Compact Example-Based Explanations for Language Models

Anonymous ACL submission

Abstract

Training data influence estimation methods quantify the contribution of training documents to a model's output, making them a promising source of information for example-based explanations. As humans cannot interpret thousands of documents, only a small subset of the training data can be presented as an explanation. Although the choice of which documents to include directly affects explanation quality, previous evaluations of such systems have largely ignored any selection strategies. To address this, we propose a novel *selection relevance score*, a retraining-free metric that quantifies how useful a set of examples is for explaining a model's output. We validate this score through fine-tuning experiments, confirming that it can predict whether a set of examples supports or undermines the model's predictions. Using this metric, we further show that common selection strategies often underperform random selection. Motivated by this finding, we propose a strategy that balances influence and representativeness, enabling better use of selection budgets than naively selecting the highest-ranking examples.

1 Introduction

Training data influence estimation methods, such as *influence functions* (Koh and Liang, 2017), estimate the contribution of individual training documents to a model's output. These estimates offer a promising source of information for creating example-based explanations for language models. However, training data influence estimates are not directly usable as explanations. Humans cannot meaningfully process thousands of documents, nor can retrieval-augmented generation systems designed to generate natural language explanations. To provide explanations of a human-interpretable size, existing systems rely on naive selection strategies. For example, they may choose the k highest-ranking examples from the influence estimate. This is problematic for two reasons. First, examples in the upper tail of the influence distribution tend to be globally influential outliers. These outliers are not necessarily the most relevant for the test instance (Barshan et al., 2020). Second, the highest-ranking training examples often contain redundant information (Bhatt et al., 2021). Strictly selecting only the most influential documents can yield diminishing returns, as these examples may offer little new information to users. Similarly, when selecting documents as a basis for generating natural language explanations, choosing outlier or redundant examples can undermine explanation faithfulness.

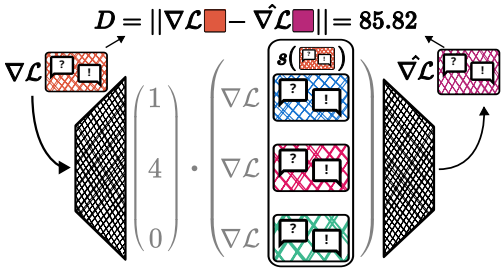


Figure 1: **Simplified scoring setup**. We score example-based explanations based on how well the test instance's loss gradient (left) can be reconstructed (right) as a linear combination of the selected examples' gradients.

Although example selection directly affects explanation quality, existing evaluations focus on the correctness of raw estimates or on testing the usefulness of the full explanation system with users. This is problematic as it can obscure flaws in the selection logic: a system may appear helpful even if the selected examples are not truly informative.

In this paper, we propose an evaluation score that **specifically targets the selection logic**. For this, we frame example-based explanation as a reconstruction task (Figure 1): specifically, we propose the measure of *selection relevance*, which quantifies how well the gradients of the selected examples reconstruct the gradient of the test instance they are meant to explain. Intuitively, selections truly

relevant to the test instance should yield lower reconstruction errors than sets of globally influential outliers or redundant examples. Notably, this score considers the relevance of the selected examples **in combination**. In contrast, existing approaches either assess individual examples in isolation or evaluate groups only in costly retraining-based experiments. Our setup is task- and estimation-method agnostic. It can also be easily adapted to additional explanation-specific selection constraints.

We validate our proposed selection relevance score by comparing it to an alternative notion of relevance derived from fine-tuning experiments: two fine-tuning-based metrics capture how training on the selection affects the original prediction. As neither notion of relevance provides a definitive ground truth, we assess their alignment through correlation analysis. We further demonstrate the utility of our score by evaluating various combinations of influence estimation methods and selection strategies. We show that common selection strategies often underperform random baselines and introduce a new strategy that more effectively leverages human-interpretable selection budgets than simply choosing the most influential examples. Our main contributions are as follows:[1]

(1) We propose the *selection relevance score*, a retraining-free selection quality score that evaluates the utility of a set of training documents for providing example-based explanations.
(2) We benchmark 3 influence estimation and 4 selection strategies, providing insights into why naive selection strategies are ineffective if paired with popular estimation methods.
(3) Using our score, we introduce a new strategy that better leverages small selection budgets.

2 Background and Related Work

Evaluating influence estimates. For influence estimates to be a useful source of information for explanations, they must accurately represent how individual training examples affect the model's behavior during training. Existing work on evaluating the *correctness* of influence rankings has largely followed two approaches: testing how well rankings correlate with results from leave-one-out (LOO) retraining (Park et al., 2023; Bae et al., 2022; Basu et al., 2021; K and Søgaard, 2021; Choe et al., 2024), or assessing performance in downstream data discovery tasks, such as data pruning (Koh and Liang, 2017). However, for our task, assessing correctness is insufficient, we target the *selection mechanism*, which is difficult to isolate from other components in downstream task evaluations.

Evaluating explanations. We observe that existing example-based explanation methods share a common motivation: they assume that an explanation can only faithfully represent model behavior if the selected examples are sufficiently relevant to the output being explained. For example, some approaches define relevance in terms of embedding similarity between the test instance and explanation elements (e.g., Zhang et al., 2021; Nematov et al., 2024b), or by checking if the selected instances share the label of the test instance (Hanawa et al., 2021). Still others verify whether training only on the selected instances can reproduce the outputs obtained from training on the full dataset (e.g., Gu et al., 2023), which also requires sufficient relevance to test instances. However, these measures either operate in embedding space (while ranking typically occurs in gradient space), rely on class labels (making them unsuitable for text generation tasks), or require retraining (which is intractable for LLMs). Our score avoids all of these limitations.

Locally vs. globally influential training data. Barshan et al. (2020) observe that influence functions tend to assign high influence to a small, consistent set of training examples across test instances, which frequently consists of outliers or mislabeled examples. Similarly, Hammoudeh and Lowd (2022) argue that popular influence estimation methods underestimate the influence of highly influential instances because confidently predicted examples have small gradient magnitudes. Additionally, Nematov et al. (2024b) find that examples that consistently rank highly across many test instances tend to have high loss. Based on the idea that showing outliers to users does not constitute good example-based explanations (Barshan et al., 2020), some works have attempted to adjust for the influence of globally influential examples in new influence estimation methods (Barshan et al., 2020; Nematov et al., 2024a; Hammoudeh and Lowd, 2022). However, none explicitly account for the difference between globally and locally relevant examples in their evaluations. In contrast, our work directly evaluates the *local relevance* of *selections*.

Reducing redundancy within local explanations. Observing that the most influential examples of-

[1] We release our code at anonymous.4open.science/r/ceelm.

ten form clusters in feature space, both Bhatt et al. (2021) and Nematov et al. (2024a) consider example diversity alongside influence. Both utilize objectives of the form $\max_{S \in \mathcal{D}, |S|=k} \; I(S) + D(S)$, where D quantifies diversity and I measures influence. However, this strategy can favor outliers (as Bhatt et al. note), which we argue is problematic given the behavior discussed above. In this work, we instead aim to increase *representativeness*.

Prediction-constrained influence. Influence functions sometimes poorly predict re-training effects for deep models not trained to convergence (Bae et al., 2022; Basu et al., 2021; Grosse et al., 2023). Bae et al. (2022) partially attribute this to the fact that the re-trained model's parameters and predictions are not guaranteed to remain sufficiently close to those of the original model after LOO retraining. They propose an alternative measure to LOO-influence, which measures the effect of removing examples while penalizing deviations from the model's original predictions.They find that influence functions correlate more strongly with this *prediction-constrained* measure than with the LOO ground truth, and argue that it is more useful in practice than measuring LOO effects directly. Prediction-constrained influence appears compatible with our task: local explanations are meaningful only as long as the model is not altered so much that it deviates from the original prediction (Ribeiro et al., 2016; Guidotti et al., 2018). However, there is no guarantee that this is considered during example selection: the set may instead support an entirely different decision. We therefore also evaluate how our score aligns with this alternative view of relevance (Section 3.3).

3 Methodology

We study the task of selecting training examples for example-based explanation, where we aim to explain a model's prediction ($\hat{y} = f_\theta(x)$) through a small set of training examples. The criteria for a good example-based explanation are application-specific (Barshan et al., 2020). In this paper, we focus on explanations that present *training examples supporting the prediction of interest* (e.g., Barshan et al., 2020; Hanawa et al., 2021; Yeh et al., 2018), rather than alternative setups such as selecting counterfactual examples that would change the model's prediction. The training data influence estimation methods we use produce a ranking over the entire training dataset ($\phi(f, x, \hat{y}) \in \mathbb{R}^{||D||}$). However, we evaluate selections S of human-interpretable sizes $k = \{1, 5, 10, 25\}$. In this section, we introduce a *selection relevance score* $\xi^{\mathcal{SR}}$. For illustration, we consider a simple selection method s that chooses the k highest-ranked training examples according to some influence estimate.

3.1 Evaluation Setup

Our score quantifies how *relevant* the chosen training examples are for explaining the model output $\hat{y}$ of interest. We specifically consider relevance at the selection level: Given the small selection budget, we argue that the most influential examples from the training data are not necessarily the most useful from the user's perspective. Providing users with a document that is highly influential but largely redundant with other examples is less likely to improve their understanding of the model. The selection must represent a sufficiently diverse set of training behaviors to adequately explain the output. We operationalize this notion of relevance as a gradient reconstruction task:

Encoding. Our scoring approach uses model loss gradients, in line with popular influence estimators that also rely exclusively on gradients. We aim to reconstruct the loss gradient of a given test instance, $\nabla\mathcal{L}'$, using a linear combination, $\hat{\nabla\mathcal{L}'} = At$, where $t \in \mathbb{R}^k$ denotes a set of learned coefficients to be introduced in Section 3.2, and A is a matrix of gradients of the k selected training examples: $A = [\, \nabla\mathcal{L}_1 \; \nabla\mathcal{L}_2 \; \cdots \; \nabla\mathcal{L}_k \,]$, $\nabla\mathcal{L}_i \in \mathbb{R}^d$.

Selection relevance score. We interpret the reconstruction error as a measure of the relevance of the selected examples for a given test instance. Let $\mathcal{D} = \{\mathbf{y}^{(n)}\}_{n=1}^N \subset \Sigma^*$ be the dataset used in training, drawn from the ground-truth distribution p_{θ^*}: $\mathbf{y}^{(n)} \sim p_{\theta^*}, \quad n = 1, \ldots, N$. Let Y be a random variable representing training examples uniformly sampled from $\mathcal{D}$. We define a random vector $G : \Omega \to \mathbb{R}^d$ representing gradients of the model's loss function $G = \nabla_\theta \mathcal{L}(Y, \hat{Y}; \theta)$. Each realization $G(\omega_0)$, $\omega_0 \in \Omega$ corresponds to the gradient of the loss with respect to the model parameters for a single training example. A single $G(\omega_0)$ can be reconstructed by $\hat{G}(\omega_0) = At_0$. The reconstruction error is then defined as $D(\omega_0) \in \mathbb{R}^d$: $D(\omega_0) = G(\omega_0) - \hat{G}(\omega_0) = G(\omega_0) - At_0$. Extending this reconstruction to all realizations of $G(\omega)$, $\omega \in \Omega$, we approximate $G(\omega), \omega \in \Omega$ as $\hat{G}(\omega) = At_\omega, \quad D(\omega) = G(\omega) - \hat{G}(\omega)$. The coefficients t_ω are optimized individually for each

$G(\omega), \omega \in \Omega$. We define the *selection quality score* $\xi^{\mathcal{SR}}$ as the ratio of the expected squared gradient norm to expected squared reconstruction error:

$$\xi^{\mathcal{SR}} = \frac{\mathbb{E}_\omega[\|G(\omega)\|^2]}{\mathbb{E}_\omega[\|D(\omega)\|^2]} = \frac{\mathbb{E}_\omega[\|G(\omega)\|^2]}{\mathbb{E}_\omega[\|G(\omega) - At_\omega\|^2]} \quad (1)$$

Interpretation. We report this score in decibels ($10 \log_{10} \xi^{\mathcal{SR}}$) throughout to ease visualisation. Our score $\xi^{\mathcal{SR}}$ quantifies how well the chosen training examples can reconstruct the gradient of a test instance, capturing their informativeness. Values below 0 dB (i.e., $\xi^{\mathcal{SR}} < 1$ in absolute scale) indicate worse approximation than the trivial baseline $||G - \vec{0}||$, meaning the selected examples fail to convey relevant information. Values above 0 dB ($\xi^{\mathcal{SR}} > 1$) suggest non-redundant information.

3.2 Scoring Models

In practice, multiple t can minimize the approximation error, among them the least squares solution $t^* = (A^\top A)^{-1} A^\top \nabla \mathcal{L}'$. However, we impose two additional constraints on the solution: First, we impose a non-negativity constraint on the coefficients t. This ensures the weights are non-negative, preventing cancellations between irrelevant examples. More generally, we prefer that all examples either provide evidence for or evidence against the prediction (i.e., all either increase or decrease loss, or have a strong or weak effect on θ). Second, by enforcing $\sum t = 1$, we effectively normalize the importance scores across selections. This allows users to interpret t as a vector of relative importance within the selected set. To account for both constraints while preserving an analytical solution, we first compute an unconstrained least-squares solution, and then project it onto the unit simplex to obtain a normalized, non-negative vector. We consider alternative constraints in Appendix C.

3.3 Fine-Tuning-Based Validation Experiment

As discussed in Section 2, local explanations may no longer hold if the model is altered in a way that causes it to produce substantially different outputs. However, there is no guarantee that this aspect is considered during example selection: the selected set could inadvertently support a different decision. To assess whether selections with high selection relevance scores also sufficiently support the original predictions, we report Spearman correlations with two fine-tuning-based metrics:

Prediction support ξ^+. For each test instance-, model- and estimator pairing, we train for one step on the documents in the selection (LR=1e-5; one batch) and measure the impact on the original prediction. Specifically, we calculate the likelihood of the originally generated output $\log p(y|x; \cdot)$ for a model that was trained for one additional step on the selected examples S ($\log p(y|x; \theta_{+S})$), and for a model trained on a random subset R of the training data ($\log p(y|x; \theta_{+R})$; $|R| = |S|$). The intuition is that fine-tuning on truly informative examples should increase the likelihood of the original output more than training on a random set. The following score indicates whether S supports the original output more than a random selection R:

$$\xi^+(S) = \log p(y \mid x; \theta_{+S}) - \log p(y \mid x; \theta_{+R}) \quad (2)$$

Prediction shift ξ^{JSD}. Additionally, we measure Jensen-Shannon divergences to capture whether S induces a larger shift in the model's full predicted distribution $p(y \mid x; \cdot)$ than a random subset would:

$$\xi^{JSD}(S) = \frac{\mathrm{JSD}\left(p(y|x;\theta), p(y|x;\theta_{+S})\right)}{\mathrm{JSD}\left(p(y|x;\theta), p(y|x;\theta_{+R})\right)} \quad (3)$$

3.4 Experimental Setup

We use DataInf (approximates influence functions; Kwon et al., 2024), and LESS (gradient similarity; Xia et al., 2024) to obtain influence estimates. Both methods restrict influence estimation to the LoRA layers. Additionally, we include a BM25 baseline that retrieves training examples based on their token overlap with the test instance. We discuss parameter choices in detail in Appendix A.

Naive selection. Given an influence estimate $\phi(f, x, \hat{y}) \in \mathbb{R}^{||D||}$, we select k documents according to the following strategies: lowest influence scores (*most helpful:* Koh and Liang, 2017), largest influence scores (*most harmful*), largest absolute scores (*most influential*), and lowest absolute scores (*least influential*). Additionally, we report the mean performance of 5 random selections for each k.

Coverage-aware selection. To systematically evaluate our scoring method, we implement a selection strategy that optimizes for example coverage. Our goal is to select a more representative, less redundant set of training examples than the naive selection method. We therefore treat selection as a *facility location problem* (see e.g., Krause and Golovin, 2014). Facility location functions are typically defined as $F(S) = \sum_{i \in V} \max_{j \in S} \mathrm{sim}_{ij}$,

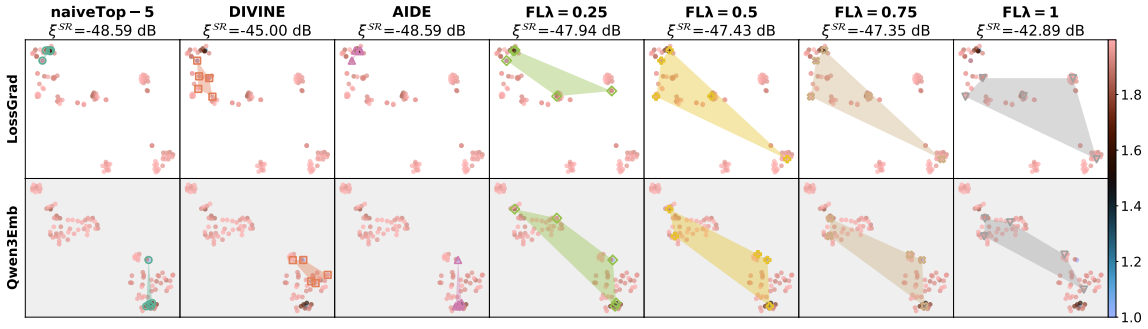


Figure 2: **Case Study**. t-SNE visualizations of the 100 *most-influential* examples by DataInf in gradient and embedding spaces (Qwen/Qwen3-Embedding-0.6B: Zhang et al., 2025). Our selections (FL) improve coverage.

where $S \subseteq V$ is the selected subset. The marginal gain of adding an element $j \in V$ to the current set S is $\Delta(j \mid S) = F(S \cup \{j\}) - F(S)$. We extend this formulation to incorporate influence scores:

$$\Delta_\lambda(j \mid S) = \frac{(\Delta(j|S)+1)^\lambda}{c_j^{1-\lambda}}, \quad (4)$$

where c_j is a normalized cost score based on the training data influence of element j. Setting $\lambda = 1$ performs a purely coverage-based selection, while $\lambda = 0$ is identical to the naive selection by influence scores. We implement selection via a custom optimizer in the `apricot` library (Schreiber et al., 2020): we greedily select the example with the highest $\Delta_\lambda(j \mid S)$ among the top-100 examples in the naive ranking until the budget is exhausted.

Diversity-based selection. We reimplement DIVINE (Bhatt et al., 2021) as its code is not publicly available, and remove elements from AIDE (Nematov et al., 2024a) that are only applicable to classification tasks. See Appendix B for details.

Datasets and models. We include models from three families: Olmo2 (allenai/OLMo-2-0425-1B: OLMo et al., 2025), Llama 3.2 (meta-llama/Llama-3.2-1B: Dubey et al., 2024), and Qwen 2.5 (Qwen/Qwen2.5-0.5B: Yang et al., 2024). We fine-tune for one epoch using LoRA (Hu et al., 2022) on the full *Tülu3* (allenai/tulu-3-sft-olmo-2-mixture-0225: Lambert et al., 2025) instruction-fine tuning dataset, see Appendix I for hyperparameters and benchmark results. For training data attribution, we randomly sample 10% (86.6k) of the examples as the set to attribute to, and a disjoint set of 1,000 test instances to explain.

4 Results

We first present a case study to illustrate the behavior of different selection methods. We then benchmark influence estimation methods and selection pairings, before finally analyzing our score in a fine-tuning-based validation experiment.

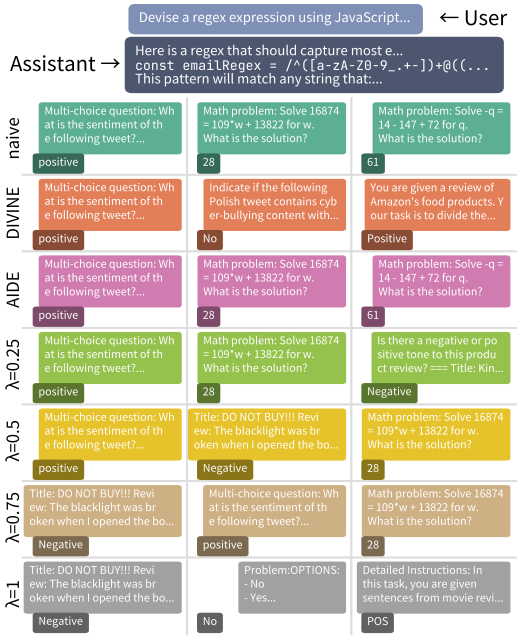


Figure 3: **Case Study**. First 3 documents per selection (k=5, DataInf, most inf.). Full figure in Appendix G.

4.1 Case Study

We present selections for a single test instance derived from a DataInf influence estimate for the Olmo2 model. In Figure 2, we plot t-SNE visualizations of the 100 most-influential examples and highlight the selections for a budget of k=5. Color indicates influence scores rescaled to selection costs in the range $[1, m]$ using min-max normalization, where higher influence corresponds to lower cost.

Facility location-based strategies improve coverage over naive selection in both gradient- and embedding spaces. Consistent with the intended behavior of our scoring setup, these strategies also

achieve higher selection relevance $\xi^{\mathcal{SR}}$. We additionally provide the first three examples for each method in text form in Figure 3 (see Appendix G for the full selection). The naive selection and AIDE include examples with highly similar prompts. Only two of these redundant examples are included when $\lambda = 0.25$, and only one appears when $\lambda = \{0.5, 0.75\}$. The purely coverage-based selection $\lambda = 1$ retains none of the examples selected by the naive selection. For this particular test instance, AIDE produces the same selection as the naive strategy. DIVINE appears to prioritize diversity more aggressively than AIDE in our setup. Note that both require manual hyperparameter selection, which makes their behavior difficult to interpret on a per-example basis (Appendix B).

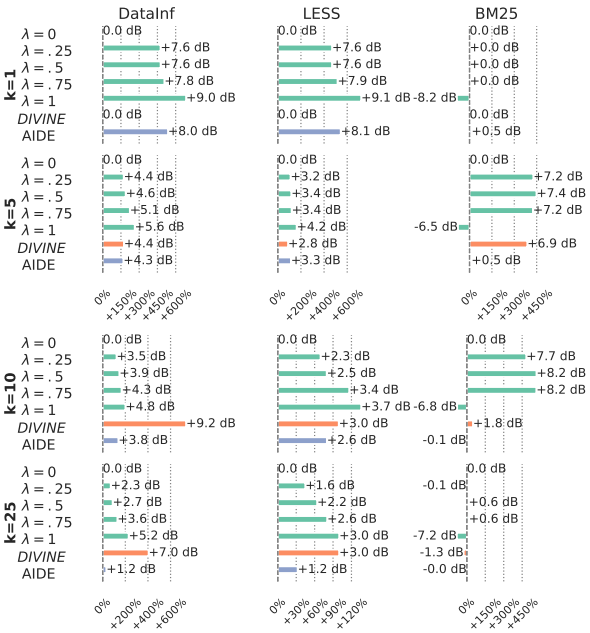


Figure 4: **Improvement over naive selection.** Relative increase over naive selection in per-cent and dB. Green: our facility location-based selections at different λ.

4.2 Selection Quality Scoring

Table 1 reports selection quality scores for different selection strategies at a fixed budget of $k = 10$. To avoid redundancy, we report only the *most* and *least influential* strategies for BM25, since BM25 scores are strictly positive. At this budget, all selections derived from gradient-based influence estimates feature an average selection relevance below 0 dB, indicating that they fail to provide sufficiently relevant examples. Only strategies that select the least influential examples outperform the baseline random selection. For selections based on BM25 rankings, only *most inf.* strategies and AIDE, which select the most influential examples (= highest token overlap for BM25) achieve selection relevance above 0 dB. Random, DIVINE and *least influential* selections do not. We provide a theoretical perspective on this in Section 5.

To assess whether this ranking holds across other selection budgets, we also compute an overall AUC score by aggregating the results for $k = \{1, 5, 10, 25\}$, reported in Appendix F. Rankings differ primarily across facility location settings that use the same selection strategy but different values of λ. Selecting the least influential examples remains the most effective strategy.

Improving over naive selection. Figure 4 shows the relative improvements achieved through facility location-based selection compared to naively selecting the highest-ranking examples (corresponding to $\lambda = 0$). We observe a clear improvement for the two gradient-based estimators: our strategy increases selection relevance, on average, for all selection methods except DIVINE at $k = 1$.

For BM25, selection relevance decreases when $\lambda = 1$, as expected, since selection is then based purely on coverage in gradient space, whereas it previously relied on token overlap. Gains are small for $k = \{1, 25\}$ compared to the gradient-based estimators; still, several strategies at $k = \{5, 10\}$ show substantial improvements.

4.3 Fine-Tuning-Based Validation

The purpose of this experiment is to examine whether our notion of *selection relevance* aligns with an alternative notion of relevance derived from fine-tuning behavior. Selection relevance $\xi^{\mathcal{SR}}$ measures how well a selected set of examples can reconstruct the gradient of a test instance, whereas the fine-tuning–based metrics *prediction support* ξ^{+} and *prediction shift* ξ^{JSD} measure whether training on the selection supports the model's original prediction or causes it to deviate from it. As neither notion constitutes a ground-truth definition of relevance, we evaluate through correlation analysis.

Sanity check. To ensure that our fine-tuning parameters are appropriate and that our measurements do not merely reflect random noise, we first run an experiment in which S contains only the test instance. We find that, in 98.49% of cases, the log-likelihood of a test instance increases more when fine-tuning on the instance itself than when fine-tuning on a random example. Similarly,

most inf.	-22.87 dB
most harmful	-22.50 dB
most helpful	-21.52 dB
AIDE	-20.70 dB
FL most inf. $\lambda = .25$	-20.65 dB
FL most harmful $\lambda = .25$	-20.34 dB
FL most inf. $\lambda = .5$	-20.26 dB
FL most inf. $\lambda = .75$	-19.99 dB
FL most harmful $\lambda = .5$	-19.94 dB
FL most helpful $\lambda = .25$	-19.75 dB
FL most harmful $\lambda = .75$	-19.63 dB
FL most inf. $\lambda = 1$	-19.54 dB
FL most helpful $\lambda = .5$	-19.49 dB
FL most helpful $\lambda = .75$	-19.13 dB
FL most harmful $\lambda = 1$	-18.91 dB
DIVINE most helpful	-18.69 dB
FL most helpful $\lambda = 1$	-18.23 dB
DIVINE most inf.	-17.31 dB
DIVINE most harmful	-17.09 dB
random	-2.57 dB
DIVINE least inf.	-0.14 dB
least inf.	-0.14 dB
FL least inf. $\lambda = .25$	-0.13 dB
FL least inf. $\lambda = .5$	-0.13 dB
FL least inf. $\lambda = 1$	-0.13 dB
FL least inf. $\lambda = .75$	-0.13 dB

DataInfEstimator

most inf.	-22.28 dB
most helpful	-21.81 dB
most harmful	-21.79 dB
FL most inf. $\lambda = .25$	-20.56 dB
AIDE	-20.39 dB
FL most inf. $\lambda = .5$	-20.21 dB
FL most harmful $\lambda = .25$	-20.16 dB
FL most helpful $\lambda = .25$	-20.09 dB
FL most helpful $\lambda = .5$	-19.78 dB
FL most inf. $\lambda = .75$	-19.64 dB
DIVINE most inf.	-19.63 dB
DIVINE most harmful	-19.38 dB
FL most harmful $\lambda = .5$	-19.30 dB
FL most inf. $\lambda = 1$	-19.28 dB
FL most harmful $\lambda = .75$	-19.25 dB
FL most helpful $\lambda = .75$	-19.22 dB
FL most harmful $\lambda = 1$	-19.07 dB
DIVINE most helpful	-18.80 dB
FL most helpful $\lambda = 1$	-18.42 dB
random	-2.57 dB
DIVINE least inf.	-0.15 dB
least inf.	-0.14 dB
FL least inf. $\lambda = .25$	-0.13 dB
FL least inf. $\lambda = .5$	-0.13 dB
FL least inf. $\lambda = 1$	-0.13 dB
FL least inf. $\lambda = .75$	-0.13 dB

LESSEstimator

random	-2.57 dB
DIVINE least inf.	-2.51 dB
FL least inf. $\lambda = 1$	-2.41 dB
FL least inf. $\lambda = .75$	-1.57 dB
FL least inf. $\lambda = .5$	-1.25 dB
FL least inf. $\lambda = .25$	-1.09 dB
least inf.	-1.01 dB
FL most inf. $\lambda = 1$	15.50 dB
AIDE	22.08 dB
most inf.	22.24 dB
DIVINE most inf.	23.77 dB
FL most inf. $\lambda = .25$	29.44 dB
FL most inf. $\lambda = .75$	29.99 dB
FL most inf. $\lambda = .5$	30.00 dB

BM25Estimator

Label	Selection	Effect
Most helpful	Most negative	Loss likely increases if selection removed.
Most harmful	Most positive	Loss likely decreases if selection removed.
Most infl.	Large absolute	Strongest impact on parameter update.
Least infl.	Small absolute	Weakest impact on parameter update.

Naive selection logics. See Appendix E for additional explanation.

Table 1: **Results with a selection budget of $k = 10$.** Facility location based selections outperform naive selection of the most influential-, harmful-, and helpful examples from the gradient-based influence estimates.

in 96.28% of cases, fine-tuning on the instance leads to a larger Jensen–Shannon divergence in the model's full predicted distribution $p(y \mid x; \cdot)$.

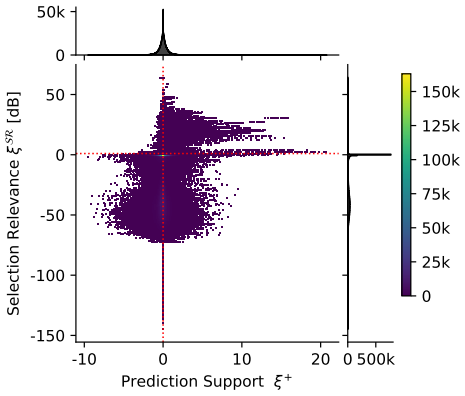


Figure 5: **Fine-tuning based validation**. Relationship between prediction support ξ^+ and selection relevance.

Correlation analysis. When including data points across the full range of selection relevance scores, we observe negligible correlation between $\xi^{\mathcal{SR}}$ and prediction support ξ^+ ($\rho = 0.09$). The correlation between $\xi^{\mathcal{SR}}$ and the prediction shift score ξ^{JSD} is also negligible ($\rho = 0.07$). However, we find that the fine-tuning behavior still aligns with expectations when examining the score distributions more closely: Figure 5 shows that when selection relevance is low, validation scores are effectively uncorrelated and centered around zero. This is to be expected, as fine-tuning on an unrelated selection should not systematically increase or decrease the likelihood of a test instance. Only when the selection is sufficiently relevant (greater than 0 dB) should one expect fine-tuning to have a reliably positive effect.

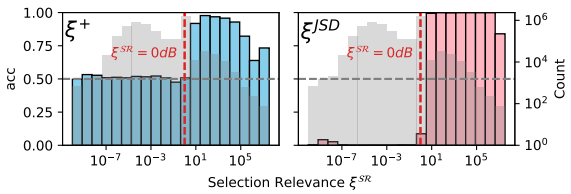


Figure 6: **Fine-tuning based validation**. Sufficiently high selection relevance predicts fine-tuning success.

We test whether this heuristic holds using a rule-based estimator: $\hat{g}(S) = \begin{smallmatrix} 1 \text{ if } \xi^{\mathcal{SR}}(S) > 0dB \\ 0 \quad \text{otherwise} \end{smallmatrix}$ which predicts if fine-tuning on the selected set S increases the likelihood of the original prediction (i.e., if $\xi^+(S) > 0$), based solely on its relevance score $\xi^{\mathcal{SR}}$. Figure 6 confirms that selection relevance is indicative of fine-tuning success. Consequently, only selections with high selection relevance can faithfully reflect fine-tuning behavior here. We also perform a correlation analysis restricted to instances with $\xi^{\mathcal{SR}} > 0$ dB to test whether higher selection relevance is associated with stronger fine-tuning effects. Correlations are then substantially stronger for both scores (ξ^+ : 0.58, ξ^{JSD} 0.37).

Impact of re-ranking. Finally, facility location–based selection significantly increases correlation for LESS $\lambda = \{0.25, 0.5, 0.75, 1.0\}$, and significantly decreases it only for the purely coverage-based strategy BM25 $\lambda = \{1.0\}$. For all other settings, the change is insignificant. This is a positive outcome. If this strategy had consistently lowered correlation, it would suggest that improving selection relevance comes at the cost of producing a less prediction-constrained selection.

5 Discussion

High relevance despite low absolute influence. For the selections based on DataInf and LESS estimates, only strategies that select the k *least influential* examples outperform random selection. This appears counterintuitive at first, but can be explained by considering how influence scores map to changes in test loss and model parameters:

these examples feature the smallest absolute influence scores and are assumed to have the weakest impact on θ with respect to the test instance in a remove-and-retrain experiment. Relative to all other selection strategies, the removal of these examples is least likely to cause the model to deviate from its original prediction, making their selection support and relevance scores higher than those of competing strategies (though not necessarily high in absolute terms). We argue that this is desirable for the type of example-based explanation we consider here (ones that present examples that support the test instance), and observe that our score aligns with this intuition, ranking the *least influential* selection above the *most helpful*, followed by *most harmful* and *most influential*. This strategy is less suitable for counterfactual explanations, where examples with large absolute influence scores are likely more useful. See Appendix E for an overview of remove-and-retrain behavior for alternative naive selection strategies.

Influence vs. similarity. The ineffectiveness of selecting the most helpful examples (i.e., those with the most negative influence scores) illustrates that training data influence (as approximated by DataInf and LESS) and similarity-based notions of relevance (BM25) are not well aligned: while the most helpful examples are assumed to support the test instance most strongly (their removal increases test loss and decreases the likelihood of the test instance), they are not necessarily also the most similar examples to the test instance. This is because a reduction in loss can also result from training on contrastive or otherwise dissimilar examples.

Selection relevance vs. explanation faithfulness. In this work, we define selection relevance as a property distinct from both explanation faithfulness and the correctness of influence estimates. This distinction is reflected in our validation experiment. Unlike traditional training-based evaluation, which aims to measure the causal impact of adding or removing data, our fine-tuning scores capture an alternative notion of relevance for validating alignment. While this separation enables the two concepts to be evaluated independently, it does not allow us to identify if the low relevance scores for the gradient-based estimators are due to low faithfulness or to other factors. Future work should include dedicated evaluations of explanation faithfulness, ideally at human-interpretable selection budgets, rather than focusing on the correctness of raw estimates.

6 Conclusion

We introduce a retraining-free score for evaluating example-based explanations derived from training data influence estimates, accounting for the example selection process rather than focusing solely on the influence estimation step. We find that naively selecting the most influential examples conflicts with the goal of providing examples that support the model's prediction as explanations, as they are, on average, less relevant to the test instance than a random selection. Additionally, we observe that selections with high selection relevance scores tend to provide stronger support for the model's outputs in fine-tuning experiments than random selections, suggesting that our score is a useful signal for predicting training dynamics.

Addressing findings from prior work that highly influential but irrelevant examples are less informative from a user's perspective, we propose a novel selection strategy that increases selection relevance and data coverage. Our results demonstrate that the choice of selection strategy can substantially affect the quality of example-based explanations, and that it should therefore be considered alongside the correctness of the influence estimates when designing explanation systems. To this end, we argue that future work should explicitly consider selection relevance, because evaluating faithfulness alone does not ensure that selected examples are also sufficiently informative from a user's perspective.

Limitations

In line with prior work, we restrict gradient-based estimators to LoRA layers introduced during fine-tuning to make influence estimation computationally feasible for large language models. To reduce computational cost, given the large number of selection parameter combinations and the inclusion of a fine-tuning-based evaluation, we restricted our experiments to models in the 0.5–1B parameter range. Nevertheless, the proposed scoring framework is general and can be applied to larger model gradients given sufficient computational resources.

As we have pointed out in the paper, selection strategies based on gradient-based estimators show overall low performance. While we propose a method to increase the relevance of their selections, investigating the exact cause of this low performance is beyond the scope of this paper. One possible explanation is that the instruction fine-tuning data we use may have limited feature redundancy, and as a result, there may not be enough truly influential examples to retrieve. However, the strong performance of BM25-based selection suggests that relevant examples do exist, at least in terms of token overlap with the test instance. Nonetheless, we cannot rule out the possibility that these examples had little or no effect on the model during training, for example, due to saturation effects.

Finally, we would like to re-emphasize that our selection relevance score measures the relevance of selected training examples, but relevance alone does not guarantee explanation faithfulness, as it is only a necessary condition. Consequently, this score should not be treated as a standalone metric for evaluating explanation faithfulness, and complementary evaluations specifically targeting faithfulness remain essential.

References

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B. Grosse. 2022. If Influence Functions are the Answer, Then What is the Question? *Advances in Neural Information Processing Systems*, 35:17953–17967.

Elnaz Barshan, Marc-Etienne Brunet, and Gintare Karolina Dziugaite. 2020. RelatIF: Identifying Explanatory Training Samples via Relative Influence. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 1899–1909. PMLR.

Samyadeep Basu, Phil Pope, and Soheil Feizi. 2021. Influence Functions in Deep Learning Are Fragile. In *9th International Conference on Learning Representations*.

Umang Bhatt, Isabel Chien, Muhammad Bilal Zafar, and Adrian Weller. 2021. DIVINE: Diverse Influential Training Points for Data Visualization and Model Refinement. *Preprint*, arXiv:2107.05978.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. 2024. What is Your Data Worth to GPT? LLM-Scale Data Valuation with Influence Functions. *Preprint*, arXiv:2405.13954.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 516 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying Large Language Model Generalization with Influence Functions. *Preprint*, arXiv:2308.03296.

Peijian Gu, Yaozong Shen, Lijie Wang, Quan Wang, Hua Wu, and Zhendong Mao. 2023. IAEval: A Comprehensive Evaluation of Instance Attribution on Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11966–11977, Singapore. Association for Computational Linguistics.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2025. OLMES: A Standard for Language Model Evaluations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5005–5033, Albuquerque, New Mexico. Association for Computational Linguistics.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5):93:1–93:42.

Zayd Hammoudeh and Daniel Lowd. 2022. Identifying a Training-Set Attack's Target Using Renormalized Influence Estimation. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, pages 1367–1381, New York, NY, USA. Association for Computing Machinery.

Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. 2021. Evaluation of Similarity-based Explanations. In *9th International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*. OpenReview.net.

Karthikeyan K and Anders Søgaard. 2021. Revisiting Methods for Finding Influential Examples. *Preprint*, arXiv:2111.04683.

Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894. PMLR.

Andreas Krause and Daniel Golovin. 2014. Submodular function maximization. In Lucas Bordeaux, Youssef Hamadi, and Kohli, editors, *Tractability: Practical Approaches to Hard Problems*, pages 71–104. Cambridge University Press, Cambridge.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2024. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. In *The Twelfth International Conference on Learning Representations*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. *Preprint*, arXiv:2411.15124.

Ikhtiyor Nematov, Dimitris Sacharidis, Katja Hose, and Tomer Sagi. 2024a. AIDE: Antithetical, Intent-based, and Diverse Example-Based Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):1051–1062.

Ikhtiyor Nematov, Dimitris Sacharidis, Tomer Sagi, and Katja Hose. 2024b. The susceptibility of example-based explainability methods to class outliers. *CoRR*, abs/2407.20678.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. 2 OLMo 2 Furious. *Preprint*, arXiv:2501.00656.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: attributing model behavior at scale. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 27074–27113. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

Jacob M. Schreiber, Jeffrey A. Bilmes, and William Stafford Noble. 2020. apricot: Submodular selection for data summarization in python. *J. Mach. Learn. Res.*, 21:161:1–161:6.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Chih-Kuan Yeh, Joon Sik Kim, Ian En-Hsu Yen, and Pradeep Ravikumar. 2018. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9311–9321.

Wei Zhang, Ziming Huang, Yada Zhu, Guangnan Ye, Xiaodong Cui, and Fan Zhang. 2021. On Sample Based Explanation Methods for NLP: Faithfulness, Efficiency and Semantic Evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5399–5411, Online. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176.

A Training Data Influence Estimation

LESS (Xia et al., 2024) uses TRAK random projections (Park et al., 2023) to obtain low-dimensional gradient representations. We distribute the total projection dimension of 2^{13} across LoRA layers proportionally to their gradient size:

$$\text{proj_dim}_\ell = \min\left(d_\ell, \frac{\text{proj_dim} \cdot d_\ell}{\sum_{k \in \text{LoRA}} d_k}\right),$$

where d_ℓ is the gradient dimension of layer ℓ. As we need to store gradients to disk for later re-use by our selection relevance score $\xi^{\mathcal{SR}}$, we also employ the same random projection strategy for DataInf, and score BM25-based selections on DataInf gradients. Additionally, we normalize gradients before computing dot products to reduce the impact of gradient magnitude in line with previous work (Xia et al., 2024; Hammoudeh and Lowd, 2022; Park et al., 2023).

B Selection by Submodular Optimization

The baseline selection methods DIVINE and AIDE feature hyperparameters, which we select as follows:

DIVINE. Bhatt et al. (2021) propose the following selection objective:

$$\max_{\mathcal{S}\in\mathcal{D},|\mathcal{S}|=m} \mathcal{I}(\mathcal{S}) + \gamma\mathcal{R}(\mathcal{S}),$$

where $\mathcal{I}(\mathcal{S})$ quantifies importance, and $\mathcal{R}(\mathcal{S})$ captures diversity of the points in $\mathcal{S}$. We use the strategy for finding γ recommended by Bhatt et al. that maximizes the average pairwise distance between examples. Specifically, we consider 20 values of γ logarithmically spaced between 10^{-4} and 10^{5}. We perform subset selection for each candidate γ and compute the average pairwise cosine distance among the selected points, choosing the one that yields the highest mean pairwise distance for the current test instance.

AIDE. Nematov et al. (2024a) define the following selection objective:

$$\arg\max_{\mathcal{S}\subseteq\mathcal{D},|\mathcal{S}|=k} \sum_{z\in\mathcal{S}} (\alpha|I(z,z_t)| + \beta P(z,z_t)) + \gamma D(\mathcal{S})$$

where I is a point-wise influence measure, P is a pairwise proximity measure (cosine-similarity in our case), and D measures selection diversity. We set $\alpha = 0.2, \beta = 0.8, \gamma = 0.5$ as empirically determined by the authors. Note that Nematov et al. also incorporate labels into their selection logic, which is only applicable to classification tasks.

C Alternative Constraints

In addition to the approach for computing t described in Section 3.2, which enforces that the coefficients are non-negative and sum to one, we also conducted experiments using alternative scoring models. The scoring models in Table 2 introduce an additional sparsity constraint (*MSEProjUSimpSparse*, *MSEProjUSimpSparse*), remove the sum-to-one constraint (*MSENNLSL2*), or find unconstrained least squares solutions. We chose the approach in Section 3.2 for simplicity, as we did not observe substantial differences in the final rankings in preliminary experiments.

Table 2: Statistics for coefficient vector t with alternative scoring models.

linear_coder	l1	l2	Prop. non-zero	Prop. negative
KLT	0.06	0.05	0.81	0.40
MSE	0.14	3558.67	0.81	0.39
MSENNLSL2	0.23	473.80	0.35	0.00
MSEProjUSimp	1.00	0.34	0.99	0.00
MSEProjUSimpSparse	1.01	0.49	0.59	0.00
MSEProjUSimpSparseSoft	1.00	0.34	1.00	0.00

D Validation Experiment

Estimator	Model	$\rho(\xi^{\mathcal{SR}},\xi^{+})$	$\rho(\xi^{\mathcal{SR}},\xi^{JSD})$
BM25Estimator	Llama-3.2-1B	0.475212	0.415524
	Olmo2-1B	0.527940	0.311160
	Qwen2.5-0.5B	0.528606	0.236657
DataInfEstimator	Llama-3.2-1B	-0.063321	0.300648
	Olmo2-1B	-0.424335	0.656108
	Qwen2.5-0.5B	-0.115234	0.521586
LESSEstimator	Llama-3.2-1B	-0.211124	0.264172
	Olmo2-1B	-0.379308	0.684368
	Qwen2.5-0.5B	-0.056097	0.490828

Table 3: **Validation Experiment.** Correlation analysis for selections with sufficient relevance ($\xi^{\mathcal{SR}} > 0$ dB)

Estimator	Model	$\rho(\xi^{\mathcal{SR}},\xi^{+})$	$\rho(\xi^{\mathcal{SR}},\xi^{JSD})$
BM25Estimator	Llama-3.2-1B	0.218482	0.289192
	Olmo2-1B	0.226630	0.220364
	Qwen2.5-0.5B	0.172134	0.151243
DataInfEstimator	Llama-3.2-1B	0.069286	0.024539
	Olmo2-1B	0.026482	0.011703
	Qwen2.5-0.5B	0.022877	0.046752
LESSEstimator	Llama-3.2-1B	0.074594	0.031241
	Olmo2-1B	0.026210	0.023588
	Qwen2.5-0.5B	0.029103	0.046321

Table 4: **Validation Experiment.** Correlation analysis for all selections

E Naive Selection Strategies

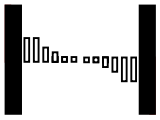
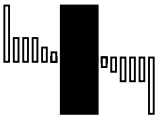
	Most helpful	**Most harmful**	**Most influential**	**Least influential**
Selection	most negative scores	most positive scores	largest absolute scores	smallest absolute scores
Re-training dataset				
Anticipated effect	$\mathcal{L}'$ increases if removed	$\mathcal{L}'$ decreases if removed	Strong impact on θ	Weak impact on θ
Nature of the selected examples	Examples supporting the test instance	Noisy examples; source of bias i.r.t test instance	Unique features or outliers; points whose removal cannot be compensated with other examples	Prototypical and redundant; other points compensate for this point's removal well
Expected average selection relevance score	**More relevant than a random selection:** Training on these examples decreases test loss; however, they are not necessarily the most relevant, as the reduction in loss can also result from training on contrastive or otherwise dissimilar examples (see Section 5).	**Less relevant than a random selection:** Training on these examples increases test loss; not necessarily the least relevant, as mislabeled or noisy examples still affect the model.	**Least relevant:** Likely unique or outlier data points that represent information not present in or not directly relevant to the test instance. Training on these makes it likely that the model will deviate from the original prediction.	**Most relevant:** Examples that are prototypical or redundant with respect to the test instance are likely also representative of training dynamics, as removing them causes little deviation from the model's original parameters.

Figure 7: Anticipated remove-and-re-train effects under naive selection strategies.

F AUC Aggregate Score

most inf.	-8.87 dB
most harmful	-8.33 dB
most helpful	-7.46 dB
AIDE	-7.04 dB
FL most inf. $\lambda = .25$	-6.95 dB
FL most inf. $\lambda = .5$	-6.62 dB
FL most harmful $\lambda = .25$	-6.55 dB
FL most inf. $\lambda = .75$	-6.18 dB
FL most harmful $\lambda = .5$	-6.10 dB
FL most helpful $\lambda = .25$	-5.90 dB
FL most helpful $\lambda = .5$	-5.64 dB
FL most inf. $\lambda = 1$	-5.38 dB
FL most helpful $\lambda = .75$	-5.23 dB
FL most harmful $\lambda = .75$	-4.95 dB
DIVINE most helpful	-4.62 dB
FL most harmful $\lambda = 1$	-4.26 dB
FL most helpful $\lambda = 1$	-4.13 dB
DIVINE most inf.	-4.03 dB
DIVINE most harmful	-3.34 dB
random	10.95 dB
DIVINE least inf.	13.66 dB
least inf.	13.66 dB
FL least inf. $\lambda = .5$	13.67 dB
FL least inf. $\lambda = .75$	13.67 dB
FL least inf. $\lambda = 1$	13.67 dB
FL least inf. $\lambda = .25$	13.67 dB

DataInfEstimator

most inf.	-8.28 dB
most harmful	-7.80 dB
most helpful	-7.67 dB
FL most inf. $\lambda = .25$	-6.73 dB
AIDE	-6.56 dB
FL most inf. $\lambda = .5$	-6.37 dB
FL most harmful $\lambda = .25$	-6.32 dB
FL most helpful $\lambda = .25$	-6.19 dB
FL most inf. $\lambda = .75$	-5.95 dB
FL most helpful $\lambda = .5$	-5.88 dB
FL most harmful $\lambda = .5$	-5.74 dB
DIVINE most inf.	-5.72 dB
FL most harmful $\lambda = .75$	-5.54 dB
DIVINE most harmful	-5.45 dB
FL most inf. $\lambda = 1$	-5.43 dB
FL most helpful $\lambda = .75$	-5.38 dB
FL most harmful $\lambda = 1$	-4.98 dB
DIVINE most helpful	-4.85 dB
FL most helpful $\lambda = 1$	-4.29 dB
random	10.95 dB
DIVINE least inf.	13.66 dB
least inf.	13.66 dB
FL least inf. $\lambda = .5$	13.67 dB
FL least inf. $\lambda = .75$	13.67 dB
FL least inf. $\lambda = 1$	13.67 dB
FL least inf. $\lambda = .25$	13.67 dB

LESSEstimator

random	10.95 dB
FL least inf. $\lambda = 1$	11.37 dB
DIVINE least inf.	11.62 dB
FL least inf. $\lambda = .75$	12.42 dB
FL least inf. $\lambda = .5$	12.68 dB
FL least inf. $\lambda = .25$	12.82 dB
least inf.	12.88 dB
FL most inf. $\lambda = 1$	29.03 dB
most inf.	38.16 dB
AIDE	38.20 dB
DIVINE most inf.	40.10 dB
FL most inf. $\lambda = .25$	42.36 dB
FL most inf. $\lambda = .75$	42.69 dB
FL most inf. $\lambda = .5$	42.73 dB

BM25Estimator

Table 5: **Aggregate results.** auc $\xi^{\mathcal{SR}}$ $k = 1, 5, 10, 25$. Per model results in Table 7.

G Case Study

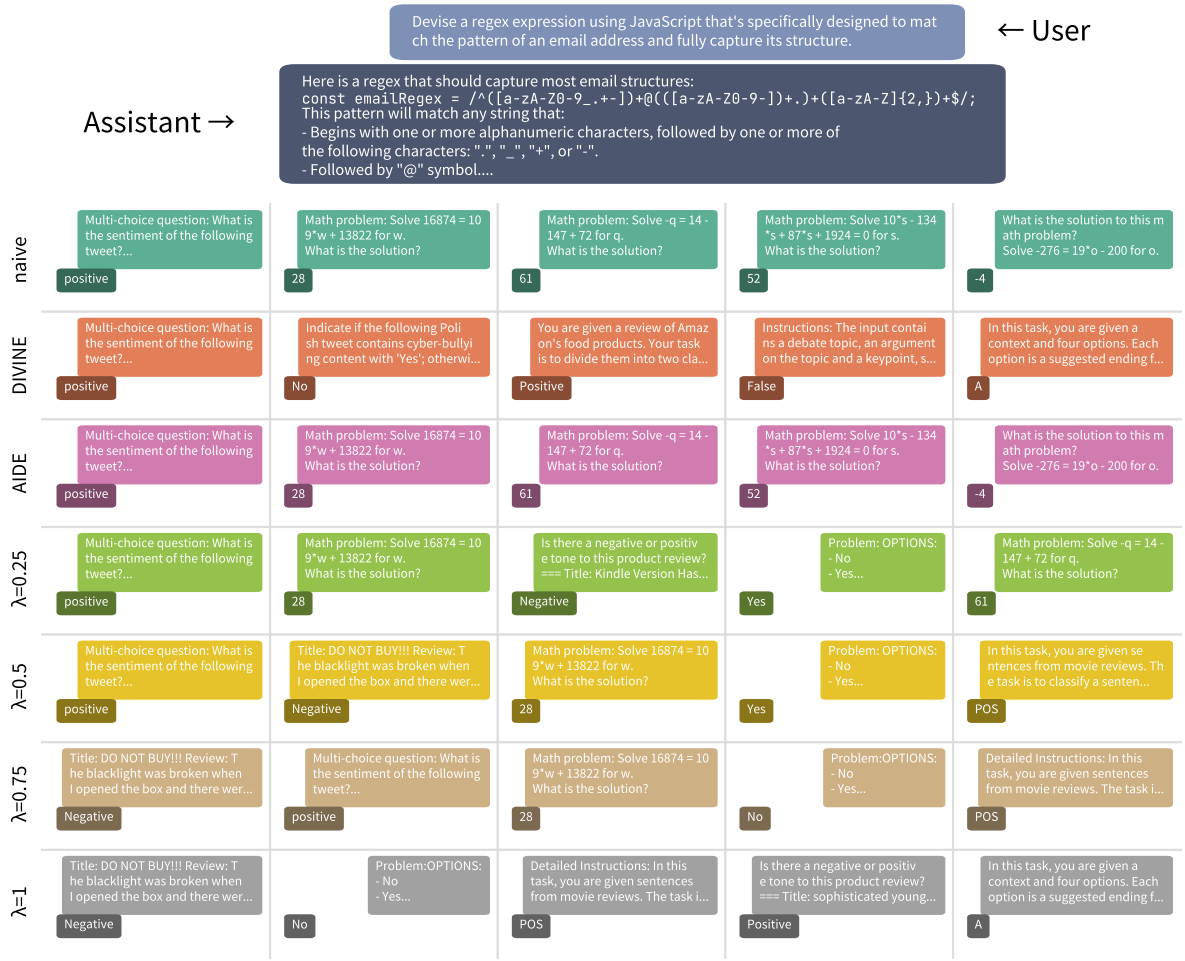


Figure 8: **Case Study**. Full selection for the test instance in Figure 3 (k=5, DataInf, most inf.).

H Per-Model Results

Olmo2-1B	most inf.	-28.72 dB
Olmo2-1B	most harmful	-25.82 dB
Olmo2-1B	most helpful	-24.76 dB
Qwen2.5-0.5B	most inf.	-24.48 dB
Olmo2-1B	FL most inf. $\lambda = .25$	-23.64 dB
Qwen2.5-0.5B	most harmful	-23.62 dB
Qwen2.5-0.5B	most helpful	-23.56 dB
Olmo2-1B	FL most inf. $\lambda = .5$	-23.21 dB
Olmo2-1B	AIDE	-23.12 dB
Olmo2-1B	FL most inf. $\lambda = .75$	-22.84 dB
Olmo2-1B	FL most harmful $\lambda = .25$	-22.41 dB
Olmo2-1B	FL most inf. $\lambda = 1$	-22.22 dB
Olmo2-1B	FL most harmful $\lambda = .5$	-22.06 dB
Olmo2-1B	FL most helpful $\lambda = .25$	-22.01 dB
Olmo2-1B	FL most helpful $\lambda = .5$	-21.81 dB
Olmo2-1B	FL most harmful $\lambda = .75$	-21.62 dB
Olmo2-1B	FL most helpful $\lambda = .75$	-21.43 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .25$	-21.19 dB
Olmo2-1B	DIVINE most helpful	-21.14 dB
Qwen2.5-0.5B	DIVINE most inf.	-21.11 dB
Qwen2.5-0.5B	DIVINE most helpful	-20.87 dB
Qwen2.5-0.5B	AIDE	-20.73 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .25$	-20.62 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .5$	-20.55 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .25$	-20.54 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .5$	-20.44 dB
Olmo2-1B	FL most harmful $\lambda = 1$	-20.32 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .75$	-20.20 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .75$	-20.20 dB
Qwen2.5-0.5B	DIVINE most harmful	-20.11 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .5$	-20.06 dB
Llama-3.2-1B	most harmful	-20.04 dB
Olmo2-1B	FL most helpful $\lambda = 1$	-19.82 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .75$	-19.80 dB
Qwen2.5-0.5B	FL most inf. $\lambda = 1$	-19.77 dB
Llama-3.2-1B	most inf.	-19.76 dB
Qwen2.5-0.5B	FL most helpful $\lambda = 1$	-19.61 dB
Qwen2.5-0.5B	FL most harmful $\lambda = 1$	-19.19 dB
Llama-3.2-1B	AIDE	-19.14 dB
Llama-3.2-1B	FL most harmful $\lambda = .25$	-18.81 dB
Llama-3.2-1B	most helpful	-18.73 dB
Llama-3.2-1B	FL most inf. $\lambda = .25$	-18.58 dB
Llama-3.2-1B	FL most harmful $\lambda = .5$	-18.45 dB
Llama-3.2-1B	FL most inf. $\lambda = .5$	-18.34 dB
Llama-3.2-1B	FL most harmful $\lambda = .75$	-18.14 dB
Llama-3.2-1B	FL most inf. $\lambda = .75$	-18.14 dB
Llama-3.2-1B	FL most inf. $\lambda = 1$	-17.74 dB
Llama-3.2-1B	FL most helpful $\lambda = .25$	-17.74 dB
Llama-3.2-1B	FL most harmful $\lambda = 1$	-17.64 dB
Llama-3.2-1B	FL most helpful $\lambda = .5$	-17.41 dB
Llama-3.2-1B	DIVINE most harmful	-17.36 dB
Llama-3.2-1B	DIVINE most inf.	-17.06 dB
Llama-3.2-1B	FL most helpful $\lambda = .75$	-17.01 dB
Llama-3.2-1B	FL most helpful $\lambda = 1$	-16.25 dB
Llama-3.2-1B	DIVINE most helpful	-16.08 dB
Olmo2-1B	DIVINE most inf.	-15.48 dB
Olmo2-1B	DIVINE most harmful	-15.15 dB
Qwen2.5-0.5B	random	-2.99 dB
Llama-3.2-1B	random	-2.39 dB
Olmo2-1B	random	-2.37 dB
Olmo2-1B	least inf.	-0.19 dB
Olmo2-1B	DIVINE least inf.	-0.19 dB
Olmo2-1B	FL least inf. $\lambda = .5$	-0.18 dB
Olmo2-1B	FL least inf. $\lambda = 1$	-0.18 dB
Olmo2-1B	FL least inf. $\lambda = .25$	-0.18 dB
Olmo2-1B	FL least inf. $\lambda = .75$	-0.18 dB
Qwen2.5-0.5B	DIVINE least inf.	-0.15 dB
Qwen2.5-0.5B	least inf.	-0.14 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .5$	-0.13 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .25$	-0.13 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .75$	-0.13 dB
Qwen2.5-0.5B	FL least inf. $\lambda = 1$	-0.13 dB
Llama-3.2-1B	DIVINE least inf.	-0.10 dB
Llama-3.2-1B	least inf.	-0.10 dB
Llama-3.2-1B	FL least inf. $\lambda = 1$	-0.08 dB
Llama-3.2-1B	FL least inf. $\lambda = .5$	-0.08 dB
Llama-3.2-1B	FL least inf. $\lambda = .75$	-0.08 dB
Llama-3.2-1B	FL least inf. $\lambda = .25$	-0.08 dB

DataInfEstimator

Olmo2-1B	most inf.	-25.95 dB
Olmo2-1B	most harmful	-24.99 dB
Olmo2-1B	most helpful	-24.08 dB
Qwen2.5-0.5B	most inf.	-23.91 dB
Qwen2.5-0.5B	most helpful	-23.53 dB
Qwen2.5-0.5B	most harmful	-22.82 dB
Olmo2-1B	FL most inf. $\lambda = .25$	-22.70 dB
Olmo2-1B	FL most inf. $\lambda = .5$	-22.61 dB
Olmo2-1B	FL most inf. $\lambda = .75$	-22.39 dB
Olmo2-1B	AIDE	-22.37 dB
Olmo2-1B	DIVINE most inf.	-22.24 dB
Olmo2-1B	FL most harmful $\lambda = .25$	-22.08 dB
Olmo2-1B	FL most inf. $\lambda = 1$	-21.91 dB
Olmo2-1B	FL most harmful $\lambda = .5$	-21.85 dB
Olmo2-1B	DIVINE most harmful	-21.65 dB
Olmo2-1B	FL most helpful $\lambda = .25$	-21.62 dB
Olmo2-1B	FL most harmful $\lambda = .75$	-21.56 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .25$	-21.54 dB
Olmo2-1B	FL most helpful $\lambda = .5$	-21.29 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .5$	-21.29 dB
Qwen2.5-0.5B	AIDE	-21.21 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .25$	-21.18 dB
Qwen2.5-0.5B	DIVINE most inf.	-21.00 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .5$	-20.85 dB
Olmo2-1B	FL most helpful $\lambda = .75$	-20.83 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .25$	-20.83 dB
Olmo2-1B	FL most harmful $\lambda = 1$	-20.60 dB
Olmo2-1B	DIVINE most helpful	-20.39 dB
Qwen2.5-0.5B	DIVINE most harmful	-20.38 dB
Qwen2.5-0.5B	DIVINE most helpful	-20.18 dB
Olmo2-1B	FL most helpful $\lambda = 1$	-20.05 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .75$	-19.97 dB
Llama-3.2-1B	most inf.	-19.53 dB
Llama-3.2-1B	most helpful	-19.42 dB
Llama-3.2-1B	most harmful	-19.41 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .75$	-19.34 dB
Qwen2.5-0.5B	FL most harmful $\lambda = 1$	-19.29 dB
Qwen2.5-0.5B	FL most helpful $\lambda = 1$	-19.27 dB
Qwen2.5-0.5B	FL most inf. $\lambda = 1$	-19.05 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .5$	-19.02 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .75$	-18.90 dB
Llama-3.2-1B	FL most inf. $\lambda = .25$	-18.55 dB
Llama-3.2-1B	AIDE	-18.53 dB
Llama-3.2-1B	FL most harmful $\lambda = .25$	-18.40 dB
Llama-3.2-1B	FL most helpful $\lambda = .25$	-18.29 dB
Llama-3.2-1B	FL most inf. $\lambda = .75$	-18.18 dB
Llama-3.2-1B	FL most inf. $\lambda = .5$	-18.05 dB
Llama-3.2-1B	FL most harmful $\lambda = .75$	-18.01 dB
Llama-3.2-1B	FL most helpful $\lambda = .5$	-17.97 dB
Llama-3.2-1B	FL most harmful $\lambda = .5$	-17.92 dB
Llama-3.2-1B	FL most inf. $\lambda = 1$	-17.81 dB
Llama-3.2-1B	FL most harmful $\lambda = 1$	-17.78 dB
Llama-3.2-1B	FL most helpful $\lambda = .75$	-17.55 dB
Llama-3.2-1B	DIVINE most harmful	-17.31 dB
Llama-3.2-1B	DIVINE most inf.	-17.26 dB
Llama-3.2-1B	DIVINE most helpful	-16.82 dB
Llama-3.2-1B	FL most helpful $\lambda = 1$	-16.68 dB
Qwen2.5-0.5B	random	-2.99 dB
Llama-3.2-1B	random	-2.39 dB
Olmo2-1B	random	-2.37 dB
Olmo2-1B	least inf.	-0.19 dB
Olmo2-1B	DIVINE least inf.	-0.19 dB
Olmo2-1B	FL least inf. $\lambda = .25$	-0.18 dB
Olmo2-1B	FL least inf. $\lambda = .75$	-0.18 dB
Olmo2-1B	FL least inf. $\lambda = 1$	-0.18 dB
Olmo2-1B	FL least inf. $\lambda = .5$	-0.18 dB
Qwen2.5-0.5B	DIVINE least inf.	-0.14 dB
Qwen2.5-0.5B	least inf.	-0.14 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .75$	-0.13 dB
Qwen2.5-0.5B	FL least inf. $\lambda = 1$	-0.13 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .25$	-0.13 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .5$	-0.13 dB
Llama-3.2-1B	DIVINE least inf.	-0.10 dB
Llama-3.2-1B	least inf.	-0.10 dB
Llama-3.2-1B	FL least inf. $\lambda = .5$	-0.09 dB
Llama-3.2-1B	FL least inf. $\lambda = 1$	-0.09 dB
Llama-3.2-1B	FL least inf. $\lambda = .75$	-0.09 dB
Llama-3.2-1B	FL least inf. $\lambda = .25$	-0.09 dB

LESSEstimator

Qwen2.5-0.5B	FL least inf. $\lambda = 1$	-3.50 dB
Qwen2.5-0.5B	DIVINE least inf.	-3.06 dB
Qwen2.5-0.5B	random	-2.99 dB
Llama-3.2-1B	random	-2.39 dB
Olmo2-1B	DIVINE least inf.	-2.38 dB
Olmo2-1B	random	-2.37 dB
Olmo2-1B	FL least inf. $\lambda = 1$	-2.32 dB
Llama-3.2-1B	DIVINE least inf.	-2.14 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .75$	-1.96 dB
Olmo2-1B	FL least inf. $\lambda = .75$	-1.76 dB
Llama-3.2-1B	FL least inf. $\lambda = 1$	-1.61 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .5$	-1.58 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .25$	-1.47 dB
Olmo2-1B	FL least inf. $\lambda = .5$	-1.44 dB
Qwen2.5-0.5B	least inf.	-1.43 dB
Olmo2-1B	FL least inf. $\lambda = .25$	-1.25 dB
Olmo2-1B	least inf.	-1.13 dB
Llama-3.2-1B	FL least inf. $\lambda = .75$	-1.03 dB
Llama-3.2-1B	FL least inf. $\lambda = .5$	-0.79 dB
Llama-3.2-1B	FL least inf. $\lambda = .25$	-0.60 dB
Llama-3.2-1B	least inf.	-0.52 dB
Olmo2-1B	FL most inf. $\lambda = 1$	12.78 dB
Qwen2.5-0.5B	FL most inf. $\lambda = 1$	15.65 dB
Llama-3.2-1B	FL most inf. $\lambda = 1$	17.05 dB
Olmo2-1B	DIVINE most inf.	20.23 dB
Olmo2-1B	FL most inf. $\lambda = .5$	21.46 dB
Olmo2-1B	FL most inf. $\lambda = .25$	21.46 dB
Qwen2.5-0.5B	AIDE	21.62 dB
Olmo2-1B	FL most inf. $\lambda = .75$	21.65 dB
Qwen2.5-0.5B	most inf.	21.69 dB
Olmo2-1B	AIDE	21.86 dB
Olmo2-1B	most inf.	21.91 dB
Llama-3.2-1B	FL most inf. $\lambda = .75$	22.01 dB
Llama-3.2-1B	FL most inf. $\lambda = .5$	22.42 dB
Llama-3.2-1B	DIVINE most inf.	22.50 dB
Llama-3.2-1B	AIDE	22.68 dB
Llama-3.2-1B	FL most inf. $\lambda = .25$	22.87 dB
Llama-3.2-1B	most inf.	22.99 dB
Qwen2.5-0.5B	DIVINE most inf.	26.35 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .25$	33.63 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .5$	34.30 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .75$	34.30 dB

BM25Estimator

Table 6: **Per-model results.** $\xi^{\mathcal{SR}}$ $k = 10$**.**

Olmo2-1B	most inf.	-14.13 dB
Olmo2-1B	most harmful	-11.32 dB
Olmo2-1B	most helpful	-10.57 dB
Qwen2.5-0.5B	most inf.	-10.48 dB
Olmo2-1B	AIDE	-9.87 dB
Qwen2.5-0.5B	most helpful	-9.42 dB
Qwen2.5-0.5B	most harmful	-9.40 dB
Olmo2-1B	FL most inf. $\lambda = .25$	-8.92 dB
Olmo2-1B	FL most inf. $\lambda = .5$	-8.54 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .25$	-7.97 dB
Olmo2-1B	FL most harmful $\lambda = .25$	-7.76 dB
Olmo2-1B	FL most inf. $\lambda = .75$	-7.76 dB
Qwen2.5-0.5B	AIDE	-7.48 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .5$	-7.46 dB
Olmo2-1B	FL most helpful $\lambda = .25$	-7.46 dB
Olmo2-1B	FL most harmful $\lambda = .5$	-7.41 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .25$	-7.28 dB
Qwen2.5-0.5B	DIVINE most inf.	-7.24 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .25$	-7.22 dB
Olmo2-1B	FL most helpful $\lambda = .5$	-7.21 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .5$	-7.03 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .75$	-7.02 dB
Olmo2-1B	FL most helpful $\lambda = .75$	-6.81 dB
Olmo2-1B	DIVINE most helpful	-6.73 dB
Qwen2.5-0.5B	DIVINE most helpful	-6.66 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .75$	-6.65 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .5$	-6.57 dB
Qwen2.5-0.5B	FL most inf. $\lambda = 1$	-6.39 dB
Olmo2-1B	FL most inf. $\lambda = 1$	-6.37 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .75$	-6.20 dB
Qwen2.5-0.5B	DIVINE most harmful	-6.10 dB
Llama-3.2-1B	most harmful	-5.98 dB
Llama-3.2-1B	most inf.	-5.84 dB
Qwen2.5-0.5B	FL most helpful $\lambda = 1$	-5.79 dB
Qwen2.5-0.5B	FL most harmful $\lambda = 1$	-5.55 dB
Olmo2-1B	FL most helpful $\lambda = 1$	-5.37 dB
Llama-3.2-1B	FL most harmful $\lambda = .25$	-5.12 dB
Llama-3.2-1B	AIDE	-5.06 dB
Llama-3.2-1B	FL most inf. $\lambda = .25$	-4.99 dB
Llama-3.2-1B	FL most inf. $\lambda = .5$	-4.76 dB
Llama-3.2-1B	FL most harmful $\lambda = .5$	-4.76 dB
Llama-3.2-1B	most helpful	-4.74 dB
Llama-3.2-1B	FL most inf. $\lambda = .75$	-4.48 dB
Olmo2-1B	FL most harmful $\lambda = .75$	-4.47 dB
Llama-3.2-1B	FL most harmful $\lambda = .75$	-4.41 dB
Llama-3.2-1B	FL most helpful $\lambda = .25$	-3.93 dB
Llama-3.2-1B	FL most inf. $\lambda = 1$	-3.88 dB
Llama-3.2-1B	FL most harmful $\lambda = 1$	-3.77 dB
Olmo2-1B	FL most harmful $\lambda = 1$	-3.69 dB
Llama-3.2-1B	FL most helpful $\lambda = .5$	-3.66 dB
Llama-3.2-1B	DIVINE most harmful	-3.58 dB
Llama-3.2-1B	DIVINE most inf.	-3.35 dB
Llama-3.2-1B	FL most helpful $\lambda = .75$	-3.23 dB
Olmo2-1B	DIVINE most inf.	-2.71 dB
Llama-3.2-1B	FL most helpful $\lambda = 1$	-2.18 dB
Llama-3.2-1B	DIVINE most helpful	-2.16 dB
Olmo2-1B	DIVINE most harmful	-1.50 dB
Qwen2.5-0.5B	random	10.61 dB
Llama-3.2-1B	random	11.09 dB
Olmo2-1B	random	11.12 dB
Olmo2-1B	least inf.	13.61 dB
Olmo2-1B	DIVINE least inf.	13.61 dB
Olmo2-1B	FL least inf. $\lambda = .75$	13.63 dB
Olmo2-1B	FL least inf. $\lambda = .5$	13.63 dB
Olmo2-1B	FL least inf. $\lambda = 1$	13.63 dB
Olmo2-1B	FL least inf. $\lambda = .25$	13.63 dB
Qwen2.5-0.5B	DIVINE least inf.	13.66 dB
Qwen2.5-0.5B	least inf.	13.66 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .75$	13.67 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .25$	13.67 dB
Qwen2.5-0.5B	FL least inf. $\lambda = 1$	13.67 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .5$	13.67 dB
Llama-3.2-1B	DIVINE least inf.	13.71 dB
Llama-3.2-1B	least inf.	13.71 dB
Llama-3.2-1B	FL least inf. $\lambda = 1$	13.72 dB
Llama-3.2-1B	FL least inf. $\lambda = .75$	13.72 dB
Llama-3.2-1B	FL least inf. $\lambda = .5$	13.72 dB
Llama-3.2-1B	FL least inf. $\lambda = .25$	13.72 dB

DataInfEstimator

Olmo2-1B	most inf.	-11.86 dB
Olmo2-1B	most harmful	-10.82 dB
Qwen2.5-0.5B	most inf.	-9.86 dB
Olmo2-1B	most helpful	-9.78 dB
Qwen2.5-0.5B	most helpful	-9.33 dB
Olmo2-1B	AIDE	-8.87 dB
Qwen2.5-0.5B	most harmful	-8.72 dB
Olmo2-1B	FL most inf. $\lambda = .25$	-8.23 dB
Olmo2-1B	DIVINE most inf.	-8.16 dB
Olmo2-1B	FL most inf. $\lambda = .5$	-8.05 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .25$	-7.96 dB
Olmo2-1B	FL most inf. $\lambda = .75$	-7.83 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .25$	-7.55 dB
Olmo2-1B	FL most harmful $\lambda = .25$	-7.54 dB
Qwen2.5-0.5B	AIDE	-7.48 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .5$	-7.37 dB
Olmo2-1B	DIVINE most harmful	-7.33 dB
Olmo2-1B	FL most harmful $\lambda = .5$	-7.30 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .5$	-7.21 dB
Olmo2-1B	FL most inf. $\lambda = 1$	-7.21 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .25$	-7.17 dB
Olmo2-1B	FL most helpful $\lambda = .25$	-7.12 dB
Qwen2.5-0.5B	DIVINE most inf.	-7.09 dB
Olmo2-1B	FL most harmful $\lambda = .75$	-7.06 dB
Olmo2-1B	FL most helpful $\lambda = .5$	-6.87 dB
Qwen2.5-0.5B	FL most helpful $\lambda = .75$	-6.62 dB
Olmo2-1B	FL most helpful $\lambda = .75$	-6.45 dB
Qwen2.5-0.5B	DIVINE most harmful	-6.38 dB
Olmo2-1B	DIVINE most helpful	-6.32 dB
Qwen2.5-0.5B	DIVINE most helpful	-6.23 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .75$	-6.19 dB
Olmo2-1B	FL most harmful $\lambda = 1$	-5.97 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .5$	-5.95 dB
Qwen2.5-0.5B	FL most inf. $\lambda = 1$	-5.89 dB
Qwen2.5-0.5B	FL most harmful $\lambda = .75$	-5.58 dB
Llama-3.2-1B	most inf.	-5.57 dB
Qwen2.5-0.5B	FL most helpful $\lambda = 1$	-5.54 dB
Llama-3.2-1B	most harmful	-5.51 dB
Qwen2.5-0.5B	FL most harmful $\lambda = 1$	-5.46 dB
Olmo2-1B	FL most helpful $\lambda = 1$	-5.42 dB
Llama-3.2-1B	most helpful	-5.37 dB
Llama-3.2-1B	FL most inf. $\lambda = .25$	-4.86 dB
Llama-3.2-1B	FL most harmful $\lambda = .25$	-4.80 dB
Llama-3.2-1B	FL most helpful $\lambda = .25$	-4.55 dB
Llama-3.2-1B	FL most inf. $\lambda = .5$	-4.54 dB
Llama-3.2-1B	AIDE	-4.52 dB
Llama-3.2-1B	FL most inf. $\lambda = .75$	-4.47 dB
Llama-3.2-1B	FL most harmful $\lambda = .5$	-4.44 dB
Llama-3.2-1B	FL most harmful $\lambda = .75$	-4.38 dB
Llama-3.2-1B	FL most helpful $\lambda = .5$	-4.22 dB
Llama-3.2-1B	FL most inf. $\lambda = 1$	-3.87 dB
Llama-3.2-1B	FL most harmful $\lambda = 1$	-3.81 dB
Llama-3.2-1B	FL most helpful $\lambda = .75$	-3.72 dB
Llama-3.2-1B	DIVINE most harmful	-3.57 dB
Llama-3.2-1B	DIVINE most inf.	-3.41 dB
Llama-3.2-1B	DIVINE most helpful	-2.92 dB
Llama-3.2-1B	FL most helpful $\lambda = 1$	-2.59 dB
Qwen2.5-0.5B	random	10.61 dB
Llama-3.2-1B	random	11.09 dB
Olmo2-1B	random	11.12 dB
Olmo2-1B	least inf.	13.61 dB
Olmo2-1B	DIVINE least inf.	13.61 dB
Olmo2-1B	FL least inf. $\lambda = .75$	13.63 dB
Olmo2-1B	FL least inf. $\lambda = .5$	13.63 dB
Olmo2-1B	FL least inf. $\lambda = .25$	13.63 dB
Olmo2-1B	FL least inf. $\lambda = 1$	13.63 dB
Qwen2.5-0.5B	DIVINE least inf.	13.66 dB
Qwen2.5-0.5B	least inf.	13.66 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .5$	13.67 dB
Qwen2.5-0.5B	FL least inf. $\lambda = 1$	13.67 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .25$	13.67 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .75$	13.67 dB
Llama-3.2-1B	DIVINE least inf.	13.70 dB
Llama-3.2-1B	least inf.	13.70 dB
Llama-3.2-1B	FL least inf. $\lambda = 1$	13.71 dB
Llama-3.2-1B	FL least inf. $\lambda = .75$	13.71 dB
Llama-3.2-1B	FL least inf. $\lambda = .5$	13.71 dB
Llama-3.2-1B	FL least inf. $\lambda = .25$	13.71 dB

LESSEstimator

Qwen2.5-0.5B	FL least inf. $\lambda = 1$	10.52 dB
Qwen2.5-0.5B	random	10.61 dB
Qwen2.5-0.5B	DIVINE least inf.	11.02 dB
Llama-3.2-1B	random	11.09 dB
Olmo2-1B	random	11.12 dB
Olmo2-1B	FL least inf. $\lambda = 1$	11.31 dB
Olmo2-1B	DIVINE least inf.	11.75 dB
Llama-3.2-1B	DIVINE least inf.	12.02 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .75$	12.06 dB
Llama-3.2-1B	FL least inf. $\lambda = 1$	12.13 dB
Olmo2-1B	FL least inf. $\lambda = .75$	12.26 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .5$	12.41 dB
Olmo2-1B	FL least inf. $\lambda = .5$	12.52 dB
Qwen2.5-0.5B	FL least inf. $\lambda = .25$	12.52 dB
Qwen2.5-0.5B	least inf.	12.54 dB
Olmo2-1B	FL least inf. $\lambda = .25$	12.66 dB
Olmo2-1B	least inf.	12.74 dB
Llama-3.2-1B	FL least inf. $\lambda = .75$	12.89 dB
Llama-3.2-1B	FL least inf. $\lambda = .5$	13.09 dB
Llama-3.2-1B	FL least inf. $\lambda = .25$	13.24 dB
Llama-3.2-1B	least inf.	13.31 dB
Olmo2-1B	FL most inf. $\lambda = 1$	27.14 dB
Qwen2.5-0.5B	FL most inf. $\lambda = 1$	28.52 dB
Llama-3.2-1B	FL most inf. $\lambda = 1$	30.69 dB
Olmo2-1B	DIVINE most inf.	33.73 dB
Olmo2-1B	FL most inf. $\lambda = .5$	34.69 dB
Olmo2-1B	FL most inf. $\lambda = .25$	34.79 dB
Olmo2-1B	FL most inf. $\lambda = .75$	34.96 dB
Olmo2-1B	most inf.	35.08 dB
Olmo2-1B	AIDE	35.38 dB
Llama-3.2-1B	FL most inf. $\lambda = .75$	37.86 dB
Llama-3.2-1B	AIDE	38.03 dB
Llama-3.2-1B	FL most inf. $\lambda = .25$	38.10 dB
Llama-3.2-1B	most inf.	38.11 dB
Llama-3.2-1B	FL most inf. $\lambda = .5$	38.43 dB
Llama-3.2-1B	DIVINE most inf.	39.91 dB
Qwen2.5-0.5B	most inf.	39.97 dB
Qwen2.5-0.5B	AIDE	40.00 dB
Qwen2.5-0.5B	DIVINE most inf.	42.68 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .25$	46.26 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .5$	46.66 dB
Qwen2.5-0.5B	FL most inf. $\lambda = .75$	46.68 dB

BM25Estimator

Table 7: **Per-model results.** auc $\xi^{\mathcal{SR}}$ $k = 1, 5, 10, 25$.

I Model Fine-tuning

We fine-tune all base models using LoRA for one epoch on the full *Tülu3* (allenai/tulu-3-sft-olmo-2-mixture-0225: Lambert et al., 2025) instruction-fine tuning dataset. We report performance on the `OLMES` evaluation suite (Gu et al., 2025) in Table 9.

Precision	bfloat16
Optimizer	AdamW (torch fused)
Learning rate	1×10^{-4}
LR scheduler	Linear
Weight decay	0.0
Max grad norm	1.0
LoRA rank (r)	16
LoRA alpha	32
LoRA dropout	0.1
LoRA bias	none
Target modules	Qwen & Llama: `q_proj, k_proj, v_proj, o_proj` Olmo2: `q_proj, c_attn, v_proj`
Trainable params	LoRA only
Train batch size / device	4
Gradient accumulation	8
Effective batch size	32
Training epochs	1
Max sequence length	1024
Gradient checkpointing	False
Seed	42

Table 8: LoRA fine-tuning hyperparameters

Task	Avg	AGIEval	ARC_C	ARC_E	BBH	BoolQ	CSQA	CoQA	DROP	GSM8K	HSwag	JPRDY	MMLU	MMLU-Pro	NatQs	OBQA	PIQA	SIQA	SQuAD	TriviaQA	WinoG
Fine-tuned Models																					
Llama-3.2-1B	.44	.24	.38	.60	.32	.67	.50	.65	.25	.08	.53	.53	.30	.16	.16	.39	.67	.47	.73	.48	.58
OLMo-2-0425-1B	.52	.34	.47	.74	.30	.69	.60	.69	**.35**	.36	.60	**.63**	.43	.19	.19	.51	.71	.56	.80	**.55**	**.61**
Qwen2.5-0.5B	.46	.37	.49	.71	.32	.66	.58	.60	.24	.34	.48	.32	**.49**	.23	.12	.54	.66	.56	.73	.19	.54
External Models																					
Llama-3.2-1B-Instruct	.51	.35	.49	.73	**.37**	.64	.63	.67	.29	.35	.54	.51	.46	**.25**	**.20**	**.54**	.71	.57	.79	.47	.59
OLMo-2-0425-1B-SFT	**.53**	.36	.48	**.75**	.32	**.75**	**.65**	**.72**	.33	**.43**	**.61**	.57	.44	.21	.17	.51	**.72**	**.58**	**.83**	.48	.60
Qwen2.5-0.5B-Instruct	.44	**.38**	**.49**	.71	.31	.68	.60	.41	.25	.30	.49	.33	.48	.23	.09	.54	.67	.56	.74	.09	.53

Table 9: **Benchmark results.** Performance on the OLMES evaluation suite. Best model in bold. Note that *OLMo-2-0425-1B-SFT* is trained for one additional epoch and without LoRA.