SUADS: SAMPLING-UNCERTAINTY-AWARE DATA SELECTION FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradient-based data selection has shown promise for data-efficient fine-tuning of large language models (LLMs). However, we find that its gains diminish or disappear on challenging coding tasks, sometimes falling below random selection. Our analysis identifies finite-reference sampling uncertainty as a key source of this unreliability: estimating the target gradient direction from a limited reference set introduces noise that scales with the incoherence of the references. Under a stylized model, we prove that this uncertainty attenuates the expected gain over random selection by a multiplicative factor determined by reference reliability, independent of the selection fraction. This degradation occurs even when population influence scores exactly capture one-step training utility, revealing a bottleneck beyond instance-level score accuracy. Guided by this analysis, we propose **SUADS**, a clustering-based method that combines within-cluster gradient aggregation with stratified selection to exploit more reliable local reference directions. Across Qwen3.5-4B, Llama-3.1-8B, and Gemma-2-9B on MBPP, HumanEval, and LiveCodeBench, SUADS is the only selection method whose average exceeds random selection in every setting and has the highest average pass@1 of any method under matched selection budgets. On Gemma-2-9B it matches full-data fine-tuning on average with $5\,\%$ of the data.

1 INTRODUCTION

Supervised fine-tuning (SFT) plays a central role in large language model (LLM) post-training, improving instruction following (Chung et al., 2024) and providing a starting point for reinforcement learning to align models with human preferences (Ouyang et al., 2022) and improve reasoning (Guo et al., 2025). Despite the remarkable LLM generalization capability to learn from a small amount of data, it remains a central challenge to determine how to best utilize the fine-tuning data to save training cost. Motivated by such a practical demand, a rich body of data selection methods has been proposed (Zhou et al., 2023; Xia et al., 2024b; Dai et al., 2025; Wang et al., 2025c; Min et al., 2026). In various instruction-following and world knowledge tasks, they show promising results: fine-tuning on as few as 5% of the data can greatly outperform fine-tuning on randomly selected 5% data, and perform comparably to or even exceed unfiltered data tuning.

Data selection methods for LLMs typically select data that improve diversity and utility (Liu et al., 2024a; Lu et al., 2024). Diversity-oriented criteria seek to reduce redundancy and promote coverage of different topics, instruction types, and capabilities (Li et al., 2024). In this work, we focus on utility, namely the expected contribution of a training example to performance on a targeted task. To measure utility, data selection methods assign example-level selection scores and use the score to guide subset selection, either through ranking or in combination with additional balancing criteria (Dai et al., 2025). Existing methods use loss-derived scores (Li et al., 2024), LLM-based quality ratings (Chen et al., 2024), or gradient-based scores. Among these methods, gradient-based data selection has emerged as a prominent and empirically effective approach to targeted LLM fine-tuning (Wang et al., 2025c; Min et al., 2026; Nayak et al., 2026). Representative methods have demonstrated substantial data efficiency (Xia et al., 2024b; Dai et al., 2025).

We study gradient-based data selection methods. In a simplified model, these methods (1) collect a small number of validation data from the task domain and the unfiltered training data corpus; (2) use a model checkpoint to compute their gradients; and (3) select the top $k\%$ training data

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

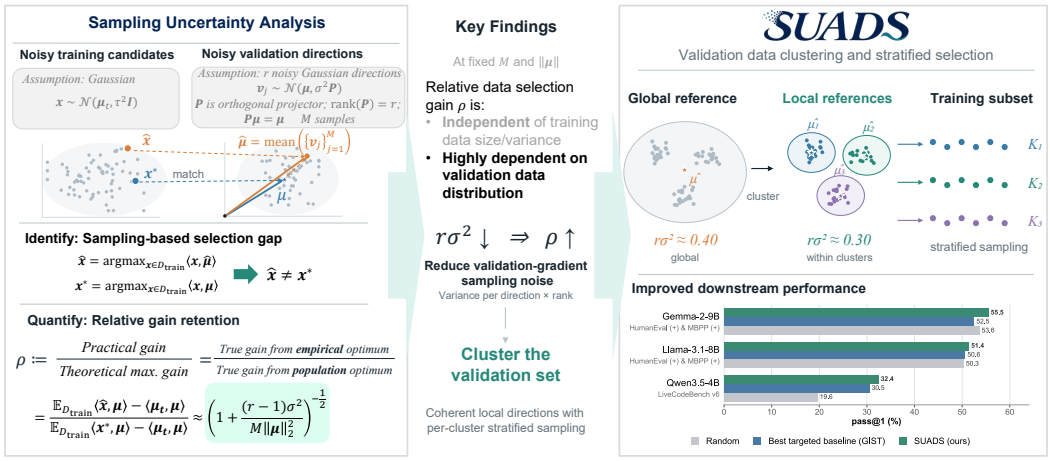

Figure 1: **Overview of our analysis and SUADS. Left:** Sampling uncertainty in validation-gradient estimates attenuates the expected gain of gradient-based selection over random selection. **Middle:** The relative selection gain ρ quantifies the fraction of maximum gain retained, suggesting that reducing validation-gradient noise $r\sigma^2$ improves gain retention. **Right:** Guided by this analysis, **SUADS** clusters validation gradients and significantly improves data selection effectiveness.

that are most aligned with the mean validation gradient measured by the inner product or cosine similarity. Existing works improve upon this simplified selection protocol by applying optimizer-aware preconditioning (Xia et al., 2024b), introducing diversity-aware balancing (Dai et al., 2025), modeling the temporal dependence of influence along training trajectories (Wang et al., 2025b), and projecting gradients onto task-specific subspaces (Min et al., 2026). However, these advances leave an elephant in the room unaddressed: **sampling uncertainty in the estimated validation gradient**.

On the theoretical side, we characterize sampling uncertainty in validation-gradient estimates and quantify its impact on data selection. Under an isotropic noise model, we prove that the extra loss reduction brought by data selection over random selection is attenuated by an approximately *multiplicative factor* caused by the sampling uncertainty. This attenuation persists even when population influence scores perfectly capture the underlying one-step training utility, revealing a statistical bottleneck beyond the accuracy of population utility scores. Interestingly, this relative gain retention, which we denote as ρ, is independent of the candidate training data pool size, its distribution, or data selection ratio, but depends heavily on validation data size and distribution. In particular, for a fixed validation-set size, reducing the gradient variance per direction or the dimension of the noise subspace mitigates this attenuation. Figure 1 provides a simplified illustration of the analysis where detailed results are in Section 3.

Motivated by our analysis, we propose to *cluster validation gradients* to identify coherent local directions and reduce within-cluster gradient variance. Our analysis specifically motivates clustering the *validation* set to address uncertainty in the estimated reference directions, whereas prior approaches cluster the *training* set (Wang et al., 2025c; Zhang et al., 2025). Building on this principle, we introduce **SUADS**, a **S**ampling-**U**ncertainty-**A**ware **D**ata **S**election method that combines validation-gradient clustering with outlier removal and coherence-based cluster reweighting.

We conduct a systematic empirical evaluation for SUADS and other state-of-the-art data selection methods on modern LLMs (Qwen3.5-4B, Llama-3.1-8B, and Gemma-2-9B) and challenging coding tasks (MBPP, HumanEval, and LiveCodeBench). We focus on coding tasks because they offer well-defined, execution-based evaluation metrics (e.g., pass@1) and remain challenging for models at the scales we study. In this challenging setting, many data selection methods offer little or no improvement over uniform random sampling, consistent with the attenuation of selection gains predicted by our sampling-uncertainty analysis. In contrast, SUADS has the highest average pass@1 in every setting, leading the best baseline by +1.9%, +0.8%, and +1.9% on Qwen3.5-4B, Llama-3.1-8B, and Gemma-2-9B, and on Gemma-2-9B matching full-data fine-tuning on average with 5 % of the data.

We summarize our key contributions in Figure 1 and as follows:

- We propose a theoretical analysis framework that theoretically characterizes sampling uncertainty in validation-gradient estimates and quantifies its impact on gradient-based data selection.

- We identify key factors that improve the data selection effectiveness from the analysis and propose SUADS, a new sampling-uncertainty-aware data selection method.
- We implement SUADS and conduct systematic empirical evaluation of five representative data selection methods on modern LLMs, revealing the effectiveness of SUADS.

Code and data are attached in the supplementary material and will be open source.

2 PRELIMINARIES

In this section, we introduce the gradient-based data selection (Xia et al., 2024b; Dai et al., 2025; Min et al., 2026) and then discuss other related work.

2.1 PRELIMINARIES AND RELATED WORK ON GRADIENT-BASED DATA SELECTION

We consider the single-task data selection setting. A *gradient-based data selection method* is given: (1) A base LLM model f_{θ_0}, where $\theta_0 \in \mathbb{R}^P$ denotes weights; (2) The training data corpus, consisting of N samples: $D_{\text{train}} = \{z_i\}_{i=1}^N$; (3) The in-domain validation data: $D_{\text{val}} = \{w_j\}_{j=1}^M$; and (4) The data selection budget, a fraction $k\%$ of the training data, that is $K = k\% \cdot N$ examples. The method is tasked to select a subset $S \subseteq D_{\text{train}}$ with $|S| = K$, such that the model tuned on S performs better on the task, typically measured on a held-out set of domain data using a task-specific metric. The training corpus is typically large, and in-domain validation data is scarce, often $100 \leq M \leq 500$ examples.

The training data and validation data samples can be of various forms. Here, we consider LLM supervised fine-tuning, so a data sample is a tuple $(\text{instruction}, \text{response})$. Given a model checkpoint f_θ where θ are tunable weights, fine-tuning on a data sample z_i incurs negative log-likelihood loss and its sample-level gradient: $\nabla_\theta \mathcal{L}(z_i) := \nabla_\theta(-\log \Pr_\theta[\text{response} \mid \text{instruction}])$.

Gradient-based data selection methods aim to select data that minimizes the empirical validation data loss $\hat{\mathcal{L}}_{\text{val}} := {}^1\!/_M \sum_j \mathcal{L}(w_j)$. The first-order approximation of training with one-step gradient descent on selected data S leads to empirical validation loss reduction

$$\theta' = \theta - \eta \sum_{z_i \in S} \nabla_\theta \mathcal{L}(z_i),\ \hat{\mathcal{L}}_{\text{val}}(\theta) - \hat{\mathcal{L}}_{\text{val}}(\theta') \approx \langle \nabla_\theta \hat{\mathcal{L}}_{\text{val}}, (\theta - \theta') \rangle = \langle \nabla_\theta \hat{\mathcal{L}}_{\text{val}}, \sum_{z_i \in S} \eta \nabla_\theta \mathcal{L}(z_i) \rangle, \quad (1)$$

where η is the learning rate. Denote $\hat{\boldsymbol{\mu}} := \nabla_\theta \hat{\mathcal{L}}_{\text{val}}$ to the empirical mean of validation gradients, and $\boldsymbol{x}_i := \eta \nabla_\theta \mathcal{L}(z_i)$ to the negative of parameter change incurred by training sample z_i. Hence, a greedy selection (without replacement) that maximizes loss reduction chooses $S = \text{top-}K\{\langle \hat{\boldsymbol{\mu}}, \boldsymbol{x}_i \rangle\}_{i=1}^N$.

The raw gradient-based data selection assigns each training data sample the score $s_i := \langle \hat{\boldsymbol{\mu}}, \boldsymbol{x}_i \rangle$ and selects data with the highest scores, reproducing this greedy selection procedure. Some methods directly use the formulation (Pruthi et al., 2020), while recent ones introduce further refinements (Killamsetty et al., 2021b; Wang et al., 2024; Xia et al., 2024b; Dai et al., 2025; Liu et al., 2024b; Min et al., 2026). For example, LESS (Xia et al., 2024b) adapts influence estimation to Adam updates, uses cosine similarity to mitigate gradient-norm bias associated with response length, and aggregates scores across model checkpoints. It further reduces computational costs through LoRA gradient features and random projections.

2.2 OTHER RELATED WORK

Gradient-based data selection is closely connected to data valuation and influence estimation, which quantify the contributions of individual training examples. Representative methods, such as influence functions (Koh & Liang, 2017; Grosse et al., 2023) and data Shapley (Ghorbani & Zou, 2019; Wang et al., 2025a), admit approximation forms that resemble the gradient-alignment formulation. These connections situate our formulation within a broader family of utility estimates, though our analysis focuses specifically on how uncertainty in the estimated validation reference propagates through data selection. Recently, Nayak et al. (2026) systematically studied targeted instruction selection and found gradient-based methods effective. They further provide generalization bounds for models trained on selected data under convexity assumptions. In contrast, our analysis provides a

characterization of selection-gain attenuation under a structured gradient model and, with an additional smoothness assumption, yields an explicit one-step population-risk bound.

Beyond gradient-based selection, other quality and diversity criteria exist for LLM data selection (Chen et al., 2024; Liu et al., 2024a; Li et al., 2024; Xie et al., 2023; Liu et al., 2024b), some proposing training-set clustering (Wang et al., 2025c; Zhang et al., 2025) while we do validation-set clustering. Data selection builds on the broader literature on data-efficient learning, particularly active learning (Cohn et al., 1994; Settles, 2009) and coreset construction (Chen, 2009; Feldman & Langberg, 2011). It also aligns with the data-centric AI perspective (Ré, 2021; Zha et al., 2025). We provide a detailed discussion of these connections and our comparison baselines in Appendix A.

3 Analysis of Gradient-based Data Selection

In this section, we present our analysis of sampling uncertainty in validation-gradient estimates for gradient-based data selection methods, which reveals key factors that impact data selection effectiveness and motivates our main method SUADS.

3.1 Main Results and Relative Gain Retention

We first illustrate our main results under an isotropic Gaussian noise model, then discuss implications and generalizations.

Problem setup. Following Section 2.1, we select a subset $S \subseteq D_{\text{train}}$, and write $\boldsymbol{x}_i$ for the negative parameter update that training on data sample z_i incurs at checkpoint $\theta \in \mathbb{R}^P$, which is $\eta \nabla_\theta \mathcal{L}(z_i)$ for gradient descent and the preconditioned update of Xia et al. (2024b) for Adam. Existing methods reduce the *empirical* validation loss $\hat{\mathcal{L}}_{\text{val}} = {}^1\!/_M \sum_j \mathcal{L}(w_j)$, whereas we model $w_j \sim \mathcal{D}_{\text{val}}$ and reduce the *population* loss $\mathcal{L}_{\text{val}} := \mathbb{E}_{w \sim \mathcal{D}_{\text{val}}} \mathcal{L}(w)$, which quantifies the test-time utility of the selected data. If all of S attend training at the current step, the reduction of population validation loss is, to first order,

$$\Delta\mathcal{L}_{\text{val}}(S;\theta) := \mathcal{L}_{\text{val}}(\theta) - \mathcal{L}_{\text{val}}\left(\theta - \sum_{z_i \in S} \boldsymbol{x}_i\right) \approx \sum_{z_i \in S} \langle \boldsymbol{x}_i, \nabla_\theta \mathcal{L}_{\text{val}} \rangle, \tag{2}$$

which assumes full-batch rather than mini-batch training and ignores in-batch interactions, for which a finer approximation is feasible (Wang et al., 2024). We analyze gradient-based selection in terms of *maximizing this reduction*. The key limitation is the sparsity of high-quality in-domain validation data: the criterion calls for alignment with $\nabla_\theta \mathcal{L}_{\text{val}}$, but in practice we can only align with the empirical mean gradient $\hat{\boldsymbol{\mu}} = \nabla_\theta \hat{\mathcal{L}}_{\text{val}} = {}^1\!/_M \sum_j \nabla_\theta \mathcal{L}(w_j)$. We thus define the relative gain retention ρ to measure how much of the optimal alignment a method retains.

Definition 1 (Relative Gain Retention ρ)**.** The training data D_{train} is i.i.d. sampled from a training corpus distribution $\mathcal{D}_{\text{train}}$. For a data selection algorithm $\mathcal{A}$ that selects a data subset $S \subseteq D_{\text{train}}$, on model f_θ, its relative gain retention ρ is its expected excess loss reduction compared to same-size random selection normalized by the maximal loss reduction:

$$\rho(\mathcal{A};\mathcal{D}_{\text{train}},\theta) := \mathop{\mathbb{E}}_{D_{\text{train}} \sim \mathcal{D}_{\text{train}}} \left(\overbrace{\Delta\mathcal{L}_{\text{val}}(S;\theta)}^{\mathcal{A}\text{'s loss reduction}} - \overbrace{\mathop{\mathbb{E}}_{\substack{S' \subseteq D_{\text{train}} \\ |S'|=|S|}} \Delta\mathcal{L}_{\text{val}}(S';\theta)}^{\text{Random selection loss reduction}} \right) \Big/ \mathop{\mathbb{E}}_{D_{\text{train}} \sim \mathcal{D}_{\text{train}}} \left(\overbrace{\max_{\substack{S' \subseteq D_{\text{train}} \\ |S'|=|S|}} \Delta\mathcal{L}_{\text{val}}(S';\theta)}^{\text{Theoretical max. loss reduction}} - \overbrace{\mathop{\mathbb{E}}_{\substack{S' \subseteq D_{\text{train}} \\ |S'|=|S|}} \Delta\mathcal{L}_{\text{val}}(S';\theta)}^{\text{Random selection loss reduction}} \right)$$

$$\approx \frac{\mathbb{E}_{D_{\text{train}} \sim \mathcal{D}_{\text{train}}} \sum_{z_i \in S} \langle \boldsymbol{x}_i, \nabla_\theta \mathcal{L}_{\text{val}} \rangle - |S| \langle \mathbb{E}_{z_i \sim D_{\text{train}}} \boldsymbol{x}_i, \nabla_\theta \mathcal{L}_{\text{val}} \rangle}{\mathbb{E}_{D_{\text{train}} \sim \mathcal{D}_{\text{train}}} \sum_{z_i \in \text{Top}|S|(D_{\text{train}})} \langle \boldsymbol{x}_i, \nabla_\theta \mathcal{L}_{\text{val}} \rangle - |S| \langle \mathbb{E}_{z_i \sim D_{\text{train}}} \boldsymbol{x}_i, \nabla_\theta \mathcal{L}_{\text{val}} \rangle}. \quad \text{(by first-order approximation)}$$

Remark. The relative gain retention ρ measures how well the selected subset S approaches the theoretical optimum under the same data budget compared to the random baseline in a scale-invariant manner: When it reaches the theoretical optimum, $\rho = 1$; when it does not outperform random downsampling, $\rho \leq 0$. It is the central metric in the following analysis.

Assumption 1 (Isotropic Gaussian Noise)**.** *We assume that the training data update $\boldsymbol{x}_i \in \mathbb{R}^P$ follows a Gaussian distribution $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_t, \tau^2 \boldsymbol{I}_P)$; and the validation data gradients $\nabla_\theta \mathcal{L}(w_j) \in \mathbb{R}^P$ follow a Gaussian distribution $\boldsymbol{v}_j \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{P})$ where $\boldsymbol{P}$ is a rank-r orthogonal projector that contains the $\boldsymbol{\mu}$ direction, i.e., $\boldsymbol{P}^\intercal = \boldsymbol{P}, \boldsymbol{P}^2 = \boldsymbol{P}, \boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{\mu}, \text{rank}(\boldsymbol{P}) = r$.*

Remark. We assume the training data *updates* and validation data *gradients* are isotropic Gaussian, not the raw data. Since the validation data is high quality, we confine its gradient noise to a rank-r subspace, whereas training-update noise spans all of $\mathbb{R}^P$. Appendix B.3 relaxes this assumption.

The following theorem provides an explicit characterization of the retention in this case.

Theorem 1 (Gain Retention Characterization (Informal))**.** *Under Assumption 1, we independently draw N training samples to form D_{train} and M validation samples to form D_{val} and compute $\hat{\boldsymbol{\mu}} = \frac{1}{M}\sum_{w_j \in D_{val}} \nabla_\theta \mathcal{L}(w_j)$. $\mathcal{A}_{\hat{\boldsymbol{\mu}}}(D_{train})$ selects K distinct training examples with largest scores $\langle \boldsymbol{x}_i, \hat{\boldsymbol{\mu}} \rangle$.*

Under the first-order approximation for ρ (see Definition 1), when validation gradient noise along the mean direction is small relative to mean $\boldsymbol{\mu}$ and sample number M, i.e., $\sigma^2/(M\|\boldsymbol{\mu}\|^2)$ is small, the relative gain retention of $\mathcal{A}$ satisfies

$$\mathop{\mathbb{E}}_{D_{val}\sim\mathcal{D}_{val}} \rho(\mathcal{A}; \mathcal{D}_{train}, \theta) \approx \mathop{\mathbb{E}}_{D_{val}\sim\mathcal{D}_{val}} \cos\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\mu} \rangle \approx \left(1 + \frac{(r-1)\sigma^2}{M\|\boldsymbol{\mu}\|^2}\right)^{-1/2}, \tag{3}$$

where the two approximations reflect loss linearization and finite-sample concentration, respectively. A formal version with error control and full proof can be found in Appendix B.1.

Proof Sketch. Under loss linearization and Assumption 1, notice that $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2/M \cdot \boldsymbol{P})$ and $\boldsymbol{x}_i = \boldsymbol{\mu}_t + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \tau^2 \boldsymbol{I}_P)$, the ratio can thus be simplified to $\rho \approx \left(\mathbb{E}\sum_{\epsilon_i \in \mathrm{TopK}\langle \epsilon_i, \hat{\boldsymbol{\mu}}\rangle} \langle \epsilon_i, \boldsymbol{\mu}\rangle\right) / \left(\mathbb{E}\sum_{\epsilon_i \in \mathrm{TopK}\langle \epsilon_i, \boldsymbol{\mu}\rangle} \langle \epsilon_i, \boldsymbol{\mu}\rangle\right)$, where $\mathrm{u} := \langle \epsilon_i, \boldsymbol{\mu}\rangle \sim \mathcal{N}(0, \tau^2\|\boldsymbol{\mu}\|^2)$, $\mathrm{v} := \langle \epsilon_i, \hat{\boldsymbol{\mu}}\rangle \sim \mathcal{N}(0, \tau^2\|\hat{\boldsymbol{\mu}}\|^2)$.

We then observe that random variables u and v follow a joint Gaussian distribution conditioned on $\hat{\boldsymbol{\mu}}$, and the correlation $\mathrm{Corr}(\mathrm{u}, \mathrm{v}) = \mathbb{E}[\mathrm{uv}]/\sqrt{\mathbb{E}\,\mathrm{u}^2\,\mathbb{E}\,\mathrm{v}^2} = (\boldsymbol{\mu}^\intercal \mathbb{E}[\epsilon_i \epsilon_i^\intercal]\hat{\boldsymbol{\mu}})/(\tau^2\|\boldsymbol{\mu}\|\|\hat{\boldsymbol{\mu}}\|) = \cos\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle$. So $\mathrm{u} = \cos\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle \|\boldsymbol{\mu}\|/\|\hat{\boldsymbol{\mu}}\|\mathrm{v} + \mathcal{N}(0, \tau^2\|\boldsymbol{\mu}\|^2 \sin^2\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle)$, and $\mathbb{E}[\mathrm{u} \mid \mathrm{v}] = \cos\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle \|\boldsymbol{\mu}\|/\|\hat{\boldsymbol{\mu}}\|\mathrm{v}$. As a result, the numerator $\left(\mathbb{E}\sum_{\epsilon_i \in \mathrm{TopK}\langle \epsilon_i, \hat{\boldsymbol{\mu}}\rangle} \langle \epsilon_i, \boldsymbol{\mu}\rangle\right) = \cos\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle \frac{\|\boldsymbol{\mu}\|}{\|\hat{\boldsymbol{\mu}}\|} \left(\mathbb{E}\sum_{\epsilon_i \in \mathrm{TopK}(\mathrm{v}_i)} \mathrm{v}_i\right)$ which is exactly the denominator multiplied by $\cos\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle$ after scaling. Hence, $\rho \approx \cos\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle$.

Now we compute $\mathbb{E}\cos\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle = \frac{\langle \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}\rangle}{\|\boldsymbol{\mu}\|\|\hat{\boldsymbol{\mu}}\|} \approx \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|\sqrt{\|\boldsymbol{\mu}\|^2 + (r-1)\sigma^2/M}} = \left(1 + \frac{(r-1)\sigma^2}{M\|\boldsymbol{\mu}\|^2}\right)^{-1/2}$. □

Implications. Theorem 1 is our central theoretical contribution in this work. It provides a closed-form characterization (rather than a conservative bound) of the gap between gradient alignment from finite validation gradient samples $\nabla_\theta \hat{\mathcal{L}}_{\text{val}}$ and true validation gradients $\nabla_\theta \mathcal{L}_{\text{val}}$. Note that the gain retention ρ is a multiplicative factor that linearly interpolates between random selection and theoretical maximal reduction: (actual loss reduction)=$(1-\rho)$·(random baseline's loss reduction) + ρ(theoretical maximal reduction). Hence, the existence of such retention ρ indicates a multiplicative attenuation to the maximal data selection performance even when training samples perfectly align with the validation population distribution, revealing a statistical bottleneck.

An investigation of Equation (3) provides principled guidance for improving gradient-based data selection effectiveness. On one hand, ρ is agnostic to the statistics (τ^2), size (N) of the training data, and selection budget K. Though they may affect reference absolute loss reductions, they do not affect the isolated sampling-induced multiplicative factor, so improving ρ may benefit data selection in general. On the other hand, what is critical in ρ are *statistics of the validation data distribution*: ρ improves with the number of validation samples M and degrades with the noise dimension r and magnitude σ^2. When in-domain validation data is limited, we should put effort into suppressing r and σ^2 (Appendix B.6). This also explains the effectiveness of LoRA training (Xia et al., 2024b) and subspace extraction (Min et al., 2026), both of which reduce r, and it implies a proportional scaling ratio $(r-1)/M$ between tunable parameter size and validation data size, discussed in Appendix B.1. We further generalize the theoretical results to the cosine similarity selection metric in Appendix B.2 and anisotropic Gaussian distribution cases in Appendix B.3.

3.2 Clustering

In Section 3.1, we reveal how sampling uncertainty induces a multiplicative retention factor ρ, where existing efforts reduce r to improve ρ for better data selection effectiveness. Besides r, the noise

magnitude σ^2 also plays a critical role. Is there any technique to reduce σ^2? Inspired by the classical variance-reduction techniques for Monte Carlo estimation, stratified sampling and importance sampling (Owen, 2013), we now study when a clustering yields reliable local references.

Rather than using the clustered validation gradients for importance sampling, we propose a protocol that applies stratified selection for each validation gradient cluster to select its corresponding best-matched training data. Formally, given a set of validation loss gradients $\{\nabla_\theta \mathcal{L}(w_j)\}_{j=1}^M$, we partition them into c non-empty clusters $V_1, \ldots, V_c$, and compute the empirical centroid of each cluster $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_c$. Then, for budget K, from each cluster V_ℓ we proportionally select $K_\ell := K|V_\ell|/M$ training samples from the pool that have the largest alignment scores with the centroid, i.e., $\langle \boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_\ell \rangle$, to form the selected subset S. We denote this cluster-based selection protocol by $\mathcal{A}_c$.

Theorem 2 (Gain Retention Characterization for Clustering (Informal))**.** *Under Assumption 1, we assume a cluster-based Gaussian mixture model from which validation gradients are drawn:* $\nabla_\theta \mathcal{L}(w_j) \sim \sum_{\ell=1}^c \frac{|V_\ell|}{M} \mathcal{N}(\boldsymbol{\mu}_\ell, \sigma_\ell^2 \boldsymbol{P}_\ell)$*, where* $\boldsymbol{P}_\ell$ *is a rank-*r *orthogonal projector satisfying* $\boldsymbol{P}_\ell \boldsymbol{\mu}_\ell = \boldsymbol{\mu}_\ell$*. The selection protocol selects* $K_\ell := K|V_\ell|/M$ *training samples from the pool for each cluster* ℓ*, counting coincident selections with multiplicity, which is exact for disjoint top sets, the regime of well-separated clusters and small* K/N *(Appendix B.4).*

Define $\gamma_{N,q}$ *as the expected sum of largest* q *among* N *independent standard Gaussian variables. When* $\sigma_\ell^2/(|V_\ell|\|\boldsymbol{\mu}_\ell\|^2)$ *is small for every component, the relative gain retention of* $\mathcal{A}_c$ *satisfies*

$$\mathop{\mathbb{E}}_{D_{val} \sim \mathcal{D}_{val}} \rho(\mathcal{A}_c; \mathcal{D}_{train}, \theta) \approx \mathop{\mathbb{E}}_{D_{val} \sim \mathcal{D}_{val}} \sum_{\ell=1}^c \frac{\gamma_{N,K_\ell}}{\gamma_{N,K}} \cos\langle \hat{\boldsymbol{\mu}}_\ell, \boldsymbol{\mu} \rangle \approx \sum_{\ell=1}^c \frac{\gamma_{N,K_\ell}}{\gamma_{N,K}} \frac{\langle \boldsymbol{\mu}_\ell, \boldsymbol{\mu} \rangle}{\|\boldsymbol{\mu}\| \sqrt{\|\boldsymbol{\mu}_\ell\|^2 + \frac{(r-1)\sigma_\ell^2}{|V_\ell|}}}, \quad (4)$$

where the two approximations reflect loss linearization and finite-sample concentration, respectively. A formal version with error control and full proof can be found in Appendix B.4.

Proof Sketch. The proof follows Theorem 1, with the denominator unchanged and the numerator stratified across clusters. We apply the same technique to each cluster's numerator term. □

Remark. The additional terms in Equation (4) hinder intuitive understanding or comparison with the non-cluster ρ in Equation (3). Hence, we simplify by observing that $\boldsymbol{\mu}$ is the weighted average of each $\boldsymbol{\mu}_\ell$ so approximately $\|\boldsymbol{\mu}\| \approx \|\boldsymbol{\mu}_\ell\|$, cluster sizes and variances are equal, and the per-cluster top sets are disjoint, which yields

$$\rho \approx \sum_{\ell=1}^c \frac{\gamma_{N,K_\ell}}{\gamma_{N,K}} \left(1 + \frac{(r-1)\sigma_\ell^2}{M/c\|\boldsymbol{\mu}\|^2}\right)^{-1/2} \overset{K/N \text{ small}}{\approx} \left(1 + \frac{\log c}{\log(N/K)}\right)^{1/2} \left(1 + \frac{(r-1)\sigma_\ell^2}{M/c\|\boldsymbol{\mu}\|^2}\right)^{-1/2}. \quad (5)$$

To qualitatively study the impact of clustering, we can divide Equation (5) by Equation (3) to investigate when clustering outperforms non-clustering in gradient-based data selection:

$$\frac{\mathbb{E}\,\rho(\mathcal{A}_c)}{\mathbb{E}\,\rho(\mathcal{A})} \approx \sqrt{\left(1 + \frac{\log c}{\log(N/K)}\right) \frac{M \mathrm{Sim_g} \mathrm{Sim_c} + \mathrm{Sim_c}}{M \mathrm{Sim_g} \mathrm{Sim_c} + c\,\mathrm{Sim_g}}}, \quad (6)$$

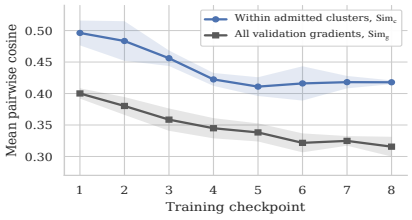

Figure 2: Within-cluster coherence $\mathrm{Sim_c}$ and global coherence $\mathrm{Sim_g}$ of the validation gradients of Gemma-2-9B on Magicoder dataset (Wei et al., 2024) across eight training checkpoints, mean and standard deviation over three runs. See main text for definitions of $\mathrm{Sim_c}$ and $\mathrm{Sim_g}$.

where c is the number of clusters, N is training corpus size, K is select budget, $\mathrm{Sim_g}$ is expected pairwise cosine similarity without clustering, and $\mathrm{Sim_c}$ is expected intra-cluster pairwise cosine similarity: $\mathrm{Sim_g} := \mathbb{E}_{\boldsymbol{v}_a, \boldsymbol{v}_b \sim \{\nabla_\theta \mathcal{L}(w_j)\}_{j=1}^M} \cos\langle \boldsymbol{v}_a, \boldsymbol{v}_b \rangle$, $\mathrm{Sim_c} := \mathbb{E}_{V_\ell} \mathbb{E}_{\boldsymbol{v}_a, \boldsymbol{v}_b \sim V_\ell} \cos\langle \boldsymbol{v}_a, \boldsymbol{v}_b \rangle$.

Equation (6) (derivation in Appendix B.5), shows that clustering helps only when the data forms *coherent, compact clusters* and the cluster count is *small*. Specifically, the coherent and compact clusters can improve $\mathrm{Sim_c}$ over $\mathrm{Sim_g}$, which holds on real validation gradients. Indeed, we observed that $\mathrm{Sim_c}$ exceeds $\mathrm{Sim_g}$ at every training checkpoint, with their ratio growing from 1.24 to 1.33 as training proceeds (Figure 2). Further, we observe that the number of clusters c needs to be larger than 1 but brings diminishing returns and will hurt if too large (empirically verified in Table 2), as it affects the ratio in $\sqrt{(1 + \Theta(\log c))/(1 + \Theta(c))}$ rate.

4 SUADS

Section 3 analyzed the bottleneck of gradient-based selection on the validation side: When validation gradient noise grows relative to its mean, by Theorem 1, the benefit from gradient-based data selection vanishes. Theorem 2 and Equation (6) show that per-cluster references $\mathcal{A}_c$ are reliable when the clusters are coherent, $\mathrm{Sim_c}$ well above $\mathrm{Sim_g}$, and serve components the global mean under-serves (Appendix B.6).

Applying $\mathcal{A}_c$ of Section 3.2 in practice raises two issues: the clusters must be found from the gradients rather than assumed, and, since k-means partitions any input, only coherent and compact clusters should be kept. Our method SUADS bridges the practical gap and follows four steps: gradient collection, cluster discovery and admission, cluster-conditional scoring, and stratified selection. Compared to $\mathcal{A}_c$, SUADS adds the coherence test and assigns each training example to the single cluster that scores it highest before taking the per-cluster top-K_ℓ. Single assignment enforces the disjoint top sets the analysis assumes, so the selection has exactly K distinct examples and splitting a coherent cluster cannot double-count its best ones.

Gradient collection. Following LESS (Xia et al., 2024b), we collect per-example gradients with LoRA (Hu et al., 2022) on a warmup model trained on a small random subset of the pool, reweight them by the Adam pre-conditioner, and compress them by random projection. Training updates and validation gradients are ℓ_2-normalized. $\{\boldsymbol{v}_j\}_{j=1}^{M}$ denotes the validation gradients.

Cluster discovery and admission. We partition the validation gradients into different candidate clusters by spherical k-means and discard singletons, which have no pairs to measure. Each remaining cluster V_ℓ is summarized by its *size-free coherence* κ_ℓ, the mean pairwise cosine among its members, computed from the norm of its centroid (details in Appendix C.2). We keep only the clusters whose coherence κ_ℓ is well above the global coherence κ_{pool}. Note that the threshold cannot simply be κ_{pool}, because k-means optimizes within-cluster coherence and yields clusters above the average even from structureless data. We therefore set it to $\kappa^\star$, a high quantile of the coherences the same clustering reaches on structureless pools matched to the validation set. In Section 5, SUADS uses at most four clusters and admits a cluster when its size-free coherence exceeds the 0.95 quantile of 25 synthetic structureless pools, with no per-task tuning. If no cluster passes, selection falls back to a single global top-K, where K is the selection budget.

Scoring and stratified selection. Each admitted cluster V_ℓ with centroid $\hat{\boldsymbol{\mu}}_\ell$ scores a training example z_i (whose incurred update is denoted by $\boldsymbol{x}_i$) by its weighted alignment score, and the training example score $s(z_i)$ is the maximum one over all clusters:

$$s(z_i) = \max_{V_\ell} w_\ell \langle \boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_\ell / \|\hat{\boldsymbol{\mu}}_\ell\| \rangle \quad \text{where} \quad w_\ell = \left(\tfrac{1}{M} + \left(1 - \tfrac{1}{M}\right)\kappa_\ell \right)^{1/2}. \tag{7}$$

The weight w_ℓ puts clusters of different sizes on a common scale (Appendix C.3). Then each example is assigned to the cluster that attains maximum score. The selection budget is divided in proportion to cluster size over admitted clusters, $K_\ell = K|V_\ell| / \sum_{\ell'} |V_{\ell'}|$, and each cluster takes the K_ℓ highest-scoring examples among those assigned to it, with quotas rounded by largest remainder. A coherent cluster covering a small fraction of the validation set thus keeps that fraction of the budget. More details in Appendix C, including the pseudocode (Algorithm 1).

5 Experimental Evaluation

Models, tasks, and data. We focus on the challenging task of code instruction tuning. On Llama-3.1-8B and Gemma-2-9B, the training pool contains $N = 35{,}000$ Magicoder samples (Wei et al., 2024), and the $M = 538$ validation samples are composed of 374 MBPP training split samples (Austin et al., 2021) and 164 EvoEval samples (Xia et al., 2024a). They are evaluated on HumanEval, HumanEval+, MBPP, and MBPP+ (Chen et al., 2021; Liu et al., 2023) under greedy decoding over three training runs. On Qwen3.5-4B, the training pool contains $N = 4{,}573$ DS_RSTAR traces (details in Appendix D.1), and the validation samples are $M = 144$ LiveCodeBench v5 problems (Jain et al., 2025). The model is evaluated on LiveCodeBench v6 problems released between Feb 1st and May 1st, 2025. Following the LiveCodeBench protocol (Jain et al., 2025), we report pass@1 averaged over independently sampled generations at temperature 1.0, training once per configuration. Validation sets, decoding, and test splits are detailed in Appendix D.

Table 1: Main results at a 5 % selection budget, except the last column. Pass@1 (%) with the standard deviation in parentheses. Magicoder: HumanEval (HE), HumanEval+ (HE+), MBPP, and MBPP+. DS_RSTAR: overall pass@1 on LiveCodeBench v6 at 5 % and 10 %, with the per-difficulty breakdown in Appendix E.2. Bold marks the best selection method. Magicoder stds are over three training runs; DS_RSTAR stds are over sampled generations of one training run.

	Gemma-2-9B (Magicoder)					Llama-3.1-8B (Magicoder)					Qwen3.5-4B (DS_RSTAR)	
Method	HE	HE+	MBPP	MBPP+	Avg.	HE	HE+	MBPP	MBPP+	Avg.	LCB v6 (5 %)	LCB v6 (10 %)
RANDOM	48.6 (1.8)	41.7 (0.9)	67.1 (1.4)	57.2 (1.4)	53.6 (1.0)	44.8 (2.8)	38.9 (1.8)	**64.2** (1.6)	53.2 (1.6)	50.3 (1.7)	19.6 (2.9)	32.1 (1.5)
LESS	43.9 (2.1)	36.4 (1.5)	65.2 (0.1)	55.1 (0.8)	50.2 (1.0)	46.3 (1.6)	39.0 (1.0)	61.2 (0.4)	52.9 (0.0)	49.9 (0.7)	28.5 (1.6)	38.2 (3.2)
BIDS	45.7 (1.6)	38.2 (1.2)	63.7 (1.7)	53.9 (1.8)	50.4 (0.6)	44.1 (1.4)	38.4 (1.2)	60.8 (1.2)	51.4 (1.5)	48.7 (0.3)	30.2 (2.7)	40.1 (2.7)
GIST	45.9 (1.3)	37.8 (1.6)	**67.8** (1.5)	**58.5** (2.2)	52.5 (0.6)	45.7 (1.3)	40.0 (1.8)	63.3 (1.5)	53.2 (1.3)	50.6 (1.4)	30.5 (2.2)	37.8 (7.0)
CLUSTERUCB	43.7 (3.4)	36.8 (3.1)	65.3 (0.9)	55.8 (0.7)	50.4 (1.4)	42.1 (0.0)	37.2 (0.6)	62.1 (0.4)	52.3 (0.2)	48.4 (0.2)	29.4 (0.5)	33.1 (1.9)
SUADS (ours)	**54.1** (1.0)	**45.3** (0.7)	66.7 (0.6)	55.8 (0.5)	**55.5** (0.5)	**47.6** (1.0)	**41.3** (1.3)	63.5 (1.0)	**53.4** (1.1)	**51.4** (0.2)	**32.4** (1.8)	**41.5** (1.3)
Base model	40.2	32.9	63.8	52.6	47.4	37.8	31.1	61.9	52.1	45.7	6.0 (0.5)	6.0 (0.5)
Full training	49.4 (1.7)	43.9 (0.8)	69.2 (0.2)	58.2 (0.0)	55.2 (0.2)	50.9 (1.3)	44.8 (0.4)	63.8 (1.1)	52.8 (0.9)	53.1 (0.9)	48.1 (0.8)	48.1 (0.8)

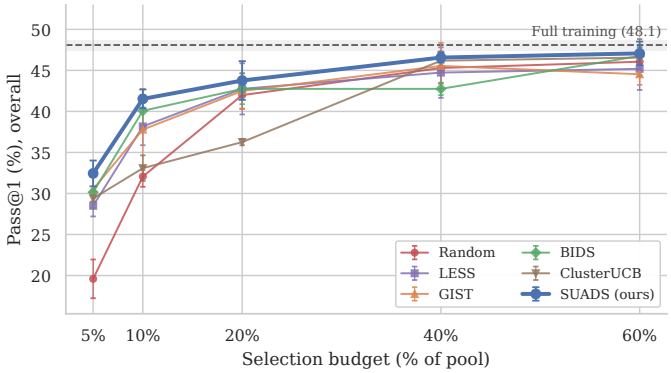

Figure 3: Mean pass@1 on LiveCodeBench v6 across different selection budgets on Qwen3.5-4B (DS_RSTAR), mean and standard deviation over independent generation runs per point. The dashed line and shaded band mark pass@1 mean and std when training on the full pool.

Fine-tuning and data selection budgets. We use the same LoRA, optimizer, training, and gradient collection protocol as LESS (Xia et al., 2024b). LESS reads gradients at all four warmup checkpoints, SUADS only the last, a quarter of the gradient computation (Appendix D.3). However, instead of picking maximum alignment score among manually divided validation data partitions in LESS (Xia et al., 2024b), SUADS uses the clustering-based algorithm introduced in Section 4. The main data selection budget is 5 % of the pool, and on DS_RSTAR we sweep 5 to 60 % (Figure 3).

Baselines and configuration. We compare against random data selection (RANDOM) and four targeted data selection methods, LESS (Xia et al., 2024b), BIDS (Dai et al., 2025), GIST (Min et al., 2026), and CLUSTERUCB (Wang et al., 2025c). Baselines use implementation and settings from their corresponding papers. More details described in Appendices A.2 and D.

5.1 MAIN RESULTS

Table 1 reports both settings, Magicoder tuning on Gemma-2-9B and Llama-3.1-8B at a 5 % budget and DS_RSTAR tuning on Qwen3.5-4B at 5 % and 10 %, and Figure 3 extends DS_RSTAR from 5 % to 60 % of the pool. SUADS has the **highest average** of any selection method in every configuration of the two tables and the **highest mean** at every budget of the sweep. It leads the strongest baseline by 1.9 points on Gemma-2-9B and by 0.8 on Llama-3.1-8B. On DS_RSTAR it leads GIST by 1.9 points at 5 % and BIDS by 1.4 at 10 %. Against LESS, which *shares its training configuration* but differs in score computation, the margins are 5.3 and 1.5 points on the two Magicoder tuning models and 3.9 and 3.3 on DS_RSTAR tuning. On Gemma-2-9B, SUADS *matches training on the full pool*, on the average, 55.5 against 55.2, and exceeds it on HumanEval and HumanEval+, by 4.7 and 1.4 points, the only method to do so on any of the three. On DS_RSTAR it approaches full training as the budget grows, coming within 1.0 point at 60 %, 47.1 against 48.1. All targeted methods select traces shorter than the pool median, and SUADS is the least length-biased (Figure 6), so its lead is not a length effect. Its results are also among the *most stable*, with deviations on average of 0.5 points on Gemma-2-9B and 0.2 on Llama-3.1-8B.

Random is a strong baseline. On Magicoder, RANDOM averages 53.6 on Gemma-2-9B, ahead of every targeted baseline. Theorem 1 anticipates part of this for per-example references (Appendix B.6): when validation gradients are few and noisy relative to their mean, the gain of top-K selection over an unbiased sample shrinks toward zero, an intrinsic bottleneck of such selection. Why several baselines fall below it, though, lies outside the first-order analysis. On DS_RSTAR every targeted baseline clears RANDOM by 8.9 to 10.9 points at 5 %, but the gap closes as the budget grows. SUADS is the only method whose average is above RANDOM in every configuration.

Headroom. Benchmarks differ in how much tuning can move them, and selection has the most to gain where that room is largest. On Llama-3.1-8B, training on the full pool improves the base model by 13.1 and 13.7 points on HumanEval and HumanEval+ but by only 1.9 and 0.7 on MBPP and MBPP+, so MBPP appears close to saturated there. On Gemma-2-9B the MBPP room is larger, 5.4 and 5.6 points, but still only half of HumanEval, 9.2 and 11.0. Differences between methods follow this headroom: On Gemma-2-9B the six methods span 10.4 on HumanEval but only 4.1 on MBPP.

5.2 ABLATION STUDIES

Table 2 varies the number of clusters c on Magicoder with Gemma-2-9B, with RANDOM and LESS as references. We observe that: (1) Clustering is generally better than non-clustering. For example, $c = 1$ (non-clustering) falls behind $c = 2$ by 4.0 points on average and by 6.5 points on HumanEval. Moreover, every clustering setting improves on a single cluster by 1.2 to 4.0 points on average, and on every targeted baseline in Table 1, aligning with our analysis in Section 3 where coherent clusters improve data selection. CLUSTERUCB, which clusters the training pool instead, reaches only 50.4 in Table 1, compared to SUADS 55.5. (2) The impact of the number of clusters roughly follows our analysis in Section 3.2, where too many clusters hurt the performance, forming an approximately unimodal pattern. However, the unimodality is not strict as shown in Table 2, $c = 3$ dips to 52.7, below RANDOM, so the best cluster number is data-dependent and the method is moderately sensitive to it. Nonetheless, every clustering setting ($c > 1$) improves on non-clustering ($c = 1$).

Table 2: Impact of the number of admitted clusters c for SUADS on Magicoder with Gemma-2-9B at 5 % budget. c is admitted from $c_{\max} = 1, 4, 8, 12, 16$; $c = 1$ is global top-K. Pass@1 (%) with the standard deviation over three training runs. Bold marks the best number of clusters in each column. RANDOM and LESS are repeated from Table 1 for reference.

# clusters c	HE	HE+	MBPP	MBPP+	Avg.
1	47.6 (1.6)	41.5 (1.6)	63.8 (0.6)	53.4 (1.6)	51.5 (1.0)
2	**54.1** (1.0)	**45.3** (0.7)	66.7 (0.6)	55.8 (0.5)	**55.5** (0.5)
3	50.4 (2.1)	41.7 (1.9)	64.6 (0.7)	54.1 (1.7)	52.7 (0.5)
6	51.2 (3.1)	44.3 (1.2)	**67.7** (0.3)	**57.3** (0.8)	55.1 (1.1)
7	49.8 (1.7)	42.5 (1.2)	67.3 (0.8)	56.1 (0.5)	53.9 (1.0)
RANDOM	48.6 (1.8)	41.7 (0.9)	67.1 (1.4)	57.2 (1.4)	53.6 (1.0)
LESS	43.9 (2.1)	36.4 (1.5)	65.2 (0.1)	55.1 (0.8)	50.2 (1.0)

Training dynamics. Figure 4 tracks SUADS and LESS on Gemma-2-9B across eight fine-tuning checkpoints, with dashed lines for random selection and full training, whose per-checkpoint curves were not recorded. After the first two checkpoints, SUADS stays between 54.4 and 56.2, above the final score of every baseline and around that of full training, 55.2. It leads LESS at every checkpoint by a large margin. The reported result is therefore neither a best-checkpoint pick nor dependent on where training stops.

In Appendix E.3, we further conducted content analysis for clustering. We observe, for example, that the data-driven clustering criteria cover a diverse set of observable data features.

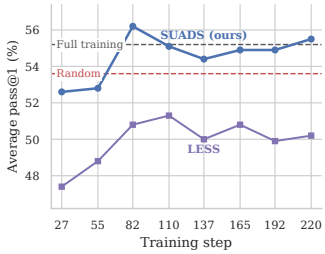

Figure 4: Training Dynamics on Gemma-2-9B. y-axis: test set average pass@1.

6 CONCLUSION

In this work, we propose an analysis framework that identifies a key factor affecting the effectiveness of gradient-based data selection methods for LLM fine-tuning. This factor is the sampling uncertainty in validation-gradient estimates. Through exact characterization of the gain retention ratio, we quantify the impact of validation-gradient noise, and propose a validation-set clustering-based method, SUADS, to improve the selection effectiveness. Through systematic evaluation on modern LLMs and challenging coding tasks, we demonstrate that SUADS has the highest average pass@1 of any selection method in every setting under matched selection budgets.

AI USE STATEMENT

We used generative AI tools to assist with checking and polishing mathematical derivations, identifying relevant literature, and editing the manuscript, notation, and figure designs. However, the key methodology and main components of theoretical analysis, method implementation, and experimental evaluation are developed and executed manually by the authors. We also used language models to generate candidate solutions for constructing validation data, as described in Section 5 and Appendix D.1. The authors take full responsibility for the correctness, originality, and integrity of all proofs, experimental results, and other content in this work.

ETHICS STATEMENT

This work studied data selection towards more efficient fine-tuning of large language models, with an empirical focus on code generation. Selection guided by a validation set evaluated in this work does not raise direct ethical concerns. However, the LLM data selection methods have direct applications in LLM safety and alignment (He et al., 2024; Li et al., 2025). Hence, the revealed statistical bottleneck and mitigation could benefit broader safety and alignment tuning of LLMs, which we leave as future work for further empirical investigation. On the other hand, practitioners should be aware that improvements on targeted benchmarks do not establish broader fairness, safety, or reliability. Generated code may also contain errors or security vulnerabilities despite passing benchmark tests. Downstream applications should therefore include appropriate and comprehensive evaluation and necessary system-level safeguards.

REPRODUCIBILITY STATEMENT

An anonymized implementation, experiment configurations, data, and instructions for reproducing the results are provided in the supplementary materials. For experiments, we report results across three runs in most experiments with different random seeds, together with their standard deviations, to characterize variability. Besides main text, experiment and implementation details are illustrated in Appendices C and D. For the theory part, we state the assumptions and provide formal theorem statements and full proofs in Appendix B.

REFERENCES

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. Alpagasus: Training a better alpaca with fewer data. In *International Conference on Learning Representations*, volume 2024, pp. 34767–34797, 2024.

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine learning*, 15(2):201–221, 1994.

Qirun Dai, Dylan Zhang, Jiaqi W. Ma, and Hao Peng. Improving influence-based instruction tuning data selection for balanced learning of diverse capabilities. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 7079–7102. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.373. URL https://aclanthology.org/2025.findings-emnlp.373/.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 569–578, 2011.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pp. 2242–2251. PMLR, 2019.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Luxi He, Mengzhou Xia, and Peter Henderson. What is in your safe data? identifying benign data that breaks safety. *arXiv preprint arXiv:2404.01099*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Naman Jain, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, volume 2025, pp. 58791–58831, 2025.

Krishnateja Killamsetty, Durga S, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5464–5474. PMLR, 2021a. URL https://proceedings.mlr.press/v139/killamsetty21a.html.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021b. doi: 10.1609/aaai.v35i9.16988. URL https://ojs.aaai.org/index.php/AAAI/article/view/16988.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL https://proceedings.mlr.press/v70/koh17a.html.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics*, 28(5):1302–1338, 2000. doi: 10.1214/aos/1015957395.

Hao Li, Lijun Li, Zhenghao Lu, Xianyi Wei, Rui Li, Jing Shao, and Lei Sha. Layer-aware representation filtering: Purifying finetuning data to preserve llm safety alignment. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8041–8061, 2025.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, 2024.

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

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 22353–22373, 2024a.

Yifei Liu, Li Lyna Zhang, Yi Zhu, Bingcheng Dong, Xudong Zhou, Ning Shang, Fan Yang, Cheng Li, and Mao Yang. rStar-Coder: Scaling competitive code reasoning with a large-scale verified dataset. *Advances in Neural Information Processing Systems*, 38:58780–58807, 2025.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. TSDS: Data selection for task-specific model finetuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. doi: 10.52202/079017-0325. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/13848b5893119ff772b69812c95914fa-Abstract-Conference.html.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. # instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 36456–36474, 2024.

Guanghui Min, Tianhao Huang, Ke Wan, and Chen Chen. GIST: Targeted data selection for instruction tuning via coupled optimization geometry. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=K2KeBuKAtu.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Nihal V. Nayak, Paula Rodriguez-Diaz, Neha Hulkund, Sara Beery, and David Alvarez-Melis. A critical look at targeted instruction selection: Disentangling what matters (and what doesn't). In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Dy5GeCd003.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Art B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in neural information processing systems*, 33: 19920–19930, 2020.

Chris Ré. The road to Software 2.0 or data-centric AI. Hazy Research blog, July 2021. URL https://hazyresearch.stanford.edu/data-centric-ai. Accessed: 2026-09-19.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Burr Settles. Active learning literature survey. 2009. URL https://burrsettles.com/pub/settles.activelearning.pdf.

Jiachen Tianhao Wang, Tong Wu, Dawn Song, Prateek Mittal, and Ruoxi Jia. GREATS: Online selection of high-quality data for llm training in every iteration. *Advances in Neural Information Processing Systems*, 37:131197–131223, 2024.

Jiachen Tianhao Wang, Prateek Mittal, Dawn Song, and Ruoxi Jia. Data shapley in one training run. In *International conference on learning representations*, volume 2025, pp. 12358–12395, 2025a.

Jiachen Tianhao Wang, Dawn Song, James Y Zou, Prateek Mittal, and Ruoxi Jia. Capturing the temporal dependence of training data influence. In *International conference on learning representations*, volume 2025, pp. 11236–11273, 2025b.

Zige Wang, Qi Zhu, Fei Mi, Minghui Xu, Ruochun Jin, and Wenjing Yang. ClusterUCB: Efficient gradient-based data selection for targeted fine-tuning of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18867–18880. Association for Computational Linguistics, 2025c. doi: 10.18653/v1/2025.findings-emnlp.1026. URL https://aclanthology.org/2025.findings-emnlp.1026/.

Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Empowering code generation with OSS-instruct. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52632–52657. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wei24h.html.

Chunqiu Steven Xia, Yinlin Deng, and Lingming Zhang. Top leaderboard ranking = top coding proficiency, always? evoeval: Evolving coding benchmarks via llm. *arXiv preprint*, 2024a.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132. PMLR, 2024b. URL https://proceedings.mlr.press/v235/xia24c.html.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36: 34201–34227, 2023.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pp. 10842–10851. PMLR, 2020.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42, 2025.

Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. Tagcos: Task-agnostic gradient clustered coreset selection for instruction tuning data. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4686–4701, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

CONTENTS

1 Introduction **1**

2 Preliminaries **3**

2.1 Preliminaries and Related Work on Gradient-based Data Selection 3

2.2 Other Related Work 3

3 Analysis of Gradient-based Data Selection **4**

3.1 Main Results and Relative Gain Retention 4

3.2 Clustering 5

4 SUADS **7**

5 Experimental Evaluation **7**

5.1 Main Results 8

5.2 Ablation Studies 9

6 Conclusion **9**

A Detailed Discussion of Related Work **15**

A.1 Influence Estimation and Analyses of Data Selection 15

A.2 Gradient-Based Data Selection Baselines 15

A.3 Other Data Selection Methods for LLMs 16

A.4 Earlier Work on Data Selection and Data Valuation 17

B Proofs **17**

B.1 Formal Statement and Proof for Gain Retention Characterization under Isotropic Gaussian Noise 17

B.1.1 Setup and notation 17

B.1.2 Formal theorem 18

B.1.3 Proof of the formal theorem 19

B.1.4 A sufficient condition for controlling loss linearization 21

B.2 Generalization to Cosine-score-based Selection 22

B.3 Generalization to Anisotropic Gaussian 24

B.4 Gain Retention Rate for Clustering 27

B.4.1 Model and selection protocol 28

B.5 Derivation of Equation (6) 31

B.6 Empirical Scale of the Retention Factors and Scope of Equation (6) 33

C Method and Implementation Details **35**

C.1 Algorithm Pseudo-code for SUADS 35

C.2 Clustering 35
C.3 Scoring 36
C.4 Quota Rounding 36

D Experiment Details **36**
D.1 Training and Validation Data 36
D.2 Decoding 37
D.3 Fine-tuning and Gradient Features 37
D.4 Baseline Settings 37

E Additional Experimental Results **37**
E.1 Discussion on Performance across Different Budgets 37
E.2 Per-difficulty Results on DS_RSTAR 38
E.3 What Clusters Capture 38

F Limitations **40**

A Detailed Discussion of Related Work

A.1 Influence Estimation and Analyses of Data Selection

Influence estimation studies how individual training examples affect model predictions or target losses. Classical influence functions approximate the effect of upweighting a training example through the local sensitivity of the fitted parameters, involving an inverse-Hessian-weighted interaction between training and target gradients (Koh & Liang, 2017). TracIn (Pruthi et al., 2020) instead traces training dynamics through gradient inner products accumulated across checkpoints. More recently, Wang et al. (2025b) study trajectory-specific leave-one-out influence, which measures the effect of removing an example from a particular training iteration on the final model. These quantities capture different notions of influence: parameter sensitivity, accumulated training interactions, and counterfactual effects along a specified trajectory. Our analysis focuses on the simpler local setting in which gradient alignment approximates the utility of a single training update.

Beyond influence estimation, Nayak et al. (2026) systematically separate the roles of data representations and subset-selection algorithms in targeted instruction tuning. Their experiments show that gradient-based representations provide informative signals of alignment with target queries, but that these signals do not uniformly translate into downstream improvements. They also find that the relative effectiveness of selection algorithms varies with the selection budget and candidate pool, and develop generalization bounds based on distances between selected subsets and target queries. Our analysis complements this perspective by isolating sampling uncertainty in the estimated validation direction. Under the assumptions of our theoretical model, we quantify how this uncertainty propagates into a loss of selection gain, even when population gradient scores faithfully represent one-step training utility.

A.2 Gradient-Based Data Selection Baselines

We compare against four methods that improve different components of gradient-based selection: influence estimation, capability balancing, computational allocation, and gradient geometry.

LESS. LESS (Xia et al., 2024b) develops an optimizer-aware influence score for targeted instruction tuning. It computes candidate training representations from Adam update directions and compares them with validation gradients using cosine similarity, mitigating the response-length bias associated with gradient norms. Scores are aggregated across checkpoints from a short warmup run,

while LoRA and random projection reduce the cost of constructing and storing gradient features. For multiple target subtasks, LESS averages validation gradient features within each subtask and ranks candidates by their maximum aggregated subtask score. It then selects the highest-ranked examples. Thus, LESS incorporates validation-gradient aggregation within predefined subtasks following the introduced framework in Section 2.1. Our method instead identifies local groups through validation-gradient clustering and uses the resulting reference directions for stratified selection, motivated by our analysis of sampling uncertainty.

BIDS. BIDS (Dai et al., 2025) addresses imbalanced capability coverage in influence-based instruction selection. Its experiments use LESS-derived influence estimates to construct a matrix relating candidate training examples to individual validation examples. BIDS first standardizes each column of this matrix across the candidate pool, making influence scores for different validation examples comparable in scale. It then iteratively selects candidates according to their normalized influence relative to the average influence of the currently selected subset. Selection priorities therefore change as the subset grows, even when the underlying influence matrix remains fixed. BIDS primarily addresses score-scale imbalance and capability allocation. Our approach focuses on the reliability of the validation reference directions themselves and uses local aggregation to inform selection. These mechanisms address different aspects of the dependence on a finite validation set.

ClusterUCB. ClusterUCB (Wang et al., 2025c) reduces the computational cost of identifying influential training examples. It initially *clusters the candidate training pool* using gradient similarity and treats each cluster as an arm in a budget-constrained sampling procedure. A modified upper-confidence-bound rule uses previously observed influence scores to allocate further scoring effort across clusters, following an initial cold-start phase. The final subset consists of high-scoring examples among the candidates evaluated within the available budget. Its clustering therefore organizes the search over *training candidates*, and its exploration concerns which clusters contain influential examples. Our clustering operates on *validation gradients* to construct coherent local reference directions. The two approaches differ in both the object being clustered and the uncertainty that motivates their design.

GIST. GIST (Min et al., 2026) improves the geometry used to compare training and validation gradients. After a lightweight LoRA warmup, it applies truncated singular value decomposition to the validation-gradient matrix to recover a task-relevant subspace. Training and validation gradients are projected into this shared subspace, and each candidate is scored by its maximum cosine similarity to an individual validation example. The highest-scoring candidates are then selected. GIST therefore already exploits the structure of validation gradients and filters directions outside the learned subspace. GIST reduces r but scores each candidate against an individual validation example, a reference with $m = 1$ for which Equation (3) gives retention $\sqrt{\kappa}$. Our method scores against cluster centroids, whose retention Theorem 2 gives at $m = |V_\ell|$ (Appendix B.6).

A.3 Other Data Selection Methods for LLMs

LLM data selection also uses signals beyond gradient alignment. AlpaGasus (Chen et al., 2024) filters instruction data using quality assessments from a strong language model. DEITA (Liu et al., 2024a) combines instruction complexity, quality, and diversity to construct compact tuning datasets. The instruction-following difficulty metric (Li et al., 2024) uses response losses with and without instruction conditioning to identify informative training examples. These methods provide alternative proxies for data usefulness, without estimating the one-step reduction in target validation loss.

Another line of work emphasizes matching a target distribution. DSIR (Xie et al., 2023) estimates importance weights in a reduced feature space and resamples candidate data accordingly. TSDS (Liu et al., 2024b) formulates selection through optimal transport. Its framework supports different representations, and its instruction-tuning experiments use LESS-style gradient features. Consequently, distribution matching and gradient-based representations are compatible design choices.

Clustering has also been used to construct representative instruction datasets. TAGCOS (Zhang et al., 2025) clusters training examples in gradient space and performs coreset selection within the resulting groups. Together with ClusterUCB, it illustrates how training-data clustering can support

representativeness or efficient search. Our use of clustering is directed at estimating local validation reference directions and organizing selection around those references.

We note that, as studied by Nayak et al. (2026), gradient-based data selection is among the most competitive for targeted LLM instruction selection, which is our study focus. On the other hand, our gradient-based selection criteria can be combined with other criteria outlined here to further improve effectiveness.

A.4 Earlier Work on Data Selection and Data Valuation

Data selection has a substantial history in active learning and coreset construction. For example, Sener & Savarese (2018) formulate batch active learning as selecting a representative subset using the geometry of learned representations. In this setting, selection determines which examples to label, whereas our setting selects from an existing instruction-tuning corpus. For efficient training, CRAIG (Mirzasoleiman et al., 2020) constructs weighted subsets that approximate the full training gradient. GRAD-MATCH (Killamsetty et al., 2021a) directly minimizes gradient-matching error using orthogonal matching pursuit and supports either training or validation gradients as the reference. GLISTER (Killamsetty et al., 2021b) instead optimizes a validation-based objective through a one-step training approximation and Taylor-based greedy selection. These objectives differ from independently scoring candidates and applying a fixed top-K rule.

Data valuation studies how to assign utility or credit to individual training examples. Data Shapley (Ghorbani & Zou, 2019) averages marginal contributions across subsets to obtain an equitable allocation of predictive utility. DVRL (Yoon et al., 2020) learns a data-value estimator jointly with a predictor, using target-task performance as a reinforcement signal. Such approaches address the definition or estimation of data value, and their scores can support data selection. Our work isolates a further statistical issue within gradient-based selection: even a faithful population utility score can produce suboptimal selections when its reference direction must be estimated from finitely many validation examples. We characterize this effect under a simplified model and use the resulting insights to motivate SUADS.

B Proofs

B.1 Formal Statement and Proof for Gain Retention Characterization under Isotropic Gaussian Noise

This appendix formalizes Theorem 1. We first establish an exact identity for the first-order loss model, then control the loss-linearization error and the uncertainty due to finite validation sampling. The relative gain retention itself remains the population-loss quantity in Definition 1.

B.1.1 Setup and notation

Fix a checkpoint θ and let $\boldsymbol{\mu} = \nabla_\theta \mathcal{L}_{\mathrm{val}}(\theta) \neq 0$. Let $N \geq 2$, $M \geq 1$, and $1 \leq K < N$ be integers. Assume that

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_t, \tau^2 \boldsymbol{I}_P), \quad \boldsymbol{v}_j = \nabla_\theta \mathcal{L}(w_j) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{P}), \tag{8}$$

where $\tau > 0$, $\sigma \geq 0$, and the training and validation collections are independent. The orthogonal projector matrix $\boldsymbol{P}$ satisfies

$$\boldsymbol{P}^\mathsf{T} = \boldsymbol{P}, \quad \boldsymbol{P}^2 = \boldsymbol{P}, \quad \boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{\mu}, \quad \mathrm{rank}(\boldsymbol{P}) = r \in \{1, \ldots, P\}. \tag{9}$$

Define

$$\hat{\boldsymbol{\mu}} = \frac{1}{M} \sum_{j=1}^{M} \boldsymbol{v}_j, \quad q = \frac{\sigma^2}{M\|\boldsymbol{\mu}\|^2}, \quad d = r - 1, \quad \rho_0 = (1 + dq)^{-1/2}. \tag{10}$$

The estimator $\hat{\boldsymbol{\mu}}$ is nonzero almost surely. We write $\mathbb{E}_{\mathrm{tr}}$ for expectation over the training pool and $\mathbb{E}_{\mathrm{val}}$ and $\Pr_{\mathrm{val}}$ for expectation and probability over the validation sample. An expectation over the training pool holds the validation reference fixed.

Identify a subset of training examples with its index set, and let $\mathcal{S}_K = \{S \subseteq \{1, \ldots, N\} : |S| = K\}$. For a nonzero reference vector $\boldsymbol{h}$, let $S_{\boldsymbol{h}} = \mathcal{A}_{\boldsymbol{h}}(D)$ denote the indices of the K largest scores

$\langle \boldsymbol{x}_i, \boldsymbol{h} \rangle$ in the training pool D. Any deterministic tie-breaking rule can be used; ties occur with probability zero for fixed $\boldsymbol{h} \neq 0$. For $S \in \mathcal{S}_K$, define the actual and linearized loss reductions

$$F_D(S) = \mathcal{L}_{\text{val}}(\theta) - \mathcal{L}_{\text{val}}\left(\theta - \sum_{i \in S} \boldsymbol{x}_i\right), \tag{11}$$

$$\ell_D(S) = \sum_{i \in S} \langle \boldsymbol{x}_i, \boldsymbol{\mu} \rangle. \tag{12}$$

Their uniform, size-K random-selection baselines are

$$\overline{F}_D = \binom{N}{K}^{-1} \sum_{S \in \mathcal{S}_K} F_D(S), \quad \overline{\ell}_D = \frac{K}{N} \sum_{i=1}^{N} \langle \boldsymbol{x}_i, \boldsymbol{\mu} \rangle. \tag{13}$$

Thus, the retention in Definition 1, conditional on a fixed reference $\boldsymbol{h}$, can be written as

$$\rho(\boldsymbol{h}) := \rho(\mathcal{A}_{\boldsymbol{h}}; \mathcal{D}_{\text{train}}, \theta) = \frac{\mathbb{E}_{\text{tr}}[F_D(S_{\boldsymbol{h}}) - \overline{F}_D]}{\mathbb{E}_{\text{tr}}[\max_{S \in \mathcal{S}_K} F_D(S) - \overline{F}_D]}. \tag{14}$$

We also introduce an auxiliary first-order ratio

$$\rho_1(\boldsymbol{h}) = \frac{\mathbb{E}_{\text{tr}}[\ell_D(S_{\boldsymbol{h}}) - \overline{\ell}_D]}{\mathbb{E}_{\text{tr}}[\max_{S \in \mathcal{S}_K} \ell_D(S) - \overline{\ell}_D]}. \tag{15}$$

This auxiliary quantity is used only to analyze the approximation to the original retention.

Let $G_1, \ldots, G_N$ be independent standard normal variables, with order statistics $G_{(1)} \leq \cdots \leq G_{(N)}$, and define

$$\gamma_{N,K} = \mathbb{E}\left[\sum_{j=N-K+1}^{N} G_{(j)}\right], \quad \Gamma_{N,K} = \tau \|\boldsymbol{\mu}\| \gamma_{N,K}. \tag{16}$$

Finally, measure the loss-linearization error by

$$R_D = \max_{S \in \mathcal{S}_K} |F_D(S) - \ell_D(S)|, \quad \varepsilon_T = \mathbb{E}_{\text{tr}} R_D. \tag{17}$$

B.1.2 Formal theorem

Now we are ready to present the formal version of Theorem 1 in the main text.

Theorem 3 (Finite-sample gain retention characterization)**.** *Under the setup above, $\Gamma_{N,K} > 0$, and the following hold.*

1. ***Exact first-order identity.*** *For every fixed $\boldsymbol{h} \neq 0$,*

$$\rho_1(\boldsymbol{h}) = c(\boldsymbol{h}) := \frac{\langle \boldsymbol{\mu}, \boldsymbol{h} \rangle}{\|\boldsymbol{\mu}\| \|\boldsymbol{h}\|}. \tag{18}$$

In particular, $\rho_1(\hat{\boldsymbol{\mu}}) = \cos\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\mu} \rangle$ almost surely over validation sampling, and

$$\mathbb{E}_{\text{val}} \rho_1(\hat{\boldsymbol{\mu}}) = \mathbb{E}_{\text{val}} c(\hat{\boldsymbol{\mu}}). \tag{19}$$

2. ***Loss-linearization error.*** *If $\varepsilon_T < \infty$ and $2\varepsilon_T < \Gamma_{N,K}$, then the denominator of Equation* (14) *is positive and, for every $\boldsymbol{h} \neq 0$,*

$$|\rho(\boldsymbol{h}) - c(\boldsymbol{h})| \leq B_T, \quad B_T := \frac{4\varepsilon_T}{\Gamma_{N,K} - 2\varepsilon_T}. \tag{20}$$

Consequently,

$$|\mathbb{E}_{\text{val}} \rho(\hat{\boldsymbol{\mu}}) - \mathbb{E}_{\text{val}} c(\hat{\boldsymbol{\mu}})| \leq B_T. \tag{21}$$

3. ***Finite-validation concentration.*** *For any $\epsilon, \delta \in (0,1)$ satisfying*

$$q = \frac{\sigma^2}{M\|\boldsymbol{\mu}\|^2} \le \frac{\epsilon^2}{16\log(4/\delta)}, \tag{22}$$

we have

$$\Pr_{\text{val}}\left(\left|\frac{c(\hat{\boldsymbol{\mu}})}{\rho_0} - 1\right| \le \epsilon\right) \ge 1-\delta. \tag{23}$$

Moreover,

$$|\mathbb{E}_{\text{val}} c(\hat{\boldsymbol{\mu}}) - \rho_0| \le \epsilon\rho_0 + 2\delta. \tag{24}$$

If the assumptions in part 2 also hold, then

$$\Pr_{\text{val}}(|\rho(\hat{\boldsymbol{\mu}}) - \rho_0| \le B_T + \epsilon\rho_0) \ge 1-\delta, \tag{25}$$

$$|\mathbb{E}_{\text{val}}\rho(\hat{\boldsymbol{\mu}}) - \rho_0| \le B_T + \epsilon\rho_0 + 2\delta. \tag{26}$$

Remark. The concentration condition is sufficient, with conservative constants, and holds for every finite $r \ge 1$; it does not require $r \to \infty$. The exact first-order ratio is independent of N, K, $\boldsymbol{\mu}_t$, and τ. The loss-linearization bound B_T can depend on these quantities through ε_T and $\Gamma_{N,K}$.

B.1.3 Proof of the formal theorem

Proof. **Step 1: exact retention for a fixed validation reference.** Fix $\boldsymbol{h} \neq 0$ and write $\boldsymbol{x}_i = \boldsymbol{\mu}_t + \tau\boldsymbol{g}_i$, where $\boldsymbol{g}_1, \dots, \boldsymbol{g}_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I}_P)$. Define the standardized scores

$$U_i = \left\langle \boldsymbol{g}_i, \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} \right\rangle, \quad V_i = \left\langle \boldsymbol{g}_i, \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|} \right\rangle, \quad c = c(\boldsymbol{h}). \tag{27}$$

Each pair (U_i, V_i) is jointly Gaussian with zero mean, unit marginal variances, and correlation c. The pairs are independent across i. Therefore, including the degenerate cases $c = \pm 1$, Gaussian regression gives

$$\mathbb{E}[U_i \mid V_1, \dots, V_N] = cV_i. \tag{28}$$

The selector $S_{\boldsymbol{h}}$ is the set of indices of the K largest V_i, so it is measurable with respect to $(V_1, \dots, V_N)$. It follows that

$$\mathbb{E}_{\text{tr}}\left[\sum_{i\in S_{\boldsymbol{h}}} U_i\right] = \mathbb{E}_{\text{tr}}\left[\sum_{i=1}^N \mathbf{1}\{i \in S_{\boldsymbol{h}}\}\mathbb{E}[U_i \mid V_1, \dots, V_N]\right] \tag{29}$$

$$= c\,\mathbb{E}_{\text{tr}}\left[\sum_{i\in S_{\boldsymbol{h}}} V_i\right] = c\gamma_{N,K}. \tag{30}$$

Because $\langle \boldsymbol{x}_i, \boldsymbol{\mu}\rangle = \langle \boldsymbol{\mu}_t, \boldsymbol{\mu}\rangle + \tau\|\boldsymbol{\mu}\|U_i$, the expected random-selection baseline is

$$\mathbb{E}_{\text{tr}}\bar{\ell}_D = K\langle \boldsymbol{\mu}_t, \boldsymbol{\mu}\rangle. \tag{31}$$

Thus the selected subset has expected excess first-order gain

$$\mathbb{E}_{\text{tr}}[\ell_D(S_{\boldsymbol{h}}) - \bar{\ell}_D] = c\tau\|\boldsymbol{\mu}\|\gamma_{N,K} = c\Gamma_{N,K}. \tag{32}$$

The first-order oracle selects the K largest U_i, yielding

$$\mathbb{E}_{\text{tr}}\left[\max_{S\in\mathcal{S}_K} \ell_D(S) - \bar{\ell}_D\right] = \tau\|\boldsymbol{\mu}\|\gamma_{N,K} = \Gamma_{N,K}. \tag{33}$$

To see that $\gamma_{N,K} > 0$, observe that almost surely

$$\sum_{j=N-K+1}^{N} G_{(j)} > \frac{K}{N}\sum_{i=1}^{N} G_i, \tag{34}$$

since $K < N$ and the normal samples are not all equal. Taking expectations gives strict positivity; finiteness follows from the integrability of $\sum_i |G_i|$. Dividing the two excess gains proves Equation (18). Independence of the two datasets permits the same argument conditionally on $\boldsymbol{h} = \hat{\boldsymbol{\mu}}$.

Step 2: comparison with the original loss-based retention. Let $A_F(\boldsymbol{h})$ and D_F denote the numerator and denominator of Equation (14). For every realized training pool,

$$|\overline{F}_D - \overline{\ell}_D| \le R_D, \quad \left| \max_{S \in \mathcal{S}_K} F_D(S) - \max_{S \in \mathcal{S}_K} \ell_D(S) \right| \le R_D. \tag{35}$$

These bounds also give, uniformly over $\boldsymbol{h} \neq 0$,

$$|A_F(\boldsymbol{h}) - c(\boldsymbol{h})\Gamma_{N,K}| \le 2\varepsilon_T, \quad |D_F - \Gamma_{N,K}| \le 2\varepsilon_T. \tag{36}$$

Hence $D_F \ge \Gamma_{N,K} - 2\varepsilon_T > 0$ and

$$|\rho(\boldsymbol{h}) - c(\boldsymbol{h})| = \frac{|A_F(\boldsymbol{h}) - c(\boldsymbol{h})D_F|}{D_F} \tag{37}$$

$$\le \frac{2\varepsilon_T + 2|c(\boldsymbol{h})|\varepsilon_T}{\Gamma_{N,K} - 2\varepsilon_T} \le B_T. \tag{38}$$

This proves Equation (20). Averaging the pointwise bound over the validation sample gives Equation (21).

Step 3: exact distribution of the validation alignment. Choose an orthonormal basis $\boldsymbol{e}_1, \dots, \boldsymbol{e}_r$ for the range of $\boldsymbol{P}$ with $\boldsymbol{e}_1 = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|$. This is possible because $\boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{\mu}$. Since $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{P}/M)$, we can write

$$\hat{\boldsymbol{\mu}} = \|\boldsymbol{\mu}\| \left[(1 + \sqrt{q}Z_1)\boldsymbol{e}_1 + \sqrt{q} \sum_{j=2}^{r} Z_j \boldsymbol{e}_j \right], \tag{39}$$

where the Z_j are independent standard normal variables. Consequently,

$$c(\hat{\boldsymbol{\mu}}) = \frac{1 + \sqrt{q}Z}{\sqrt{(1 + \sqrt{q}Z)^2 + qQ}}, \quad Z \sim \mathcal{N}(0,1), \quad Q \sim \chi^2_d, \quad Z \perp Q. \tag{40}$$

Here $Q = 0$ almost surely when $d = 0$. The mean-direction noise is represented by Z, while Q accounts for the $r - 1$ orthogonal noise directions.

Step 4: a finite-sample relative error bound. Set $t = \log(4/\delta)$. The standard Gaussian tail bound and the chi-square inequalities of Laurent & Massart (2000) give

$$\Pr(|Z| > \sqrt{2t}) \le 2e^{-t}, \tag{41}$$

$$\Pr\big(|Q - d| > 2\sqrt{dt} + 2t\big) \le 2e^{-t}. \tag{42}$$

For $d = 0$, the second event is impossible, so the bound remains valid. A union bound yields an event $\mathcal{E}$ with probability at least $1 - 4e^{-t} = 1 - \delta$ on which

$$|\sqrt{q}Z| \le a := \sqrt{2qt}, \quad \frac{|qQ - dq|}{1 + dq} \le \frac{2q\sqrt{dt} + 2qt}{1 + dq} \le b := \sqrt{qt} + 2qt. \tag{43}$$

The last inequality follows from $2\sqrt{x}/(1 + x) \le 1$ for $x \ge 0$.

Suppose $a, b \le 1/2$, as ensured below by Equation (22). Write $u = \sqrt{q}Z$ and $s = qQ \ge 0$. On $\mathcal{E}$, $1 + u > 0$, and

$$\frac{c(\hat{\boldsymbol{\mu}})}{\rho_0} = \underbrace{\frac{(1+u)\sqrt{1+s}}{\sqrt{(1+u)^2 + s}}}_{T(u,s)} \sqrt{\frac{1 + dq}{1 + s}}. \tag{44}$$

The factor $T(u, s)$ lies between 1 and $1 + u$. Indeed, if $1 + u \ge 1$, its square lies between 1 and $(1 + u)^2$; if $0 < 1 + u \le 1$, the inequalities reverse. Therefore $1 - a \le T(u, s) \le 1 + a$ and

$$\frac{1 - a}{\sqrt{1 + b}} \le \frac{c(\hat{\boldsymbol{\mu}})}{\rho_0} \le \frac{1 + a}{\sqrt{1 - b}}. \tag{45}$$

For $0 \le b \le 1/2$, $(1-b)^{-1/2} - 1 \le b$ and $1 - (1+b)^{-1/2} \le b/2$. Using these inequalities in Equation (45) gives

$$\left| \frac{c(\hat{\boldsymbol{\mu}})}{\rho_0} - 1 \right| \le \sqrt{2}\, a + b = 3\sqrt{qt} + 2qt. \tag{46}$$

Under Equation (22), $qt \le \epsilon^2/16$, so

$$a \le \frac{\epsilon}{2\sqrt{2}} < \frac{1}{2}, \quad b \le \frac{\epsilon}{4} + \frac{\epsilon^2}{8} < \frac{1}{2}, \quad 3\sqrt{qt} + 2qt \le \frac{3\epsilon}{4} + \frac{\epsilon^2}{8} \le \epsilon. \tag{47}$$

This proves Equation (23).

Step 5: expectation and combined bounds. On $\mathcal{E}$, $|c(\hat{\boldsymbol{\mu}}) - \rho_0| \le \epsilon\rho_0$. On its complement, the bound $|c(\hat{\boldsymbol{\mu}}) - \rho_0| \le 2$ follows from $c(\hat{\boldsymbol{\mu}}) \in [-1, 1]$ and $\rho_0 \in (0, 1]$. Hence

$$|\mathbb{E}_{\mathrm{val}} c(\hat{\boldsymbol{\mu}}) - \rho_0| \le \mathbb{E}_{\mathrm{val}} |c(\hat{\boldsymbol{\mu}}) - \rho_0| \le \epsilon\rho_0 + 2\delta. \tag{48}$$

Combining this inequality and the corresponding eventwise bound with Equation (20) proves Equation (25) and Equation (26). □

B.1.4 A SUFFICIENT CONDITION FOR CONTROLLING LOSS LINEARIZATION

Corollary 1 (Smooth-loss specialization)**.** *Suppose $\mathcal{L}_{\mathrm{val}}$ has an H-Lipschitz gradient on every segment between θ and $\theta - \sum_{i\in S} \boldsymbol{x}_i$, for all $S \in \mathcal{S}_K$, almost surely over the training pool. Then*

$$\varepsilon_T \le \frac{H}{2}\, \mathbb{E}_{\mathrm{tr}} \left[\max_{S\in\mathcal{S}_K} \left\| \sum_{i\in S} \boldsymbol{x}_i \right\|^2 \right]. \tag{49}$$

In particular, fix N, K, θ, and the distribution of unscaled update vectors $\widetilde{\boldsymbol{x}}_i$, and set $\boldsymbol{x}_i = \eta\widetilde{\boldsymbol{x}}_i$ for $\eta > 0$. If the same finite H applies for all sufficiently small η, then $\varepsilon_T = O(\eta^2)$, $\Gamma_{N,K} = \Theta(\eta)$, and $B_T = O(\eta)$ as $\eta \to 0$.

Proof. With $\boldsymbol{a}_S = \sum_{i\in S} \boldsymbol{x}_i$, the integral form of the Taylor remainder and the Lipschitz-gradient assumption yield

$$|F_D(S) - \langle \boldsymbol{a}_S, \boldsymbol{\mu} \rangle| = \left| \int_0^1 \langle \nabla\mathcal{L}_{\mathrm{val}}(\theta - t\boldsymbol{a}_S) - \nabla\mathcal{L}_{\mathrm{val}}(\theta), \boldsymbol{a}_S \rangle\, dt \right| \tag{50}$$

$$\le \int_0^1 Ht \|\boldsymbol{a}_S\|^2\, dt = \frac{H}{2} \|\boldsymbol{a}_S\|^2. \tag{51}$$

Maximizing over S and taking the training-pool expectation proves Equation (49). The expectation is finite: by Cauchy–Schwarz, $\max_{S\in\mathcal{S}_K} \|\sum_{i\in S} \boldsymbol{x}_i\|^2 \le K \sum_{i=1}^N \|\boldsymbol{x}_i\|^2$, whose expectation is finite under the Gaussian assumption. Scaling the updates by η scales this remainder bound by η^2 and the positive first-order oracle gain by η. The claimed scaling of B_T follows. □

Interpretation of the two approximations. The relation $\rho(\hat{\boldsymbol{\mu}}) \approx \cos\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\mu} \rangle$ is controlled by loss linearization; the corresponding first-order identity is exact after averaging over the training pool. The relation $\cos\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\mu} \rangle \approx \rho_0$ is controlled by validation sampling. Thus the averaged form in Theorem 1 has the explicit error bound Equation (26).

Scaling and the noise dimension. One may question the negative correlation between r and ρ, which may suggest larger model size hurts tuning in contrast to the classical scaling law. The dependence holds when the per-direction noise σ^2 is fixed, so that the total noise energy $r\sigma^2$ grows with the number of tunable directions; in that regime ρ suggests a proportional scaling ratio $(r-1)/M$ between tunable parameter size and validation data size. For ℓ_2-normalized gradients, as used in Section 4, $r\sigma^2$ is pinned at $1 - \kappa$ and the dependence on r disappears (Appendix B.6).

B.2 Generalization to Cosine-score-based Selection

In this subsection, we observe that some gradient-based data selection methods use cosine similarity of gradients Xia et al. (2024b) instead of inner product motivated from the Adam optimizer, or normalize the gradients before inner product (which leads to cosine similarity). Now we analyze the finite-sampling-induced gain retention in this case.

We retain the model and the population-loss retention in Appendix B.1, and assume $P \geq 2$. For a nonzero reference $\boldsymbol{h}$, define

$$S_{\boldsymbol{h}}^{\cos} := \mathrm{Top}_K\{\cos\langle \boldsymbol{x}_i, \boldsymbol{h}\rangle\}_{i=1}^N, \qquad \cos\langle \boldsymbol{x}_i, \boldsymbol{h}\rangle := \frac{\langle \boldsymbol{x}_i, \boldsymbol{h}\rangle}{\|\boldsymbol{x}_i\|\|\boldsymbol{h}\|}.$$

The selected updates remain the original vectors $\boldsymbol{x}_i$. Let $\rho^{\cos}(\boldsymbol{h})$ and $\rho_1^{\cos}(\boldsymbol{h})$ denote the ratios in Equations (14) and (15) with $S_{\boldsymbol{h}}$ replaced by $S_{\boldsymbol{h}}^{\cos}$. The random baseline and the population oracle are unchanged. In particular, the first-order oracle still maximizes $\sum_{i\in S}\langle \boldsymbol{x}_i, \boldsymbol{\mu}\rangle$ over size-K subsets. Since the training distribution has a density and $P \geq 2$, the candidate norms are nonzero and cosine-score ties have probability zero for each fixed $\boldsymbol{h} \neq 0$.

Normalizing the common reference $\boldsymbol{h}$ does not change the ranking. Normalizing each candidate $\boldsymbol{x}_i$ can change it. We first quantify this effect without imposing any restriction on the training mean $\boldsymbol{\mu}_t$.

Theorem 4 (Cosine selection with an arbitrary training mean)**.** *Under the isotropic training and validation model in Appendix B.1, fix a deterministic $L > 0$ and define*

$$e_{\cos}(L) := \sup_{\|\mathbf{u}\|=1} \mathbb{E}_{\mathrm{tr}} \max_{1\leq i\leq N} \left|\langle \boldsymbol{x}_i, \mathbf{u}\rangle - L\cos\langle \boldsymbol{x}_i, \mathbf{u}\rangle\right|, \tag{52}$$

$$B_{\cos}(L) := \min\left\{2, 2\sqrt{\frac{K e_{\cos}(L)}{\tau\gamma_{N,K}}}\right\}. \tag{53}$$

The quantity $e_{\cos}(L)$ is finite; one natural choice is $L = \sqrt{\|\boldsymbol{\mu}_t\|^2 + P\tau^2}$. For every fixed $\boldsymbol{h} \neq 0$,

$$|\rho_1^{\cos}(\boldsymbol{h}) - \cos\langle \boldsymbol{h}, \boldsymbol{\mu}\rangle| \leq B_{\cos}(L). \tag{54}$$

If $\varepsilon_T < \infty$ and $2\varepsilon_T < \Gamma_{N,K}$, then

$$|\rho^{\cos}(\boldsymbol{h}) - \rho_1^{\cos}(\boldsymbol{h})| \leq B_T := \frac{4\varepsilon_T}{\Gamma_{N,K} - 2\varepsilon_T}. \tag{55}$$

Consequently, for $\epsilon, \delta \in (0,1)$ satisfying

$$\frac{\sigma^2}{M\|\boldsymbol{\mu}\|^2} \leq \frac{\epsilon^2}{16\log(4/\delta)}, \tag{56}$$

and $\rho_0 = (1 + (r-1)\sigma^2/(M\|\boldsymbol{\mu}\|^2))^{-1/2}$,

$$\Pr_{\mathrm{val}}(|\rho^{\cos}(\hat{\boldsymbol{\mu}}) - \rho_0| \leq B_T + B_{\cos}(L) + \epsilon\rho_0) \geq 1 - \delta, \tag{57}$$

$$|\mathbb{E}_{\mathrm{val}}\rho^{\cos}(\hat{\boldsymbol{\mu}}) - \rho_0| \leq B_T + B_{\cos}(L) + \epsilon\rho_0 + 2\delta. \tag{58}$$

For the first-order retention alone, the same conclusions hold with $B_T = 0$, without any smoothness or Taylor-error assumption.

Remark. As we can observe, the cosine selection complicates the analysis, and a closed-form retention rate ρ like Equation (3) is hard to obtain. Hence, we only prove that ρ in this setting still approximately equal to the expectation of cosine similarity up to some error terms. In Theorem 5, when the training update is centered, we further derive the closed-form rate ρ in Equation (3) form.

Proof. **Step 1: a bound on the expected selected update.** Fix $\boldsymbol{h} \neq 0$ and write $\mathbf{e} = \boldsymbol{h}/\|\boldsymbol{h}\|$. Define

$$\mathbf{w} := \frac{1}{\tau\gamma_{N,K}}\mathbb{E}_{\mathrm{tr}} \sum_{i\in S_{\boldsymbol{h}}^{\cos}} (\boldsymbol{x}_i - \boldsymbol{\mu}_t).$$

For every fixed unit vector $\mathbf{u}$, pointwise maximization over size-K subsets gives

$$\langle \mathbf{u}, \mathbf{w}\rangle \leq \frac{1}{\tau\gamma_{N,K}}\mathbb{E}_{\mathrm{tr}} \max_{S\in\mathcal{S}_K} \sum_{i\in S}\langle \boldsymbol{x}_i - \boldsymbol{\mu}_t, \mathbf{u}\rangle = 1.$$

Taking the supremum over $\mathbf{u}$ yields $\|\mathbf{w}\| \le 1$. Since the expected random baseline is $K\langle \boldsymbol{\mu}_t, \boldsymbol{\mu}\rangle$ and the oracle excess gain is $\Gamma_{N,K} = \tau\|\boldsymbol{\mu}\|\gamma_{N,K}$,

$$\rho_1^{\cos}(\boldsymbol{h}) = \left\langle \mathbf{w}, \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|} \right\rangle, \qquad |\rho_1^{\cos}(\boldsymbol{h})| \le 1. \tag{59}$$

Step 2: control of the change in ranking. Let $s_i = \langle \boldsymbol{x}_i, \mathbf{e}\rangle$ and $t_i = L\cos\langle \boldsymbol{x}_i, \mathbf{e}\rangle$. The dot-product selector $S_{\boldsymbol{h}}$ maximizes the sum of the s_i, whereas $S_{\boldsymbol{h}}^{\cos}$ maximizes the sum of the t_i. Therefore, for each training pool,

$$0 \le \sum_{i\in S_{\boldsymbol{h}}} s_i - \sum_{i\in S_{\boldsymbol{h}}^{\cos}} s_i \le 2K \max_i |s_i - t_i|.$$

After taking expectations and subtracting $K\langle \boldsymbol{\mu}_t, \mathbf{e}\rangle$ from each selected sum,

$$\tau\gamma_{N,K}(1 - \langle \mathbf{e}, \mathbf{w}\rangle) \le 2Ke_{\cos}(L).$$

Using $\|\mathbf{w}\| \le 1$ gives

$$\|\mathbf{w} - \mathbf{e}\|^2 \le 2(1 - \langle \mathbf{e}, \mathbf{w}\rangle) \le \frac{4Ke_{\cos}(L)}{\tau\gamma_{N,K}}.$$

Cauchy–Schwarz and Equation (59) now imply Equation (54); the additional bound by 2 follows because both terms lie in $[-1, 1]$. Finiteness follows, for example, from $e_{\cos}(L) \le \mathbb{E}_{\mathrm{tr}} \max_i \|\boldsymbol{x}_i\| + L$.

Step 3: loss linearization and validation sampling. Let A_F, A_1 be the actual and linearized numerators for $S_{\boldsymbol{h}}^{\cos}$, and let D_F be the actual oracle denominator. The definition of R_D implies

$$|A_F - A_1| \le 2\varepsilon_T, \qquad |D_F - \Gamma_{N,K}| \le 2\varepsilon_T.$$

Since $|A_1| \le \Gamma_{N,K}$ by Equation (59),

$$\left|\frac{A_F}{D_F} - \frac{A_1}{\Gamma_{N,K}}\right| \le \frac{2\varepsilon_T + 2\varepsilon_T|A_1|/\Gamma_{N,K}}{\Gamma_{N,K} - 2\varepsilon_T} \le B_T.$$

Finally, training and validation independence allows the fixed-reference bounds to be applied conditionally on $\boldsymbol{h} = \hat{\boldsymbol{\mu}}$. Under Equation (56), Theorem 3 gives $|\cos\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\mu}\rangle - \rho_0| \le \epsilon\rho_0$ with probability at least $1 - \delta$, and an expected absolute error at most $\epsilon\rho_0 + 2\delta$. Combining the three errors proves the claims. □

An exact characterization for centered training updates. The following specialization has a simpler multiplicative form. The additional assumption $\boldsymbol{\mu}_t = \mathbf{0}$ is substantive.

Theorem 5 (Exact cosine retention under spherical symmetry)**.** *In addition to the preceding isotropic model, suppose $\boldsymbol{\mu}_t = \mathbf{0}$. Let $R \sim \chi_P$, let $\mathbf{U}_1, \dots, \mathbf{U}_N$ be independent and uniform on the unit sphere in $\mathbb{R}^P$, and set $T_i = \langle \mathbf{U}_i, \mathbf{e}_1\rangle$ for any fixed unit vector $\mathbf{e}_1$. With $T_{(1)} \le \cdots \le T_{(N)}$, define*

$$\beta_{N,K,P} := \frac{\mathbb{E}R}{\gamma_{N,K}} \mathbb{E}\left[\sum_{j=N-K+1}^{N} T_{(j)}\right]. \tag{60}$$

Then, for every $\boldsymbol{h} \neq 0$,

$$\rho_1^{\cos}(\boldsymbol{h}) = \beta_{N,K,P}\cos\langle \boldsymbol{h}, \boldsymbol{\mu}\rangle, \qquad 0 < \beta_{N,K,P} \le 1. \tag{61}$$

Moreover, the following finite-dimensional bound holds:

$$0 \le 1 - \beta_{N,K,P} \le \min\left\{1, \frac{K}{\gamma_{N,K}}\sqrt{\frac{2N}{P}}\right\}. \tag{62}$$

In particular, $\beta_{N,K,P} \to 1$ as $P \to \infty$ for fixed N and K.

If $2\varepsilon_T < \Gamma_{N,K}$ and Equation (56) holds, then

$$\Pr_{\mathrm{val}}(|\rho^{\cos}(\hat{\boldsymbol{\mu}}) - \beta_{N,K,P}\rho_0| \le B_T + \beta_{N,K,P}\epsilon\rho_0) \ge 1 - \delta, \tag{63}$$

$$|\mathbb{E}_{\mathrm{val}}\rho^{\cos}(\hat{\boldsymbol{\mu}}) - \beta_{N,K,P}\rho_0| \le B_T + \beta_{N,K,P}(\epsilon\rho_0 + 2\delta). \tag{64}$$

Thus the explicit approximation is

$$\mathbb{E}_{\mathrm{val}}\rho^{\cos}(\hat{\boldsymbol{\mu}}) \approx \beta_{N,K,P}\,\mathbb{E}_{\mathrm{val}}\cos\langle\hat{\boldsymbol{\mu}},\boldsymbol{\mu}\rangle \approx \beta_{N,K,P}\left(1+\frac{(r-1)\sigma^2}{M\|\boldsymbol{\mu}\|^2}\right)^{-1/2}, \tag{65}$$

where the first error is at most B_T and the second is controlled by Equation (64) *with $B_T = 0$.*

Remark. Comparing Equation (65) and Equation (3), we observe that the cosine similarity selection introduces an independent but additional attenuation factor $\beta_{N,K,P}$ that introduces some dependency on the dimensionality of the training update noise distribution. Besides this, our previous key factor analysis in Section 3.1 still holds.

Proof. **Step 1: polar decomposition and selection.** Write $\boldsymbol{x}_i = \tau R_i \mathbf{U}_i$, where $R_i \overset{\text{i.i.d.}}{\sim} \chi_P$, the directions $\mathbf{U}_i$ are independent and uniform on the sphere, and all radii are independent of all directions. This follows directly by expressing the centered Gaussian density in polar coordinates. For $\mathbf{e} = \boldsymbol{h}/\|\boldsymbol{h}\|$, put $T_i = \langle \mathbf{U}_i, \mathbf{e}\rangle$. The selector $S^{\cos}_{\boldsymbol{h}}$ depends only on the T_i. Rotational symmetry around $\mathbf{e}$ gives $\mathbb{E}[\mathbf{U}_i \mid T_i] = T_i\mathbf{e}$. Consequently,

$$\mathbb{E}_{\mathrm{tr}}\sum_{i\in S^{\cos}_{\boldsymbol{h}}}\boldsymbol{x}_i = \tau\mathbb{E}R\,\mathbb{E}\left[\sum_{j=N-K+1}^{N} T_{(j)}\right]\mathbf{e}.$$

Taking its inner product with $\boldsymbol{\mu}$ and dividing by $\tau\|\boldsymbol{\mu}\|\gamma_{N,K}$ proves Equation (61). The expected sum of the largest $K < N$ centered, continuous T_i is positive, so $\beta_{N,K,P} > 0$. At $\boldsymbol{h} = \boldsymbol{\mu}$, the cosine-selected subset cannot exceed the original first-order oracle, which proves $\beta_{N,K,P} \le 1$.

Step 2: a finite-dimensional bound. Let $\mathcal{T}_K(\mathbf{z})$ be the sum of the largest K coordinates of $\mathbf{z} \in \mathbb{R}^N$. It satisfies

$$|\mathcal{T}_K(\mathbf{z}) - \mathcal{T}_K(\mathbf{z}')| \le K\max_i |z_i - z_i'|.$$

Each R_iT_i is standard Gaussian, independently across i. Applying this inequality to $z_i = R_iT_i$ and $z_i' = (\mathbb{E}R)T_i$ gives

$$\begin{aligned}\gamma_{N,K}(1-\beta_{N,K,P}) &\le K\mathbb{E}\max_i |(R_i - \mathbb{E}R)T_i| \\ &\le K\sqrt{\sum_{i=1}^{N}\mathbb{E}[(R_i-\mathbb{E}R)^2T_i^2]} \\ &= K\sqrt{\frac{N\,\mathrm{Var}(R)}{P}}.\end{aligned}$$

Here we used independence and $\mathbb{E}T_i^2 = 1/P$. Since $R^2 \sim \chi^2_P$,

$$\mathrm{Var}(R) \le \mathbb{E}(R-\sqrt{P})^2 = \mathbb{E}\frac{(R^2-P)^2}{(R+\sqrt{P})^2} \le \frac{\mathbb{E}(R^2-P)^2}{P} = 2.$$

This proves Equation (62) and its stated limit.

Step 3: validation sampling and actual loss. The pointwise bound Equation (55) and the exact identity Equation (61) give

$$|\rho^{\cos}(\hat{\boldsymbol{\mu}}) - \beta_{N,K,P}\cos\langle\hat{\boldsymbol{\mu}},\boldsymbol{\mu}\rangle| \le B_T.$$

Multiply the validation concentration and expectation bounds in Theorem 3 by $\beta_{N,K,P}$ and apply the triangle inequality. This proves the remaining conclusions. □

B.3 Generalization to Anisotropic Gaussian

We now generalize dot-product-based selection to arbitrary Gaussian covariances for both training updates and validation gradients. This is a separate extension of the dot-product theorem; it does not assume that raw cosine selection has the same ranking. Retain F_D, ℓ_D, $\mathcal{S}_K$, the random baselines, and the ratios $\rho(\boldsymbol{h}), \rho_1(\boldsymbol{h})$ from Appendix B.1, but assume

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_{\mathrm{tr}}), \qquad \boldsymbol{v}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathrm{val}}). \tag{66}$$

The two collections are independent. Both covariance matrices are symmetric positive semidefinite, and $\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu} > 0$. In particular, singular covariances are permitted. All training-pool expectations and the Taylor remainder $\varepsilon_T = \mathbb{E}_{\mathrm{tr}} R_D$ are evaluated under Equation (66).

Define the transformed mean and validation covariance by

$$\mathbf{a} := \boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}\boldsymbol{\mu}, \qquad \mathbf{C} := \boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}\boldsymbol{\Sigma}_{\mathrm{val}}\boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}, \tag{67}$$

where the square root is the symmetric positive semidefinite one. Let

$$\mathbf{u} := \frac{\mathbf{a}}{\|\mathbf{a}\|}, \qquad \mathbf{Q} := \boldsymbol{I}_P - \mathbf{u}\mathbf{u}^\mathsf{T}, \qquad q_{\mathrm{A}} := \frac{\|\mathbf{C}\|_{\mathrm{op}}}{M\|\mathbf{a}\|^2}. \tag{68}$$

The deterministic retention approximation is

$$b_{\mathrm{A}} := \frac{\mathrm{tr}(\mathbf{QCQ})}{M\|\mathbf{a}\|^2}, \qquad \rho_{0,\mathrm{A}} := (1 + b_{\mathrm{A}})^{-1/2}, \tag{69}$$

$$b_{\mathrm{A}} = \frac{1}{M\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}} \left(\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\Sigma}_{\mathrm{val}}) - \frac{\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\Sigma}_{\mathrm{val}}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}}{\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}} \right). \tag{70}$$

Theorem 6 (Gain retention under anisotropic Gaussian noise)**.** *Under Equation* (66)*, let $N \geq 2$, $M \geq 1$, $1 \leq K < N$, and $\hat{\boldsymbol{\mu}} = M^{-1}\sum_{j=1}^{M} \boldsymbol{v}_j$. Let $S_{\boldsymbol{h}}$ select the K largest dot-product scores $\langle \boldsymbol{x}_i, \boldsymbol{h} \rangle$. Define*

$$\Gamma_{N,K}^{\mathrm{A}} := \sqrt{\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}}\,\gamma_{N,K} > 0. \tag{71}$$

Then the following conclusions hold.

1. ***Exact first-order identity.*** *For every fixed $\boldsymbol{h}$ with $\boldsymbol{h}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{h} > 0$,*

$$\rho_1(\boldsymbol{h}) = \frac{\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{h}}{\sqrt{\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}}\sqrt{\boldsymbol{h}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{h}}} = \cos\left\langle \boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}\boldsymbol{h}, \boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}\boldsymbol{\mu} \right\rangle. \tag{72}$$

   In particular, this identity applies to $\boldsymbol{h} = \hat{\boldsymbol{\mu}}$ almost surely over validation sampling.

2. ***Loss-linearization error.*** *If $\varepsilon_T < \infty$ and $2\varepsilon_T < \Gamma_{N,K}^{\mathrm{A}}$, then*

$$|\rho(\boldsymbol{h}) - \rho_1(\boldsymbol{h})| \leq B_{T,\mathrm{A}}, \qquad B_{T,\mathrm{A}} := \frac{4\varepsilon_T}{\Gamma_{N,K}^{\mathrm{A}} - 2\varepsilon_T}. \tag{73}$$

3. ***Finite-validation concentration.*** *For any $\epsilon, \delta \in (0,1)$ satisfying*

$$q_{\mathrm{A}} \leq \frac{\epsilon^2}{16\log(4/\delta)}, \tag{74}$$

   we have

$$\Pr_{\mathrm{val}}\left(\left| \frac{\rho_1(\hat{\boldsymbol{\mu}})}{\rho_{0,\mathrm{A}}} - 1 \right| \leq \epsilon \right) \geq 1 - \delta, \tag{75}$$

$$|\mathbb{E}_{\mathrm{val}}\rho_1(\hat{\boldsymbol{\mu}}) - \rho_{0,\mathrm{A}}| \leq \epsilon\rho_{0,\mathrm{A}} + 2\delta. \tag{76}$$

   If the assumptions in part 2 also hold, then

$$\Pr_{\mathrm{val}}(|\rho(\hat{\boldsymbol{\mu}}) - \rho_{0,\mathrm{A}}| \leq B_{T,\mathrm{A}} + \epsilon\rho_{0,\mathrm{A}}) \geq 1 - \delta, \tag{77}$$

$$|\mathbb{E}_{\mathrm{val}}\rho(\hat{\boldsymbol{\mu}}) - \rho_{0,\mathrm{A}}| \leq B_{T,\mathrm{A}} + \epsilon\rho_{0,\mathrm{A}} + 2\delta. \tag{78}$$

Proof. **Step 1: conditional Gaussian regression.** Fix a reference $\boldsymbol{h}$ with $\boldsymbol{h}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{h} > 0$ and set

$$U_i := \frac{\langle \boldsymbol{x}_i - \boldsymbol{\mu}_t, \boldsymbol{\mu} \rangle}{\sqrt{\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}}}, \qquad V_i := \frac{\langle \boldsymbol{x}_i - \boldsymbol{\mu}_t, \boldsymbol{h} \rangle}{\sqrt{\boldsymbol{h}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{h}}}.$$

The pairs (U_i, V_i) are independent across i and jointly standard Gaussian within each pair, with correlation

$$c_{\mathrm{A}}(\boldsymbol{h}) := \frac{\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{h}}{\sqrt{\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}}\sqrt{\boldsymbol{h}^\mathsf{T}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{h}}}.$$

Thus $\mathbb{E}[U_i \mid V_1, \ldots, V_N] = c_{\mathrm{A}}(\boldsymbol{h})V_i$, including the degenerate cases $c_{\mathrm{A}}(\boldsymbol{h}) = \pm 1$. Because $S_{\boldsymbol{h}}$ selects the largest K values of V_i,

$$\mathbb{E}_{\mathrm{tr}}[\ell_D(S_{\boldsymbol{h}}) - \overline{\ell}_D] = \sqrt{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}}\, c_{\mathrm{A}}(\boldsymbol{h})\gamma_{N,K},$$
$$\mathbb{E}_{\mathrm{tr}}[\max_{S\in\mathcal{S}_K} \ell_D(S) - \overline{\ell}_D] = \sqrt{\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{tr}}\boldsymbol{\mu}}\,\gamma_{N,K} = \Gamma^{\mathrm{A}}_{N,K}.$$

The common term $K\langle \boldsymbol{\mu}_t, \boldsymbol{\mu}\rangle$ cancels against the expected random-selection baseline. Dividing the two excess gains proves Equation (72).

Step 2: loss-linearization error. Write A_F, A_1, D_F for the actual numerator, linearized numerator, and actual oracle denominator. The uniform remainder R_D implies

$$|A_F - A_1| \le 2\varepsilon_T, \qquad |D_F - \Gamma^{\mathrm{A}}_{N,K}| \le 2\varepsilon_T.$$

By Equation (72), $|A_1| \le \Gamma^{\mathrm{A}}_{N,K}$. Hence $D_F \ge \Gamma^{\mathrm{A}}_{N,K} - 2\varepsilon_T > 0$ and

$$\left|\frac{A_F}{D_F} - \frac{A_1}{\Gamma^{\mathrm{A}}_{N,K}}\right| \le \frac{2\varepsilon_T + 2\varepsilon_T|A_1|/\Gamma^{\mathrm{A}}_{N,K}}{\Gamma^{\mathrm{A}}_{N,K} - 2\varepsilon_T} \le B_{T,\mathrm{A}}.$$

Step 3: transformed validation noise. The sample mean has distribution $\hat{\boldsymbol{\mu}} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathrm{val}}/M)$. Therefore,

$$\boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}\hat{\boldsymbol{\mu}} = \mathbf{a} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}/M).$$

Since $\mathbf{a} \neq \mathbf{0}$, this transformed reference is nonzero almost surely: a Gaussian vector with nonzero mean is either deterministic and nonzero, or assigns probability zero to each singleton in its positive-dimensional affine support. Thus Equation (72) applies almost surely. Define

$$U := \frac{\langle \mathbf{u}, \boldsymbol{\xi}\rangle}{\|\mathbf{a}\|}, \qquad W := \frac{\|\mathbf{Q}\boldsymbol{\xi}\|^2}{\|\mathbf{a}\|^2}.$$

Orthogonal decomposition gives the exact representation

$$\rho_1(\hat{\boldsymbol{\mu}}) = \frac{1+U}{\sqrt{(1+U)^2 + W}}, \qquad \mathbb{E}W = b_{\mathrm{A}}, \qquad \mathrm{Var}(U) \le q_{\mathrm{A}}. \tag{79}$$

In general, U and W are not independent; the argument below does not require independence.

Step 4: concentration of the orthogonal noise energy. Let

$$\mathbf{B} := \frac{\mathbf{QCQ}}{M\|\mathbf{a}\|^2}.$$

Diagonalizing this positive semidefinite matrix shows that W has the distribution of $\sum_j \lambda_j Z_j^2$, where λ_j are its eigenvalues and the Z_j are independent standard Gaussian variables. Moreover,

$$\mathrm{tr}(\mathbf{B}) = b_{\mathrm{A}}, \qquad \|\mathbf{B}\|_{\mathrm{op}} \le q_{\mathrm{A}}, \qquad \mathrm{tr}(\mathbf{B}^2) \le q_{\mathrm{A}}b_{\mathrm{A}}.$$

The weighted chi-square inequality of Laurent & Massart (2000, Lemma 1) gives, for $t > 0$,

$$\Pr\Big(|W - b_{\mathrm{A}}| > 2\sqrt{q_{\mathrm{A}}b_{\mathrm{A}}t} + 2q_{\mathrm{A}}t\Big) \le 2e^{-t}.$$

The Gaussian tail bound also gives

$$\Pr\Big(|U| > \sqrt{2q_{\mathrm{A}}t}\Big) \le 2e^{-t}.$$

These statements include the zero-variance cases by continuity or by direct evaluation. Set $t = \log(4/\delta)$. By a union bound, with probability at least $1-\delta$,

$$|U| \le A := \sqrt{2q_{\mathrm{A}}t}, \qquad \frac{|W - b_{\mathrm{A}}|}{1 + b_{\mathrm{A}}} \le B := \sqrt{q_{\mathrm{A}}t} + 2q_{\mathrm{A}}t. \tag{80}$$

For the second inequality we used $2\sqrt{b_{\mathrm{A}}}/(1+b_{\mathrm{A}}) \le 1$.

Step 5: a relative retention bound. Under Equation (74), $A, B < 1/2$. On the event in Equation (80), $1 + U > 0$ and

$$\frac{\rho_1(\hat{\boldsymbol{\mu}})}{\rho_{0,\mathrm{A}}} = \frac{(1+U)\sqrt{1+W}}{\sqrt{(1+U)^2+W}} \sqrt{\frac{1+b_{\mathrm{A}}}{1+W}}.$$

For $W \geq 0$, the first factor lies between 1 and $1 + U$. The second factor lies between $(1 + B)^{-1/2}$ and $(1 - B)^{-1/2}$. It follows that

$$\frac{1-A}{\sqrt{1+B}} \leq \frac{\rho_1(\hat{\boldsymbol{\mu}})}{\rho_{0,\mathrm{A}}} \leq \frac{1+A}{\sqrt{1-B}}.$$

For $0 \leq B \leq 1/2$, $(1 - B)^{-1/2} - 1 \leq B$ and $1 - (1 + B)^{-1/2} \leq B/2$. Consequently,

$$\begin{aligned}\left| \frac{\rho_1(\hat{\boldsymbol{\mu}})}{\rho_{0,\mathrm{A}}} - 1 \right| &\leq \sqrt{2}A + B = 3\sqrt{q_{\mathrm{A}} t} + 2q_{\mathrm{A}} t \\ &\leq \frac{3\epsilon}{4} + \frac{\epsilon^2}{8} \leq \epsilon.\end{aligned}$$

This proves Equation (75).

Step 6: expectation and combined bounds. On the good event the absolute error is at most $\epsilon\rho_{0,\mathrm{A}}$; on its complement it is at most 2, since $\rho_1(\hat{\boldsymbol{\mu}}) \in [-1, 1]$ and $\rho_{0,\mathrm{A}} \in (0, 1]$. Thus

$$|\mathbb{E}_{\mathrm{val}}\rho_1(\hat{\boldsymbol{\mu}}) - \rho_{0,\mathrm{A}}| \leq \epsilon\rho_{0,\mathrm{A}} + 2\delta.$$

Combining this with Equation (73) gives Equation (78) and its probability bound. Finally, $\mathbf{Q}^2 = \mathbf{Q}$ implies

$$\mathrm{tr}(\mathbf{QCQ}) = \mathrm{tr}(\mathbf{C}) - \mathbf{u}^{\mathsf{T}}\mathbf{Cu}.$$

Substituting Equation (67) gives Equation (70). □

Corollary 2 (Anisotropic validation noise with isotropic training updates)**.** *If $\boldsymbol{\Sigma}_{\mathrm{tr}} = \tau^2 \boldsymbol{I}_P$, then*

$$\rho_1(\boldsymbol{h}) = \cos\langle \boldsymbol{h}, \boldsymbol{\mu} \rangle, \qquad q_{\mathrm{A}} = \frac{\|\boldsymbol{\Sigma}_{\mathrm{val}}\|_{\mathrm{op}}}{M\|\boldsymbol{\mu}\|^2}, \tag{81}$$

and

$$\rho_{0,\mathrm{A}} = \left(1 + \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{val}}) - \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Sigma}_{\mathrm{val}}\boldsymbol{\mu}/\|\boldsymbol{\mu}\|^2}{M\|\boldsymbol{\mu}\|^2}\right)^{-1/2}. \tag{82}$$

All bounds in Theorem 6 apply with these substitutions. In particular, if $\boldsymbol{\Sigma}_{\mathrm{val}} = \sigma^2 \boldsymbol{P}$, $\boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{\mu}$, and $\mathrm{rank}(\boldsymbol{P}) = r$, then Equation (82) *reduces to*

$$\rho_{0,\mathrm{A}} = \left(1 + \frac{(r-1)\sigma^2}{M\|\boldsymbol{\mu}\|^2}\right)^{-1/2}, \qquad q_{\mathrm{A}} = \frac{\sigma^2}{M\|\boldsymbol{\mu}\|^2}.$$

Proof. Substituting $\boldsymbol{\Sigma}_{\mathrm{tr}} = \tau^2 \boldsymbol{I}_P$ into Equations (70) and (72) cancels all factors of τ. For the projector specialization, $\mathrm{tr}(\sigma^2 \boldsymbol{P}) = r\sigma^2$ and $\boldsymbol{\mu}^{\mathsf{T}}\sigma^2\boldsymbol{P}\boldsymbol{\mu}/\|\boldsymbol{\mu}\|^2 = \sigma^2$, which proves the claimed reduction. □

Interpretation. Under anisotropic training updates, the relevant reference alignment is measured after transformation by $\boldsymbol{\Sigma}_{\mathrm{tr}}^{1/2}$. The leading attenuation depends on the validation noise energy orthogonal to the transformed population gradient, as quantified by b_{A}. With isotropic training updates, this is exactly the orthogonal validation-noise energy in the original coordinates. The original factor $(r - 1)\sigma^2$ is a special case of this projected covariance trace. No eigenvector relationship between $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_{\mathrm{val}}$ is required.

B.4 Gain Retention Rate for Clustering

In this section, we present the formal theorem and full proof for retention gain analysis with clustering, corresponding to Section 3.2.

B.4.1 MODEL AND SELECTION PROTOCOL

Fix a checkpoint θ. The training updates are independently drawn as

$$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_t, \tau^2 \boldsymbol{I}_P), \quad \tau > 0.$$

Independently, the validation gradients follow a Gaussian mixture with latent labels:

$$\Pr(L_j = \ell) = \pi_\ell, \quad \boldsymbol{v}_j \mid L_j = \ell \sim \mathcal{N}(\boldsymbol{\mu}_\ell, \sigma_\ell^2 \boldsymbol{P}_\ell),$$

where the pairs $(L_j, \boldsymbol{v}_j)$ are independent, $\pi_\ell > 0$, and $\sum_{\ell=1}^{c} \pi_\ell = 1$. For each component, assume $\boldsymbol{\mu}_\ell \neq 0$ and

$$\boldsymbol{P}_\ell^\top = \boldsymbol{P}_\ell, \quad \boldsymbol{P}_\ell^2 = \boldsymbol{P}_\ell, \quad \operatorname{rank}(\boldsymbol{P}_\ell) = r_\ell, \quad \boldsymbol{P}_\ell \boldsymbol{\mu}_\ell = \boldsymbol{\mu}_\ell.$$

The population validation gradient is

$$\boldsymbol{\mu} := \nabla_\theta \mathcal{L}_{\text{val}}(\theta) = \sum_{\ell=1}^{c} \pi_\ell \boldsymbol{\mu}_\ell \neq 0.$$

We analyze the latent component partition. Condition on positive component counts $\mathbf{m} = (m_1, \ldots, m_c)$, where $m_\ell = |\{j : L_j = \ell\}|$, and define

$$\hat{\boldsymbol{\mu}}_\ell := \frac{1}{m_\ell} \sum_{j : L_j = \ell} \boldsymbol{v}_j.$$

Let $1 \leq K < N$ and assume the proportional quotas $K_\ell = K m_\ell / M$ are integers. For each component, select

$$T_\ell := \operatorname{Top}_{K_\ell}\{\langle \boldsymbol{x}_i, \hat{\boldsymbol{\mu}}_\ell \rangle\}_{i=1}^{N}.$$

Selections are distinct within each T_ℓ, but overlap across components is permitted and counted with multiplicity. Writing

$$n_i^c := \sum_{\ell=1}^{c} \mathbf{1}\{i \in T_\ell\},$$

the selected update is $\sum_{i=1}^{N} n_i^c \boldsymbol{x}_i$, with $\sum_i n_i^c = K$. When the sets $T_1, \ldots, T_c$ are pairwise disjoint, every $n_i^c \in \{0, 1\}$, so $\sum_i n_i^c \boldsymbol{x}_i$ is the update of a size-K subset and Equation (83) is the retention of that subset. When they overlap, the same sample is credited more than once and $\rho_1(\mathcal{A}_c)$ is not bounded by 1. Disjointness holds when the cluster directions $\hat{\boldsymbol{\mu}}_\ell$ are well separated and K/N is small, and SUADS enforces it directly (Section 4).

Loss reduction and normalization. For a training pool $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ and a multiplicity vector

$$\mathbf{n} \in \mathcal{W}_K := \left\{ \mathbf{n} \in \mathbb{Z}_{\geq 0}^N : \sum_{i=1}^{N} n_i = K \right\},$$

define the actual and first-order loss reductions by

$$F_X(\mathbf{n}) := \mathcal{L}_{\text{val}}(\theta) - \mathcal{L}_{\text{val}}\left(\theta - \sum_{i=1}^{N} n_i \boldsymbol{x}_i \right), \quad G_X(\mathbf{n}) := \sum_{i=1}^{N} n_i \langle \boldsymbol{x}_i, \boldsymbol{\mu} \rangle.$$

For clusterwise selection, the selected loss reduction is $F_X(\mathbf{n}^c)$. Both the random baseline and the oracle remain restricted to subsets of K distinct training examples, as in Definition 1. Let ρ_1 denote the corresponding quantity obtained by replacing F_X with G_X throughout.

Now we are ready to present the formal version of Theorem 2 in the main text.

Theorem 7 (Finite-Sample Gain Retention for Clustering)**.** *Under the preceding model and protocol, define*

$$\gamma_{N,q} := \mathbb{E}\left[\sum_{j=N-q+1}^{N} Z_{(j)} \right], \quad Z_1, \ldots, Z_N \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1),$$

where $Z_{(1)} \le \cdots \le Z_{(N)}$. *Write*

$$a_\ell := \frac{\gamma_{N,K_\ell}}{\gamma_{N,K}}, \quad q_\ell := \frac{\sigma_\ell^2}{m_\ell \|\boldsymbol{\mu}_\ell\|^2}, \quad \Lambda_c := \sum_{\ell=1}^{c} a_\ell,$$

and define

$$R_c(\mathbf{m}) := \sum_{\ell=1}^{c} a_\ell \frac{\langle \boldsymbol{\mu}_\ell, \boldsymbol{\mu} \rangle}{\|\boldsymbol{\mu}\| \sqrt{\|\boldsymbol{\mu}_\ell\|^2 + (r_\ell - 1)\sigma_\ell^2/m_\ell}}.$$

First, conditional on the validation data, the exact first-order identity is

$$\rho_1(\mathcal{A}_c) = \sum_{\ell=1}^{c} a_\ell \cos\langle \hat{\boldsymbol{\mu}}_\ell, \boldsymbol{\mu} \rangle. \tag{83}$$

Moreover,

$$|\mathbb{E}_{\mathrm{val}}[\rho_1(\mathcal{A}_c) \mid \mathbf{m}] - R_c(\mathbf{m})| \le \mathcal{E}_{\mathrm{G}}, \quad \mathcal{E}_{\mathrm{G}} := 2\sqrt{3} \sum_{\ell=1}^{c} a_\ell \sqrt{\frac{q_\ell}{1 + (r_\ell - 1)q_\ell}}. \tag{84}$$

To control loss linearization, let

$$\varepsilon_{\mathrm{T}} := \mathbb{E}_X \max_{\mathbf{n} \in \mathcal{W}_K} |F_X(\mathbf{n}) - G_X(\mathbf{n})|, \quad \Gamma_{N,K} := \tau \|\boldsymbol{\mu}\| \gamma_{N,K}.$$

If $2\varepsilon_{\mathrm{T}} < \Gamma_{N,K}$, *then*

$$|\rho(\mathcal{A}_c) - \rho_1(\mathcal{A}_c)| \le \mathcal{E}_{\mathrm{T}}, \quad \mathcal{E}_{\mathrm{T}} := \frac{2(1 + \Lambda_c)\varepsilon_{\mathrm{T}}}{\Gamma_{N,K} - 2\varepsilon_{\mathrm{T}}}. \tag{85}$$

Consequently,

$$|\mathbb{E}_{\mathrm{val}}[\rho(\mathcal{A}_c) \mid \mathbf{m}] - R_c(\mathbf{m})| \le \mathcal{E}_{\mathrm{T}} + \mathcal{E}_{\mathrm{G}}. \tag{86}$$

Proof. **Step 1: Exact first-order gain conditional on validation.** Fix the validation data and a nonzero reference $\boldsymbol{h}$. Define

$$U_i := \frac{\langle \boldsymbol{x}_i - \boldsymbol{\mu}_t, \boldsymbol{h} \rangle}{\tau \|\boldsymbol{h}\|}.$$

The variables $U_1, \ldots, U_N$ are independent standard Gaussians, and ranking the original alignment scores is equivalent to ranking the U_i. Gaussian regression gives

$$\mathbb{E}[\boldsymbol{x}_i - \boldsymbol{\mu}_t \mid U_1, \ldots, U_N] = \tau \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|} U_i.$$

Hence, if $T_{\boldsymbol{h}}$ consists of the top q examples according to $\boldsymbol{h}$,

$$\mathbb{E}_X \sum_{i \in T_{\boldsymbol{h}}} (\boldsymbol{x}_i - \boldsymbol{\mu}_t) = \tau \gamma_{N,q} \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}.$$

A uniformly selected size-K subset has expected first-order gain $K\langle \boldsymbol{\mu}_t, \boldsymbol{\mu} \rangle$. The population first-order oracle selects the top K examples according to $\boldsymbol{\mu}$, so its expected excess gain over this baseline is

$$\Gamma_{N,K} = \tau \|\boldsymbol{\mu}\| \gamma_{N,K} > 0.$$

Applying the regression identity to each T_ℓ, the clusterwise selector's expected excess gain is

$$\tau \sum_{\ell=1}^{c} \gamma_{N,K_\ell} \frac{\langle \hat{\boldsymbol{\mu}}_\ell, \boldsymbol{\mu} \rangle}{\|\hat{\boldsymbol{\mu}}_\ell\|}.$$

Dividing by $\Gamma_{N,K}$ proves Equation (83). Overlap requires no correction because repeated selections are counted with multiplicity.

Step 2: Averaging over validation-centroid noise. Conditional on the component counts,

$$\hat{\boldsymbol{\mu}}_\ell \sim \mathcal{N}\left(\boldsymbol{\mu}_\ell, \frac{\sigma_\ell^2}{m_\ell}\boldsymbol{P}_\ell\right).$$

Choose orthonormal coordinates in the range of $\boldsymbol{P}_\ell$, with the first coordinate along $\boldsymbol{\mu}_\ell$. Rotational symmetry in the orthogonal coordinates implies

$$\mathbb{E}_{\mathrm{val}}\left[\frac{\hat{\boldsymbol{\mu}}_\ell}{\|\hat{\boldsymbol{\mu}}_\ell\|}\,\middle|\,\mathbf{m}\right] = \Psi_{r_\ell}(q_\ell)\frac{\boldsymbol{\mu}_\ell}{\|\boldsymbol{\mu}_\ell\|},$$

where

$$\Psi_r(q) := \mathbb{E}\frac{1+\sqrt{q}Z}{\sqrt{(1+\sqrt{q}Z)^2+qQ}}, \quad Z\sim\mathcal{N}(0,1), \quad Q\sim\chi^2_{r-1}, \quad Z\perp Q.$$

For $r=1$, interpret Q as the constant zero. Writing

$$\alpha_\ell := \frac{\langle\boldsymbol{\mu}_\ell,\boldsymbol{\mu}\rangle}{\|\boldsymbol{\mu}_\ell\|\|\boldsymbol{\mu}\|},$$

we obtain the exact representation

$$\mathbb{E}_{\mathrm{val}}[\rho_1(\mathcal{A}_c)\mid\mathbf{m}] = \sum_{\ell=1}^{c} a_\ell\alpha_\ell\Psi_{r_\ell}(q_\ell). \tag{87}$$

Step 3: Explicit approximation of $\Psi_r(q)$. Set $d=r-1$ and introduce

$$\mathbf{a} = \begin{pmatrix}1+\sqrt{q}Z\\ \sqrt{qQ}\end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix}1\\ \sqrt{qd}\end{pmatrix}.$$

Their normalized first coordinates are the integrand defining $\Psi_r(q)$ and $(1+dq)^{-1/2}$. For nonzero vectors,

$$\left\|\frac{\mathbf{a}}{\|\mathbf{a}\|}-\frac{\mathbf{b}}{\|\mathbf{b}\|}\right\| \le \frac{2\|\mathbf{a}-\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Taking expectations and applying Cauchy–Schwarz yields

$$\left|\Psi_r(q)-(1+dq)^{-1/2}\right| \le \frac{2\sqrt{\mathbb{E}\|\mathbf{a}-\mathbf{b}\|^2}}{\sqrt{1+dq}}.$$

If $d>0$, then

$$\mathbb{E}(\sqrt{Q}-\sqrt{d})^2 = \mathbb{E}\frac{(Q-d)^2}{(\sqrt{Q}+\sqrt{d})^2} \le \frac{\mathbb{E}(Q-d)^2}{d} = 2.$$

When $d=0$, the same expectation is zero. Thus

$$\mathbb{E}\|\mathbf{a}-\mathbf{b}\|^2 = q\left(\mathbb{E}Z^2+\mathbb{E}(\sqrt{Q}-\sqrt{d})^2\right) \le 3q,$$

and therefore

$$\left|\Psi_r(q)-(1+(r-1)q)^{-1/2}\right| \le 2\sqrt{3}\sqrt{\frac{q}{1+(r-1)q}}. \tag{88}$$

Since

$$R_c(\mathbf{m}) = \sum_{\ell=1}^{c} a_\ell\alpha_\ell(1+(r_\ell-1)q_\ell)^{-1/2},$$

subtracting this expression from Equation (87), applying Equation (88), and using $|\alpha_\ell|\le 1$ proves Equation (84).

Step 4: Controlling loss linearization. Define

$$T_X := \max_{\mathbf{n}\in\mathcal{W}_K}|F_X(\mathbf{n})-G_X(\mathbf{n})|, \quad \mathbb{E}_X T_X = \varepsilon_{\mathrm{T}}.$$

This class contains the selected multiset and every size-K subset used by the baseline and oracle. Consequently, the selected gain and random baseline each change by at most T_X under linearization. Also,

$$\left| \max_{|T|=K} F_X(\mathbf{1}_T) - \max_{|T|=K} G_X(\mathbf{1}_T) \right| \le T_X.$$

Let A and B denote the numerator and denominator of the actual retention, after expectation over X. Let A_{lin} be its linearized numerator. The linearized denominator is $\Gamma_{N,K}$. The preceding inequalities give

$$|A - A_{\mathrm{lin}}| \le 2\varepsilon_{\mathrm{T}}, \quad |B - \Gamma_{N,K}| \le 2\varepsilon_{\mathrm{T}}.$$

Furthermore, Equation (83) implies

$$|A_{\mathrm{lin}}| \le \Gamma_{N,K}\Lambda_c.$$

If $2\varepsilon_{\mathrm{T}} < \Gamma_{N,K}$, then $B \ge \Gamma_{N,K} - 2\varepsilon_{\mathrm{T}} > 0$, and

$$\begin{aligned}\left| \frac{A}{B} - \frac{A_{\mathrm{lin}}}{\Gamma_{N,K}} \right| &\le \frac{|A - A_{\mathrm{lin}}|}{B} + \frac{|A_{\mathrm{lin}}||B - \Gamma_{N,K}|}{B\Gamma_{N,K}} \\ &\le \frac{2(1+\Lambda_c)\varepsilon_{\mathrm{T}}}{\Gamma_{N,K} - 2\varepsilon_{\mathrm{T}}}.\end{aligned}$$

This proves Equation (85). Taking conditional expectations over validation data and combining with Equation (84) proves Equation (86). □

B.5 Derivation of Equation (6)

We derive Equation (6) as a qualitative approximation by combining the gain-retention formulas with an additional approximation relating gradient coherence to noise strength. The resulting expression is not an exact identity for an arbitrary Gaussian mixture.

Simplifying assumptions. We consider c equally sized clusters, with $m = M/c$ validation examples and selection quota K/c per cluster, assuming integer sizes and quotas. Cross-cluster repetitions are counted with multiplicity, while the original size-K subset oracle is retained.

We additionally adopt the aligned-reference approximation $\|\boldsymbol{\mu}_\ell\| \approx \|\boldsymbol{\mu}\|$, and assume a common within-cluster noise variance σ_ℓ^2 and noise rank r. These are additional simplifying assumptions: the identity $\boldsymbol{\mu} = \sum_\ell \pi_\ell \boldsymbol{\mu}_\ell$ alone does not imply that the component means are close. Let σ^2 denote the effective global noise variance in the non-clustered Gaussian-reference approximation.

Step 1: Comparing the retention formulas. Define the dimensionless orthogonal noise levels

$$b_{\mathrm{g}} := \frac{(r-1)\sigma^2}{\|\boldsymbol{\mu}\|^2}, \quad b_\ell := \frac{(r-1)\sigma_\ell^2}{\|\boldsymbol{\mu}\|^2}.$$

Under the loss-linearization and centroid-direction approximations used in Equations (3) and (4),

$$\mathbb{E}\rho(\mathcal{A}) \approx \left(1 + \frac{b_{\mathrm{g}}}{M}\right)^{-1/2},$$

whereas balanced clusterwise selection gives

$$\mathbb{E}\rho(\mathcal{A}_c) \approx \kappa_c \left(1 + \frac{c\, b_\ell}{M}\right)^{-1/2}, \quad \kappa_c := \frac{c\,\gamma_{N,K/c}}{\gamma_{N,K}}.$$

Consequently, the ratio of expected gain retentions satisfies

$$R := \frac{\mathbb{E}\rho(\mathcal{A}_c)}{\mathbb{E}\rho(\mathcal{A})} \approx \kappa_c \sqrt{\frac{M + b_{\mathrm{g}}}{M + c\, b_\ell}}. \tag{89}$$

The factor c in the denominator accounts for each cluster centroid using only M/c validation examples.

Step 2: Approximating the order-statistic factor. For N independent standard Gaussian variables, the expected sum of the largest q values satisfies

$$\gamma_{N,q} \sim q\sqrt{2\log(N/q)}$$

as $N \to \infty$ with $q/N \to 0$. Applying this leading-order approximation to $q = K/c$ and $q = K$ yields

$$\kappa_c \approx \frac{c(K/c)\sqrt{2\log(cN/K)}}{K\sqrt{2\log(N/K)}} = \left(1 + \frac{\log c}{\log(N/K)}\right)^{1/2}. \tag{90}$$

Step 3: Relating coherence to noise strength. We interpret pairwise coherence using independent gradient draws. Its empirical estimator should therefore use distinct sample pairs, excluding self-pairs. For independent gradients $\boldsymbol{v}, \boldsymbol{v}'$ from the same reference distribution, let

$$C := \mathbb{E}\left[\frac{\langle \boldsymbol{v}, \boldsymbol{v}'\rangle}{\|\boldsymbol{v}\|\|\boldsymbol{v}'\|}\right].$$

Independence gives the exact identity

$$C = \left\|\mathbb{E}\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}\right\|^2.$$

To obtain a scalar noise proxy, we additionally approximate the mean normalized gradient by

$$\mathbb{E}\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|} \approx \frac{\mathbb{E}\boldsymbol{v}}{\sqrt{\mathbb{E}\|\boldsymbol{v}\|^2}}.$$

This requires a sufficiently accurate approximation of the random gradient norms; it does not follow from independence alone.

For the isotropic Gaussian-reference model $\boldsymbol{v} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\boldsymbol{P})$, with $\mathrm{rank}(\boldsymbol{P}) = r$, this gives

$$C \approx \frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + r\sigma^2}.$$

Thus,

$$\frac{(r-1)\sigma^2}{\|\boldsymbol{\mu}\|^2} \approx \eta(C^{-1} - 1), \quad \eta := \frac{r-1}{r}.$$

Applying this calibration to the effective global and within-cluster reference models yields

$$b_{\mathrm{g}} \approx \eta({\mathrm{Sim_g}}^{-1} - 1), \quad b_\ell \approx \eta({\mathrm{Sim_c}}^{-1} - 1).$$

Here $\mathrm{Sim_c}$ represents the common within-cluster coherence under the equal-component-statistics approximation. In general, averaging unequal cluster coherences before applying this nonlinear conversion is an additional approximation.

Substituting into Equation (89) gives

$$R \approx \kappa_c\sqrt{\frac{M + \eta({\mathrm{Sim_g}}^{-1} - 1)}{M + c\eta({\mathrm{Sim_c}}^{-1} - 1)}}. \tag{91}$$

For large effective rank, $\eta \approx 1$, and hence

$$R \approx \kappa_c\sqrt{\frac{M - 1 + {\mathrm{Sim_g}}^{-1}}{M - c + c{\mathrm{Sim_c}}^{-1}}} = \kappa_c\sqrt{\frac{(M-1)\mathrm{Sim_g}\mathrm{Sim_c} + \mathrm{Sim_c}}{(M-c)\mathrm{Sim_g}\mathrm{Sim_c} + c\mathrm{Sim_g}}}. \tag{92}$$

Step 4: Recovering the expression in the main text. When $M \gg c$, we further approximate $M - 1$ and $M - c$ by M in Equation (92). For $0 < \mathrm{Sim_g}, \mathrm{Sim_c} \le 1$, this replacement changes the ratio by a relative factor $1 + O(c/M)$. Combining it with Equation (90) gives

$$R \approx \sqrt{\left(1 + \frac{\log c}{\log(N/K)}\right)\frac{M\mathrm{Sim_g}\mathrm{Sim_c} + \mathrm{Sim_c}}{M\mathrm{Sim_g}\mathrm{Sim_c} + c\mathrm{Sim_g}}}, \tag{93}$$

which is Equation (6).

B.6 Empirical Scale of the Retention Factors and Scope of Equation (6)

This section calibrates the retention factors of Section 3 on our validation gradients and refines their reading. We use the projected, ℓ_2-normalized validation gradients of the final warmup checkpoint for Gemma-2-9B on Magicoder ($M = 538$, $d = 8192$) and for Qwen3.5-4B on DS_RSTAR ($M = 144$, $d = 8192$), i.e., exactly the vectors SUADS clusters.

From coherence to reference reliability. For unit vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m$ with mean pairwise cosine κ, the squared norm of their mean is $R^2 = 1/m + (1 - 1/m)\kappa$ (Appendix C.2), so $\|\boldsymbol{\mu}\|^2 \approx \kappa$ and the total noise energy per example is $1 - \kappa$. Substituting into Equation (3), a reference built from m examples with coherence κ retains

$$\rho(m, \kappa) \;\approx\; \Big(1 + \frac{1 - \kappa}{m\,\kappa}\Big)^{-1/2} \tag{94}$$

of the first-order gain of the population reference.[1] This depends on m and κ only: neither the identity nor the error of the mean, $\mathbb{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = (1 - \kappa)/m$, involves the dimension, and the latter holds for any i.i.d. distribution of unit vectors.

Measured reliability. Table 3 compares Equation (94) with direct measurements that use no model. For $m = 1$ we report the mean cosine between a validation gradient and the leave-one-out mean of the others. For larger m we report the cosine between the means of two random halves of the set (2,000 splits), which estimates the reliability of a reference of size $m/2$ and is compared with $(1 + (1 - \kappa)/((m/2)\kappa))^{-1}$.

Table 3: Reliability of references built from m validation examples, measured against Equation (94), on the Gemma-2-9B validation gradients for Magicoder and the Qwen3.5-4B validation gradients for DS_RSTAR, both at the final warmup checkpoint. Clusters are from spherical k-means with five restarts on the same gradients; on Magicoder the two-cluster split coincides with the MBPP/EvoEval source split of Appendix E.3 ($\kappa = 0.551/0.428$), and on DS_RSTAR the 13-member cluster is cluster A1 of Figure 5 (centroid cosine -0.64 to A0).

Setting	Reference	m	κ	measured	Equation (94)
Magicoder	single example (leave-one-out)	1	0.470	0.685	0.686
Magicoder	cluster (half-split)	35	0.347	0.905	0.903
Magicoder	cluster (half-split)	47	0.457	0.954	0.952
Magicoder	EvoEval cluster (half-split)	164	0.428	0.984	0.984
Magicoder	MBPP cluster (half-split)	374	0.551	0.996	0.996
Magicoder	global mean (half-split)	538	0.470	0.996	0.996
DS_RSTAR	single example (leave-one-out)	1	0.296	0.542	0.544
DS_RSTAR	incoherent cluster (half-split)	25	0.084	0.532	0.534
DS_RSTAR	cluster A1 (half-split)	13	0.338	0.788	0.768
DS_RSTAR	cluster (half-split)	22	0.390	0.879	0.876
DS_RSTAR	cluster A0 (half-split)	84	0.773	0.994	0.993
DS_RSTAR	global mean (half-split)	144	0.296	0.976	0.968

The measurements follow Equation (94) to within 0.01 at every size except the 13-member cluster (0.02), where each half has only six or seven members. Two readings follow. First, the attenuation of Theorem 1 is material for references built from few examples: a single validation gradient retains 0.54–0.69 of the achievable gain, which is the operating point of methods that score candidates against individual validation examples (GIST, BIDS), and a small or incoherent cluster is not much better (0.53 for 25 examples at $\kappa = 0.08$). Second, it is small for the global mean: 0.996 on Magicoder and 0.976 on DS_RSTAR from half the data, and by Equation (94) 0.999 and 0.992 for the full set. The gate of Section 4 is the part of SUADS that acts on this factor: it admits only clusters whose size and coherence place them in the reliable regime.

[1] Theorem 1 uses only the noise orthogonal to $\boldsymbol{\mu}$, $(r - 1)\sigma^2$; the total energy $1 - \kappa$ is an upper bound on it, so Equation (94) is slightly conservative.

Scope of Equations (4) to (6). Theorem 2 and Equation (6) concern $\mathcal{A}_c$, which counts coincident selections with multiplicity, so $\sum_\ell \gamma_{N,K_\ell}/\gamma_{N,K} > 1$ (by concavity of $q \mapsto \gamma_{N,q}$) and $\rho(\mathcal{A}_c)$ may exceed 1. Under single assignment, as in SUADS, the selected set is a size-K subset and, by Definition 1, cannot retain more of the reduction of $\mathcal{L}_{\text{val}}$ than the population direction $\boldsymbol{\mu}$, which is the oracle for that average and which the global reference already estimates to retention above 0.97 at our M (Table 3). Equations (4) to (6) should therefore be read as characterizing the reliability of local references, i.e., the conditions (coherent, compact clusters; small c) under which $\rho(m_\ell, \kappa_\ell)$ is close to one, rather than as a gain in ρ. The benefit of local references appears on individual components, which we quantify next.

Per-component retention. Let $\mathcal{L}_\ell$ be the loss on component ℓ of the mixture in Theorem 7, so $\nabla_\theta \mathcal{L}_\ell = \boldsymbol{\mu}_\ell$ and $\mathcal{L}_{\text{val}} = \sum_\ell \pi_\ell \mathcal{L}_\ell$. For a selection S define $G_\ell(S) := \sum_{i\in S}\langle \boldsymbol{x}_i, \boldsymbol{\mu}_\ell\rangle$, and let ρ_ℓ be the first-order retention of Definition 1 with G_ℓ in place of ℓ_D: the expected excess of $G_\ell(S)$ over the uniform size-K baseline, divided by the expected excess of $\max_{S\in\mathcal{S}_K} G_\ell(S)$.

Theorem 8 (Per-component retention of a reference)**.** *Under the model of Appendix B.4, let $\boldsymbol{h} \neq \mathbf{0}$ be fixed given the validation data and $S_{\boldsymbol{h}}$ the indices of the K largest $\langle \boldsymbol{x}_i, \boldsymbol{h}\rangle$. Then:*

(i) $\rho_\ell(S_{\boldsymbol{h}}) = \cos\langle \boldsymbol{h}, \boldsymbol{\mu}_\ell\rangle$ *for every ℓ.*

(ii) For the global reference $\boldsymbol{h} = \hat{\boldsymbol{\mu}}$, $\mathbb{E}_{\text{val}}\rho_\ell = \mathbb{E}_{\text{val}} \cos\langle \hat{\boldsymbol{\mu}}, \boldsymbol{\mu}_\ell\rangle \to \cos\langle \boldsymbol{\mu}, \boldsymbol{\mu}_\ell\rangle$ as $M \to \infty$. In particular a component with $\langle \boldsymbol{\mu}, \boldsymbol{\mu}_\ell\rangle \leq 0$ receives no first-order gain over random selection from the global reference, for any validation-set size.

(iii) For the clusterwise selection of Theorem 7, $\rho_\ell = \sum_j (\gamma_{N,K_j}/\gamma_{N,K}) \cos\langle \hat{\boldsymbol{\mu}}_j, \boldsymbol{\mu}_\ell\rangle$, and the own-quota term satisfies $\mathbb{E}_{\text{val}}[\cos\langle \hat{\boldsymbol{\mu}}_\ell, \boldsymbol{\mu}_\ell\rangle \mid \mathbf{m}] = \Psi_{r_\ell}(q_\ell) \approx (1 + (r_\ell - 1)q_\ell)^{-1/2}$, i.e., Theorem 1 at $m = m_\ell$.

Proof. Step 1 of the proof of Theorem 7 gives, for any fixed $\boldsymbol{h}$, $\mathbb{E}_{\text{tr}} \sum_{i\in S_{\boldsymbol{h}}}(\boldsymbol{x}_i - \boldsymbol{\mu}_t) = \tau\gamma_{N,K}\, \boldsymbol{h}/\|\boldsymbol{h}\|$. Taking the inner product with $\boldsymbol{\mu}_\ell$, the expected excess of $G_\ell(S_{\boldsymbol{h}})$ over the baseline $K\langle \boldsymbol{\mu}_t, \boldsymbol{\mu}_\ell\rangle$ is $\tau\gamma_{N,K}\|\boldsymbol{\mu}_\ell\| \cos\langle \boldsymbol{h}, \boldsymbol{\mu}_\ell\rangle$; the same identity at $\boldsymbol{h} = \boldsymbol{\mu}_\ell$ gives the component-ℓ oracle $\tau\gamma_{N,K}\|\boldsymbol{\mu}_\ell\|$. Dividing proves (i). Independence of the training and validation data lets (i) hold conditionally on $\boldsymbol{h} = \hat{\boldsymbol{\mu}}$, giving the first claim of (ii); the limit follows from $\hat{\boldsymbol{\mu}} \to \boldsymbol{\mu}$ almost surely and continuity of the cosine at $\boldsymbol{\mu} \neq \mathbf{0}$. For (iii), apply the identity to each T_ℓ with K_ℓ in place of K and sum, as in Step 1 of that proof; the conditional expectation is Step 2 of the same proof, $\mathbb{E}_{\text{val}}[\hat{\boldsymbol{\mu}}_\ell/\|\hat{\boldsymbol{\mu}}_\ell\| \mid \mathbf{m}] = \Psi_{r_\ell}(q_\ell)\, \boldsymbol{\mu}_\ell/\|\boldsymbol{\mu}_\ell\|$, followed by the inner product with $\boldsymbol{\mu}_\ell/\|\boldsymbol{\mu}_\ell\|$, and the approximation is Equation (88). □

Corollary 3 (When a local reference is worth admitting)**.** *A local reference improves on the global reference on component ℓ if and only if $\Psi_{r_\ell}(q_\ell) > \cos\langle \boldsymbol{\mu}, \boldsymbol{\mu}_\ell\rangle$: its reliability (Theorem 1 at m_ℓ) must exceed the alignment of the global reference with the component. A large coherent cluster satisfies the left side; a component far from the global mean fails the right side; a small or incoherent cluster fails the left, which motivates the coherence gate and the exclusion of singletons.*

Measured per-component alignment. Table 4 reports $\cos\langle \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_\ell\rangle$, the empirical value of the global reference's retention on each component by Theorem 8(ii), together with the local reference's own-quota retention from Table 3. It also reports the heterogeneity index $x := \text{Sim}_\text{c}/\text{Sim}_\text{g} - 1$, which by the variance decomposition $\sum_\ell \pi_\ell\|\boldsymbol{\mu}_\ell\|^2 = \|\boldsymbol{\mu}\|^2 + \sum_\ell \pi_\ell\|\boldsymbol{\mu}_\ell - \boldsymbol{\mu}\|^2$ equals the between-component spread relative to $\|\boldsymbol{\mu}\|^2$; for equal component norms, $\sum_\ell \pi_\ell \cos\langle \boldsymbol{\mu}, \boldsymbol{\mu}_\ell\rangle = (1 + x)^{-1/2}$. This is the calibrated reading of the ratio $\text{Sim}_\text{c}/\text{Sim}_\text{g}$ in Figure 2: it measures how much better local centroids summarize the validation gradients than the global mean, not a reduction of the sampling noise of the mean.

On DS_RSTAR the global reference has cosine -0.52 with the admitted cluster A1: by Theorem 8, global top-K selection reduces that component's loss by less than random selection at first order, whereas SUADS's local reference retains 0.93 of its gain and its quota reserves 13.4% of the budget for it (Appendix E.3). On Magicoder every component lies within cosine 0.86 of the global reference (at $k \leq 4$), so the first-order analysis predicts only small differences between references there.

Table 4: Per-component retention of the global reference, $\cos\langle\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}_\ell\rangle$, versus the own-quota retention of the local reference, for the clusters of Table 3 (Gemma-2-9B gradients on Magicoder; Qwen3.5-4B gradients on DS_RSTAR).

Setting	Component	m_ℓ	global ref.	local ref.
Magicoder ($x \approx 0.09$)	MBPP	374	0.984	0.999
Magicoder	EvoEval	164	0.888	0.996
DS_RSTAR ($x \approx 0.88$)	A0	84	0.969	0.998
DS_RSTAR	A1	13	$\mathbf{-0.521}$	0.932

What the analysis does and does not predict. It predicts that per-example references are attenuated by $\sqrt{\kappa}$; that the global reference is reliable at M in the hundreds; that a component pointing away from the global mean receives nothing from global selection however large M; that the benefit of local references is therefore a coverage effect, which should not shrink as M grows; and that too many clusters hurt through reliability once clusters become small or incoherent. It does not predict the size of the gain on the validation average, nor the size of the gap between $c = 1$ and $c = 2$ in Table 2, nor $c = 1$ falling below random; on Magicoder in particular, where $x \approx 0.09$, the gap in Table 2 lies outside the first-order model (Appendix F).

C Method and Implementation Details

C.1 Algorithm Pseudo-code for SUADS

Algorithm 1 summarizes the procedure of Section 4, with at most $c_{\max}$ candidate clusters ($c_{\max} = 4$, $R = 5$, $n_{\text{null}} = 25$, $\alpha = 0.05$ in all experiments). Table 2 sweeps $c_{\max} \in \{1, 4, 8, 12, 16\}$ with the same gate, which admitted $c = 1, 2, 3, 6, 7$ clusters. Everything after gradient collection is deterministic given a seed and uses no labels or downstream evaluation. Beyond gradient collection, scoring costs $O(Ndc)$ for $c \le c_{\max}$ admitted clusters, and clustering and the null calibration touch only the M validation gradients.

Algorithm 1 Coherence-gated stratified selection

Require: pool D_{train}, validation gradients $\{\boldsymbol{v}_j\}_{j=1}^{M}$, budget K, at most $c_{\max}$ candidate clusters, null level α, restarts R

1: $\{\boldsymbol{g}_i\}, \{\boldsymbol{v}_j\} \leftarrow$ projected, ℓ_2-normalized features
2: $\kappa^\star \leftarrow (1-\alpha)$-quantile of coherences from the same clustering run on synthetic structureless pools matched to $\{\boldsymbol{v}_j\}$
3: $\{V_\ell\} \leftarrow \textsc{SphericalKMeans}(\{\boldsymbol{v}_j\}, c_{\max})$, $\mathcal{A} \leftarrow \{\ell : m_\ell \ge 2,\ \kappa_\ell > \kappa^\star\}$, or the whole set if empty // global top-K
4: $w_\ell \leftarrow (1/M + (1-1/M)\kappa_\ell)^{1/2}$, $s(z_i) \leftarrow \max_\ell w_\ell \langle \boldsymbol{g}_i, \boldsymbol{u}_\ell\rangle$, $a(z_i) \leftarrow \arg\max_\ell w_\ell \langle \boldsymbol{g}_i, \boldsymbol{u}_\ell\rangle$
5: $\{K_\ell\} \leftarrow$ largest-remainder allocation of K with targets $Km_\ell / \sum_{\ell'} m_{\ell'}$, capped at $|D_\ell|$, overflow redistributed
6: **return** $S = \bigcup_{\ell \in \mathcal{A}} \textsc{Top}(K_\ell, \{z_i : a(z_i) = \ell\}, s)$

C.2 Clustering

Clustering. We partition the validation gradients by spherical k-means. Assignment is by maximum inner product and each center is the normalized mean of its members, so the objective is the total resultant length that scoring later exploits. Initialization is cosine k-means++. A single initialization frequently misses a real but borderline cluster, so we keep the best of $R = 5$ restarts by that objective. The number of candidate clusters is clamped to $\min(c_{\max}, M)$, and clusters of size one are discarded, having no pairs and hence no measurable coherence.

Size-free coherence. For m unit vectors with mean pairwise cosine κ, the squared norm of their centroid is $R^2 = 1/m + (1 - 1/m)\kappa$, since the m self-pairs each contribute 1 and the $m(m-1)$

distinct pairs contribute their cosines. The $1/m$ term makes R itself hard to compare across clusters of different sizes. At $m = 2$ it contributes $1/2$ whatever the two members look like, so an unrelated pair already reports $R \approx 0.7$. Inverting the identity, $\kappa_\ell = (R_\ell^2 - 1/m_\ell)/(1 - 1/m_\ell)$ recovers the mean pairwise cosine exactly, at the compute cost of a centroid rather than cosine similarity of all pairs.

Null calibration. The threshold $\kappa^\star$ is calibrated against what the clustering itself manufactures. We synthesise n_{null} structureless pools matched to the validation set in size, dimension, and background coherence. Each pool consists of unit vectors built as a shared direction plus isotropic noise, mixed so that their expected pairwise cosine equals κ_{pool}. We run the identical clustering on each, with the same number of restarts, and take the $1 - \alpha$ quantile of the resulting cluster coherences as $\kappa^\star$. Matching the restarts matters, since a restart-boosted real coherence compared against a single-initialization null would over-admit.

C.3 Scoring

Weights. Weighting each admitted cluster by its resultant length $R_\ell = \|\hat{\boldsymbol{\mu}}_\ell\|$ would favor small clusters, because $R_\ell^2 = 1/m_\ell + (1 - 1/m_\ell)\kappa_\ell$ contains a size term $1/m_\ell$ that is large for a small cluster whatever its coherence (Appendix C.2). The weight w_ℓ of equation 7 instead re-inflates the size-free coherence κ_ℓ at a common reference size M. This removes the size bias while keeping w_ℓ on the scale of a resultant length, so the fallback cluster, whose weight is $\|\hat{\boldsymbol{\mu}}\|$, stays commensurate with admitted ones.

Maximum and assignment. The score takes the maximum over clusters rather than a sum, which keeps clusters from competing. An example is credited for serving one cluster well and is not penalized for being orthogonal to the others. Each example is assigned to the cluster that attains the maximum, $a(z_i) = \arg\max_{\ell \in \mathcal{A}} w_\ell \langle \boldsymbol{g}_i, \boldsymbol{u}_\ell \rangle$, which partitions the pool into groups $D_\ell = \{z_i : a(z_i) = \ell\}$. Because $s(z_i)$ is attained at $a(z_i)$, an example is always ranked by the alignment of the cluster it is assigned to. And because each example belongs to one group, no example is selected twice, whereas in $\mathcal{A}_c$ the same example can be chosen by several clusters.

C.4 Quota Rounding

The quotas K_ℓ of Section 4 are real-valued. We round them to integers by largest remainder, which keeps $\sum_\ell K_\ell = K$ while giving each group the whole number of slots closest to its share. A quota larger than its group can fill, $K_\ell > |D_\ell|$, is capped at $|D_\ell|$, and the remainder is passed on to groups with examples to spare, repeating until the budget is exhausted or no capacity remains, so that exactly $\min(K, N)$ examples are returned.

D Experiment Details

D.1 Training and Validation Data

- **Gemma-2-9B and Llama-3.1-8B**: As mentioned in Section 5, our **training data** is sourced from (Wei et al., 2024)[2]. Our **validation data** contains two parts: EvoEval (Xia et al., 2024a) and MBPP training split (Austin et al., 2021). HumanEval has no training split, so we use the verbose variant of EvoEval, which rephrases the HumanEval problems, as an in-domain proxy. The HumanEval and MBPP test prompts and solutions appear in neither the pool nor the validation set, and all baselines share the same validation set.
- **Qwen3.5-4B**: The **validation data** is built from LiveCodeBench v5 (Jain et al., 2025). For each problem we generate a candidate solution with GLM-5.2, falling back to DeepSeek-v4-Pro and Kimi-K2.7-Code when GLM fails, all served through third party API services. Problems that none of the three models solves are discarded, and the solved problems are resampled to match the original v5 distribution over its three difficulty levels, easy, medium, and hard, leaving $M = 144$ validation examples. Our **training data** is DS_RSTAR. Below we present its generation details.

[2] https://huggingface.co/datasets/ise-uiuc/Magicoder-OSS-Instruct-75K

Problem source. The problems come from rSTAR-CODER (Liu et al., 2025), which collects competitive-programming problems from eleven sources, AIZU, AtCoder, CodeChef, CodeWars, GeeksforGeeks, HackerEarth, HackerRank, LeetCode, Codeforces, the International Olympiad in Informatics (IOI), and the USA Computing Olympiad (USACO). It holds 418K verified problems, 37.7K designed by experts and 380K synthesised, which its authors filter by the number of verified responses to 20K verified and 12K synthetic problems.

Trace generation. We generated reasoning traces for these problems with DeepSeek-v3.2-Speciale in December 2025. Test generation samples at temperature 0.1. Solution generation samples at temperature 0.3 when canonical tests are available from the seed prompt, and at 0.7 when tests had to be generated. We set no explicit top-p, seed, or number of samples per request, so the generated-test path issues four separate solution calls per problem and the canonical path allows up to four rounds.

Pool and contamination. We keep only traces whose final solution passes its tests, and the training pool is a uniform random sample of $N = 4{,}573$ of them. We checked the $4{,}573$ pool problems against LiveCodeBench v6 and found no overlap, so the evaluation is not contaminated by the pool.

D.2 DECODING

On LiveCodeBench v6 we sample at temperature 1.0, top-p 0.95, top-k 20, min-p 0.0, presence penalty 1.5, and repetition penalty 1.0, with generations of up to 128k tokens. The code benchmarks use greedy decoding.

D.3 FINE-TUNING AND GRADIENT FEATURES

We adopt the LoRA (Hu et al., 2022) configuration and optimizer settings of LESS (Xia et al., 2024b) unchanged, namely rank 128, $\alpha = 512$, dropout 0.1, learning rate 2×10^{-5}, and four epochs. The warmup run trains a LoRA model on a random $5\,\%$ subset of the pool for four epochs and retains a checkpoint after each epoch. LESS extracts Adam-preconditioned gradients at all four checkpoints and combines the per-checkpoint scores with the learning-rate weights of Xia et al. (2024b), while our method uses the final checkpoint alone, a quarter of the gradient computation. All gradients are projected to $d = 8192$ dimensions with a fixed random projection. In the code setting the official test splits of MBPP and HumanEval are held out of both the pool and the validation set, and every configuration is trained three times from different seeds.

Baseline details are introduced in Appendix A.2.

D.4 BASELINE SETTINGS

LESS and BIDS use the settings of their papers. GIST sets its subspace rank to the smallest r capturing $95\,\%$ of the explained variance of the validation-gradient spectrum, following Min et al. (2026). CLUSTERUCB sets its cluster count, cold-start ratio, and computing budget per setting by the recall-based sweep of Wang et al. (2025c). Baselines are described in Appendix A.2.

E ADDITIONAL EXPERIMENTAL RESULTS

E.1 DISCUSSION ON PERFORMANCE ACROSS DIFFERENT BUDGETS

In the sweep of Figure 3, the margin of SUADS over the strongest baseline at each budget shrinks as the budget grows, from 1.9 points at $5\,\%$ to 1.4, 1.1, 0.4, and 0.3 at $60\,\%$. Each margin on its own is comparable to or smaller than the run-to-run deviations, so the evidence is the consistent lead across all five budgets. The shrinking margin is expected even though retention itself does not depend on the budget (Theorem 1). The gain of any method over random selection is its retention times the gap between random and the best possible selection, and that gap closes as the budget approaches the full pool. At $60\,\%$, RANDOM is already within 1.0 points of full training.

Table 5: Per-difficulty results on LiveCodeBench v6 with Qwen3.5-4B, selecting 5 % and 10 % of DS_RSTAR. Pass@1 (%) with the standard deviation over independent generation runs in parentheses. Bold marks the best selection method in each column. The overall columns are repeated from Table 1.

Method	Top 5 %				Top 10 %			
	Easy (64 pblms.)	Medium (45 pblms.)	Hard (35 pblms.)	Overall (144 pblms.)	Easy (64 pblms.)	Medium (45 pblms.)	Hard (35 pblms.)	Overall (144 pblms.)
RANDOM	40.9 (3.7)	17.9 (7.7)	9.8 (1.6)	19.6 (2.9)	63.4 (6.7)	33.3 (6.8)	15.3 (5.0)	32.1 (1.5)
LESS	64.5 (6.5)	23.1 (6.8)	**13.7** (5.0)	28.5 (1.6)	67.7 (4.6)	35.9 (3.6)	**24.6** (7.0)	38.2 (3.2)
BIDS	**71.0** (4.6)	24.4 (1.8)	13.1 (4.6)	30.2 (2.7)	72.6 (2.3)	**43.6** (0.0)	21.3 (4.6)	40.1 (2.7)
GIST	69.4 (2.3)	28.2 (0.0)	12.3 (3.5)	30.5 (2.2)	72.6 (6.8)	33.3 (14.5)	23.0 (2.3)	37.8 (7.0)
CLUSTERUCB	67.7 (9.1)	26.9 (5.4)	11.5 (2.3)	29.4 (0.5)	65.6 (8.1)	34.2 (3.9)	15.8 (2.5)	33.1 (1.9)
SUADS (ours)	70.2 (3.1)	**34.0** (2.5)	12.3 (5.1)	**32.4** (1.8)	**78.7** (5.9)	41.5 (4.2)	22.6 (2.1)	**41.5** (1.3)
Base model	24.2 (2.3)	2.6 (0.0)	0.0 (0.0)	6.0 (0.5)	24.2 (2.3)	2.6 (0.0)	0.0 (0.0)	6.0 (0.5)
Full training	78.5 (9.9)	59.0 (2.6)	25.7 (3.8)	48.1 (0.8)	78.5 (9.9)	59.0 (2.6)	25.7 (3.8)	48.1 (0.8)

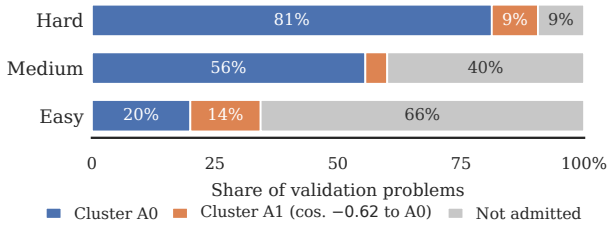

Figure 5: Cluster structure on DS_RSTAR under warmup gradients at a 5 % budget. Each bar shows the share of one difficulty level's validation problems that falls in each admitted cluster or is rejected by the gate. Both clusters mix all three levels, and cluster A1 points almost opposite to A0.

E.2 PER-DIFFICULTY RESULTS ON DS_RSTAR

Table 5 gives the breakdown behind the overall scores of Table 1, by the three difficulty levels of LiveCodeBench v6.

The lead of SUADS comes from different levels at the two budgets. At 5 % it comes from medium problems, 34.0 against 28.2 for the next best method, while SUADS trails LESS on hard ones, 12.3 against 13.7, a gap that lies well inside the deviations of both methods, 5.1 and 5.0. At 10 % the lead comes from easy problems instead, 78.7 against 72.6, while BIDS leads on medium ones, 43.6 against 41.5, and LESS again on hard ones, 24.6 against 22.6. *Per-difficulty deviations are much larger than those on the overall score, since each level holds a third of the* 144 *problems, so single-cell comparisons carry little weight on their own.*

E.3 WHAT CLUSTERS CAPTURE

On Magicoder with Gemma-2-9B, we cross-tabulate the validation examples of each admitted cluster against their source. With two, three, and six clusters, no admitted cluster mixes MBPP with EvoEval, so clustering recovers the source boundary without being given it. The gate also judges the two sources differently. MBPP's size-free coherence clears the gate's threshold and EvoEval's does not, $\kappa = 0.551$ and 0.428 against $\kappa^\star \approx 0.47$. In the terms of Equation (6), MBPP is coherent enough for its own centroid to help and EvoEval is not, so in the main configuration both admitted clusters lie within MBPP and every EvoEval candidate is rejected; with three or six admitted clusters, only a small EvoEval core passes. EvoEval still shapes selection, since it takes part in the clustering and in the background coherence that sets the threshold. Nor does the gain of SUADS come from redirecting budget between the sources, as it gives examples aligned with EvoEval almost the same share as the subtask-max rule on the same features, 5.9 % against 6.6 %.

DS_RSTAR, with Qwen3.5-4B, shows the opposite case, labels that mark little difference in the gradients. Its three difficulty levels point nearly the same way, and here the clusters cut across the labels.

Table 6: Predicting which cluster a validation problem of DS_RSTAR falls into, with six admitted clusters. Accuracy is averaged over ten repeats of five-fold cross validation. The baseline always predicts the largest cluster, which holds 68 of the 144 problems, giving accuracy $47.2\,\%$, which is $68/144$, and balanced accuracy $14.3\,\%$, which is $1/7$.

Features	Accuracy (%)	Balanced accuracy (%)
target composition alone	75.8	37.3
target length alone	69.9	27.4
difficulty alone	59.0	21.8
algorithm type alone	50.4	17.7
generating model alone	48.5	27.1
source and answer format alone	43.4	13.7
the five features other than the type	**77.6**	52.1
all six features	71.1	46.4
majority baseline	47.2	14.3

Both admitted clusters mix all three levels, and the gate rejects most easy problems, whose gradients are the least coherent (Figure 5). One cluster, 13 problems drawn from all three levels, points almost opposite to the other, with a centroid cosine of -0.62. The subtask-max rule scores the training examples aligned with it low, but SUADS reserves them $13.4\,\%$ of its budget. Together, the two settings show that the clusters follow the labels where the labeled groups differ in their gradients and depart from them where they do not.

That reading of clustering can be tested directly. We run an analysis on a configuration with a larger cap that admits six clusters rather than the two of our main experiments, because the finer partition shows the structure more clearly. We label each validation problem by hand with one of eleven algorithm types, in decreasing frequency implementation and simulation (25), dynamic programming (20), maths, number theory and bit manipulation (19), string processing (17), binary search, two pointers and prefix sums (13), graphs and trees (12), combinatorics and counting (11), data structures (9), greedy and sorting (8), brute force and backtracking (7), and game theory (3).

We then ask how well cluster membership can be predicted from properties that need no gradients, by multinomial logistic regression under ten repeats of five-fold cross validation (Table 6). Besides the algorithm type we use the *response length* in tokens, where the response is the reasoning trace the example trains on, the *response composition*, which is the fraction of the response inside `<think>`, the fraction that is code, how often backtracking markers such as *wait* occur per thousand words, the fraction of distinct lines and the mean line length, the problem's source and answer format, its difficulty, and the model that produced the response.

Membership is largely predictable, at $77.6\,\%$ against a majority baseline of $47.2\,\%$, so the gate is selecting on properties we can name. The algorithm type is not among them. It is the weakest feature we tried, $50.4\,\%$ on its own against $75.8\,\%$ for response composition and $69.9\,\%$ for response length, and adding it to any other combination costs 4.2 points of accuracy on average, helping in 15 of the 127 combinations. What predicts a cluster is how the answer is written and where the problem comes from, not the technique it needs. The two-cluster configuration gives the same answer. There the five features reach $87.3\,\%$ against a baseline of $58.3\,\%$, while the algorithm type alone reaches $58.5\,\%$, which is the baseline. Per-cluster recall is uneven in a way that matches the rest of this section, $97\,\%$ for the largest cluster, which holds 68 of the 144 problems, and $89\,\%$ for the rejected group, but near zero for the cluster whose centroid points away from it, which no feature we measured reproduces.

Figure 6 compares the length of the DS_RSTAR training examples with the length of the $5\,\%$ that each method selects. Every targeted method takes far shorter examples than a random draw, whose median of 20,640 tokens matches the pool at 23,255. SUADS is the least extreme of them, at 2,756 tokens against 1,702 for BIDS, 1,436 for GIST, 1,176 for LESS and 1,072 for CLUSTERUCB, and it keeps the largest fraction of long examples. Alignment is measured by cosine, so length earns no credit, and a short focused trace aligns more closely with a cluster centroid than a long one that wanders. Selecting per cluster tempers this, since each admitted cluster spends its own quota.

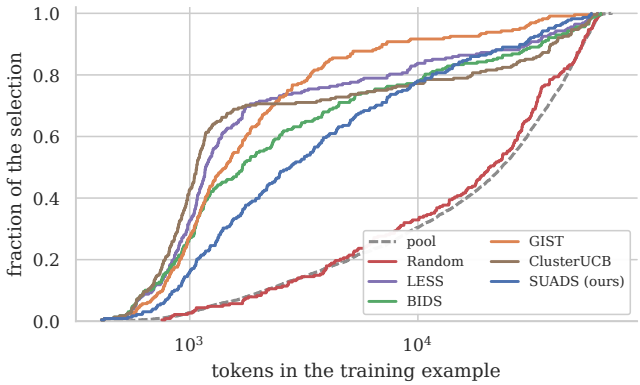

Figure 6: Length of the DS_RSTAR training examples in tokens, instruction and response together, as the fraction of each selection shorter than a given length. The pool is dashed.

F LIMITATIONS

Our analysis is first-order and Gaussian: it explains why targeted selection can gain little over random sampling but not why it sometimes falls below it, and the clustering result analyzes an idealized protocol with known clusters, multiplicity-counted selection, and cluster means close to the global mean (Appendices B.4 and B.5), which SUADS approximates rather than instantiates. Appendix B.6 measures the retention factors on our validation gradients: the attenuation is material for references built from few examples and small for the global mean, and the analysis does not predict the size of the gain of local references over the global one (Table 2). On the method side, the coherence test rejects clusters without real structure but cannot reject the pieces of a coherent one, so the cap on the number of clusters remains a hyperparameter to which the method is moderately sensitive (Table 2).