TRAINING-AWARE TARGET COVERAGE FOR SYNTHETIC DATA SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic data are increasingly used to scale LLM training, yet more synthetic data do not necessarily produce better models. Useful synthetic data must add information relevant to the target task without introducing errors that offset their benefit, and the value of an example can change as the training set grows. We develop a linear theory that characterizes this tradeoff and determines where synthetic data are useful, how much should be added, and the marginal value of adding one example to an existing set. The analysis shows the conditions when input coverage alone is sufficient and when synthetic errors must also be considered. Guided by these results, we introduce *Training-Aware Target Coverage* (TATC), a synthetic data selection method for LLM fine-tuning. TATC identifies candidates whose training effects are beneficial to the target task and selects among them to expand coverage of target-relevant directions not already represented by the available data. Experiments on text and image data verify the linear theory. With mathematical reasoning tasks, TATC selects synthetic solutions for fine-tuning Qwen2.5-Math-1.5B-Instruct and outperforms alternative synthetic-data selection methods on GSM8K across selection budgets. In summary, we provide a principled approach to synthetic data selection by quantifying and maximizing its value to the target task.

1 INTRODUCTION

Synthetic data have become an integral part of large language model training, spanning multiple stages of the training pipeline, from pretraining and instruction tuning to reasoning distillation and model evaluation (Gunasekar et al., 2023; Wang et al., 2023b; Zelikman et al., 2022). Models can generate large volumes of synthetic data, such as mathematical solutions, code, instructions, critiques, and preference feedback, with little additional human annotation, making it a scalable way to construct training data. This scalability, however, does not imply that more synthetic data are always better. The value of synthetic examples can differ substantially: adding redundant examples may provide little new information, while examples containing errors can actively harm downstream model performance; these effects can become particularly consequential as synthetic data are incorporated at larger scales or reused across generations of models (Shumailov et al., 2024; Alemohammad et al., 2024; Yi et al., 2026). The central question is therefore no longer simply how to generate more synthetic data, but *which synthetic examples are useful for training, and how many of them should be incorporated into the training dataset?*

Existing approaches to synthetic data selection emphasize complementary properties of training examples. Quality-based methods filter generated responses using correctness or verification signals and model-based quality estimates (Chen et al., 2024; Liu et al., 2024), while diversity-based methods favor collections that span distinct regions of representation or model-induced gradient space (Friedman & Dieng, 2023; Jung et al., 2025). Other approaches select examples by matching the synthetic and target distributions or by optimizing information-based design objectives (Ash et al., 2021; Rezaei et al., 2026). A separate line of work estimates training utility more directly through gradient similarity or influence on a target objective (Xia et al., 2024; Fan et al., 2026). These perspectives capture different aspects of useful synthetic data, but they do not by themselves answer the question regarding a synthetic example's contribution to the existing training set and its impact on the target task. A synthetic input may be able to add information in an underrepresented area of data distribution, but its generated response may introduce a harmful training effect at the same time.

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.

Conversely, an example with a favorable training effect may add little information if its local area is already well represented by the real data or previously selected synthetic examples. This motivates us to develop a synthetic data selection method that jointly considers each example's training effect and its marginal information contribution.

We first develop a linear theory to understand when and how much synthetic data are beneficial for learning. Rather than treating each synthetic example as having merely an intrinsic value, the theory characterizes the value of a selected synthetic set through the tradeoff between the information it adds and the training errors it introduces. This perspective reveals three key properties of synthetic data selection. First, value is *directional*: synthetic data are most useful when they add information in directions that matter for the target task but are insufficiently represented by the existing data. Second, value depends on *amount*: along a given direction, the useful amount of synthetic data can be zero, finite, or continue to increase, depending on the balance between additional information and error. Third, value is *marginal*: the benefit of adding a candidate depends on what the real data and previously selected synthetic examples already provide. Together, these results explain why neither data quality nor coverage alone is sufficient to determine which synthetic examples, or how many of them, should be used. We experiment on text and image data to validate these predictions: accounting for synthetic error distinguishes beneficial examples from harmful examples, while the theory also identifies the amount of synthetic data that can be most useful.

Guided by these results, we introduce **Training-Aware Target Coverage (TATC)**, a practical method for synthetic data selection in LLM fine-tuning. TATC asks two complementary questions for each candidate: whether training on the synthetic example is expected to benefit the target task, and whether its input adds information in target-relevant directions that are not yet well represented. Using a small labeled real probe set, TATC first retains candidates with beneficial training effects and then selects those that provide complementary target coverage. Coverage is evaluated relative to the real data and previously selected synthetic examples, allowing selection to adapt as the training set grows rather than scoring each candidate in isolation. We evaluate TATC by fine-tuning Qwen2.5-Math-1.5B-Instruct (Yang et al., 2024) with synthetic mathematical solutions and evaluating on GSM8K (Cobbe et al., 2021). Our approach improves exact-match accuracy over both the real data checkpoint and other competing selection methods across a range of selection budgets.

Our contributions can be summarized below:

- We develop a theoretical framework for characterizing the value of synthetic data, showing how downstream benefit depends jointly on the information added by synthetic inputs and the errors they introduce. The analysis shows *where* synthetic data are useful across feature directions, *how much* data should be added, and *why* the value of each additional example depends on the existing training dataset.
- Guided by this theory, we introduce **Training-Aware Target Coverage (TATC)**, a synthetic data selection method in LLM fine-tuning. TATC identifies examples with beneficial training effects and selects candidates that provide complementary coverage of target-relevant directions, updating their value as the selected set grows.
- We empirically validate the proposed theory on text and image data and show that TATC outperforms existing synthetic data selection methods on GSM8K mathematical reasoning.

2 Related Work

Synthetic inputs and generated response error. Recent pipelines construct synthetic training data through rationale distillation, instruction evolution, mathematical problem synthesis, error-driven generation, and diversified instruction generation (Hsieh et al., 2023; Xu et al., 2024; Yu et al., 2024; Wang et al., 2023a; Li et al., 2024; Xu et al., 2025; Ge et al., 2024); synthetic augmentation has also been studied for calibration and generalization (Ba et al., 2024). Recursive synthetic training can amplify model errors (Shumailov et al., 2024; Alemohammad et al., 2024), while verification can mitigate degradation when generated responses can be checked (Yi et al., 2026). High-dimensional regression analyses identify covariance shift as a factor in synthetic-data utility and motivate covariance matching (Rezaei et al., 2026). We study selection from a candidate pool of LLM-generated input-response pairs, jointly accounting for input information, systematic response error, and excess response noise.

Quality and validation-guided data selection. AlpaGasus filters instruction data using model-based quality scores (Chen et al., 2024), while DEITA combines complexity, quality, and diversity (Liu et al., 2024). GLISTER selects subsets for a held-out validation objective (Killamsetty et al., 2021), and RHO-LOSS prioritizes examples by reducible loss (Mindermann et al., 2022). Our setting instead separates the predicted training effect of a complete example from the set-dependent target coverage of its features.

Diversity and experimental design. The Vendi Score measures set diversity through the spectrum of a similarity matrix (Friedman & Dieng, 2023); G-Vendi applies the same idea to model-induced gradients for synthetic reasoning data generation (Jung et al., 2025). These methods reward dispersion within a set. In contrast, experimental design targets reductions in estimation error; BAIT makes a Fisher inverse-trace criterion practical for neural active learning (Ash et al., 2021). Our target coverage term follows this design principle and is evaluated as an ablation rather than a new objective. Unlike active learning, however, synthetic data selection observes a generated response rather than acquiring a trusted label after selecting an input. Because that response can be systematically wrong, its value must also depend on response error.

Gradient and influence selection. Influence functions and Data Shapley scores quantify the contribution of individual training examples (Koh & Liang, 2017; Ghorbani & Zou, 2019). LESS uses optimizer-aware gradient similarity for targeted instruction tuning (Xia et al., 2024), while OptimSyn uses estimated training influence to guide synthetic-data generation (Fan et al., 2026). SPICE combines a Fisher-information objective with gradient conflict to select informative and compatible instruction data (Chang et al., 2026). Our analysis instead asks how input information and generated-response error jointly determine the marginal value of adding a synthetic example to the existing training set.

3 Downstream Risk Theory in Synthetic Data Selection

3.1 Risk of Training with Synthetic Data

We study when synthetic data improve downstream performance by accounting for both the information provided by synthetic inputs and errors in their labels. Let $\phi(X) \in \mathbb{R}^d$ denote a feature representation of input X, let Y denote the corresponding label, and consider a linear downstream model $f_{\boldsymbol{\theta}}(X) = \phi(X)^\top \boldsymbol{\theta}, \boldsymbol{\theta} \in \mathbb{R}^d$. We assume that there exists a true parameter $\boldsymbol{\theta}^\star \in \mathbb{R}^d$ such that

$$\mathbb{E}[Y \mid X] = \phi(X)^\top \boldsymbol{\theta}^\star.$$

For the real and synthetic samples, let $\{(\boldsymbol{x}_{r,i}, y_{r,i})\}_{i=1}^n, \{(\boldsymbol{x}_{s,j}, y_{s,j})\}_{j=1}^m$, denote their feature-label pairs, where $\boldsymbol{x}_{r,i}, \boldsymbol{x}_{s,j} \in \mathbb{R}^d$ are obtained from the representation ϕ defined above (i.e., $\boldsymbol{x} = \phi(X)$).

Let

$$\boldsymbol{X}_r = [\boldsymbol{x}_{r,1}, \ldots, \boldsymbol{x}_{r,n}]^\top \in \mathbb{R}^{n \times d}, \qquad \boldsymbol{X}_s = [\boldsymbol{x}_{s,1}, \ldots, \boldsymbol{x}_{s,m}]^\top \in \mathbb{R}^{m \times d}$$

be feature matrices. We model their labels as

$$\boldsymbol{y}_r = \boldsymbol{X}_r \boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}_r, \qquad \boldsymbol{y}_s = \boldsymbol{X}_s \boldsymbol{\theta}^\star + \boldsymbol{d}_s + \boldsymbol{\varepsilon}_s,$$

where $\boldsymbol{d}_s \in \mathbb{R}^m$ represents systematic error in the synthetic labels. Conditional on the feature matrices, $\boldsymbol{\varepsilon}_r$ and $\boldsymbol{\varepsilon}_s$ are independent noises with zero means and covariances $\boldsymbol{\Omega}_r$ and $\boldsymbol{\Omega}_s$, respectively.

We estimate $\boldsymbol{\theta}$ from the combined real and synthetic data by least squares:

$$\widehat{\boldsymbol{\theta}}_s = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \frac{1}{2} \|\boldsymbol{X}_r \boldsymbol{\theta} - \boldsymbol{y}_r\|_2^2 + \frac{1}{2} \|\boldsymbol{X}_s \boldsymbol{\theta} - \boldsymbol{y}_s\|_2^2 \right\}.$$

We evaluate the fitted model on a downstream distribution P_{test} and define its second moment matrix as $\boldsymbol{Q} := \mathbb{E}_{X \sim P_{\text{test}}} \left[\phi(X) \phi(X)^\top \right] \succeq 0$. We define the expected squared error relative to the conditional mean on P_{test} as the *downstream risk*:

$$\mathcal{R}_Q(\widehat{\boldsymbol{\theta}}_s) := \mathbb{E}_{X \sim P_{\text{test}}} \left[\left(\phi(X)^\top (\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}^\star) \right)^2 \right] = (\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}^\star)^\top \boldsymbol{Q} (\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}^\star).$$

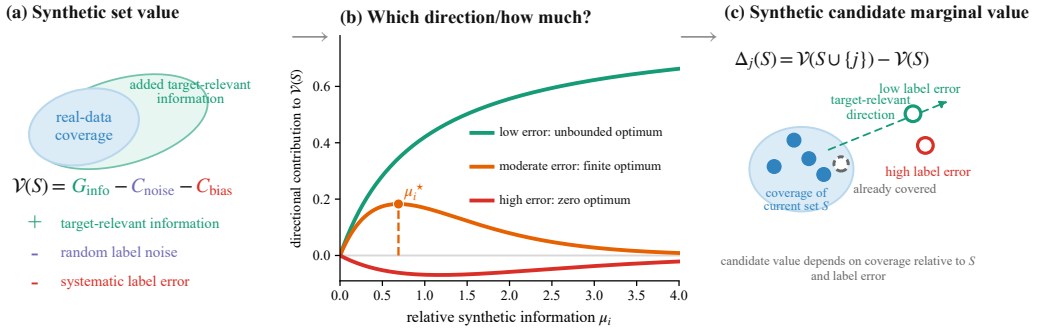

Figure 1: **Risk and value of synthetic data.** The theory in one view. (a) A synthetic set is valuable when target-relevant input information outweighs random label noise and systematic label error. (b) This tradeoff determines whether the optimal amount along a direction is zero, finite, or unbounded. (c) Candidate value depends on both input coverage relative to the current set and label error.

To characterize the downstream risk, define

$$\boldsymbol{A} = \boldsymbol{X}_r^\top \boldsymbol{X}_r + \boldsymbol{X}_s^\top \boldsymbol{X}_s, \qquad \boldsymbol{u} = \boldsymbol{X}_s^\top \boldsymbol{d}_s, \qquad \boldsymbol{N} = \boldsymbol{X}_r^\top \boldsymbol{\Omega}_r \boldsymbol{X}_r + \boldsymbol{X}_s^\top \boldsymbol{\Omega}_s \boldsymbol{X}_s.$$

Here, $\boldsymbol{A}$ is the feature Gram matrix, $\boldsymbol{u}$ represents the contribution of systematic label error, and $\boldsymbol{N}$ represents the contribution of random label noise.

Proposition 1 (Downstream risk decomposition)**.** *Assume that $\boldsymbol{A}$ is nonsingular. Conditional on $(\boldsymbol{X}_r, \boldsymbol{X}_s)$, under the noise assumptions above,*

$$\mathbb{E}\left[\mathcal{R}_Q(\widehat{\boldsymbol{\theta}}_s) \mid \boldsymbol{X}_r, \boldsymbol{X}_s\right] = \underbrace{(\boldsymbol{A}^{-1}\boldsymbol{u})^\top \boldsymbol{Q} (\boldsymbol{A}^{-1}\boldsymbol{u})}_{\text{bias from systematic label error}} + \underbrace{\operatorname{tr}\left(\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{N}\boldsymbol{A}^{-1}\right)}_{\text{variance from label noise}} \tag{1}$$

Proposition 1 gives a fundamental decomposition used throughout this analysis. Synthetic inputs affect downstream risk through the information matrix $\boldsymbol{A}$, while errors in synthetic labels affect the risk through both systematic error and random variation. Thus, the benefit of adding synthetic data cannot be decided by input coverage alone. See proof in Appendix A.1.

3.2 Value of Synthetic Data

We now move from analyzing a given synthetic dataset to selecting examples from a synthetic candidate pool. Let $\mathcal{P}_s = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^N$ denote the candidate pool, and let $S \subseteq \{1, \dots, N\}$ index a selected synthetic subset. We treat S as fixed with respect to the realized label noise and use $\boldsymbol{X}_S, \boldsymbol{y}_S, \boldsymbol{d}_S$, and $\boldsymbol{\Omega}_S$ for the corresponding quantities restricted to S. We define the value of S as the reduction in expected downstream risk relative to training on the real data alone:

$$\mathcal{V}(S) := \mathbb{E}[\mathcal{R}_Q(\widehat{\boldsymbol{\theta}}_r) \mid \boldsymbol{X}_r] - \mathbb{E}[\mathcal{R}_Q(\widehat{\boldsymbol{\theta}}_S) \mid \boldsymbol{X}_r, \boldsymbol{X}_S].$$

Thus, $\mathcal{V}(S) > 0$ means that adding S improves expected downstream performance. To separate the sources of this added value, we consider isotropic label noise,

$$\boldsymbol{\Omega}_r = \sigma^2 \boldsymbol{I}, \qquad \boldsymbol{\Omega}_S = (\sigma^2 + \tau^2)\boldsymbol{I},$$

where $\tau^2 \geq 0$ is the excess random noise in the synthetic labels. Let

$$\boldsymbol{A}_r = \boldsymbol{X}_r^\top \boldsymbol{X}_r \succ 0, \qquad \boldsymbol{A}_S = \boldsymbol{X}_S^\top \boldsymbol{X}_S, \qquad \boldsymbol{A} = \boldsymbol{A}_r + \boldsymbol{A}_S$$

Corollary 2 (Value of a selected synthetic set)**.** *Under the specifications above,*

$$\begin{aligned}\mathcal{V}(S) = &\underbrace{\sigma^2 \operatorname{tr}[\boldsymbol{Q}(\boldsymbol{A}_r^{-1} - \boldsymbol{A}^{-1})]}_{G_{\text{info}}(S)} - \underbrace{\tau^2 \operatorname{tr}(\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{A}_S\boldsymbol{A}^{-1})}_{C_{\text{noise}}(S)} \\ &- \underbrace{(\boldsymbol{A}^{-1}\boldsymbol{X}_S^\top \boldsymbol{d}_S)^\top \boldsymbol{Q} (\boldsymbol{A}^{-1}\boldsymbol{X}_S^\top \boldsymbol{d}_S)}_{C_{\text{bias}}(S)}.\end{aligned} \tag{2}$$

Equation (2) expresses the value of a selected synthetic set as a tradeoff between the information provided by its inputs and errors in its labels. $G_{\text{info}}(S)$ measures the information relevant to the downstream task that is added beyond the real data, $C_{\text{noise}}(S)$ is the downstream cost of excess random label noise, and $C_{\text{bias}}(S)$ is the downstream cost induced by systematic error in the synthetic labels. Thus, adding S improves downstream performance when its information gain exceeds the two error terms. Figure 1 summarizes this tradeoff and the directional and candidate-level results developed below. See Appendix A.2 for the proof.

Which directions are useful? To understand where the value in Equation (2) comes from, we compare the information provided by the selected synthetic inputs with what was already provided by the real data. By spectral analysis, define

$$\boldsymbol{A}_r^{-1/2}\boldsymbol{A}_S\boldsymbol{A}_r^{-1/2} = \boldsymbol{R}\,\mathrm{diag}(\mu_1,\dots,\mu_d)\boldsymbol{R}^\top, \quad \boldsymbol{D} = \boldsymbol{A}_r^{-1/2}\boldsymbol{Q}\boldsymbol{A}_r^{-1/2},$$

where $\boldsymbol{R} = [\boldsymbol{r}_1,\dots,\boldsymbol{r}_d]$ is orthogonal. Let $w_i = \boldsymbol{r}_i^\top \boldsymbol{D}\boldsymbol{r}_i$. The eigenvalue $\mu_i \geq 0$ measures the synthetic information added along direction $\boldsymbol{r}_i$ relative to the real data, while $w_i \geq 0$ measures the downstream relevance of that direction.

Proposition 3 (Directional value of synthetic information)**.** *Assume $\boldsymbol{A}_r \succ 0$. The information gain and excess noise terms in Corollary 2 decompose as*

$$G_{\text{info}}(S) = \sigma^2 \sum_{i=1}^{d} w_i \frac{\mu_i}{1+\mu_i}, \qquad C_{\text{noise}}(S) = \tau^2 \sum_{i=1}^{d} w_i \frac{\mu_i}{(1+\mu_i)^2}.$$

These decompositions hold without further assumptions. If, in addition, the systematic label error has the form $\boldsymbol{d}_S = \boldsymbol{X}_S\boldsymbol{\delta}, \boldsymbol{\delta} \in \mathbb{R}^d$, and $\boldsymbol{R}^\top\boldsymbol{D}\boldsymbol{R}$ is diagonal, define $\boldsymbol{\eta} = \boldsymbol{R}^\top\boldsymbol{A}_r^{1/2}\boldsymbol{\delta}$. Then the full set value decomposes across directions:

$$\mathcal{V}(S) = \sum_{i=1}^{d} w_i \left[\sigma^2 \frac{\mu_i}{1+\mu_i} - \tau^2 \frac{\mu_i}{(1+\mu_i)^2} - \eta_i^2 \left(\frac{\mu_i}{1+\mu_i}\right)^2\right]. \tag{3}$$

Equation (3) simplifies the tradeoff for the set in Equation (2) direction by direction. Each direction contributes according to the information added by the synthetic input, its downstream relevance, and the random and systematic errors in the synthetic label.

Corollary 4 (Beneficial directions)**.** *Under the conditions of Proposition 3, a direction with $\mu_i > 0$ and $w_i > 0$ contributes positively to $\mathcal{V}(S)$ if and only if*

$$\eta_i^2 < \sigma^2 + \frac{\sigma^2 - \tau^2}{\mu_i}.$$

Thus, a direction relevant to the downstream task is not necessarily beneficial. Its information gain must be sufficient to offset the random and systematic errors associated with the synthetic labels. See Appendix A.4 for the proof.

Corollary 5 (Optimal amount of synthetic information)**.** *Under the conditions of Proposition 3, let $w_i > 0$. If*

$$\sigma^2 > \tau^2, \qquad 2\eta_i^2 > \sigma^2 + \tau^2,$$

the directional contribution is maximized at

$$\mu_i^\star = \frac{\sigma^2 - \tau^2}{2\eta_i^2 - \sigma^2 - \tau^2}. \tag{4}$$

Otherwise, the optimum lies at a boundary: $\mu_i^\star = 0$ when $\sigma^2 < \tau^2$ and $\eta_i^2 \geq \sigma^2$, and the supremum is approached as $\mu_i \to \infty$ in the complementary nondegenerate regime. Equality and degenerate cases are given in Appendix A.5.

Corollary 5 shows that more synthetic information is not always beneficial: the optimal amount along a direction can be zero, finite, or unbounded. Equation (4) varies μ_i while keeping w_i and η_i unchanged, and therefore characterizes the optimal amount along a given direction rather than which synthetic examples should be selected.

The directional analysis above assumes that the synthetic label error can be assessed. This requirement cannot generally be met in directions that are not covered by the real data. It also assumes $\boldsymbol{A}_r \succ 0$, so the real data identify every modeled direction. We next consider the complementary case in which a downstream-relevant direction is not covered by the real training data.

Proposition 6 (When synthetic label error is not identifiable)**.** *Let $\boldsymbol{v} \in \mathbb{R}^d$ satisfy $\boldsymbol{X}_r\boldsymbol{v} = \boldsymbol{0}, \boldsymbol{X}_S\boldsymbol{v} \neq \boldsymbol{0}, \boldsymbol{v}^\top \boldsymbol{Q}\boldsymbol{v} > 0$. Without additional labeled information identifying $\boldsymbol{v}^\top\boldsymbol{\theta}^\star$, the observations cannot distinguish $(\boldsymbol{\theta}^\star, \boldsymbol{d}_S)$ from $(\boldsymbol{\theta}^\star + c\boldsymbol{v}, \boldsymbol{d}_S - c\boldsymbol{X}_S\boldsymbol{v})$.*

Thus, when a downstream-relevant direction is absent from the real training data, synthetic labels alone cannot distinguish target signal from systematic label error. Additional labeled real data covering such directions can resolve this ambiguity; Section 4 uses a real probe set for this purpose.

3.3 Marginal Value of a Synthetic Example

The value in Equation (2) determines whether a selected synthetic set is beneficial as a whole. We now return to the general risk formulation and ask how this value changes when one candidate is added to the current set S. For a synthetic candidate $j \notin S$, define

$$\Delta_j(S) := \mathcal{V}(S \cup \{j\}) - \mathcal{V}(S).$$

Thus, $\Delta_j(S) > 0$ means that adding candidate j further reduces expected downstream risk. Let $(\boldsymbol{A}, \boldsymbol{u}, \boldsymbol{N})$ denote the information, error, and noise quantities defined before Proposition 1, now formed from the real data and selected set S. Let $\boldsymbol{x}_j$, δ_j, and ω_j denote the candidate's feature vector, systematic label error, and random label noise variance, respectively, with its noise independent of the real and selected-sample noises. Define

$$\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{u}, \qquad \boldsymbol{p}_j = \boldsymbol{A}^{-1}\boldsymbol{x}_j, \qquad c_j = 1 + \boldsymbol{x}_j^\top \boldsymbol{p}_j, \qquad e_j = \delta_j - \boldsymbol{x}_j^\top \boldsymbol{b}, \qquad q_j = \boldsymbol{p}_j^\top \boldsymbol{Q}\boldsymbol{p}_j.$$

Here, $\boldsymbol{b}$ is the bias induced by the selected set, e_j is the candidate's systematic error relative to that bias, and q_j measures the downstream relevance of its induced direction.

Theorem 7 (Conditional marginal value)**.** *The marginal value of candidate j is*

$$\boxed{\Delta_j(S) = \frac{2\boldsymbol{p}_j^\top \boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{N}\boldsymbol{p}_j}{c_j} - \frac{q_j(\boldsymbol{p}_j^\top \boldsymbol{N}\boldsymbol{p}_j + \omega_j)}{c_j^2} - \frac{2e_j\boldsymbol{b}^\top \boldsymbol{Q}\boldsymbol{p}_j}{c_j} - \frac{e_j^2 q_j}{c_j^2}.} \tag{5}$$

The first two terms describe the change in the variance component of downstream risk, while the last two describe the impact of systematic label error and its interaction with the bias already induced by S. Therefore, the value of a candidate is conditional on the selected set: the same candidate can have different marginal value depending on the examples already included.

A useful special case isolates the contribution of input coverage. If $\boldsymbol{\Omega}_r = \boldsymbol{\Omega}_S = \sigma^2\boldsymbol{I}, \boldsymbol{b} = \boldsymbol{0}, \delta_j = 0, \omega_j = \sigma^2$, then

$$\Delta_j(S) = \frac{\sigma^2 q_j}{c_j}. \tag{6}$$

The more general specializations with systematic label error or excess random noise are given in Appendix A.7. Equation (6) isolates the coverage relevant to the downstream task that is contributed by the candidate input. Theorem 7 shows that this coverage term is sufficient only when labels are clean; more generally, candidate value also depends on label error and its interaction with the current bias. This insight motivates the development of TATC in Section 4.

Figure 2 tests two implications of the linear theory using semi-synthetic constructions that keep candidate inputs real while changing their labels. This lets us examine label error without changing input information, which would change simultaneously if both inputs and labels were generated. In panel (a), controlled label errors are assigned to SetFit/subj texts (Pang & Lee, 2004); coverage alone fails to distinguish beneficial from harmful candidates, whereas the complete marginal closely predicts the risk reduction observed after refitting with each candidate. In panel (b), the same CIFAR-10 (Krizhevsky, 2009) candidate images receive labels from models with different error levels; the

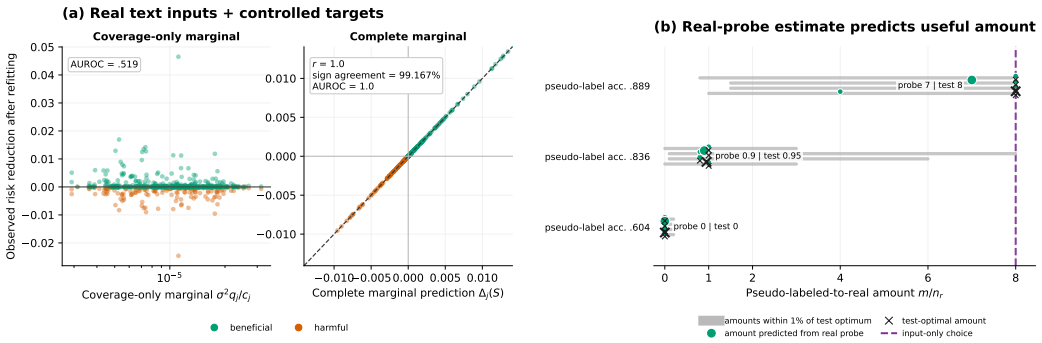

Figure 2: **Candidate value and synthetic-data amount with real inputs and controlled targets.** **(a)** Candidate-level test on SetFit/subj. Coverage alone does not distinguish beneficial from harmful candidates, whereas the complete marginal closely predicts the risk reduction from explicit refitting. **(b)** Amount selection with ResNet-50 features on CIFAR-10. The amount predicted from a labeled real probe closely tracks the test-optimal amount and decreases as pseudo-label reliability falls, while the input-only rule selects the full candidate pool. Gray segments indicate amounts within 1% of the test minimum.

probe-estimated useful amount closely tracks the test optimum as label error increases, while the input-only criterion selects the full candidate pool. Together, these tests show that changing labels alone can change both which examples are useful and how much data should be used. Section 4 extends this distinction to nonlinear fine-tuning; details are in Appendix B.

4 Training-Aware Target Coverage in Supervised Fine-Tuning

In this section, we translate the linear analysis in Section 3 into a practical synthetic-data selection method for nonlinear models. The theory shows that input coverage alone is insufficient because synthetic responses can introduce error. Using LLM fine-tuning as a concrete setting, we combine target coverage with a local estimate of each complete synthetic example's training effect.

We start from a model checkpoint $\boldsymbol{\theta}_0$ obtained by fine-tuning on real data. Let $\mathcal{D}_{\text{syn}} = \{(x_j, y_j)\}_{j=1}^{N}$ denote the synthetic candidate pool. Let $\boldsymbol{\phi}_{\boldsymbol{\theta}_0}(x) \in \mathbb{R}^p$ denote the representation of input x extracted at $\boldsymbol{\theta}_0$, and write $\boldsymbol{\phi}_j = \boldsymbol{\phi}_{\boldsymbol{\theta}_0}(x_j)$. We use the real training set $\mathcal{D}_r$ together with a disjoint real probe set $\mathcal{D}_q$. The real training set determines the checkpoint and initial representation geometry, while the probe set provides target-task inputs and labels for measuring downstream relevance and the predicted effect of synthetic examples.

Target coverage. We first measure which representation directions are covered by the real training data and which directions occur in the labeled probe set. Using the representations defined above, let

$$\widehat{\boldsymbol{A}}_0 = \rho \boldsymbol{I} + \sum_{x \in \mathcal{D}_r} \boldsymbol{\phi}_{\boldsymbol{\theta}_0}(x) \boldsymbol{\phi}_{\boldsymbol{\theta}_0}(x)^\top, \qquad \widehat{\boldsymbol{Q}} = \frac{1}{|\mathcal{D}_q|} \sum_{x \in \mathcal{D}_q} \boldsymbol{\phi}_{\boldsymbol{\theta}_0}(x) \boldsymbol{\phi}_{\boldsymbol{\theta}_0}(x)^\top.$$

Here, $\rho > 0$ ensures numerical stability of the empirical Gram matrix. For a selected set of candidate indices $S \subseteq \{1, \dots, N\}$, define

$$\widehat{\boldsymbol{A}}_S = \widehat{\boldsymbol{A}}_0 + \sum_{j \in S} \boldsymbol{\phi}_j \boldsymbol{\phi}_j^\top.$$

For a candidate $j \notin S$, its marginal coverage gain is

$$a_j(S) = \frac{\boldsymbol{\phi}_j^\top \widehat{\boldsymbol{A}}_S^{-1} \widehat{\boldsymbol{Q}} \widehat{\boldsymbol{A}}_S^{-1} \boldsymbol{\phi}_j}{1 + \boldsymbol{\phi}_j^\top \widehat{\boldsymbol{A}}_S^{-1} \boldsymbol{\phi}_j}.$$

By Sherman–Morrison, $a_j(S)$ is exactly the reduction in $\operatorname{tr}(\widehat{\boldsymbol{Q}} \widehat{\boldsymbol{A}}_S^{-1})$ from adding x_j, which we call its *target coverage*. It measures how much candidate input x_j adds coverage in directions represented by the real probe set but not already covered by the existing training data. Because $a_j(S)$ depends only on x_j, it does not account for the effect of its synthetic response y_j.

Predicted local training effect. Input coverage alone does not determine whether a synthetic example is useful: two candidates with similarly informative inputs can induce different parameter updates because their synthetic responses differ. We estimate this effect by predicting how the candidate's optimizer update changes loss on the real probe.

For a synthetic sample (x_j, y_j), let $\Delta\boldsymbol{\theta}_j$ denote the difference between the AdamW parameter displacements obtained using the candidate gradient and a zero gradient, both computed from the same checkpoint $\boldsymbol{\theta}_0$ and saved optimizer state (Loshchilov & Hutter, 2019). The construction is detailed in Appendix D.2. Let $\boldsymbol{g}_q$ and $\boldsymbol{H}_q$ denote the gradient and a local curvature approximation of the loss on the probe set $\mathcal{D}_q$, evaluated at the same checkpoint. In our implementation, H_q is a diagonal empirical-Fisher approximation (Kunstner et al., 2019). A second-order Taylor approximation gives

$$h_j = -\boldsymbol{g}_q^\top \Delta\boldsymbol{\theta}_j - \frac{1}{2}\Delta\boldsymbol{\theta}_j^\top \boldsymbol{H}_q \Delta\boldsymbol{\theta}_j. \tag{7}$$

A larger h_j indicates a larger predicted reduction in loss on the real probe. Unlike $a_j(S)$, h_j depends on the complete synthetic example through its induced parameter update. It therefore provides a fixed-checkpoint score for the response-dependent effect induced by the candidate's local optimizer update; this effect is absent from the coverage criterion.

Training Aware Target Coverage (TATC). The two quantities play complementary roles in Algorithm 1: h_j ranks candidates by their predicted local training effect, while $a_j(S)$ selects among them based on the additional target coverage of x_j given the examples already selected.

Given a synthetic-data budget B, TATC starts from $S_0 = \varnothing$ and adds one candidate at a time. At step s, it considers the $\min\{\kappa s, N\}$ highest-ranked candidates under h_j, excluding selected examples, and selects the one with the largest coverage gain $a_j(S_{s-1})$. Thus, the shortlist expands with s, while the coverage score changes with the selected set. We use $\kappa = 2$ in all experiments. The procedure stops after B selections and returns S_B for subsequent fine-tuning. Implementation details are provided in Appendix D.3.

Algorithm 1: Training-Aware Target Coverage

Require: Synthetic candidates $\{(x_j, y_j)\}_{j=1}^N$, predicted local training effects $\{h_j\}_{j=1}^N$, budget B, expansion factor κ

$S_0 \leftarrow \varnothing$; rank candidates by decreasing h_j
for $s = 1, \ldots, B$ **do**
  $\mathcal{E}_s \leftarrow$ top $\min\{\kappa s, N\}$ candidates by h_j, excluding S_{s-1}
  $j_s \leftarrow \arg\max_{j \in \mathcal{E}_s} a_j(S_{s-1})$
  $S_s \leftarrow S_{s-1} \cup \{j_s\}$
end for

Ensure: Selected synthetic set S_B

Relation to the linear theory and scope.
TATC carries the theory's distinction between input information and response error into nonlinear fine-tuning: $a_j(S)$ measures target-weighted prediction-variance reduction, while h_j evaluates the local training effect of the complete example on a labeled real probe. Note that the two scores are not on a common scale: h_j defines the shortlist and $a_j(S)$ selects within it. Thus, we provide a fixed-checkpoint approximation, where h_j is computed once at $\boldsymbol{\theta}_0$ and $a_j(S)$ changes with the selected set. See details in Appendix D.1.

5 Experiments

Experimental setup. We evaluate Qwen2.5-Math-1.5B-Instruct on GSM8K. A LoRA adapter (Hu et al., 2022) with rank 16 is trained for two epochs on 2,048 real examples, producing one checkpoint $\boldsymbol{\theta}_0$ shared by every method. A disjoint set of 512 labeled real examples serves as the target probe, and a further 1,024 examples form the holdout set used only for evaluation. The candidate pool contains $N = 4{,}096$ deduplicated trajectories from OpenR1-Math-220k (Hugging Face, 2025), with at most one trajectory per source problem; 65% of the responses are verified as correct. Each selected subset is trained for one epoch from $\boldsymbol{\theta}_0$ with an effective batch size of 16.

For TATC, we average the hidden states from the final transformer layer over prompt tokens to obtain 1,536-dimensional representations, then reduce them to 64 dimensions using principal component analysis (PCA) fitted on the 2,048 real-training examples. The same projection is applied to the real probe and synthetic candidates.

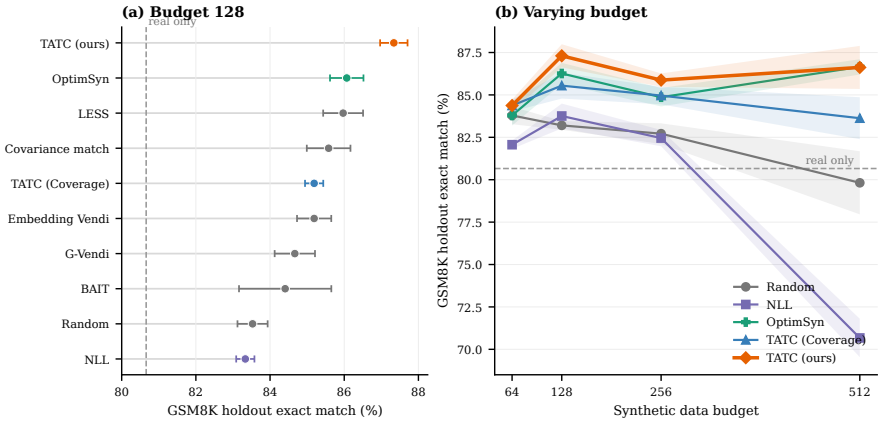

Figure 3: **(a)** Exact match at $B = 128$ on the GSM8K holdout. The dashed line marks the checkpoint trained on real data. **(b)** Exact match across selection budgets. Each method selects nested subsets, with each subset trained independently from the same checkpoint. Error bars show one standard deviation over three independent training runs.

Setup and baselines. We evaluate synthetic data selection at budget $B \in \{64, 128, 256, 512\}$, fine-tuning each selected subset from the same checkpoint and evaluating exact match on the GSM8K holdout. Baselines include Random, negative log-likelihood (NLL), Embedding Vendi (Friedman & Dieng, 2023), G-Vendi (Jung et al., 2025), LESS (Xia et al., 2024), OptimSyn (Fan et al., 2026), Covariance Matching (Rezaei et al., 2026), and BAIT (Ash et al., 2021). We additionally include a variant of TATC that removes the training-effect component, TATC (Coverage), serving as an ablation to test whether target coverage alone is sufficient. Baseline implementations and further TATC details are provided in Appendices C and D.

Results. At $B = 128$, TATC achieves 87.34% exact match accuracy, exceeding the strongest competitor, OptimSyn (86.07%), and the TATC (Coverage) ablation (85.19%); the 2.15-point ablation gap demonstrates the added value of the training-effect component (Figure 3a). This advantage over coverage alone persists across larger budgets and reaches 2.99 points at $B = 512$. TATC is higher than OptimSyn at budgets 64, 128, and 256; at 512, their means are effectively tied within run-to-run variation (Figure 3b). These results show that the benefit of the training-aware component emerges beyond the smallest budget and persists as more synthetic data are selected, while the gap between TATC and influence-based selection varies across budgets.

6 Conclusion

This work provides a novel perspective in understanding the value of synthetic data: the usefulness of an additional data point is not merely an intrinsic distribution property of this data point, but also depends on what the learner already has and what additional information this example can contribute. The linear model analysis makes this dependence explicit, showing that the same synthetic data can be valuable, redundant, or even harmful depending on the information and errors already present. This perspective also clarifies why continuously increasing the amount of synthetic data will eventually provide diminishing benefit, or even become detrimental.

Our findings also provide a practical basis for moving from understanding synthetic data to selecting data for LLM fine-tuning. TATC follows this insight by considering both how an example affects training and how it complements the data already available. Our experiments demonstrate its effectiveness for selecting synthetic mathematical reasoning data. More broadly, synthetic data pipelines should move beyond evaluating examples in isolation and consider their values in the context of the existing data and the downstream tasks for which they are used.

AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and editing, literature organization, code development and debugging, and preparation of figures and tables. They were also used to review the exposition of theoretical arguments and experimental methodology. The authors independently checked the mathematical derivations, cited sources, experimental code and results. Generative AI output was not treated as evidence or independent validation, and the authors take responsibility for the final content.

ETHICS STATEMENT

This work uses public mathematical reasoning and standard classification datasets and released model weights, and involves no human subjects or intentionally collected private or personally identifying data. Generated responses may contain systematic errors, and the proposed selection method does not replace factual verification, safety evaluation, or appropriate data governance. Our experiments do not establish the suitability of the method for high-stakes or safety-critical applications.

REPRODUCIBILITY STATEMENT

The code and data used in our experiments are available at https://anonymous.4open.science/r/TATC-syn-F61E/.

REFERENCES

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations*, 2024.

Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8927–8939, 2021.

Yang Ba, Michelle V. Mancenido, and Rong Pan. Fill in the gaps: Model calibration and generalization with synthetic data. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17211–17225. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.955. URL https://aclanthology.org/2024.emnlp-main.955/.

Powei Chang, Jinpeng Zhang, Bowen Chen, Chenyu Wang, Chenlu Guo, Yixing Zhang, Yukang Gao, JianXiang Xiang, Yue Gao, Chaoqun Sun, Yiyi Chen, and Dongying Kong. SPICE: Submodular penalized information–conflict selection for efficient large language model training. In *International Conference on Learning Representations*, 2026.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. AlpaGasus: Training a better alpaca with fewer data. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FdVXgSJhvz.

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

Zhiting Fan, Ruizhe Chen, Tianxiang Hu, Ru Peng, Zenan Huang, Haokai Xu, Yixin Chen, Jian Wu, Junbo Zhao, and Zuozhu Liu. OptimSyn: Influence-guided rubrics optimization for synthetic data generation. In *International Conference on Learning Representations*, 2026.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*, 2024. URL https://arxiv.org/abs/2406.20094.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023. URL https://arxiv.org/abs/2306.11644.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

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 *International Conference on Learning Representations*, 2022.

Hugging Face. Open r1: A fully open reproduction of DeepSeek-R1, January 2025. URL https://github.com/huggingface/open-r1.

Jaehun Jung, Seungju Han, Ximing Lu, Skyler Hallinan, David Acuna, Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, and Yejin Choi. Prismatic synthesis: Gradient-based data diversification boosts generalization in LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3032.

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, 2021. doi: 10.1609/aaai.v35i9.16988.

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, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*, 2024. URL https://arxiv.org/abs/2402.13064.

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*, 2024. URL https://openreview.net/forum?id=BTKAeLqLMw.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Sören Mindermann, Jan M. Brauner, Muhammed T. Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N. Gomez, Adrien Morisot, Sebastian Farquhar, and Yarin Gal. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15630–15649, 2022. URL https://proceedings.mlr.press/v162/mindermann22a.html.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 271–278, 2004. doi: 10.3115/1218955.1218990. URL https://aclanthology.org/P04-1035/.

Parham Rezaei, Filip Kovačević, Francesco Locatello, and Marco Mondelli. High-dimensional analysis of synthetic data selection. In *International Conference on Learning Representations*, 2026.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024.

Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 11817–11831. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.findings-emnlp.791. URL https://aclanthology.org/2023.findings-emnlp.791/.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

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, 2024.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CfXh93NDgH.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned LLMs with nothing. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Pnk7vMbznK.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*, 2024. URL https://arxiv.org/abs/2409.12122.

Bingji Yi, Qiyuan Liu, Yuwei Cheng, and Haifeng Xu. Escaping model collapse via synthetic data verification: Near-term improvements and long-term convergence. In *International Conference on Learning Representations*, 2026.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. MetaMath: Bootstrap your own mathematical questions for large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=N8N0hgNDRt.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1126.

A Mathematical Proofs in Linear Theory

Let $\boldsymbol{X}_r \in \mathbb{R}^{n\times d}$ and $\boldsymbol{X}_s \in \mathbb{R}^{m\times d}$ be the real and synthetic feature matrices. All feature and noise variables have finite second moments. The label model is

$$\boldsymbol{y}_r = \boldsymbol{X}_r\boldsymbol{\theta}^\star + \boldsymbol{\varepsilon}_r, \qquad \boldsymbol{y}_s = \boldsymbol{X}_s\boldsymbol{\theta}^\star + \boldsymbol{d}_s + \boldsymbol{\varepsilon}_s,$$

where, conditional on the feature matrices,

$$\mathbb{E}[\boldsymbol{\varepsilon}_r] = \mathbf{0}, \quad \mathbb{E}[\boldsymbol{\varepsilon}_s] = \mathbf{0}, \quad \mathrm{Cov}(\boldsymbol{\varepsilon}_r) = \boldsymbol{\Omega}_r, \quad \mathrm{Cov}(\boldsymbol{\varepsilon}_s) = \boldsymbol{\Omega}_s, \quad \mathrm{Cov}(\boldsymbol{\varepsilon}_r, \boldsymbol{\varepsilon}_s) = \mathbf{0}.$$

The systematic label error $\boldsymbol{d}_s$ is nonrandom. Selected synthetic subsets and candidate indices are fixed, or chosen independently of the realized noise innovations; expectations below are conditional on the relevant feature matrices. The least-squares estimator and downstream second-moment matrix are

$$\widehat{\boldsymbol{\theta}}_s = \arg\min_{\boldsymbol{\theta}} \left\{ \tfrac{1}{2}\|\boldsymbol{X}_r\boldsymbol{\theta} - \boldsymbol{y}_r\|_2^2 + \tfrac{1}{2}\|\boldsymbol{X}_s\boldsymbol{\theta} - \boldsymbol{y}_s\|_2^2 \right\},$$
$$\boldsymbol{Q} = \mathbb{E}_{X\sim P_{\mathrm{test}}}[\boldsymbol{\phi}(X)\boldsymbol{\phi}(X)^\top] \succeq \mathbf{0}.$$

For any parameter estimate $\widehat{\boldsymbol{\theta}}$, its downstream risk is

$$\mathcal{R}_Q(\widehat{\boldsymbol{\theta}}) = \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_{\boldsymbol{Q}}^2 := (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\top \boldsymbol{Q}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star).$$

Finally, define

$$\boldsymbol{A} = \boldsymbol{X}_r^\top\boldsymbol{X}_r + \boldsymbol{X}_s^\top\boldsymbol{X}_s, \qquad \boldsymbol{u} = \boldsymbol{X}_s^\top\boldsymbol{d}_s, \qquad \boldsymbol{N} = \boldsymbol{X}_r^\top\boldsymbol{\Omega}_r\boldsymbol{X}_r + \boldsymbol{X}_s^\top\boldsymbol{\Omega}_s\boldsymbol{X}_s.$$

A.1 Proof of Proposition 1

Because $\boldsymbol{A}$ is nonsingular, the normal equations for the least-squares problem above have the unique solution

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_s &= \boldsymbol{A}^{-1}(\boldsymbol{X}_r^\top\boldsymbol{y}_r + \boldsymbol{X}_s^\top\boldsymbol{y}_s)\\ &= \boldsymbol{\theta}^\star + \boldsymbol{A}^{-1}\boldsymbol{u} + \boldsymbol{A}^{-1}(\boldsymbol{X}_r^\top\boldsymbol{\varepsilon}_r + \boldsymbol{X}_s^\top\boldsymbol{\varepsilon}_s).\end{aligned}$$

Write $\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{u}$ and $\boldsymbol{z} = \boldsymbol{A}^{-1}(\boldsymbol{X}_r^\top\boldsymbol{\varepsilon}_r + \boldsymbol{X}_s^\top\boldsymbol{\varepsilon}_s)$. The conditional noise assumptions give

$$\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}_r, \boldsymbol{X}_s] = \mathbf{0}, \qquad \mathrm{Cov}(\boldsymbol{z} \mid \boldsymbol{X}_r, \boldsymbol{X}_s) = \boldsymbol{A}^{-1}\boldsymbol{N}\boldsymbol{A}^{-1}.$$

Substituting $\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}^\star = \boldsymbol{b} + \boldsymbol{z}$ into the downstream-risk definition and taking the conditional expectation yields

$$\begin{aligned}\mathbb{E}[\mathcal{R}_Q(\widehat{\boldsymbol{\theta}}_s) \mid \boldsymbol{X}_r, \boldsymbol{X}_s] &= \boldsymbol{b}^\top\boldsymbol{Q}\boldsymbol{b} + 2\boldsymbol{b}^\top\boldsymbol{Q}\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}_r, \boldsymbol{X}_s] + \mathbb{E}[\boldsymbol{z}^\top\boldsymbol{Q}\boldsymbol{z} \mid \boldsymbol{X}_r, \boldsymbol{X}_s]\\ &= \boldsymbol{b}^\top\boldsymbol{Q}\boldsymbol{b} + \mathrm{tr}(\boldsymbol{Q}\,\mathrm{Cov}(\boldsymbol{z} \mid \boldsymbol{X}_r, \boldsymbol{X}_s))\\ &= (\boldsymbol{A}^{-1}\boldsymbol{u})^\top\boldsymbol{Q}(\boldsymbol{A}^{-1}\boldsymbol{u}) + \mathrm{tr}(\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{N}\boldsymbol{A}^{-1}).\end{aligned}$$

The cross term vanishes because $\mathbb{E}[\boldsymbol{z} \mid \boldsymbol{X}_r, \boldsymbol{X}_s] = \mathbf{0}$, and the last equality uses $\mathbb{E}[\boldsymbol{z}^\top\boldsymbol{Q}\boldsymbol{z} \mid \boldsymbol{X}_r, \boldsymbol{X}_s] = \mathrm{tr}(\boldsymbol{Q}\,\mathrm{Cov}(\boldsymbol{z} \mid \boldsymbol{X}_r, \boldsymbol{X}_s))$ for a conditionally zero-mean vector. This proves Equation (1). □

A.2 Proof of Corollary 2

For a selected set S, write

$$\boldsymbol{A}_r = \boldsymbol{X}_r^\top \boldsymbol{X}_r, \qquad \boldsymbol{A}_S = \boldsymbol{X}_S^\top \boldsymbol{X}_S, \qquad \boldsymbol{A} = \boldsymbol{A}_r + \boldsymbol{A}_S.$$

Under $\boldsymbol{\Omega}_r = \sigma^2 \boldsymbol{I}$ and $\boldsymbol{\Omega}_S = (\sigma^2 + \tau^2)\boldsymbol{I}$, the noise matrix in the downstream-risk decomposition becomes

$$\boldsymbol{N} = \sigma^2 \boldsymbol{A}_r + (\sigma^2 + \tau^2)\boldsymbol{A}_S = \sigma^2 \boldsymbol{A} + \tau^2 \boldsymbol{A}_S.$$

The real-only estimator is unbiased and, because $\boldsymbol{A}_r \succ \boldsymbol{0}$, its conditional risk is

$$\mathbb{E}[\mathcal{R}_Q(\widehat{\boldsymbol{\theta}}_r) \mid \boldsymbol{X}_r, \boldsymbol{X}_S] = \sigma^2 \operatorname{tr}(\boldsymbol{Q}\boldsymbol{A}_r^{-1}).$$

Proposition 1, the expression for $\boldsymbol{N}$ above, and $\boldsymbol{u} = \boldsymbol{X}_S^\top \boldsymbol{d}_S$ give the selected-set risk

$$\begin{aligned}\mathbb{E}[\mathcal{R}_Q(\widehat{\boldsymbol{\theta}}_S) \mid \boldsymbol{X}_r, \boldsymbol{X}_S] = (\boldsymbol{A}^{-1}\boldsymbol{X}_S^\top \boldsymbol{d}_S)^\top \boldsymbol{Q}(\boldsymbol{A}^{-1}\boldsymbol{X}_S^\top \boldsymbol{d}_S) \\ + \sigma^2 \operatorname{tr}(\boldsymbol{Q}\boldsymbol{A}^{-1}) + \tau^2 \operatorname{tr}(\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{A}_S\boldsymbol{A}^{-1}).\end{aligned}$$

Subtracting the selected-set risk from the real-only risk gives Equation (2). □

A.3 Proof of Proposition 3

Let

$$\boldsymbol{B} = \boldsymbol{A}_r^{-1/2}\boldsymbol{A}_S\boldsymbol{A}_r^{-1/2} = \boldsymbol{R}\boldsymbol{M}\boldsymbol{R}^\top, \quad \boldsymbol{M} = \operatorname{diag}(\mu_1, \ldots, \mu_d), \quad \boldsymbol{D} = \boldsymbol{A}_r^{-1/2}\boldsymbol{Q}\boldsymbol{A}_r^{-1/2}.$$

Then

$$\boldsymbol{A} = \boldsymbol{A}_r^{1/2}(\boldsymbol{I} + \boldsymbol{B})\boldsymbol{A}_r^{1/2}, \qquad \boldsymbol{A}^{-1} = \boldsymbol{A}_r^{-1/2}(\boldsymbol{I} + \boldsymbol{B})^{-1}\boldsymbol{A}_r^{-1/2}.$$

Using cyclic invariance of the trace,

$$\begin{aligned}\operatorname{tr}[\boldsymbol{Q}(\boldsymbol{A}_r^{-1} - \boldsymbol{A}^{-1})] &= \operatorname{tr}\big[\boldsymbol{D}\{\boldsymbol{I} - (\boldsymbol{I} + \boldsymbol{B})^{-1}\}\big] \\ &= \operatorname{tr}\left[\boldsymbol{R}^\top \boldsymbol{D}\boldsymbol{R} \operatorname{diag}\left(\frac{\mu_i}{1+\mu_i}\right)_{i=1}^{d}\right] \\ &= \sum_{i=1}^{d} (\boldsymbol{r}_i^\top \boldsymbol{D}\boldsymbol{r}_i)\frac{\mu_i}{1+\mu_i}.\end{aligned}$$

Thus the information term equals $\sigma^2 \sum_i w_i\mu_i/(1+\mu_i)$, where $w_i = \boldsymbol{r}_i^\top \boldsymbol{D}\boldsymbol{r}_i$.

The excess-noise term is handled in the same basis:

$$\begin{aligned}\operatorname{tr}(\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{A}_S\boldsymbol{A}^{-1}) &= \operatorname{tr}\big[\boldsymbol{D}(\boldsymbol{I} + \boldsymbol{B})^{-1}\boldsymbol{B}(\boldsymbol{I} + \boldsymbol{B})^{-1}\big] \\ &= \sum_{i=1}^{d} w_i \frac{\mu_i}{(1+\mu_i)^2}.\end{aligned}$$

These two trace decompositions require no alignment between $\boldsymbol{D}$ and $\boldsymbol{B}$, because the matrices multiplying $\boldsymbol{R}^\top \boldsymbol{D}\boldsymbol{R}$ are diagonal and a trace uses only its diagonal.

It remains to evaluate systematic error. Suppose $\boldsymbol{d}_S = \boldsymbol{X}_S\boldsymbol{\delta}$ and define $\boldsymbol{\eta} = \boldsymbol{R}^\top \boldsymbol{A}_r^{1/2}\boldsymbol{\delta}$. Then

$$\begin{aligned}\boldsymbol{A}^{-1}\boldsymbol{X}_S^\top \boldsymbol{d}_S &= \boldsymbol{A}^{-1}\boldsymbol{A}_S\boldsymbol{\delta} \\ &= \boldsymbol{A}_r^{-1/2}\boldsymbol{R} \operatorname{diag}\left(\frac{\mu_i}{1+\mu_i}\right)_{i=1}^{d} \boldsymbol{\eta}.\end{aligned}$$

Let $\boldsymbol{T} = \operatorname{diag}(\mu_i/(1+\mu_i))$. The squared $\boldsymbol{Q}$-norm of this vector is

$$\boldsymbol{\eta}^\top \boldsymbol{T}\boldsymbol{R}^\top \boldsymbol{D}\boldsymbol{R}\boldsymbol{T}\boldsymbol{\eta}.$$

When $\boldsymbol{R}^\top \boldsymbol{D}\boldsymbol{R}$ is diagonal, its i-th diagonal entry is w_i, so this expression reduces to

$$\sum_{i=1}^{d} w_i \eta_i^2 \left(\frac{\mu_i}{1+\mu_i}\right)^2.$$

Combining the information, excess-noise, and systematic-error expressions in Corollary 2 proves Equation (3). □

A.4 Proof of Corollary 4

For $\mu_i > 0$ and $w_i > 0$, the sign of the i-th contribution in Equation (3) is the sign of

$$\sigma^2 \frac{\mu_i}{1+\mu_i} - \tau^2 \frac{\mu_i}{(1+\mu_i)^2} - \eta_i^2 \frac{\mu_i^2}{(1+\mu_i)^2}.$$

Multiplying by the positive quantity $(1+\mu_i)^2/\mu_i$ shows that this contribution is positive exactly when

$$\sigma^2(1+\mu_i) - \tau^2 - \eta_i^2 \mu_i > 0.$$

Rearranging and dividing by $\mu_i > 0$ gives

$$\eta_i^2 < \sigma^2 + \frac{\sigma^2 - \tau^2}{\mu_i},$$

which proves the corollary. □

A.5 Proof of Corollary 5

Hold w_i and η_i unchanged while varying $\mu_i \geq 0$. For $w_i > 0$, it is enough to maximize the bracketed contribution

$$f(\mu) = \sigma^2 \frac{\mu}{1+\mu} - \tau^2 \frac{\mu}{(1+\mu)^2} - \eta_i^2 \left(\frac{\mu}{1+\mu}\right)^2.$$

Direct differentiation gives

$$f'(\mu) = \frac{(\sigma^2 - \tau^2) + \mu(\sigma^2 + \tau^2 - 2\eta_i^2)}{(1+\mu)^3}.$$

The denominator is positive. If $\sigma^2 > \tau^2$ and $2\eta_i^2 > \sigma^2 + \tau^2$, the numerator starts positive and decreases linearly, crossing zero once at

$$\mu_i^\star = \frac{\sigma^2 - \tau^2}{2\eta_i^2 - \sigma^2 - \tau^2}.$$

Consequently, f increases before this point and decreases after it, so the point is the unique global maximizer on $[0, \infty)$.

For the remaining cases, write $s = \sigma^2$, $t = \tau^2$, and $r = \eta_i^2$. Equivalently, $f(\mu) = \mu[(s-t) + \mu(s-r)]/(1+\mu)^2$, with $f(0) = 0$ and $\lim_{\mu\to\infty} f(\mu) = s - r$. If $s < t$ and $r \geq s$, or if $s = t$ and $r > s$, this numerator is strictly negative for $\mu > 0$, so zero is the unique maximizer. If $s = t = r$, then $f \equiv 0$. If $s > t$ and $2r \leq s + t$, the derivative above is positive for every finite μ, so the supremum is approached at infinity. If $s \leq t$ and $r < s$, the derivative is positive for $\mu > 0$ when $s = t$; when $s < t$, it changes once from negative to positive, after which f increases to the positive limit $s - r$. Thus this case also has no finite maximizer. These conclusions concern the actual directional contribution when $w_i > 0$; if $w_i = 0$, that contribution is identically zero and every amount is optimal. □

A.6 Proof of Proposition 6

Take a direction $\boldsymbol{v}$ satisfying the conditions in the proposition. For any scalar c, define a second parameterization

$$\widetilde{\boldsymbol{\theta}}^\star = \boldsymbol{\theta}^\star + c\boldsymbol{v}, \qquad \widetilde{\boldsymbol{d}}_S = \boldsymbol{d}_S - c\boldsymbol{X}_S\boldsymbol{v}.$$

The real conditional mean is unchanged because

$$\boldsymbol{X}_r\widetilde{\boldsymbol{\theta}}^\star = \boldsymbol{X}_r\boldsymbol{\theta}^\star + c\boldsymbol{X}_r\boldsymbol{v} = \boldsymbol{X}_r\boldsymbol{\theta}^\star.$$

The synthetic conditional mean is also unchanged:

$$\boldsymbol{X}_S\widetilde{\boldsymbol{\theta}}^\star + \widetilde{\boldsymbol{d}}_S = \boldsymbol{X}_S(\boldsymbol{\theta}^\star + c\boldsymbol{v}) + \boldsymbol{d}_S - c\boldsymbol{X}_S\boldsymbol{v} = \boldsymbol{X}_S\boldsymbol{\theta}^\star + \boldsymbol{d}_S.$$

The noise laws are identical, so the two parameterizations induce exactly the same distribution for all observed real and synthetic labels. Nevertheless, they assign different target conditional means in direction $\boldsymbol{v}$, which matters downstream because $\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v} > 0$.

For completeness, an admissible one-direction instance shows that this ambiguity can reverse the sign of synthetic-data value. Take the model space to be the span of $\boldsymbol{v}$, let $\boldsymbol{X}_r\boldsymbol{v} = \mathbf{0}$, $\boldsymbol{X}_S\boldsymbol{v} \neq \mathbf{0}$, and set both noise covariances to zero. Define the real-only estimate by the Moore–Penrose, or equivalently minimum-norm, rule; its coefficient in this unobserved direction is zero. Since the scalar selected-set Gram matrix is positive, the selected-set least-squares estimate is unique. Give the observed synthetic label the conditional mean $a\boldsymbol{X}_S\boldsymbol{v}$, with $a \neq 0$. The two observationally equivalent worlds are

$$\begin{aligned} \mathcal{W}_+ : &\quad \boldsymbol{\theta}^\star = a\boldsymbol{v}, \quad \boldsymbol{d}_S = \mathbf{0}, \\ \mathcal{W}_- : &\quad \boldsymbol{\theta}^\star = -a\boldsymbol{v}, \quad \boldsymbol{d}_S = 2a\boldsymbol{X}_S\boldsymbol{v}. \end{aligned}$$

They have the same real and synthetic conditional means. Omitting common orthogonal components and random-noise terms, the real-only and selected-set risks in $\mathcal{W}_+$ are respectively $a^2\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v}$ and zero, so the value is positive. In $\mathcal{W}_-$, the corresponding risks are $a^2\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v}$ and $4a^2\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v}$, so the value is $-3a^2\boldsymbol{v}^\top\boldsymbol{Q}\boldsymbol{v} < 0$. No statistic of the observed data can know which world generated them.

The sign-reversal construction is existential and uses the additional conditions—a declared minimum-norm real-only rule, a nonsingular combined Gram matrix in the one-direction instance, and zero random noise—that make $\mathcal{V}(S)$ unambiguous. The observational-equivalence result itself needs none of these extra conditions. Observational equivalence alone does not assert a sign reversal for every fixed observed distribution. The proposition is complementary to the directional value results, which assume $\boldsymbol{A}_r \succ \mathbf{0}$; it is not an application of Equation (2) with an ordinary inverse. Its information-theoretic conclusion is that without labeled real evidence in a downstream-relevant direction, target signal and systematic label error cannot be separated, and admissible worlds can assign opposite signs to value. □

A.7 Proof of Theorem 7

The scalar variance ω_j in the theorem represents a fresh, conditionally zero-mean noise innovation. Formally, the rank-one update below requires this innovation to be uncorrelated, conditional on all feature matrices, with the real and already selected label noises. It also requires $\boldsymbol{A}$ to be positive definite, which makes $c_j > 0$. If a candidate noise is correlated with existing noises, the augmented $\boldsymbol{N}$ contains cross-covariance terms and Equation (5) needs corresponding additional terms.

For the real data together with the current selected set S, let

$$\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{u}, \qquad \boldsymbol{p}_j = \boldsymbol{A}^{-1}\boldsymbol{x}_j, \qquad c_j = 1 + \boldsymbol{x}_j^\top\boldsymbol{p}_j, \qquad e_j = \delta_j - \boldsymbol{x}_j^\top\boldsymbol{b}, \qquad q_j = \boldsymbol{p}_j^\top\boldsymbol{Q}\boldsymbol{p}_j.$$

Adding candidate j changes the three sufficient statistics to

$$\boldsymbol{A}_+ = \boldsymbol{A} + \boldsymbol{x}_j\boldsymbol{x}_j^\top, \qquad \boldsymbol{u}_+ = \boldsymbol{u} + \boldsymbol{x}_j\delta_j, \qquad \boldsymbol{N}_+ = \boldsymbol{N} + \omega_j\boldsymbol{x}_j\boldsymbol{x}_j^\top.$$

The Sherman–Morrison identity gives

$$\boldsymbol{A}_+^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{p}_j\boldsymbol{p}_j^\top}{c_j}, \qquad \boldsymbol{A}_+^{-1}\boldsymbol{x}_j = \frac{\boldsymbol{p}_j}{c_j}.$$

Using $\boldsymbol{A}^{-1}\boldsymbol{u} = \boldsymbol{b}$ and $\boldsymbol{p}_j^\top\boldsymbol{u} = \boldsymbol{x}_j^\top\boldsymbol{b}$, the new bias vector is

$$\begin{aligned} \boldsymbol{b}_+ &= \boldsymbol{A}_+^{-1}\boldsymbol{u}_+ \\ &= \left(\boldsymbol{A}^{-1} - \frac{\boldsymbol{p}_j\boldsymbol{p}_j^\top}{c_j}\right)(\boldsymbol{u} + \boldsymbol{x}_j\delta_j) \\ &= \boldsymbol{b} + \frac{\delta_j - \boldsymbol{x}_j^\top\boldsymbol{b}}{c_j}\boldsymbol{p}_j = \boldsymbol{b} + \frac{e_j}{c_j}\boldsymbol{p}_j. \end{aligned}$$

Hence the reduction in the squared-bias component of risk is

$$\boldsymbol{b}^\top\boldsymbol{Q}\boldsymbol{b} - \boldsymbol{b}_+^\top\boldsymbol{Q}\boldsymbol{b}_+ = -\frac{2e_j\boldsymbol{b}^\top\boldsymbol{Q}\boldsymbol{p}_j}{c_j} - \frac{e_j^2 q_j}{c_j^2}.$$

For the variance component, first expand the contribution from the existing noise matrix. Substituting the Sherman–Morrison update and using cyclic invariance of the trace gives

$$\begin{aligned} &\operatorname{tr}(\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{N}\boldsymbol{A}^{-1}) - \operatorname{tr}(\boldsymbol{Q}\boldsymbol{A}_+^{-1}\boldsymbol{N}\boldsymbol{A}_+^{-1}) \\ &\quad = \frac{2\boldsymbol{p}_j^\top\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{N}\boldsymbol{p}_j}{c_j} - \frac{q_j\boldsymbol{p}_j^\top\boldsymbol{N}\boldsymbol{p}_j}{c_j^2}. \end{aligned}$$

The candidate's own random-noise contribution to the new risk is

$$\omega_j \operatorname{tr}(\boldsymbol{Q}\boldsymbol{A}_+^{-1}\boldsymbol{x}_j\boldsymbol{x}_j^\top\boldsymbol{A}_+^{-1}) = \frac{\omega_j q_j}{c_j^2}.$$

Because marginal value is old risk minus new risk, adding the squared-bias change and the two variance changes above proves Equation (5).

To verify the clean homoscedastic specialization in Equation (6), suppose $\boldsymbol{N} = \sigma^2\boldsymbol{A}$, $\boldsymbol{b} = \boldsymbol{0}$, $\delta_j = 0$, and $\omega_j = \sigma^2$. Then

$$\boldsymbol{p}_j^\top\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{N}\boldsymbol{p}_j = \sigma^2 q_j, \qquad \boldsymbol{p}_j^\top\boldsymbol{N}\boldsymbol{p}_j = \sigma^2\boldsymbol{x}_j^\top\boldsymbol{A}^{-1}\boldsymbol{x}_j = \sigma^2(c_j - 1).$$

Substitution into Equation (5) leaves

$$\Delta_j(S) = \frac{2\sigma^2 q_j}{c_j} - \frac{\sigma^2 q_j(c_j - 1 + 1)}{c_j^2} = \frac{\sigma^2 q_j}{c_j}.$$

Thus the stated coverage marginal follows exactly.

More general specializations. Retain $\boldsymbol{N} = \sigma^2\boldsymbol{A}$ and $\boldsymbol{b} = \boldsymbol{0}$, so that $e_j = \delta_j$, but now let the candidate's systematic error δ_j and noise variance ω_j be arbitrary rather than zero and σ^2 respectively. As computed above,

$$\boldsymbol{p}_j^\top\boldsymbol{Q}\boldsymbol{A}^{-1}\boldsymbol{N}\boldsymbol{p}_j = \sigma^2 q_j, \qquad \boldsymbol{p}_j^\top\boldsymbol{N}\boldsymbol{p}_j = \sigma^2(c_j - 1).$$

Substituting into Equation (5), the two variance terms combine as

$$\frac{2\sigma^2 q_j}{c_j} - \frac{q_j\big(\sigma^2(c_j - 1) + \omega_j\big)}{c_j^2} = \frac{q_j\big(\sigma^2 c_j + \sigma^2 - \omega_j\big)}{c_j^2},$$

and the two bias terms reduce to $-\delta_j^2 q_j/c_j^2$ since $\boldsymbol{b} = \boldsymbol{0}$. Adding the two gives the general specialization

$$\Delta_j(S) = \frac{q_j}{c_j^2}\Big(\sigma^2 c_j + \sigma^2 - \omega_j - \delta_j^2\Big).$$

Three cases follow by substitution. A candidate carrying only systematic error ($\omega_j = \sigma^2$) contributes $\frac{q_j}{c_j^2}(\sigma^2 c_j - \delta_j^2)$; a candidate carrying only the excess noise of Corollary 2 ($\delta_j = 0,\ \omega_j = \sigma^2 + \tau^2$) contributes $\frac{q_j}{c_j^2}(\sigma^2 c_j - \tau^2)$; and a candidate carrying both ($\delta_j \neq 0,\ \omega_j = \sigma^2 + \tau^2$) contributes $\frac{q_j}{c_j^2}(\sigma^2 c_j - \tau^2 - \delta_j^2)$, mirroring the information-gain/noise-cost/bias-cost tradeoff of Corollary 2 at the level of a single candidate. Setting $\delta_j = 0$ and $\omega_j = \sigma^2$ in the first case recovers Equation (6). □

B Controlled Tests of the Linear Theory

Figure 2 uses fixed representations of real text and image inputs paired with controlled targets. Panel (a) compares the exact rank-one marginal with before–after least-squares fits under shared existing-data noise and fresh candidate noise. Panel (b) uses labeled probe data to select a pseudo-labeled-to-real ratio and compares that choice with the held-out test optimum.

In both panels, *real* identifies the source of an input, whereas *synthetic* identifies its training target. Panel (a) assigns controlled scalar targets to real text inputs; panel (b) assigns labels predicted by a separate linear model to real images. Qwen2.5 and ResNet-50 provide frozen representations, and all downstream models are ordinary least squares. Holding the candidate inputs fixed allows label error to change without also changing the information provided by the inputs.

B.1 Candidate value in text data

Data and representation. We use the 8,000-example training portion of the `SetFit/subj` subjectivity-classification dataset. Qwen2.5-0.5B is frozen and used only as a feature extractor: each text is truncated to 96 tokens, the final-layer hidden states of its nonpadding tokens are averaged, and the resulting representation has 896 coordinates.

For each of three random partitions, examples are assigned without replacement to the five disjoint roles in Table 1. PCA is fitted only on D_{rep}, retains 64 components, and is applied unchanged to every other role; an appended intercept gives $\boldsymbol{x} \in \mathbb{R}^{65}$.

Table 1: Data partitions for the text experiment.

Partition	Examples	Role
D_{rep}	1,500	Fit PCA and the reference linear target
D_{cal}	500	Estimate the residual variance scale
D_r	1,200	Define the real training information
D_q	1,000	Estimate the target moment $\widehat{\boldsymbol{Q}}$
$\mathcal{P}_s$	3,800	Supply candidate inputs

We map the binary labels to $y^{\pm} = 2y - 1$, fit the reference linear target $\widehat{\boldsymbol{\theta}}^{\star}$ on D_{rep}, and use its residual MSE on the disjoint D_{cal} set to define $\widehat{\sigma}^2$. This scale includes both residual variation and linear approximation error. Observed labels in the candidate pool are not used as candidate targets.

Controlled target construction. The 1,200 real inputs are always present. We sample a current set S of size $32, 64, 128, 256, 512$, or 768, together with one candidate $j \notin S$, from the 3,800-input pool. The real training targets follow $\boldsymbol{y}_r = \boldsymbol{X}_r\widehat{\boldsymbol{\theta}}^{\star} + \boldsymbol{\varepsilon}_r$, with zero-mean variance $\widehat{\sigma}^2$.

To construct systematic label error, we set the intercept coordinate of $\widehat{\boldsymbol{\theta}}^{\star}$ to zero and normalize the resulting direction $\boldsymbol{g}$ so that $\boldsymbol{g}^{\top}\widehat{\boldsymbol{Q}}\boldsymbol{g} = 1$. For $a \in \{0.5, 1, 2, 4, 8\}$, define

$$\boldsymbol{\gamma} = a\sqrt{\widehat{\sigma}^2}\boldsymbol{g}, \qquad \delta_j = \boldsymbol{x}_j^{\top}\boldsymbol{\gamma}.$$

Removing the intercept and normalizing under $\widehat{\boldsymbol{Q}}$ makes the systematic error feature-dependent, task-relevant, and comparable across partitions. The two matched contexts use

$$\boldsymbol{d}_S^{(-)} = -\boldsymbol{X}_S\boldsymbol{\gamma}, \quad \boldsymbol{u}_- = -\boldsymbol{A}_S\boldsymbol{\gamma}, \qquad \boldsymbol{d}_S^{(+)} = +\boldsymbol{X}_S\boldsymbol{\gamma}, \quad \boldsymbol{u}_+ = +\boldsymbol{A}_S\boldsymbol{\gamma}.$$

Their controlled training targets are $\boldsymbol{y}_S^{(\pm)} = \boldsymbol{X}_S\widehat{\boldsymbol{\theta}}^{\star} + \boldsymbol{d}_S^{(\pm)} + \boldsymbol{\varepsilon}_S$ and $y_j = \boldsymbol{x}_j^{\top}\widehat{\boldsymbol{\theta}}^{\star} + \delta_j + \xi_j$. A matched pair has the same $\boldsymbol{X}_r, \boldsymbol{X}_S, \boldsymbol{x}_j, \delta_j$, and therefore the same coverage score; only the sign of the error accumulated in S changes. The candidate can therefore cancel existing bias in the minus context and reinforce it in the plus context, directly varying $e_j = \delta_j - \boldsymbol{x}_j^{\top}\boldsymbol{A}^{-1}\boldsymbol{u}$ in Theorem 7.

The matched contexts use the same random-noise variance ($\tau^2 = 0$), so their difference comes only from systematic label error. For each partition, set size, and error magnitude, we sample four input sets and evaluate both signs, yielding 360 matched pairs.

Evaluation and results. For each context, the input-only comparator is the clean coverage marginal

$$s_{\mathrm{cov}}(j \mid S) = \widehat{\sigma}^2\frac{\boldsymbol{p}_j^{\top}\widehat{\boldsymbol{Q}}\boldsymbol{p}_j}{1 + \boldsymbol{x}_j^{\top}\boldsymbol{p}_j}, \qquad \boldsymbol{p}_j = \boldsymbol{A}^{-1}\boldsymbol{x}_j,$$

and the error-aware score is Equation (5). Independently of either score, we solve the least-squares model before and after adding j, using shared existing-data noise and fresh candidate noise, and record the reduction in $\widehat{\boldsymbol{Q}}$-weighted risk. Each context averages 3,000 independent label-noise trials.

Coverage alone barely distinguishes beneficial from harmful candidates, whereas the complete marginal predicts both the sign and magnitude of the observed risk reduction (Table 2).

Table 2: Agreement with observed risk reductions in the text experiment.

Score	Metric	Value
Coverage	AUROC	0.519
Complete marginal	AUROC	0.99
Complete marginal	Correlation	0.99
Complete marginal	Sign agreement	99.2%

B.2 SYNTHETIC-DATA AMOUNT IN IMAGE DATA

Data and representation. We use CIFAR-10 images (Krizhevsky, 2009) represented by frozen ResNet-50 penultimate-layer features (He et al., 2016). Images are resized to 224×224, normalized with ImageNet channel statistics, and passed through an ImageNet-1K-V2 pretrained ResNet-50, producing 2,048-dimensional features.

For each of four random partitions, the 50,000 training images are assigned without replacement to D_r (3,000 real-training images), D_q (3,000 labeled probe images), D_g (12,000 images used to fit label models), $\mathcal{P}_s$ (24,000 candidate images), and 8,000 unused images. The official 10,000-image test split remains separate. PCA is fitted only on D_r, retains 128 components, and is applied unchanged to $D_q, D_g, \mathcal{P}_s$, and the test set. At each amount, an unregularized multiclass least-squares model is fitted jointly to the real and pseudo-labeled examples.

Pseudo-label reliability. Each label model assigns its argmax class as a deterministic pseudo-label. We vary reliability using five models fitted on D_g: three use all 12,000 labeled examples or the first 400 or 150 examples, and two use all 12,000 examples after removing four or 12 task-relevant PCA coordinates. Figure 2(b) displays the model fitted on all examples, the model fitted on 400 examples, and the model with 12 coordinates removed. No candidate or test labels are used to fit these models. For the coordinate-removed condition, directions are ranked by explained variance weighted by their reference-classifier coefficients, concentrating errors in task-relevant directions. Conditional on the fitted model, pseudo-labels are deterministic, so their errors enter the systematic-error term rather than adding random label noise.

For each pseudo-label model, the downstream model is fitted at pseudo-labeled-to-real ratios

$$0,\ 0.05,\ 0.10,\ 0.20,\ 0.40,\ 0.80,\ 1.00,\ 1.50,\ 2.00,\ 3.00,\ 4.00,\ 6.00,\ 8.00.$$

All conditions use the same candidate order, and larger amounts extend smaller subsets. The black cross in Figure 2(b) marks the test-MSE-minimizing grid point, and the gray interval contains ratios within 1% of that minimum. The test split is used only for this retrospective reference.

Probe-based risk estimate. The labeled probe set is divided into two disjoint halves. Let $\widehat{\boldsymbol{\beta}}_a$ and $\widehat{\boldsymbol{\beta}}_b$ be 128-by-10 least-squares class-score models fitted to the two 1,500-example halves. The target moment $\widehat{\boldsymbol{Q}}$ uses all 3,000 probe inputs, and $\widehat{\sigma}^2$ is the average per-coordinate residual MSE with the $n/(n-d)$ correction, where $n = 1{,}500$ and $d = 128$. For a proposed pseudo-labeled set, define

$$\boldsymbol{B}_S = \boldsymbol{X}_S^\top \boldsymbol{Y}_S, \qquad \boldsymbol{U}_a = \boldsymbol{B}_S - \boldsymbol{A}_S \widehat{\boldsymbol{\beta}}_a, \qquad \boldsymbol{U}_b = \boldsymbol{B}_S - \boldsymbol{A}_S \widehat{\boldsymbol{\beta}}_b.$$

With $\boldsymbol{A} = \boldsymbol{A}_r + \boldsymbol{A}_S$, the estimated systematic-error contribution is

$$\widehat{R}_{\text{bias}} = \max\left\{0, \operatorname{tr}\left[(\boldsymbol{A}^{-1}\boldsymbol{U}_a)^\top \widehat{\boldsymbol{Q}} (\boldsymbol{A}^{-1}\boldsymbol{U}_b)\right]\right\},$$

and the random-noise contribution is

$$\widehat{R}_{\text{var}} = 10\widehat{\sigma}^2 \operatorname{tr}(\widehat{\boldsymbol{Q}}\boldsymbol{A}^{-1}\boldsymbol{A}_r\boldsymbol{A}^{-1}).$$

The cross product avoids the positive inflation from squaring a single noisy probe estimate; truncation enforces nonnegativity. The variance term assumes isotropic residual variation in the trusted real labels; deterministic pseudo-labels add no conditional random-noise covariance. We select the ratio minimizing $\widehat{R}_{\text{bias}} + \widehat{R}_{\text{var}}$. Dropping the bias term gives the input-only comparator, which selects the full 24,000-candidate pool.

Amount-selection results. The probe estimate tracks the decrease in useful data amount as label reliability falls (Table 3). Across all five conditions, predicted and test-optimal amounts have the same ordering (Kendall $\tau = 1.00$), and every prediction across the four partitions falls within the 1% test-optimal plateau. For the most reliable label model, the test optimum on the evaluated grid occurs at the largest ratio.

Table 3: Mean label accuracy and ratios of synthetic to real data for the three conditions in Figure 2(b).

Label model	Label accuracy	Predicted ratio	Test optimum ratio
12,000 training examples	0.889	7.0	8.0
400 training examples	0.836	0.90	0.95
12 coordinates removed	0.604	0	0

This experiment evaluates amount selection when multiple feature directions contribute, complementing the characterization of a single direction in Corollary 5.

C Baseline Implementations

C.1 Representations and scores used by the selectors

Prompt features. For every real, probe, and candidate prompt, we average the 1,536-dimensional final-layer hidden states across prompt tokens to obtain one prompt representation. Principal-component analysis is fitted on the real-training representations and projects all partitions to 64 dimensions. The same projection is used by Embedding Vendi, Covariance Matching, TATC (Coverage), and the coverage stage of TATC.

Gradient features. For every complete example, we compute the gradient of the response loss with respect to the trainable LoRA parameters. To make per-example storage and matrix operations feasible, the gradient is mapped by a signed CountSketch to 512 dimensions and then projected to 64 dimensions by principal-component analysis fitted on the real-training gradients. G-Vendi and BAIT use these projected gradient features. LESS uses the 512-dimensional sketched gradients before PCA.

Quality score. Candidate negative log-likelihood is the mean token loss of its response under the checkpoint trained on real data, excluding prompt tokens. Lower values indicate responses that the checkpoint considers more likely.

C.2 Selection rules

Random. Random samples candidate indices uniformly without replacement.

NLL. NLL sorts candidates by increasing response negative log-likelihood and returns the first B.

Embedding Vendi. Let $\boldsymbol{z}_j$ be the unit-normalized prompt feature. For a nonempty set S, the Vendi score is

$$\mathrm{VS}(S) = \exp\left[-\sum_k \lambda_k \log \lambda_k\right], \qquad \{\lambda_k\}_k = \mathrm{eig}\left(\frac{1}{|S|}\sum_{j\in S} \boldsymbol{z}_j \boldsymbol{z}_j^\top\right),$$

where zero eigenvalues make zero contribution. A farthest-first traversal initializes a set of size B. We then run 32 one-swap rounds; each round tests 256 proposed replacements and accepts the proposal with the largest positive Vendi-score improvement.

G-Vendi. G-Vendi applies the same objective and swap procedure to unit-normalized projected LoRA-gradient features. It replaces prompt representations with candidate-induced training gradients while retaining the same set optimization.

LESS. Let $\boldsymbol{g}_j$ be the candidate gradient feature and let $\overline{\boldsymbol{g}}_q$ be the mean gradient on the labeled real probe set. The score is

$$s_{\mathrm{LESS}}(j) = \frac{\boldsymbol{g}_j^\top \overline{\boldsymbol{g}}_q}{\|\boldsymbol{g}_j\|_2 \|\overline{\boldsymbol{g}}_q\|_2}.$$

Candidates are ranked by this target-gradient alignment score at the shared checkpoint.

OptimSyn. For the fixed candidate pool, we adapt OptimSyn's optimizer-aware influence score. For each candidate, we form the AdamW update difference $\Delta\boldsymbol{\theta}_j$ described in Appendix D.2 and score its first-order effect on the real probe:

$$s_{\mathrm{OptimSyn}}(j) = -\frac{\boldsymbol{g}_q^\top \Delta\boldsymbol{\theta}_j}{\|\Delta\boldsymbol{\theta}_j\|_2}.$$

Covariance Matching. Normalize every prompt feature to unit Euclidean norm and define the real-training second moment

$$\boldsymbol{C}_r = \frac{1}{|\mathcal{D}_r|}\sum_{x\in\mathcal{D}_r} \boldsymbol{z}_x\boldsymbol{z}_x^\top.$$

Starting from an empty set, the method greedily adds the candidate that minimizes

$$\left\| \frac{1}{|S|+1}\left(\sum_{i\in S}\boldsymbol{z}_i\boldsymbol{z}_i^\top + \boldsymbol{z}_j\boldsymbol{z}_j^\top\right) - \boldsymbol{C}_r \right\|_F^2.$$

This baseline matches the second moment of the real-training inputs and does not use labeled probe responses.

BAIT. BAIT operates on the 64-dimensional projected gradient features. It initializes its information matrix with the real-training gradients and uses the candidate gradient distribution as the A-optimal target. A diagonal stabilizer equal to 10^{-3} times the average information-matrix diagonal is added before inversion. Following the forward–backward construction, BAIT greedily grows to $2B$ candidates and then removes examples until B remain, choosing each addition or removal by its A-optimal trace change.

TATC (Coverage). This ablation initializes $\widehat{\boldsymbol{A}}_0$ and $\widehat{\boldsymbol{Q}}$ exactly as TATC but removes the training-effect filter. At every step it searches the full remaining pool and adds

$$\arg\max_{j\notin S} \frac{\boldsymbol{\phi}_j^\top \widehat{\boldsymbol{A}}_S^{-1}\widehat{\boldsymbol{Q}}\widehat{\boldsymbol{A}}_S^{-1}\boldsymbol{\phi}_j}{1+\boldsymbol{\phi}_j^\top \widehat{\boldsymbol{A}}_S^{-1}\boldsymbol{\phi}_j}.$$

It isolates the contribution of the response-aware training-effect component beyond target coverage.

D Implementation of TATC in Supervised Fine-tuning

D.1 From the Linear Analysis to TATC

The linear results motivate two components of selection: the target coverage added by a candidate input and the training score of its complete prompt and response. TATC computes both at a checkpoint trained on real data. It ranks candidates using a quadratic score of their update differences and updates coverage as examples are selected.

For each prompt, we average the hidden states from the model's final transformer layer over prompt tokens to obtain a 1,536-dimensional representation. PCA is fitted only to the 2,048 real-training representations and then applied, without refitting, to the probe, holdout, and candidate prompts. Let $\boldsymbol{\phi}_j \in \mathbb{R}^{64}$ be the resulting candidate feature and define

$$\widehat{\boldsymbol{A}}_0 = \rho\boldsymbol{I} + \sum_{x\in\mathcal{D}_r}\boldsymbol{\phi}(x)\boldsymbol{\phi}(x)^\top,$$

$$\widehat{\boldsymbol{Q}} = \frac{1}{|\mathcal{D}_q|}\sum_{x\in\mathcal{D}_q}\boldsymbol{\phi}(x)\boldsymbol{\phi}(x)^\top,$$

$$\widehat{\boldsymbol{A}}_S = \widehat{\boldsymbol{A}}_0 + \sum_{i\in S}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top.$$

The numerical stabilizer is

$$\rho = 10^{-3}\frac{\mathrm{tr}(\sum_{x\in\mathcal{D}_r}\boldsymbol{\phi}(x)\boldsymbol{\phi}(x)^\top)}{64}.$$

The target coverage gain

$$a_j(S) = \frac{\boldsymbol{\phi}_j^\top \widehat{\boldsymbol{A}}_S^{-1} \widehat{\boldsymbol{Q}} \widehat{\boldsymbol{A}}_S^{-1} \boldsymbol{\phi}_j}{1 + \boldsymbol{\phi}_j^\top \widehat{\boldsymbol{A}}_S^{-1} \boldsymbol{\phi}_j}$$

is exactly the Sherman–Morrison decrease in $\mathrm{tr}(\widehat{\boldsymbol{Q}} \widehat{\boldsymbol{A}}_S^{-1})$ within the fixed prompt representation. This algebraic identity is exact; using the criterion as a proxy for downstream performance after LLM fine-tuning is an experimental-design approximation.

The quadratic training score is

$$h_j = -\boldsymbol{g}_q^\top \Delta\boldsymbol{\theta}_j - \tfrac{1}{2} \Delta\boldsymbol{\theta}_j^\top \boldsymbol{H}_q \Delta\boldsymbol{\theta}_j,$$

In Equation (7), $\Delta\boldsymbol{\theta}_j = \boldsymbol{U}_j - \boldsymbol{U}_\varnothing$ is the difference between the candidate and zero gradient AdamW displacements from the same checkpoint and saved optimizer state. The probe gradient $\boldsymbol{g}_q$ measures alignment with decreasing loss, while the diagonal matrix $\boldsymbol{H}_q$ supplies a quadratic penalty on displacement. The score provides a fixed ranking for the expanding shortlist. Appendix D.2 defines the quadratic model and its computation.

Table 4 summarizes the correspondence and its limitations.

Table 4: Correspondence between the linear analysis and TATC.

Linear quantity or role	TATC quantity	Status	Correct interpretation
Real information $\boldsymbol{A}_r$	$\widehat{\boldsymbol{A}}_0$	Approximation	Summarizes prompt directions already represented by real data; $\rho\boldsymbol{I}$ is numerical stabilization.
Target moment $\boldsymbol{Q}$	$\widehat{\boldsymbol{Q}}$	Sample estimate	Estimates target relevance only in the checkpoint prompt representation.
Selected information $\boldsymbol{A}_S$	$\sum_{i \in S} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top$	Exact in representation	Each selected input adds a rank-one prompt-feature contribution.
Clean coverage marginal q_j/c_j	$a_j(S)$	Exact in representation	Exact decrease in the downstream-weighted inverse-trace criterion.
Systematic label error $(\boldsymbol{d}_S, \boldsymbol{u})$	Response-dependent gradient and AdamW update	Approximation	For squared loss, response error changes the gradient by $-(\boldsymbol{x}^\top \boldsymbol{\delta})\boldsymbol{x}$; in an LLM the full response changes a nonlinear gradient.
Exact marginal value $\Delta_j(S)$	Quadratic training score h_j	Surrogate	Ranks update differences using a quadratic model of probe loss at the shared checkpoint.
Residual interaction e_j with current bias	No explicit scalar counterpart	Omitted interaction	The response score is computed once, while set dependence enters through $a_j(S)$.

D.2 Quadratic training score

Let $\boldsymbol{g}_j$ be the per-example gradient of the response-token loss and let $b_{\text{eff}} = 16$ be the effective batch size used to scale the candidate gradient. We insert $\boldsymbol{g}_j / b_{\text{eff}}$ into the saved AdamW state after fine-tuning on real data. For coordinate k, with saved moments $(m_{0,k}, v_{0,k})$, the candidate and zero-gradient counterfactual moments are

$$m_{j,k} = \beta_1 m_{0,k} + (1-\beta_1) g_{j,k}/b_{\text{eff}}, \qquad v_{j,k} = \beta_2 v_{0,k} + (1-\beta_2)(g_{j,k}/b_{\text{eff}})^2,$$
$$m_{\varnothing,k} = \beta_1 m_{0,k}, \qquad v_{\varnothing,k} = \beta_2 v_{0,k}.$$

After the usual AdamW bias corrections, let

$$U_k(m, v) = -\alpha \frac{\widehat{m}_k}{\sqrt{\widehat{v}_k} + \epsilon} - \alpha\lambda\theta_{0,k}$$

denote the full parameter displacement at $\boldsymbol{\theta}_0$, where α is the learning rate and λ is the weight decay coefficient. Let $\boldsymbol{U}_j$ and $\boldsymbol{U}_\varnothing$ collect these coordinatewise updates under the candidate and zero-gradient moments. Their difference is

$$\Delta\theta_{j,k} = U_k(m_j, v_j) - U_k(m_\varnothing, v_\varnothing).$$

The common weight decay displacement $-\alpha\lambda\boldsymbol{\theta}_0$ cancels in $\Delta\boldsymbol{\theta}_j$. The saved optimizer moments still affect this difference through the adaptive updates.

We compute the probe statistics using batches of four examples. Let $\{\mathcal{B}_b\}_{b=1}^M$ partition the probe examples across all scoring workers, let $n_b = |\mathcal{B}_b|$, and let $\ell_b(\boldsymbol{\theta})$ be the model's response token loss for batch $\mathcal{B}_b$. Set $w_b = n_b/|\mathcal{D}_q|$ and define

$$L_q(\boldsymbol{\theta}) = \sum_{b=1}^M w_b \ell_b(\boldsymbol{\theta}), \qquad \boldsymbol{g}_{q,b} = \nabla_{\boldsymbol{\theta}} \ell_b(\boldsymbol{\theta}_0),$$

$$\boldsymbol{g}_q = \sum_{b=1}^M w_b \boldsymbol{g}_{q,b}, \qquad [\boldsymbol{H}_q]_{kk} = \sum_{b=1}^M w_b [\boldsymbol{g}_{q,b}]_k^2.$$

The implementation squares each batch gradient before averaging across batches. We use the resulting positive semidefinite diagonal matrix as a curvature surrogate in the scoring model (Kunstner et al., 2019).

For a displacement $\boldsymbol{d}$ from $\boldsymbol{\theta}_0$, define the quadratic model

$$\widetilde{L}_q(\boldsymbol{d}) = L_q(\boldsymbol{\theta}_0) + \boldsymbol{g}_q^\top \boldsymbol{d} + \tfrac{1}{2}\boldsymbol{d}^\top \boldsymbol{H}_q \boldsymbol{d}.$$

Evaluating this model along the update difference gives

$$\begin{aligned} h_j &= \widetilde{L}_q(\mathbf{0}) - \widetilde{L}_q(\Delta\boldsymbol{\theta}_j) \\ &= -\boldsymbol{g}_q^\top \Delta\boldsymbol{\theta}_j - \tfrac{1}{2}\Delta\boldsymbol{\theta}_j^\top \boldsymbol{H}_q \Delta\boldsymbol{\theta}_j. \end{aligned}$$

The first term measures alignment of the update difference with decreasing probe loss. The second is a nonnegative penalty on displacement, since $\boldsymbol{H}_q \succeq 0$. The quadratic model scores $\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}_j$; the full AdamW step reaches $\boldsymbol{\theta}_0 + \boldsymbol{U}_j$. The proposition below bounds the loss change along $\Delta\boldsymbol{\theta}_j$.

Both probe statistics and candidate update differences are computed once at the shared checkpoint. Each candidate therefore has one fixed score during selection. The implementation accumulates both terms directly from the update difference and the probe statistics. CountSketch is used only for gradient features stored for the baselines.

Local justification of the score. The displacement $\Delta\boldsymbol{\theta}_j$ used for scoring is the update difference defined above. The following result relates the score in Equation (7) to the change in probe loss along this displacement and separates the error in the curvature matrix from the Taylor remainder.

Proposition 8 (Local loss approximation)**.** *Suppose L_q is twice continuously differentiable in a neighborhood of the segment from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}_j$, and its Hessian is M-Lipschitz there in spectral norm. Let*

$$\boldsymbol{H}_* = \nabla^2 L_q(\boldsymbol{\theta}_0), \qquad \epsilon_H = \|\boldsymbol{H}_* - \boldsymbol{H}_q\|_2,$$

and use $\boldsymbol{g}_q = \nabla L_q(\boldsymbol{\theta}_0)$ in the score h_j. Then

$$|L_q(\boldsymbol{\theta}_0) - L_q(\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}_j) - h_j| \le \frac{\epsilon_H}{2}\|\Delta\boldsymbol{\theta}_j\|_2^2 + \frac{M}{6}\|\Delta\boldsymbol{\theta}_j\|_2^3. \tag{8}$$

In particular, the displacement reduces probe loss whenever h_j exceeds the bound in Equation (8).

Proof. Write $\boldsymbol{d} = \Delta\boldsymbol{\theta}_j$. Taylor's theorem with a Lipschitz Hessian gives

$$L_q(\boldsymbol{\theta}_0 + \boldsymbol{d}) = L_q(\boldsymbol{\theta}_0) + \boldsymbol{g}_q^\top \boldsymbol{d} + \tfrac{1}{2}\boldsymbol{d}^\top \boldsymbol{H}_* \boldsymbol{d} + R(\boldsymbol{d}), \qquad |R(\boldsymbol{d})| \le \tfrac{M}{6}\|\boldsymbol{d}\|_2^3.$$

Subtracting the definition of h_j from the loss reduction yields

$$L_q(\boldsymbol{\theta}_0) - L_q(\boldsymbol{\theta}_0 + \boldsymbol{d}) - h_j = -\tfrac{1}{2}\boldsymbol{d}^\top(\boldsymbol{H}_* - \boldsymbol{H}_q)\boldsymbol{d} - R(\boldsymbol{d}).$$

The bound follows from $|\boldsymbol{d}^\top(\boldsymbol{H}_* - \boldsymbol{H}_q)\boldsymbol{d}| \le \epsilon_H\|\boldsymbol{d}\|_2^2$. □

The quadratic model therefore gives a local scoring rule whose accuracy depends on both the displacement size and the curvature approximation. The first term in the bound measures the error from the curvature surrogate supplied by the batch gradient construction; the cubic term controls the Taylor remainder.

D.3 SEQUENTIAL TATC SELECTION

We implement Algorithm 1 with $\kappa = 2$ and update $\widehat{\boldsymbol{A}}_S^{-1}$ after each selection using Sherman–Morrison:

$$\widehat{\boldsymbol{A}}_{S\cup\{j\}}^{-1} = \widehat{\boldsymbol{A}}_S^{-1} - \frac{\widehat{\boldsymbol{A}}_S^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top\widehat{\boldsymbol{A}}_S^{-1}}{1+\boldsymbol{\phi}_j^\top\widehat{\boldsymbol{A}}_S^{-1}\boldsymbol{\phi}_j}.$$

One run produces an ordering whose prefixes are the selected sets for every budget. Each prefix is then trained independently from the same checkpoint trained on real data and the same saved optimizer state; training does not continue from the preceding budget.

D.4 GSM8K PROTOCOL

Data. The task is GSM8K mathematical reasoning with Qwen2.5-Math-1.5B-Instruct. The disjoint real partitions contain 2,048 examples used to train the shared checkpoint, 512 labeled examples for the target probe, and 1,024 examples for the holdout set. The synthetic pool contains 4,096 OpenR1-Math-220k trajectories after deduplication to at most one trajectory per source problem. Of these trajectories, 2,662 pass the provided verifier and 1,434 do not; TATC never uses verifier status.

Model and training. LoRA adapters are attached to the attention q, k, v, o projections and the MLP gate, up, and down projections. Rank is 16, scaling is 32, and adapter dropout is zero. Fine-tuning on real data uses two epochs, learning rate 2×10^{-4}, AdamW weight decay .01, sequence length 2,304, and effective batch size 16. Every selected subset is trained for one epoch with the same learning rate, weight decay, effective batch size, and initial model and optimizer state. We evaluate budgets $B \in \{64, 128, 256, 512\}$ using nested prefixes of the selected ordering. Results are the mean and standard deviation over three independent training repetitions.

Evaluation. Exact match is computed after extracting the final numeric answer. Outputs that do not terminate within the generation allowance count as incorrect. The $B = 128$ comparison allows 1,536 generated tokens, whereas the budget sweep allows 3,072; absolute exact-match values are therefore not comparable across panels. We report the nontermination rate to distinguish answer accuracy from changes in response length.