfulQA validation token F1 (right) Training uses LoRA rank 16, $\alpha = 32$.

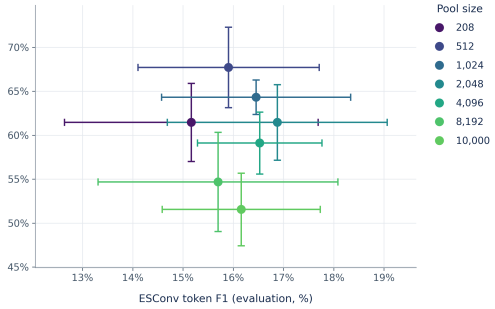


Figure 11: Pool-size effects on ESConv validation score versus GSM8K validation accuracy. Training uses LoRA rank 64, $\alpha = 128$.

D.2.5 Setting 4: FLAN instruction tuning

The goal is to improve instruction-following behavior while preserving the base model's performance on an unrelated knowledge-and-reasoning reference task. This setting is deliberately off-target: the selector is asked to identify useful instruction-tuning examples from a broad FLAN-v2 [44] pool using a small IFEval-derived target set, while the reference geometry is estimated from MMLU. Thus, unlike settings where the target, candidate, and reference distributions are closely related, this experiment tests whether a selector can extract target-relevant instruction-following signal without unnecessarily perturbing models' factual knowledge. The reported damping value for SAFE-DRIFT is $\alpha = 2 \times 10^{-4}$.

Evaluation metrics. For the target task, we evaluate instruction following on the official IFEval test set [45], which contains 491 prompts. Evaluation is performed with using OLMES [67]. We report the four standard IFEval accuracies: prompt-level strict, prompt-level loose, instruction-level strict, and instruction-level loose. Prompt-level accuracy requires every constraint in a prompt to be satisfied simultaneously, whereas instruction-level accuracy scores each individual constraint separately and averages over constraints. These aggregate accuracies use different denominators. When the number of constraints varies across prompts, neither aggregate score must exceed the other; for example, one correct single-constraint prompt and one entirely incorrect three-constraint prompt give prompt accuracy $1/2$ but instruction accuracy $1/4$. Strict scoring applies the original verifier directly, while loose scoring applies light normalization such as lowercasing, whitespace cleanup, and simple punctuation handling before verification. The OLMES primary scores for IFEval is prompt-level strict and instruction strict accuracy, which we use as the main target metric in Table 2.

For the reference task, we evaluate MMLU on a fixed held-out 200-item test slice. The chat-format mode evaluates MMLU as open-ended generation under the same chat template used during fine-tuning, making it sensitive to whether the fine-tuned model still emits clean single-letter answers

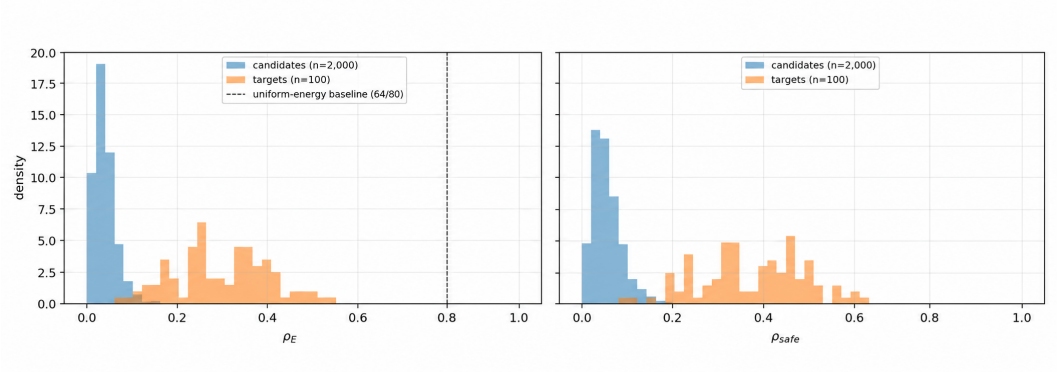


Figure 12: Reference-subspace entanglement diagnostics for Setting 3. We draw the density of ρ_E and ρ_{safe} from (26) with $K_R = 64$, and $K = 80$. Values close to 0 indicate smaller reference-block energy per coordinate relative to the residual block, and values close to 1 indicate the reverse, under the respective weighting in (26).

such as `A`, `B`, `C`, or `D`. The loglikelihood mode instead compares

$$P_\theta(\texttt{A} \mid \text{prompt}),\ P_\theta(\texttt{B} \mid \text{prompt}),\ P_\theta(\texttt{C} \mid \text{prompt}),\ P_\theta(\texttt{D} \mid \text{prompt})$$

and selects the answer with maximum probability. This mode is less sensitive to chat-template formatting and provides a cleaner estimate of whether the model's underlying multiple-choice knowledge has changed. Since the purpose of the reference task is preservation rather than improvement. Values closest to zero indicate the least off-target movement from the base model.

D.3 Entanglement Diagnostics

We show histograms of the reference-entanglement diagnostics from Section C.3.2 for Settings 3 and 4. These ratios normalize each block by its dimension, so they compare energy per coordinate rather than the unnormalized fraction of gradient energy in the reference span. In Setting 3, the concentration near zero indicates weak reference-block energy per coordinate under the respective diagnostic weighting (Figure 12). Neither these ratios nor $\pi_{Q_{\text{ent}}} = 0.08$ bounds the absolute reference cost of a selected update. Target gradients have larger diagnostic values than candidate gradients; one possible explanation is that a basis estimated from many more candidates than targets underrepresents target-specific residual directions.

Setting 4 is substantially more entangled, as displayed in Figure 13 and the entanglement scores of $\pi_{Q_{\text{ent}}} = 0.26$ and $\pi_{Q_{\text{clean}}} = 0.22$. The quadrant frequencies indicate positive residual-task alignment for many candidates with positive reference-block alignment. They do not determine total target gain or establish a necessary loss in MMLU accuracy.

D.4 Comparison Against Behavior-Preservation Baselines

We provide the complementary results of replay-based methods of DSIR, LESS, and Prismatic of Section 3.1 in Table 3. We additionally experiment with augmenting the reference KL-divergence penalty to the SFT loss:

$$\mathcal{L}_{\text{SFT}}(\theta) + \lambda\, \mathbb{E}_{u \sim \widehat{\mathcal{R}}_{\text{seq}}} \Big[\frac{1}{|u|} \sum_{t=1}^{|u|} D_{\text{KL}}\big(p_{\theta_0}(\cdot|u_{<t}) \| p_\theta(\cdot|u_{<t})\big) \Big],$$

where $\widehat{\mathcal{R}}_{\text{seq}}$ denotes the empirical distribution of nonempty observed reference token sequences, distinct from the prompt distribution in (2). The displayed regularizer averages per-sequence token means; for unequal sequence lengths this differs from a corpus-wide token-weighted mean. A fixed observed-prefix average is not generally sequence-level KL. The sequence-KL chain rule instead requires conditional KL terms averaged under the base model's prefix distribution, without per-sequence length normalization and with termination accounted for when sequence lengths vary. We

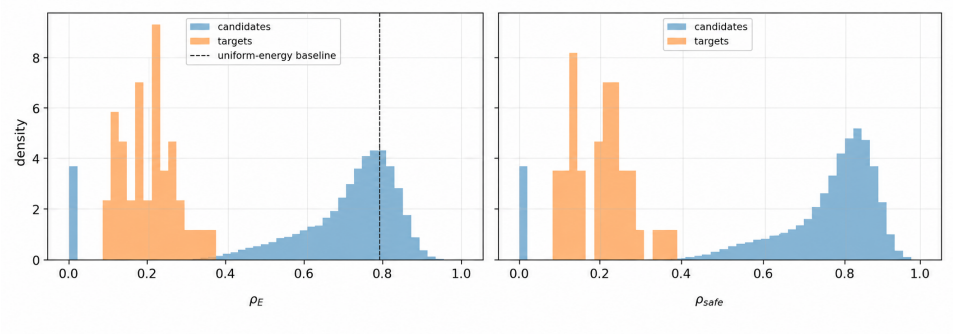


Figure 13: Reference-subspace entanglement diagnostics for Setting 4. We draw the density of ρ_E and ρ_{safe} from (26) with $K_R = 128$, and $K = 160$. The displayed dashed line marks $K_R/K = 128/160 = 0.8$, which is not the equal-coordinate-energy baseline for these dimension-normalized diagnostics. Those baselines are $1/2$ for ρ_E and $\bar{\lambda}_R/(\bar{\lambda}_R + \alpha)$ for ρ_{safe}, where $\bar{\lambda}_R$ is the mean retained reference eigenvalue (Section C.3.2).

(a) Rehearsal (replay) baselines — Setting 1

Method	Replay %	Δ HumanEval@1 ↑	Δ HumanEval@10 ↑	Held-out KL (D_{KL}) ↓
DSIR	10	0.0516 ± 0.0096	0.1362 ± 0.0040	0.0190 ± 0.0030
DSIR	20	0.0429 ± 0.0108	0.1260 ± 0.0090	0.0223 ± 0.0007
LESS	10	0.0217 ± 0.0056	0.0870 ± 0.0090	0.0152 ± 0.0017
LESS	20	0.0243 ± 0.0082	0.0995 ± 0.0153	0.0181 ± 0.0028
Prismatic	10	0.0185 ± 0.0009	0.0793 ± 0.0060	0.0197 ± 0.0020
Prismatic	20	0.0134 ± 0.0038	0.0750 ± 0.0130	0.0230 ± 0.0040
SAFE-DRIFT	0	$\mathbf{0.132 \pm 0.004}$	$\mathbf{0.183 \pm 0.023}$	$\mathbf{0.0127 \pm 0.0006}$

(b) KL-regularized selection baselines — Settings 2 and 3

Setting	Method	Selected λ	Δ test target ↑	Δ test reference ↑	Held-out KL (D_{KL}) ↓
2	DSIR	—	$+3.59 \pm 2.04$	-8.10 ± 3.58	4.210 ± 0.528
	LESS	$\lambda = 0.1$	$+2.08 \pm 0.60$	$+0.73 \pm 1.26$	1.974 ± 0.224
	Prismatic	$\lambda = 1$	$+2.99 \pm 1.63$	$+0.89 \pm 2.07$	$\mathbf{1.534 \pm 0.025}$
	SAFE-DRIFT	—	$+3.36 \pm 1.42$	-3.50 ± 3.24	4.003 ± 0.467
3	DSIR	—	$+4.65 \pm 1.18$	–	$\mathbf{0.317 \pm 0.025}$
	LESS	$\lambda = 0$	$\mathbf{+6.77 \pm 1.18}$	–	0.368 ± 0.039
	Prismatic	$\lambda = 0$	$+4.55 \pm 0.97$	–	0.364 ± 0.023
	SAFE-DRIFT	—	$+4.51 \pm 1.10$	–	0.391 ± 0.011

Table 3: *Comparison against behavior-preservation baselines.* **(a)** Rehearsal: reference examples, drawn from the Fisher-estimation split and disjoint from the drift-evaluation set, are mixed into training at 10% and 20% of the budget. SAFE-DRIFT never trains on the reference set. **(b)** LESS and Prismatic augmented with a reference KL penalty with λ chosen on validation set by the geometric mean of target and reference accuracy. We show means and standard deviations of differences from each setting's base model. For Setting 3, reference-score changes are reported in Table 2; panel (b) reports target changes and KL for this comparison.

sweep over $\lambda \in \{0, 0.01, 0.1, 1\}$ and select the best λ on a validation set using the geometric mean of target and reference performance.

When replay increases from 10% to 20%, mean target performance falls for DSIR and Prismatic but rises modestly for LESS; held-out KL rises for all three. For the KL penalty, the reported Spearman correlations between positive λ and held-out KL are -0.949 and -0.896 for LESS and Prismatic in Setting 2, and -0.949 for both in Setting 3. In Setting 2, regularized Prismatic has the lowest mean KL, whereas SAFE-DRIFT has comparable target gain to DSIR with less reference-score degradation; its KL is 4.003 versus DSIR's 4.210, a modest reduction. In Setting 3, validation selects $\lambda = 0$ for LESS and Prismatic. SAFE-DRIFT has the largest mean reference-score improvement and smallest reference-score standard deviation, but not the lowest KL. These metrics therefore capture distinct aspects of reference behavior.

Setting	Optimizer	Hyperparam. configs	Seeds per cell	ρ vs. target	ρ vs. D_{task}	Target vs. D_{task}	D_{F} vs. D_{KL}
Setting 2	Adam	9	5	0.753	0.483	0.561	0.967
Setting 3	Adam	18	5	0.672	-0.735	-0.820	0.670

Table 4: *Spearman rank correlations for the drift budget ρ.* D_{task} denotes reference-score degradation, measured as base minus fine-tuned score. The negative Setting 3 correlations indicate that target and reference accuracy increase together—consistent with the low target–reference entanglement of that setting. The final column correlates the predicted drift D_{F} with the realized drift D_{KL} over the same runs.

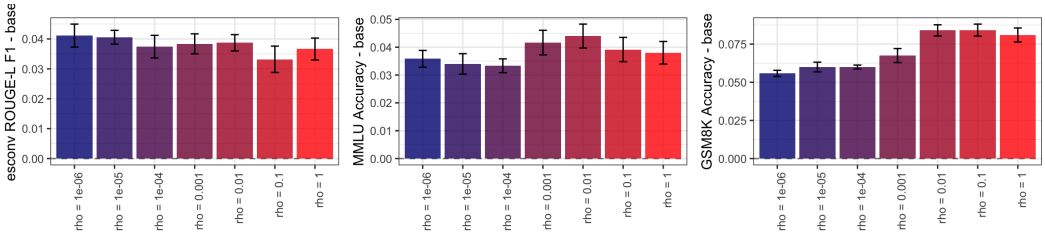


Figure 14: Effect of ρ in ESconv (Setting 3)

D.5 Choosing the Drift Budget ρ

The drift budget ρ is the main user-facing control in SAFE-DRIFT: tighter budgets avoid high-cost reference directions, looser ones let the selector pursue target gain. Section 3.1 sweeps ρ over $\{10^{-5}, 10^{-2}, 10^{0}\}$ in Setting 2; Table 4 reports rank correlations over a larger sweep of nine hyperparameter configurations in Setting 2 and eighteen in Setting 3, five seeds each. Increasing ρ is positively associated with target performance (Spearman correlation 0.753 and 0.672), but its association with reference drift reverses sign between settings: 0.483 in Setting 2, where a looser budget costs reference accuracy, against -0.735 in Setting 3, where it improves both axes at once (Figure 14). ρ is thus a trade-off parameter, not a universally monotone regularizer. The sweep also shows an empirical association between the predicted Fisher cost and held-out KL: D_{F} and D_{KL} correlate at $\rho_s = 0.967$ and 0.670. Figures 16 and 18 provide complementary candidate-level diagnostics; they do not certify the KL of the final model.

We recommend tuning ρ on validation target and reference-drift sets. For a local oracle update satisfying the population quadratic budget and $\|\Delta\theta\| \le \epsilon < r_0$, the assumptions of Proposition 1 give $D_{\mathcal{R}} \le \rho + M\epsilon^3/6$. Pinsker's inequality and Jensen's inequality then bound the reference-input-averaged change in any conditional event probability by $\sqrt{(\rho + M\epsilon^3/6)/2}$. This is neither a pointwise guarantee for every prompt nor a guarantee for the selected subset or final SFT model. Empirical Fisher estimation, truncation, subset-matching error, and the training trajectory require separate control; hence validation determines the practical meaning of a chosen ρ.

D.6 Robustness across base models, composite references, and fine-tuning methods

D.6.1 Robustness across base models

We repeat Setting 1 across two model families and five configurations—Qwen2.5 at 1.5B, 3B and 7B, and OLMo-2 at 7B and 13B—at a 15% selection budget with three seeds. Table 5 reports the full numbers summarized in Figure 3(a–c), including fine-tuning on the full candidate pool. Safe-Drift attains the lowest reported log Fisher cost in every configuration, reducing D_{KL} by 31.1–60.0% relative to full-data fine-tuning. Realized drift D_{KL} is less uniform: SAFE-DRIFT is lowest at Qwen2.5-1.5B and OLMo-2-13B, and competitive but not best elsewhere. Target performance shows a similar split, with SAFE-DRIFT leading on the smaller Qwen models but not the larger ones—trading target gain for drift control where the target directions are entangled with the reference geometry.

Model	Evaluation	Random	DSIR	LESS	Prismatic	SAFE-DRIFT	Full Data
Qwen2.5-1.5B	HumanEval pass@1 ↑	0.128 ± 0.023	0.133 ± 0.003	0.157 ± 0.083	0.165 ± 0.018	$\mathbf{0.283 \pm 0.020}$	0.327 ± 0.010
	$\log_{10} D_{\mathrm{F}} \downarrow$	-5.58 ± 0.04	-5.06 ± 0.06	-5.48 ± 0.06	-5.64 ± 0.04	$\mathbf{-6.29 \pm 0.04}$	-5.60 ± 0.05
	$D_{\mathrm{KL}} \downarrow$	0.0049 ± 0.0001	0.0055 ± 0.0006	0.0057 ± 0.0005	$\mathbf{0.0044 \pm 0.0001}$	$\mathbf{0.0044 \pm 0.0003}$	0.0110 ± 0.0001
Qwen2.5-3B	HumanEval pass@1 ↑	0.162 ± 0.028	0.175 ± 0.032	0.193 ± 0.017	0.129 ± 0.049	$\mathbf{0.242 \pm 0.053}$	0.248 ± 0.026
	$\log_{10} D_{\mathrm{F}} \downarrow$	-6.16 ± 0.03	-5.65 ± 0.06	-6.13 ± 0.03	-6.22 ± 0.01	$\mathbf{-6.89 \pm 0.03}$	-6.25 ± 0.06
	$D_{\mathrm{KL}} \downarrow$	$\mathbf{0.0031 \pm 0.0003}$	0.0038 ± 0.0004	0.0042 ± 0.0008	$\mathbf{0.0031 \pm 0.0001}$	0.0036 ± 0.0002	0.0091 ± 0.0008
Qwen2.5-7B	HumanEval pass@1 ↑	0.117 ± 0.041	0.089 ± 0.045	$\mathbf{0.180 \pm 0.055}$	0.103 ± 0.053	0.159 ± 0.044	0.269 ± 0.040
	$\log_{10} D_{\mathrm{F}} \downarrow$	-6.89 ± 0.02	-6.41 ± 0.04	-6.81 ± 0.02	-6.89 ± 0.02	$\mathbf{-7.62 \pm 0.03}$	-6.88 ± 0.13
	$D_{\mathrm{KL}} \downarrow$	0.0038 ± 0.0003	0.0038 ± 0.0002	0.0039 ± 0.0006	$\mathbf{0.0030 \pm 0.0004}$	0.0037 ± 0.0001	0.0076 ± 0.0010
OLMo-2-7B	HumanEval pass@1 ↑	0.130 ± 0.023	$\mathbf{0.148 \pm 0.021}$	0.117 ± 0.026	0.135 ± 0.013	0.125 ± 0.033	0.123 ± 0.010
	$\log_{10} D_{\mathrm{F}} \downarrow$	-5.98 ± 0.08	-5.38 ± 0.06	-5.72 ± 0.13	-5.90 ± 0.08	$\mathbf{-7.05 \pm 0.15}$	-5.95 ± 0.07
	$D_{\mathrm{KL}} \downarrow$	0.0106 ± 0.0021	0.0110 ± 0.0005	$\mathbf{0.0098 \pm 0.0006}$	0.0129 ± 0.0031	0.0127 ± 0.0006	0.0209 ± 0.0021
OLMo-2-13B	HumanEval pass@1 ↑	0.180 ± 0.043	0.182 ± 0.029	0.154 ± 0.025	$\mathbf{0.194 \pm 0.036}$	0.161 ± 0.031	0.185 ± 0.029
	$\log_{10} D_{\mathrm{F}} \downarrow$	-6.88 ± 0.17	-6.23 ± 0.08	-6.71 ± 0.06	-6.69 ± 0.10	$\mathbf{-7.95 \pm 0.07}$	-6.83 ± 0.07
	$D_{\mathrm{KL}} \downarrow$	0.0097 ± 0.0013	0.0097 ± 0.0027	0.0087 ± 0.0011	0.0095 ± 0.0010	$\mathbf{0.0084 \pm 0.0013}$	0.0122 ± 0.0012

Table 5: *Base model ablation* (Setting 1). Two model families and five configurations; 15% selection budget and following configurations in Table 1. We measure reference Fisher cost D_{F} and the held-out reference KL D_{KL}, each corresponding to predicted and realized off-target drift. Bold denotes the best selection method per row; "Full Data" is shown for reference and not ranked.

D.6.2 Composite Reference Distributions

SAFE-DRIFT constructs its local drift surrogate from a reference distribution that the practitioner specifies. A natural question is whether the method degrades when the reference is made more complex. We therefore rerun Setting 2 with the reference constructed from several benchmarks at once: writing B for BBQ, T for TruthfulQA and K for MMLU (non-medical knowledge), we set the composition as B+T, B+K, T+K and all three jointly. Every composition optimizes the same target (MedQA accuracy).

Composition	Method	Δ Target	Δ B	Δ T	Δ K	Δ Ref GM	Δ All-ref GM	$D_{\mathrm{KL}} \downarrow$
B+T	SAFE-DRIFT	$\mathbf{+0.090}$	$\mathbf{+0.085}$	-0.105	$+0.070$	$\mathbf{-0.062}$	$\mathbf{-0.032}$	$\mathbf{5.17}$
	Random	$+0.063$	-0.049	-0.107	$+0.061$	-0.100	-0.065	5.47
	DSIR	$+0.064$	-0.054	-0.125	**+0.078**	-0.117	-0.078	5.49
	LESS	$+0.022$	$+0.028$	**-0.100**	$+0.033$	-0.072	-0.049	5.54
	Prismatic	$+0.078$	-0.042	-0.107	$+0.055$	-0.098	-0.065	5.46
B+K	SAFE-DRIFT	$\mathbf{+0.086}$	**+0.066**	**-0.093**	$+0.051$	$\mathbf{+0.058}$	$\mathbf{-0.031}$	$\mathbf{5.16}$
	Random	$+0.063$	-0.049	-0.107	$+0.061$	$+0.004$	-0.065	5.47
	DSIR	$+0.064$	-0.054	-0.125	**+0.078**	$+0.009$	-0.078	5.49
	LESS	$+0.022$	$+0.028$	-0.100	$+0.033$	$+0.030$	-0.049	5.54
	Prismatic	$+0.078$	-0.042	-0.107	$+0.055$	$+0.005$	-0.065	5.46
T+K	SAFE-DRIFT	$\mathbf{+0.094}$	**+0.046**	-0.103	$+0.063$	$\mathbf{-0.066}$	$\mathbf{-0.041}$	$\mathbf{5.29}$
	Random	$+0.063$	-0.049	-0.107	$+0.061$	-0.069	-0.065	5.47
	DSIR	$+0.064$	-0.054	-0.125	$+0.078$	-0.083	-0.078	5.49
	LESS	$+0.022$	$+0.028$	-0.100	$+0.033$	-0.070	-0.049	5.54
	Prismatic	$+0.078$	-0.042	-0.107	$+0.055$	-0.071	-0.065	5.46
B+T+K	SAFE-DRIFT	$\mathbf{+0.078}$	$+0.038$	-0.098	$+0.020$	$\mathbf{-0.048}$	$\mathbf{-0.048}$	$\mathbf{5.13}$
	Random	$+0.063$	-0.049	-0.107	$+0.061$	-0.065	-0.065	5.47
	DSIR	$+0.064$	-0.054	-0.125	$+0.078$	-0.078	-0.078	5.49
	LESS	$+0.022$	$+0.028$	-0.100	$+0.033$	-0.049	-0.049	5.54
	Prismatic	$\mathbf{+0.078}$	-0.042	-0.107	$+0.055$	-0.065	-0.065	5.46

Table 6: *Composite reference distributions* (Setting 2). B = BBQ, T = TruthfulQA, K = MMLU (non-medical knowledge). Each composition names the reference tasks used to construct the SAFE-DRIFT constraint; omitted tasks are still evaluated but not protected. All other baselines optimize the same target (MedQA accuracy). Δ Ref GM is the fine-tuned-minus-base change in the geometric mean of scores over the *protected* reference tasks; Δ All-ref GM is the corresponding change over all three. The last column reports held-out reference KL D_{KL}, whose base-model value is zero.

Table 6 reports the results. Across all four compositions, SAFE-DRIFT has the lowest held-out reference KL and the largest change in the geometric mean of protected reference scores. It matches or exceeds the best baseline target gain, with gains ranging from $+0.078$ to $+0.094$ across compositions.

Update	Test MedQA accuracy	Test reference accuracy
LoRA	$44.30\% \pm 1.42\%$	$36.76\% \pm 3.58\%$
Full SFT	$37.50\% \pm 3.87\%$	$\mathbf{38.92\%} \pm \mathbf{2.42\%}$

Table 7: *LoRA versus full supervised fine-tuning* (Setting 2).

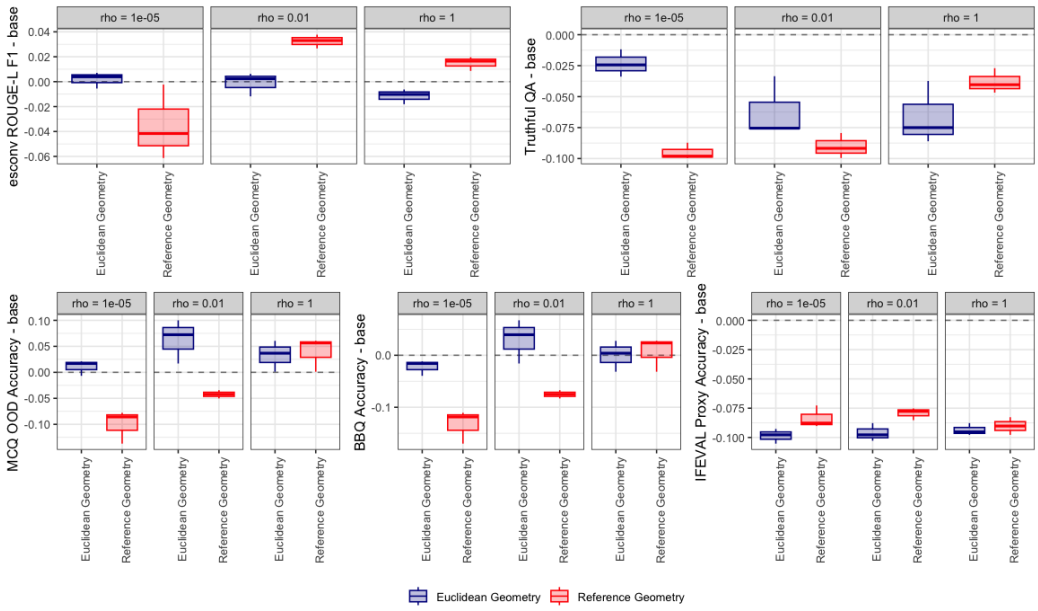


Figure 15: Target gain and off-target drift D_{task} of SAFE-DRIFT using Euclidean vs reference Fisher-induced norm objective in MedQA experiment (Setting 2). We plot the difference from the base model across drift budget ρ.

These results do not establish monotonic improvement as more references are added or minimum absolute movement on every reference task. The geometric mean over all three reference tasks also improves relative to the baselines, suggesting some shared preservation benefit, although individual benchmark scores exhibit trade-offs.

D.6.3 LoRA Versus Full Supervised Fine-Tuning

LoRA defines the trainable parameter space throughout our experiments, because it has become the dominant approach to practical parameter-efficient instruction tuning. There is no conceptual obstacle to running SAFE-DRIFT over a full SFT parameterization—only a computational one, since per-example gradients must then be formed in the full parameter space.

Table 7 compares the two on Setting 2. LoRA has the higher reported mean target accuracy, whereas full SFT has the higher reported mean reference accuracy. Higher reference accuracy alone establishes neither smaller distributional drift nor smaller movement from the base model. The table does not isolate the effect of update-subspace size or establish a general ordering between the two parameterizations.

D.7 Ablations across method design

D.7.1 Using reference-induced norm vs Euclidean norm in the selector

We evaluate the effect of using the reference Fisher-induced norm and the Euclidean norm in the data selection objective in Eq. (7). We present results for the MedQA and Esconv datasets corresponding to Settings 2 and 3. We report metrics on the target and reference tasks with varying off-target drifts ρ (Figures 17 and 15), and visualize the selected gradients (Figures 18 and 16). Overall, we see that

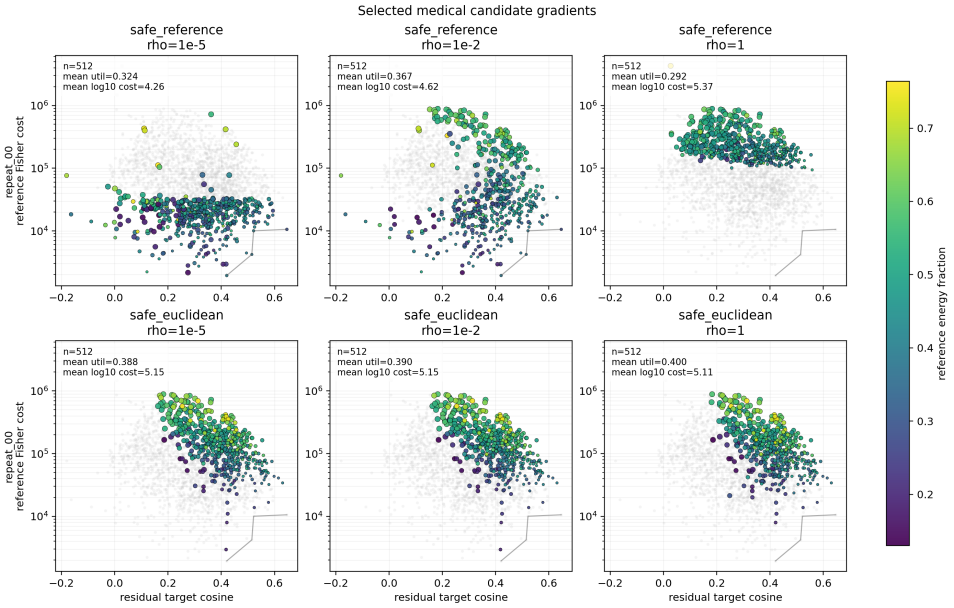


Figure 16: Selected points plotted on target gain versus realized off-target drift D_{KL} from SAFE-DRIFT using Euclidean vs reference Fisher-induced norm objective in MedQA experiment (Setting 2).

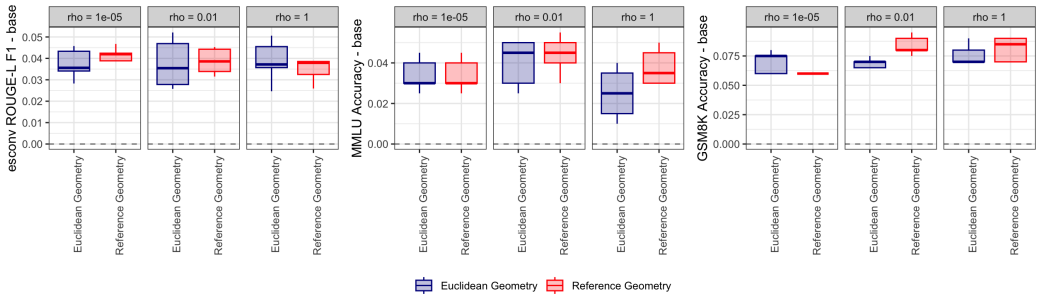


Figure 17: Target gain and off-target drift D_{task} of SAFE-DRIFT using Euclidean vs reference Fisher-induced norm objective in Esconv experiment (Setting 3). We plot the difference from the base model across drift budget ρ.

selecting data points in the Euclidean geometry yields worse performance in controlling the off-target drift: we consistently see that, in the Euclidean geometry, selected gradients have high projections along the reference geometry, irrespective of the value of ρ. By contrast, selecting points in the reference geometry yields a stronger sensitivity to higher off-target costs.

D.7.2 Effect of the optimizer preconditioner

We evaluate the effect of the optimizer-induced preconditioner $\mathbf{P}$. Our framework allows different choices of $\mathbf{P}$, including the identity preconditioner corresponding to SGD-style updates and an Adam-style diagonal preconditioner estimated from warm-up gradients. This ablation tests whether the observed behavior of SAFE-DRIFT depends critically on using the Adam preconditioner.

Figures 4 and 19-21 show that the best preconditioner varies across settings. SGD performs better in Setting 2 while Adam performs better in Setting 3. This is expected because the preconditioner changes the effective update induced by each selected example. Importantly, both variants retain the same qualitative target-reference trade-off, suggesting that SAFE-DRIFT is not tied to a single optimizer approximation. In practice, we recommend using the preconditioner that matches the downstream fine-tuning optimizer.

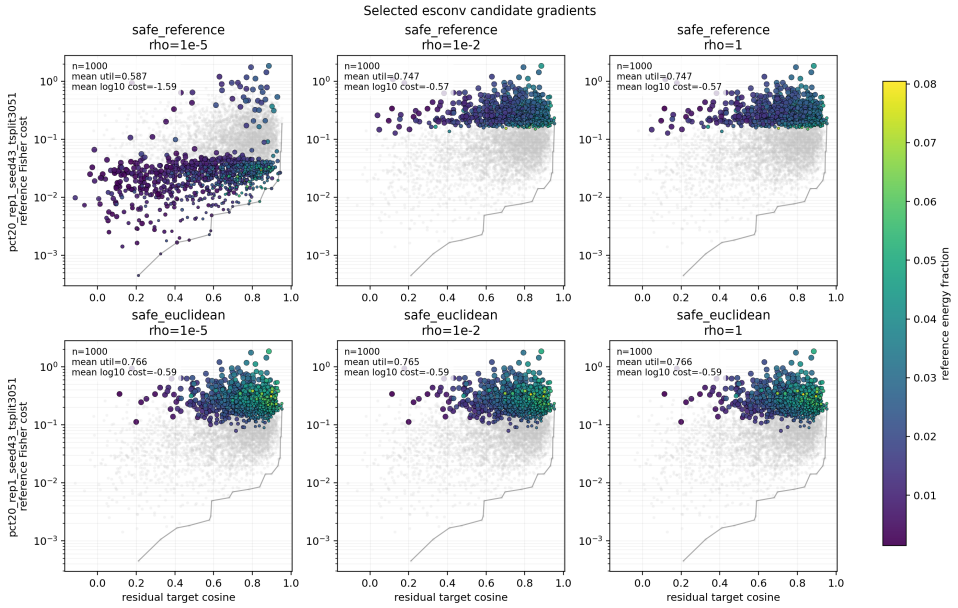


Figure 18: Selected points plotted on target gain versus realized off-target drift D_{KL} from SAFE-DRIFT using Euclidean vs reference Fisher-induced norm objective in Esconv experiment (Setting 3).

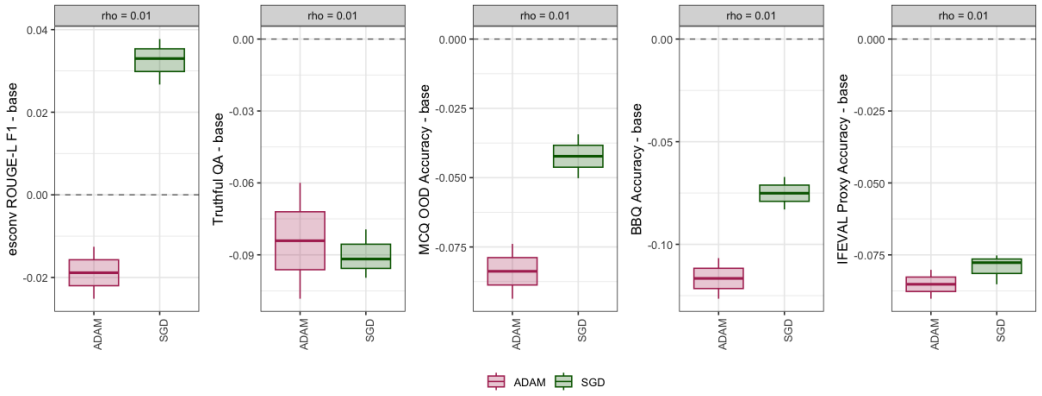


Figure 19: Comparison of using preconditioner P from Adam versus SGD optimizer in MedQA experiment (Setting 2)

D.7.3 Different choices of rank K_R and K_T

We ablate over three low-rank settings $(K_R, K_T) \in \{(128, 32),\ (256, 64),\ (512, 128)\}$ for SAFE-DRIFT on Setting 1, with results shown in Figure 22 for selection budget 15%. SAFE-DRIFT is robust to the choice of (K_R, K_T): all three variants achieve competitive HumanEval pass@1 and pass@10 rates, and all produce smaller changes in StereoSet score than most baselines. One thing to note is that the basis $(K_R, K_T) = (512, 128)$ achieves the smallest off-target drift. In this comparison, the largest basis gives the smallest mean StereoSet change; the smaller bases do not outperform every baseline on that metric.

Figure 23 further shows the Fisher cost of the displacement from initialization and the projected first-order target-gain proxy over training steps for all methods at the displayed 15% budget. SAFE-DRIFT variants have lower plotted reference cost at $(K_R, K_T) = (128, 32), (512, 128)$ than baselines at comparable or higher plotted gain proxy. These are diagnostics of the fixed local model, not guarantees on the fine-tuning trajectory. In particular, the quadratic cost of a total displacement

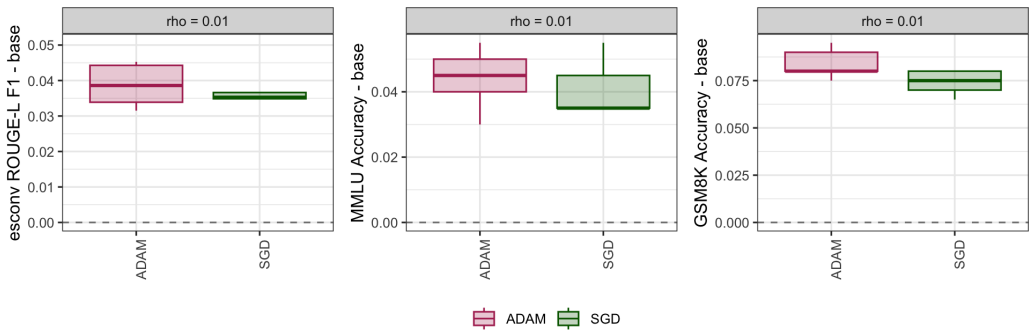


Figure 20: Comparison of using preconditioner P from Adam versus SGD optimizer in Esconv experiment (Setting 3)

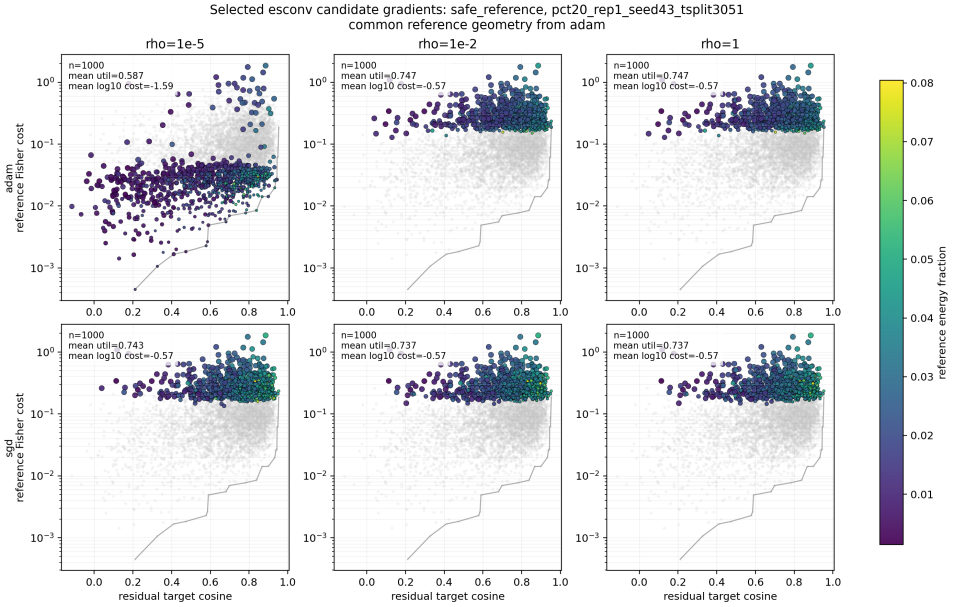


Figure 21: Selected points plotted on target gain versus realized off-target drift D_{KL} from SAFE-DRIFT using preconditioner P from Adam versus SGD optimizer in Esconv experiment (Setting 3)

contains cross terms between steps and need not increase with training time; two opposite steps with nonzero Fisher cost have zero final displacement but positive summed step costs.

Quality of the Low-Rank Approximation The reference Fisher and gradient coordinates are represented in a shared low-rank basis used for selection; this does not restrict the subsequent fine-tuning trajectory to that span. For an update of norm at most ϵ, the omitted empirical reference cost is at most $\epsilon^2 \widehat{\lambda}_{K_R+1}/2$ (Appendix C.1). A retained energy fraction alone cannot supply this bound without the spectral scale and update norm, and retained task/candidate energy need not preserve the mean target-gradient direction.

Table 8 reports cumulative squared singular-value energy, averaged over three seeds. Setting 1 retains 92.4% of reference energy at $K_R = 512$ and 74.8% of task/candidate energy at $K_T = 128$. In Setting 3, the diagnostic matrices use 128 reference and 512 task/candidate examples, explaining their rank limits. Flatter spectra generally require larger ranks for comparable matrix approximation; target alignment additionally depends on the projection error of the target gradient.

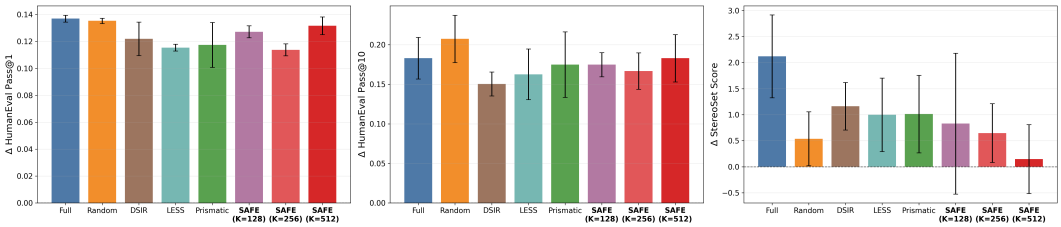


Figure 22: Ablation on the low-rank basis size (K_R, K_T) for SAFE-DRIFT on Setting 1. Left and center panels show change in HumanEval pass@1 and pass@10 relative to the base model; right panel shows change in StereoSet score relative to the base model.

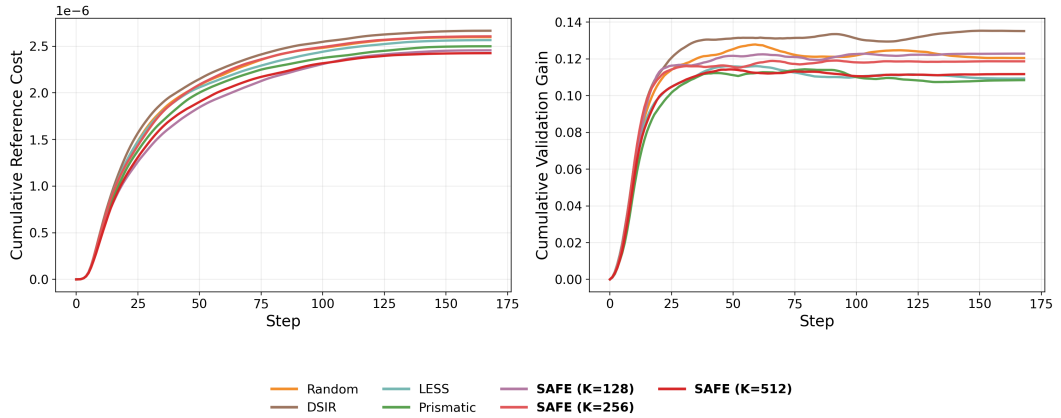


Figure 23: Reference Fisher cost $D_{\mathrm{F}} = \Delta\theta_t^\top \hat{\mathbf{F}}_R^{(K_R)} \Delta\theta_t$ and projected first-order target-gain proxy $-(\mathbf{U}_K^\top \Delta\theta_t)^\top \hat{b}_{\mathrm{tar}}^{(K)}$ over training steps, where $\Delta\theta_t = \theta_t - \theta_0$. These use displacement from initialization and fixed initial geometry; they are not a sum of per-step Fisher costs or an exact validation-risk decrease.

D.7.4 Different selection budgets

We provide the results of Setting 4 with two selection budgets: 5% and 15% of the full candidate pool in Figure 24 for IFEval (target task) and Figure 25 for MMLU (off-target task). Because the target pool is extremely small, with 42 IFEval-derived examples in Table 1, DSIR is particularly brittle in this setting. From Table 1, we know only 22% of candidates fall in Q_{clean} and 26% in Q_{ent}, so 48% have positive residual-task-block alignment d_T. This is not the fraction with positive total target alignment, which also includes the reference block. Figure 13 provides complementary block-energy diagnostics, without proving a necessary trade-off in benchmark performance.

Under this setting, SAFE-DRIFT concentrates on the highest-scoring target-beneficial examples at the 5% budget, while at 15%, SAFE-DRIFT selects some weakly target-aligned or reference-entangled examples, diluting the target update and preserving its behavior on the reference task. Comparing with LESS, SAFE-DRIFT gives a more favorable target–reference trade-off and attains more target gains. Random and Prismatic selects a broader subset from the candidate distribution, and they also select many entangled directions that result in MMLU preservation comparable to SAFE-DRIFT.

D.8 Computational Cost

SAFE-DRIFT has two computational bottlenecks. The first is per-example gradient construction over the candidate pool, which scales linearly in N_C and is equally required by LESS and Prismatic; the second is the low-rank estimation of the reference Fisher and target gradient, whose marginal cost is small by comparison. For large candidate pools the pool can be subsampled before selection, and for large models, LoRA keeps the per-example gradients tractable.

Setting	Block	1	8	32	64	128	256	512	1024
Setting 1	Reference	0.133	0.257	0.414	0.525	0.654	0.794	0.924	1.000
	Task/candidate	0.074	0.258	0.488	0.627	0.748	0.823	0.885	0.937
Setting 3	Reference	0.234	0.443	0.685	0.847	1.000	—	—	—
	Task/candidate	0.118	0.279	0.448	0.564	0.704	0.862	0.9999	—

Table 8: *Cumulative spectral energy* (squared singular-value energy) as a function of rank, mean over three seeds. These are energy fractions of the reference-score and task/candidate-gradient matrices. Bounding omitted reference cost additionally requires the spectral scale and update norm; preserving the mean target gradient requires its projection error to be small (Appendix C.1).

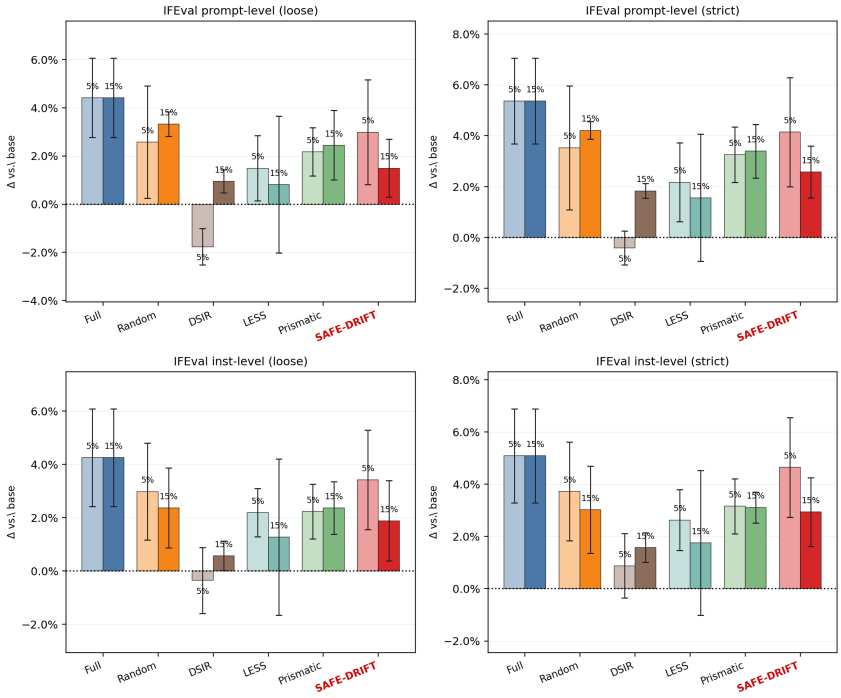


Figure 24: Relative change of IFEval performance compared to the base model. Bar heights are the means and error bars indicate one standard deviation across three seeds. The lighter bar corresponds to a 5% selection budget and the darker bar to a 15% budget.

Table 9 reports FLOP counts for Settings 1–3. LESS, Prismatic, and SAFE-DRIFT have equal reported selection counts within each base model, reflecting the shared computation for gradient construction over the candidate pool.

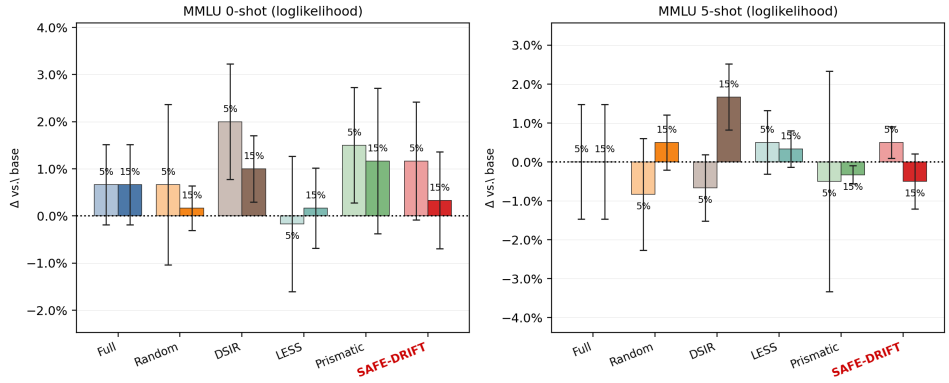


Figure 25: Relative change of MMLU log likelihood of correct answers on the held-out 200 questions compared to the base model: Panels show the 0-shot setting (left) and 5-shot setting (right). The lighter bar corresponds to a 5% selection budget and the darker bar to a 15% budget.

Setting	Model	Method	Select FLOPs	Train FLOPs	Select + Train	Select share	Runs
Setting 1	OLMo-2-7B	Random	0	7.24×10^{15}	7.24×10^{15}	0.0%	3
		DSIR	$< 10^{8}$	4.13×10^{15}	4.13×10^{15}	$< 0.1\%$	3
		LESS	5.54×10^{16}	6.05×10^{15}	6.15×10^{16}	90.2%	3
		Prismatic	5.54×10^{16}	7.74×10^{15}	6.31×10^{16}	87.7%	3
		SAFE-DRIFT	5.54×10^{16}	1.16×10^{16}	6.70×10^{16}	82.7%	3
	OLMo-2-13B	Random	0	1.39×10^{16}	1.39×10^{16}	0.0%	3
		DSIR	$< 10^{8}$	7.92×10^{15}	7.92×10^{15}	$< 0.1\%$	3
		LESS	1.06×10^{17}	1.30×10^{16}	1.19×10^{17}	89.1%	3
		Prismatic	1.06×10^{17}	1.48×10^{16}	1.21×10^{17}	87.7%	3
		SAFE-DRIFT	1.06×10^{17}	2.19×10^{16}	1.28×10^{17}	82.9%	3
Setting 2	OLMo-2-7B-Instruct	Random	0	3.056×10^{15}	3.06×10^{15}	0.0%	5
		DSIR	4.865×10^{7}	3.560×10^{15}	3.56×10^{15}	$< 0.1\%$	5
		LESS	1.219×10^{17}	3.568×10^{15}	1.25×10^{17}	97.2%	5
		Prismatic	1.219×10^{17}	3.035×10^{15}	1.25×10^{17}	97.6%	5
		SAFE-DRIFT	1.219×10^{17}	1.625×10^{15}	1.24×10^{17}	98.7%	5
Setting 3	OLMo-2-7B-Instruct	Random	0	4.420×10^{15}	4.42×10^{15}	0.0%	5
		DSIR	1.054×10^{8}	2.951×10^{15}	2.95×10^{15}	$< 0.1\%$	5
		LESS	4.988×10^{17}	5.405×10^{15}	5.04×10^{17}	98.9%	5
		Prismatic	4.988×10^{17}	4.371×10^{15}	5.03×10^{17}	99.1%	5
		SAFE-DRIFT	4.988×10^{17}	3.006×10^{15}	5.02×10^{17}	99.4%	5

Table 9: *Selection and training compute.* All runs use Adam with LoRA. "Select + Train" sums the reported selection and training FLOP estimates and excludes unitemized overhead. "Select share" is selection FLOPs divided by this subtotal, calculated before display rounding. These subtotals do not represent end-to-end compute.