ON REPRESENTATION REDUNDANCY IN TARGETED DATA SELECTION FOR LARGE-SCALE INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data quality is a crucial factor in large language models training. While prior work has shown that models trained on smaller, high-quality datasets can outperform those trained on much larger but noisy or low-quality corpora, systematic methods for industrial-scale data selection in instruction tuning remain underexplored. In this work, we study instruction-tuning data selection through the lens of semantic representation similarity and identify a key limitation of state-of-the-art LLM encoders: they produce highly redundant semantic embeddings. To mitigate this redundancy, we propose Compressed Representation Data Selection (CRDS), a novel framework with two variants. CRDS-R applies Rademacher random projection followed by concatenation of transformer hidden-layer representations, while CRDS-W employs Whitening-based dimensionality reduction to improve representational quality. Experimental results demonstrate that both variants substantially enhance data quality and almost always outperform state-of-the-art representation-based selection methods. Notably, CRDS-W achieves strong performance using only 5% of the training data, surpassing the full-data baseline by an average of 0.88% across five datasets with the 16B base model. Our code will be released.

1 INTRODUCTION

Instruction tuning, also known as supervised fine-tuning (SFT) Ouyang et al. (2022), is a crucial post-training stage for large language models (LLMs), as it endows base models with instruction-following and dialogue capabilities. Unlike the vast and heterogeneous corpora used during pre-training, SFT relies on a substantially more focused dataset. Nevertheless, many recent LLMs Team (2024); Yang et al. (2024a); Team & AI (2025) still require millions of SFT examples to effectively activate the capabilities of the underlying base model, making data curation both critical and challenging. On the one hand, equipping the model with strong conversational abilities while ensuring robustness against safety risks Bai et al. (2022); Touvron et al. (2023); Yang et al. (2025b); Qu et al. (2025) demands that the SFT dataset be both comprehensive and of high quality. On the other hand, increasing data volume does not necessarily improve performance, as additional examples may introduce redundant or inconsistent supervision. Balancing data coverage and quality therefore requires careful selection of instruction-tuning data. In this work, we study this selection problem in a targeted setting, where the goal is to select a subset of instruction-tuning data that improves model performance on a specified task.

Even before the advent of LLMs, data selection—also referred to as data pruning or data condensation—had already attracted substantial attention from researchers Mirzasoleiman et al. (2020); Sorscher et al. (2022); Zhao et al. (2021), driven by the growing recognition that data quality plays a crucial role in model training. With the rise of LLMs, numerous studies have further shown that training datasets can often be substantially reduced in scale without compromising performance Zhou et al. (2023); Ye et al. (2025). However, we observe that most existing methods focus on selecting data from a single instruction-tuning dataset, which typically contains only on the order of 10^4 to 10^5 examples. Moreover, the majority of current techniques rely on human-designed heuristic metrics Lu et al. (2024); Chen et al. (2024); Liu et al. (2024b;a;c); He et al. (2024). Even among similarity-based approaches, most adopt gradient-based criteria as their primary signal Xia et al. (2024a); Lin et al. (2024), which can be computationally expensive. In contrast, semantic representations, despite being

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

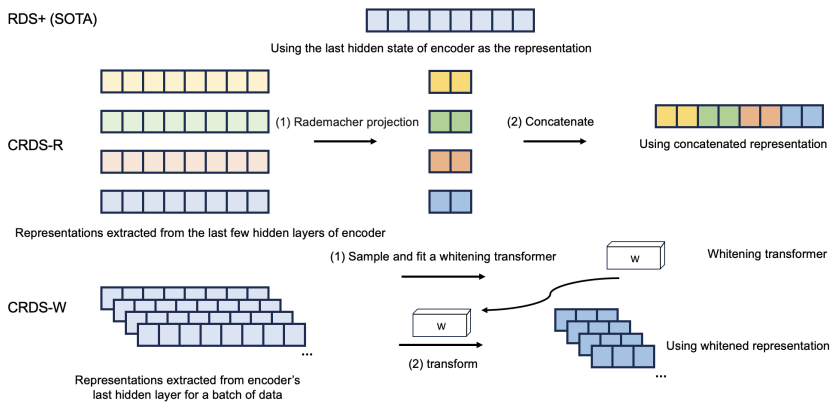


Figure 1: Illustration of our proposed methods and the SOTA baseline RDS+. RDS+ constructs data representations by directly using the last hidden state of the encoder. CRDS-R constructs representations by extracting the last several hidden states of the encoder, applying a Rademacher projection to each layer, and concatenating the projected features to obtain the final representation, which retains the same dimensionality as the original representation. CRDS-W constructs representations by first fitting a Whitening transformer on a large subset of the data using the last hidden state of the encoder, and then applying this transformer to whiten all data representations for similarity computation.

a fundamental and readily accessible property of data, remain insufficiently explored as a criterion for instruction-tuning data selection, and their effectiveness has been largely underestimated.

In this paper, we systematically investigate semantic representations for instruction-tuning data selection, focusing on whether their original dimensionality is necessary for effective selection. We begin with exploratory binarization experiments that replace each embedding component with its sign before computing similarity. Retaining only sign information preserves, and can improve, downstream performance, suggesting that LLM embeddings contain useful selection signals alongside substantial redundancy. Motivated by this observation, we propose Compressed Representation Data Selection (CRDS), with two variants. CRDS-R applies Rademacher projections to representations from the last several Transformer layers and concatenates the projected features, keeping the final dimensionality equal to that of the original single-layer embedding. CRDS-W learns a whitening transformation from a subset of the data pool, then centers, whitens, and reduces all embeddings to a target dimension. These variants explore complementary ways to construct compact representations: aggregating information across layers and correcting the geometry of final-layer embeddings.

Extensive experiments demonstrate that both variants improve the quality of the selected datasets. In particular, our results extend the use of Rademacher projection from high-dimensional gradient representations to lightweight semantic representations, where projection–concatenation can improve selection even when compression is not required for computational feasibility. They also show that whitening-based dimensionality reduction is effective for representation-based selection with million-example data pools. Together, these findings establish representation compression as a promising approach to improving targeted instruction-tuning data selection.

Our contributions include:

- We identify the strengths of semantic representation similarity as a metric for instruction-tuning data selection, as well as the redundancy present in state-of-the-art LLM encoder representations.
- We propose CRDS-R and CRDS-W, two methods based on Rademacher projection–concatenation and Whitening-based dimensionality reduction, respectively. Both methods effectively improve semantic embedding–based instruction-tuning data selection.
- We conduct extensive experiments across three base models, five datasets, and four encoders, demonstrating the broad effectiveness of CRDS-R and CRDS-W. In particular, CRDS-W surpasses the full-data setting on Ling-mini-2.0.

2 Related Work

Instruction tuning data selection. A variety of simple, effective, and human-intuitive methods have proven highly valuable for instruction tuning data selection. Examples include perplexity (Yin & Rush, 2025; Antonello et al., 2021), response length (Zhao et al., 2024), and even random sampling (Xia et al., 2024b). Moreover, high-quality in-context examples have been shown to also serve as strong instruction-tuning data. Building on this insight, many effective data quality metrics have been developed, by which the data quality can be easily evaluated using LLMs. This insight has inspired a line of related research (Li et al., 2024b; Yang et al., 2025c; Li et al., 2024c; Jiang et al., 2025; Li et al., 2024a; Wu et al., 2025).

Among data curation criteria, similarity has received particular attention. Gradient information is widely regarded as a powerful proxy for measuring data similarity in the context of model training, though different approaches utilize gradients in distinct ways. Some methods compute the cosine similarity between gradients induced by training and query examples to assess their alignment (Xia et al., 2024a; Lin et al., 2024; Wang et al., 2025a), while others cluster the data pool based on gradient representations (Sridharan et al., 2025; Pan et al., 2024; Zhang et al., 2025b). Beyond gradients, other similarity-based metrics have also demonstrated strong performance in instruction tuning data selection, including semantic representations (Ivison et al., 2025; Nagaraj et al., 2025), loss trajectories (Yang et al., 2024b), influence scores (Nikdan et al., 2025), and neural tangent kernels (Wang et al., 2025b), among others. This work falls within the line of research on semantic representation and is most closely related to RDS+ (Ivison et al., 2025).

Another common strategy involves employing models to rate data to reduce dataset scale. The most straightforward approach is using LLMs to directly evaluate data quality (Chen et al., 2024). Building on these assessments, data quality can be effectively integrated with complementary metrics—such as complexity (Liu et al., 2024b; Zhang et al., 2025a) and diversity (Ge et al., 2024; Bukharin et al., 2024; Ni et al., 2024; Pang et al., 2025)—to support more generalizable data selection. Furthermore, LLMs can be leveraged to annotate data (Lu et al., 2024; Chen et al., 2025), and these annotations can be readily transformed into metrics to uncover relationships within the data.

Semantic representation correction. Some early work identified the anisotropy in representations produced by advanced encoders (Ethayarajh, 2019). To address this issue, a variety of methods have been proposed. Many approaches rely on contrastive learning (Gao et al., 2021; Neelakantan et al., 2022; Chuang et al., 2022), while others focus on training normalization models (Li et al., 2020) or correcting the representations through Whitening (Su et al., 2021).

3 Preliminary

3.1 Problem Formulation

We denote the finite vocabulary by $\mathcal{V} \triangleq \{w_1, w_2, \ldots, w_n\}$ and define $\mathcal{V}^* \triangleq \bigcup_{i=0}^{\infty} \mathcal{V}^i$, where $\mathcal{V}^i$ denotes the i-fold Cartesian product of $\mathcal{V}$ (with $\mathcal{V}^0 = \{\varepsilon\}$, where ε is the empty string). Without loss of generality, we consider a large language mapping $\phi(\cdot \mid \theta) : \mathcal{V}^* \to \mathcal{V}^*$, which is usually composed by a large language model and its post training procedure. Typically, a mature and commercially deployed language model is trained on a dataset $\mathcal{P}$; in this work, we refer to $\mathcal{P}$ as the *data pool*.

To reduce the size of the data pool $\mathcal{P}$ while simultaneously improving the performance of the corresponding language model, we aim to select an optimal subset $S \subseteq \mathcal{P}$ such that the model exhibits stronger capability in solving tasks from a specified query dataset

$$Q \triangleq \{(t_1, a_1), (t_2, a_2), \ldots, (t_m, a_m)\}, \tag{1}$$

where $t_j, a_j \in \mathcal{V}^*$ denote the j-th query input and its ground-truth answer, respectively. Formally, we seek

$$\arg\max_{\substack{S \subseteq \mathcal{P} \\ |S|=k}} \sum_{j=1}^{|Q|} \mathbb{I}\{\phi_S(t_j) = a_j\}, \tag{2}$$

where ϕ_S denotes the language model trained (or adapted) on the selected subset S, and $\mathbb{I}\{\cdot\}$ is the indicator function.

3.2 Selection Strategy

In this work, we select data based on representation similarity using representation engineering—specifically, by manipulating the hidden states of an LLM encoder. By convention, all representations in this paper are denoted as row vectors. Let the encoder be a mapping

$$f : \mathcal{V}^* \to \mathbb{R}^v, \tag{3}$$

where v denotes the embedding dimension. Given two embeddings $e_1, e_2 \in \mathbb{R}^v$, the cosine similarity is defined as

$$\cos(e_1, e_2) \;=\; \frac{e_1 e_2^\top}{\|e_1\|_2 \, \|e_2\|_2}. \tag{4}$$

Without loss of generality, for a query example $(t_j, a_j) \in Q$, we retrieve the most similar instances from the data pool $\mathcal{P}$ by maximizing the cosine similarity between their representations and that of the query input t_j. In practice, when forming a dataset at the scale k for a set of query examples $\{t_1, t_2, \ldots, t_m\}$, a round-robin algorithm is implemented to maintain the diversity of representations and ensure balanced coverage.

Formally, let S denote the set of selected instances, initialized as $S = \emptyset$. The selection process proceeds iteratively by cycling through the query indices $j \in \{1, \ldots, m\}$. In each step, we identify the optimal candidate x^* such that:

$$x^* = \underset{x_i \in \mathcal{P} \setminus S}{\arg\max} \cos\big(f(t_j), f(x_i)\big). \tag{5}$$

The selected instance is then appended to the set, $S \leftarrow S \cup \{x^*\}$. This cyclic procedure continues until the total number of retrieved instances reaches the target capacity, i.e., $|S| = k$. By alternating between query inputs, this strategy prevents the retrieved set from being disproportionately biased toward the nearest neighbors of a specific subset of query examples, thereby achieving a more uniform representation across the entire query distribution.

4 Compressed Representation Data Selection

4.1 A Pilot Study: Binarized Experiments

Though semantic representations extracted by state-of-the-art encoders or LLMs exhibit strong alignment with human semantic judgments and demonstrate impressive performance in downstream applications, the underlying reasons for their effectiveness and optimization remain insufficiently understood. To this end, we design a simple binarized experiment that removes the normalization step after embedding extraction and instead retains only the sign information of each vector component.

Formally, for an input z, let $e = f(z) \in \mathbb{R}^v$ and define its binarized representation $b(z) \in \{-1, +1\}^v$ element-wise as

$$b_i(z) \triangleq \begin{cases} +1, & e_i \geq 0, \\ -1, & e_i < 0. \end{cases} \tag{6}$$

Under this setting, we evaluate all four combinations of original and binarized embeddings for the query set and data pool (Table 1). The two asymmetric configurations retain performance close to the original–original baseline, while binarizing both sides yields the strongest result.

Table 1: GSM8K accuracy for all query/pool embedding combinations in the binarization study. The base model is Ling-mini-2.0 and the encoder is LLaMA-2-7B. Each value represents the average of three parallel experiments. See Table 23 for the remaining settings.

Query embedding \ Pool embedding	Original	Binarized
Original	87.64	87.72
Binarized	87.64	**88.35**

The best result is obtained when both the query set and data pool use binarized embeddings, improving GSM8K accuracy from 87.64 to 88.35. This indicates that the selected datasets remain well aligned

with the GSM8K data type even after retaining only the sign of each representation component. The result strongly suggests that the representations used for similarity computation are redundant and leave substantial room for optimization. These observations motivate our subsequent methods.

4.2 CRDS-R: Rademacher Projection-concatenation Based Data Selection

We now formalize CRDS-R at the level of Transformer hidden states. Let $z = (z_1, \ldots, z_T) \in \mathcal{V}^*$ be a tokenized input and let an L-block Transformer encoder produce token-level hidden-state matrices $Z^{(0)}(z), \ldots, Z^{(L)}(z) \in \mathbb{R}^{T \times v}$. The initial state $Z^{(0)}(z)$ consists of the token embeddings and the encoder's positional representation. Suppressing architecture-specific choices such as the normalization type, attention mask, and positional encoding, a Transformer block has the residual form

$$A^{(\ell)}(z) \triangleq Z^{(\ell-1)}(z) + \mathrm{MHA}_\ell\Big(\mathrm{Norm}_\ell^{\mathrm{att}}\Big(Z^{(\ell-1)}(z)\Big)\Big), \tag{7}$$

$$Z^{(\ell)}(z) \triangleq A^{(\ell)}(z) + \mathrm{FFN}_\ell\Big(\mathrm{Norm}_\ell^{\mathrm{ffn}}\Big(A^{(\ell)}(z)\Big)\Big), \qquad \ell \in [L]. \tag{8}$$

Let $\rho : \mathbb{R}^{T \times v} \to \mathbb{R}^v$ be the readout operator used by the representation encoder to map token-level states to a single row-vector representation. CRDS-R uses exactly the same tokenization and readout as the baseline; it changes only which Transformer layers contribute to the representation. Hence, the layerwise representation is

$$e_\ell(z) \triangleq \rho\Big(Z^{(\ell)}(z)\Big) \in \mathbb{R}^v, \qquad \ell \in [L]. \tag{9}$$

Under the baseline of Section 3.2, $f(z) = e_L(z)$, i.e., only the final Transformer layer is used. In contrast, CRDS-R retains representations from an ordered layer set $\mathcal{L} = (\ell_1, \ldots, \ell_H) \subseteq [L]$. In the consecutive-final-layer setting used in our main experiments, $\ell_h = L - H + h$; other choices of $\mathcal{L}$, such as systematic layer sampling, are covered by the same definition.

CRDS-R allocates a reduced dimension $w_h \in \mathbb{N}$ to each selected layer such that

$$\sum_{h=1}^{H} w_h = v. \tag{10}$$

When H divides v, we use the balanced allocation $w_h = v/H$ for every h. Otherwise, the w_h's are chosen to differ by at most one, which preserves the same total dimensionality without requiring divisibility. For each layer ℓ_h, independently sample a Rademacher projection matrix $P_h \in \mathbb{R}^{v \times w_h}$ with entries

$$(P_h)_{ab} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1, +1\}, \qquad a \in [v],\ b \in [w_h], \tag{11}$$

where $[q] \triangleq \{1, \ldots, q\}$. The projected representation from layer ℓ_h is then

$$r_h(z) \triangleq e_{\ell_h}(z) P_h \in \mathbb{R}^{w_h}. \tag{12}$$

The CRDS-R representation is the concatenation of the independently projected layerwise representations:

$$f_{\mathrm{R}}(z) \triangleq \mathrm{Concat}\big(r_1(z), \ldots, r_H(z)\big) = \big[\, e_{\ell_1}(z)P_1 \mid \cdots \mid e_{\ell_H}(z)P_H \,\big] \in \mathbb{R}^v. \tag{13}$$

Equation (10) guarantees that $f_{\mathrm{R}}(z)$ has exactly the same dimension as the original representation $f(z)$. We use f_{R} in place of f in the cosine-similarity selection rule of Section 3.2. In particular, for query t_j, CRDS-R selects the candidate maximizing $\cos\big(f_{\mathrm{R}}(t_j), f_{\mathrm{R}}(x_i)\big)$. The construction therefore preserves the selection pipeline and its representation dimension, while allowing the similarity score to combine information from multiple encoder layers.

4.3 CRDS-W: Whitening-based Dimensionality-reduced Data Selection

CRDS-W, where *W* denotes Whitening (Su et al., 2021), corrects the second-order geometry of the final-layer Transformer representation before similarity-based selection. Unlike CRDS-R, which aggregates information from multiple Transformer layers, CRDS-W keeps the encoder readout fixed and learns a single linear post-processing map from the data pool. This design targets the anisotropy and correlation that may remain in the final-layer representations used by the baseline.

Table 2: Experimental results on Ling 2.0 data pool (2.0M) using different data scales. The base model is Ling-mini-2.0. **Bold** indicates the highest accuracy excluding the full-dataset setting, while underlining indicates the second-highest accuracy. Results are averaged over five runs, except for the full-data baselines, which are reported from a single run. Sample standard deviations are shown in parentheses.

Scale	Method	GSM8K	MMLU	MBPP	BBH	GPQA	Average
70k	Full (100%)	90.75	78.50	84.07	79.68	43.24	75.25
	Random (3.5%)	$91.28_{(0.43)}$	$77.28_{(1.17)}$	$83.37_{(1.15)}$	$78.39_{(0.34)}$	$\underline{44.09}_{(0.67)}$	74.88
	Mid PPL (3.5%)	$91.22_{(0.66)}$	$76.21_{(1.54)}$	$83.28_{(1.43)}$	$76.76_{(2.96)}$	$44.05_{(1.15)}$	74.30
	Length (3.5%)	$89.02_{(1.72)}$	$75.35_{(2.04)}$	$74.85_{(1.06)}$	$57.74_{(3.52)}$	$\mathbf{49.86}_{(0.51)}$	69.36
	RDS+ (3.5%)	$91.10_{(0.31)}$	$\underline{77.84}_{(0.27)}$	$82.58_{(1.20)}$	$78.87_{(0.93)}$	$41.03_{(0.89)}$	74.28
	CRDS-R (3.5%)	$\underline{91.40}_{(1.06)}$	$77.36_{(1.08)}$	$\underline{83.75}_{(0.84)}$	$\underline{80.54}_{(0.36)}$	$42.79_{(0.40)}$	$\underline{75.17}$
	CRDS-W (3.5%)	$\mathbf{92.55}_{(0.56)}$	$\mathbf{77.99}_{(0.48)}$	$\mathbf{84.26}_{(1.09)}$	$\mathbf{81.04}_{(0.37)}$	$\underline{44.09}_{(0.69)}$	**75.99**
100k	Full (100%)	90.75	78.50	84.07	79.68	43.24	75.25
	Random (5%)	$\underline{91.48}_{(0.42)}$	$76.44_{(1.01)}$	$83.28_{(1.07)}$	$79.01_{(0.56)}$	$44.05_{(0.62)}$	74.85
	Mid PPL (5%)	$91.01_{(0.43)}$	$76.50_{(0.30)}$	$83.98_{(1.37)}$	$78.48_{(0.56)}$	$43.79_{(1.21)}$	74.75
	Length (5%)	$85.05_{(4.04)}$	$74.77_{(0.99)}$	$78.08_{(2.32)}$	$60.22_{(3.01)}$	$\mathbf{49.28}_{(1.86)}$	69.48
	RDS+ (5%)	$90.55_{(0.41)}$	$76.44_{(0.95)}$	$83.42_{(1.66)}$	$78.72_{(0.46)}$	$40.25_{(1.32)}$	73.88
	CRDS-R (5%)	$91.02_{(1.14)}$	$\underline{77.54}_{(0.66)}$	$\underline{84.26}_{(1.09)}$	$\mathbf{80.44}_{(0.38)}$	$42.78_{(1.18)}$	$\underline{75.21}$
	CRDS-W (5%)	$\mathbf{92.16}_{(0.33)}$	$\mathbf{77.99}_{(0.30)}$	$\mathbf{85.20}_{(0.82)}$	$\underline{80.10}_{(0.31)}$	$\underline{45.20}_{(0.93)}$	**76.13**

Let $\mathcal{F} = \{z_q\}_{q=1}^{F} \subseteq \mathcal{P}$ be a randomly sampled fitting subset of the data pool, and let $u_q \triangleq f(z_q) \in \mathbb{R}^v$ be its final-layer representation, where $f(z) = e_L(z)$ follows the Transformer notation in Section 4.2. We estimate the empirical mean and covariance on $\mathcal{F}$ as

$$\mu \triangleq \frac{1}{F}\sum_{q=1}^{F} u_q, \tag{14}$$

$$C \triangleq \frac{1}{F}\sum_{q=1}^{F}(u_q-\mu)^\top(u_q-\mu) \in \mathbb{R}^{v\times v}. \tag{15}$$

Let $C = U\Lambda U^\top$ be the eigendecomposition of C, where U is orthogonal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_v)$ with $\lambda_1 \geq \cdots \geq \lambda_v \geq 0$. For a target dimension $\beta \leq v$ satisfying $\lambda_\beta > 0$, let $U_\beta \in \mathbb{R}^{v\times\beta}$ contain the first β eigenvectors and let $\Lambda_\beta \in \mathbb{R}^{\beta\times\beta}$ contain the corresponding eigenvalues. The truncated Whitening map is

$$W_\beta \triangleq U_\beta \Lambda_\beta^{-1/2} \in \mathbb{R}^{v\times\beta}. \tag{16}$$

For any pool or query instance z, CRDS-W uses the transformed representation

$$f_{\mathrm{W}}(z) \triangleq \big(f(z)-\mu\big)W_\beta \in \mathbb{R}^\beta. \tag{17}$$

Consequently, the fitting subset has identity empirical covariance in the retained subspace: $\frac{1}{F}\sum_{q=1}^{F} f_{\mathrm{W}}(z_q)^\top f_{\mathrm{W}}(z_q) = I_\beta$. The same fitted pair (μ, W_β) is then applied to all data-pool and query representations. We replace f by f_{W} in the selection rule of Section 3.2. Thus, CRDS-W preserves the original selection procedure while comparing examples in a β-dimensional, decorrelated representation space. Theoretical analysis is provided in Appendix C.

5 Experiments

We evaluate CRDS-R and CRDS-W against four data-selection baselines: random selection, data length, perplexity, and semantic representation similarity. Experiments use Ling-mini-2.0 (Team & AI, 2025), LLaMA-3.1-8B (Team, 2024), and Qwen-2.5-7B (Yang et al., 2024a), with the distributed framework in Appendix B supporting asymmetric pool and query scales.

Table 3: Results on Tulu3-SFT-Mixture (939K), with 30K examples selected for LLaMA-3.1-8B and Qwen-2.5-7B. Results are averaged over three runs, except for the full-data baselines, which are reported from a single run. Sample standard deviations are shown in parentheses.

Model	Method	GSM8K	MMLU	MBPP	BBH	GPQA	Average
LLaMA-3.1-8B	Full (100%)	82.71	65.74	62.65	65.14	30.30	61.31
	Random (3.2%)	70.71$_{(0.27)}$	65.93$_{(0.32)}$	59.66$_{(1.62)}$	66.49$_{(0.12)}$	25.42$_{(0.58)}$	57.64
	Mid PPL (3.2%)	60.40$_{(0.72)}$	**66.05**$_{(0.32)}$	60.83$_{(1.37)}$	**67.31**$_{(0.30)}$	27.61$_{(1.62)}$	56.44
	Length (3.2%)	53.58$_{(0.61)}$	64.91$_{(0.11)}$	60.31$_{(1.41)}$	60.37$_{(1.63)}$	26.59$_{(0.32)}$	53.15
	RDS+ (3.2%)	62.80$_{(8.85)}$	65.92$_{(0.25)}$	59.40$_{(1.25)}$	66.99$_{(0.43)}$	28.79$_{(0.51)}$	56.78
	CRDS-R (3.2%)	**75.34**$_{(0.55)}$	66.00$_{(0.33)}$	61.87$_{(1.35)}$	66.87$_{(0.27)}$	29.12$_{(1.27)}$	**59.84**
	CRDS-W (3.2%)	72.71$_{(0.53)}$	65.59$_{(0.12)}$	**63.29**$_{(0.22)}$	67.11$_{(0.31)}$	**29.13**$_{(0.58)}$	59.57
Qwen-2.5-7B	Full (100%)	88.70	74.56	74.71	63.97	39.39	68.27
	Random (3.2%)	83.62$_{(1.82)}$	74.70$_{(0.10)}$	73.54$_{(0.39)}$	63.39$_{(0.83)}$	32.99$_{(1.17)}$	65.65
	Mid PPL (3.2%)	83.32$_{(0.43)}$	**75.12**$_{(0.12)}$	72.76$_{(1.03)}$	64.11$_{(0.16)}$	33.84$_{(1.01)}$	65.83
	Length (3.2%)	77.21$_{(1.27)}$	74.48$_{(0.03)}$	**75.62**$_{(0.23)}$	**69.22**$_{(0.38)}$	34.68$_{(2.54)}$	66.24
	RDS+ (3.2%)	87.06$_{(0.99)}$	74.68$_{(0.14)}$	72.76$_{(0.68)}$	64.10$_{(0.35)}$	35.52$_{(0.29)}$	66.82
	CRDS-R (3.2%)	**88.45**$_{(0.31)}$	75.06$_{(0.09)}$	73.93$_{(1.56)}$	64.88$_{(0.39)}$	35.19$_{(2.33)}$	**67.50**
	CRDS-W (3.2%)	87.57$_{(0.93)}$	74.77$_{(0.07)}$	73.67$_{(0.90)}$	64.35$_{(0.82)}$	**36.03**$_{(0.77)}$	67.28

5.1 Experimental Results

We select 70K and 100K examples from Ling 2.0 for Ling-mini-2.0 (Table 2), and 30K from Tulu3-SFT-Mixture for LLaMA-3.1-8B and Qwen-2.5-7B (Table 3). Both CRDS variants outperform the selection baselines in average accuracy across these settings.

Ling-mini-2.0. CRDS-W ranks first and CRDS-R second among selection methods at both budgets. Across five benchmarks, CRDS-W exceeds full-data training by 0.74 and 0.88 percentage points at 70K and 100K, respectively, while CRDS-R nearly matches it. Both improve on RDS+ in average accuracy.

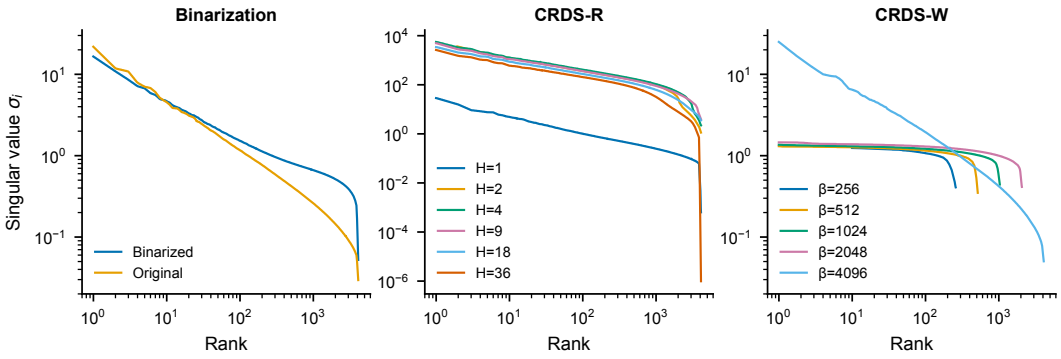
Random selection ranks behind only CRDS. Length-based selection varies considerably across tasks, yet exceeds full-data training on GPQA by 6.62 percentage points at 70K. CRDS-W exceeds the full-data baseline on GSM8K, MBPP, and BBH at both budgets. These gains show that optimizing representations can improve selection even when random subsets already provide a competitive baseline.

Small LLMs. CRDS-R ranks above CRDS-W on both smaller models. Relative to RDS+, CRDS-R improves average accuracy by 3.06 points on LLaMA-3.1-8B and 0.68 on Qwen-2.5-7B; CRDS-W gains 2.79 and 0.46 points, respectively. This comparison reflects different model–data combinations and a broader search for CRDS-R over two encoders and three layer configurations per encoder.

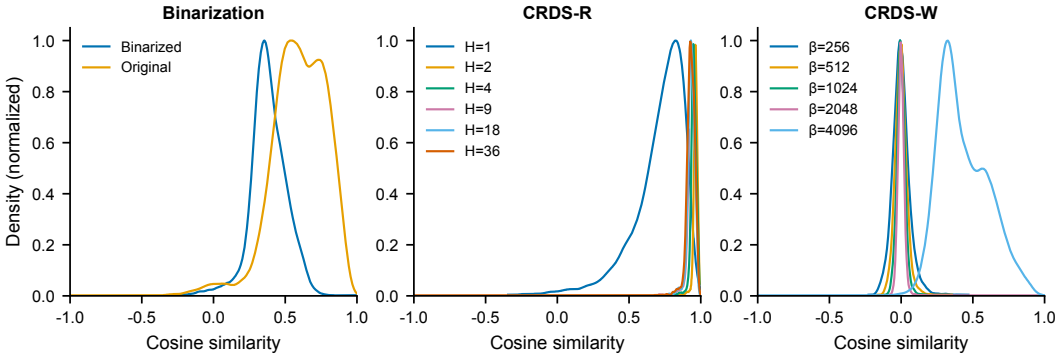
The strongest method varies by benchmark, but CRDS achieves the best average among selection methods. CRDS-R leads on GSM8K for both base models, while CRDS-W leads on GPQA and, for LLaMA-3.1-8B, MBPP. These complementary strengths suggest that representation compression benefits different task-specific selection signals. Both smaller models remain below full-data training on average; this gap may reflect model capacity, data composition, and the smaller selection fraction (3.2%).

5.2 Mechanistic Analysis of Representation Geometry

Figure 2 compares singular-value spectra and pairwise cosine similarities. Relative to $H = 1$, multilayer CRDS-R exhibits larger singular values and similarities concentrated near one, consistent with stronger anisotropy. CRDS-W instead produces a flatter retained spectrum and similarities centered near zero at moderate target dimensions, reducing the dominance of a few principal directions. Binarization also smooths the spectrum and shifts similarities downward, suggesting more evenly distributed representational energy. Removing coordinate magnitudes can therefore act as a simple regularizer, weaker than Whitening, that limits their influence on similarity. These observations suggest distinct geometric mechanisms for the three transformations.



(a) Singular-value spectra for Binarization, CRDS-R, and CRDS-W (left to right).



(b) Cosine-similarity distributions for Binarization, CRDS-R, and CRDS-W (left to right).

Figure 2: Representation geometry under three transformations. Binarization smooths the spectrum and lowers cosine similarity; multilayer CRDS-R concentrates similarities near one; CRDS-W flattens the retained spectrum and centers similarities near zero at moderate target dimensions.

6 Ablation Study

We examine layer aggregation and Whitening dimensionality in Tables 4 and 5. Ling-mini-2.0 uses a narrow set of configurations across tasks, whereas smaller-model experiments select from broader hyperparameter sweeps. Additional results appear in Appendix J.

6.1 Effect of Hidden Layers in CRDS-R

We study the number of extracted hidden states on Ling-mini-2.0 using Qwen-2.5-7B, Qwen-3-8B, LLaMA-2-7B, and LLaMA-3.1-8B. We sweep $H \in \{1, 2, 4, 7, 14, 28\}$ for Qwen-2.5-7B, $H \in \{1, 2, 4, 9, 18, 36\}$ for Qwen-3-8B, and $H \in \{1, 2, 4, 8, 16, 32\}$ for the LLaMA encoders. Figure 5 shows that performance is sensitive to this choice: Qwen-2.5-7B and Qwen-3-8B peak at $H = 14$ and $H = 18$, while gains are weaker for the LLaMA encoders. Since Qwen-3-8B also outperforms Qwen-2.5-7B (Appendix G.2), we use Qwen-3-8B with $H = 18$ in Table 2.

Table 4 shows that aggregating final layers generally improves over the final layer alone: Qwen-3-8B peaks at $H = 18$, and the LLaMA encoders at $H = 16$ or $H = 32$. This supports a *strong-tail* interpretation: useful selection signals extend across the encoder's later layers. The best configurations improve GSM8K by 2.83 and 2.33 percentage points for the Qwen encoders, compared with 1.34 and 0.79 for LLaMA-2-7B and LLaMA-3.1-8B. Table 24 shows similar benefits across datasets with LLaMA-3.1-8B as the base model.

6.2 Selection of Whitening Target Dimensionality β

Because effective representation dimensionality varies across tasks (Su et al., 2021), we study the effect of the Whitening target dimension on downstream performance. This controls the tradeoff

Table 4: Effect of layer aggregation in CRDS-R on GSM8K with Ling-mini-2.0 as the base model. Each value represents the average of three parallel experiments, with the sample standard deviation shown in parentheses. Δ denotes the accuracy gain in percentage points relative to $H = 1$ for the same encoder. The best mean per encoder is bolded.

LLaMA-2-7B			**LLaMA-3.1-8B**			**Qwen-3-8B**			**Qwen-2.5-7B**		
H	Acc. (%)	Δ	H	Acc. (%)	Δ	H	Acc. (%)	Δ	H	Acc. (%)	Δ
1	$86.86_{(1.13)}$	0.00	1	$87.41_{(0.86)}$	0.00	1	$86.53_{(1.56)}$	0.00	1	$86.98_{(0.36)}$	0.00
4	$86.20_{(0.99)}$	−0.66	4	$87.79_{(0.13)}$	+0.38	4	$87.52_{(0.46)}$	+0.99	4	$87.85_{(0.12)}$	+0.87
8	$86.40_{(0.76)}$	−0.46	8	$87.57_{(0.94)}$	+0.16	9	$88.25_{(1.01)}$	+1.72	7	$87.72_{(0.93)}$	+0.74
16	$\mathbf{88.20}_{(1.00)}$	+1.34	16	$87.31_{(0.76)}$	−0.10	18	$\mathbf{89.36}_{(0.96)}$	+2.83	14	$\mathbf{89.31}_{(0.40)}$	+2.33
32	$87.31_{(0.38)}$	+0.45	32	$\mathbf{88.20}_{(0.36)}$	+0.79	36	$89.24_{(0.60)}$	+2.71	28	$87.64_{(0.68)}$	+0.66

Table 5: Ablation study of CRDS-W. The base model is LLaMA-3.1-8B, and the encoder is LLaMA-3.1-8B. Each value represents the average of three parallel experiments, with the sample standard deviation shown in parentheses. Δ denotes the gain in percentage points from $\beta = 4096$ to the highest mean in each row. The highest mean is bolded.

Dataset	β					Δ
	512	1024	1536	2048	4096 (original)	
GSM8K	$71.57_{(0.93)}$	$\mathbf{72.71}_{(0.53)}$	$72.08_{(1.93)}$	$72.45_{(1.35)}$	$69.72_{(2.49)}$	+2.99
MMLU	$65.59_{(0.12)}$	$65.54_{(0.04)}$	$65.56_{(0.33)}$	$65.47_{(0.49)}$	$\mathbf{65.64}_{(0.06)}$	0.00
MBPP	$61.48_{(1.35)}$	$\mathbf{63.29}_{(0.22)}$	$62.13_{(1.62)}$	$62.00_{(0.90)}$	$60.83_{(0.98)}$	+2.46
BBH	$66.38_{(0.15)}$	$66.89_{(0.43)}$	$66.72_{(0.32)}$	$\mathbf{67.11}_{(0.31)}$	$66.51_{(0.72)}$	+0.60
GPQA	$28.79_{(2.53)}$	$\mathbf{29.13}_{(0.58)}$	$28.95_{(0.58)}$	$26.94_{(1.05)}$	$27.27_{(1.75)}$	+1.86

between retaining useful information and suppressing redundant or noisy directions; retaining more dimensions need not improve selection.

For Ling-mini-2.0, we use LLaMA-2-7B (dimension 4096) and sweep $\beta \in \{128, 256, 512, 768, 1024, 2048, 4096\}$. Performance peaks at $\beta = 768$ in Figure 5, retaining about 20% of the original dimension. Table 2 reports the better of 512 and 1024: BBH uses 1024, and all other tasks use 512 (Appendix G.3).

For smaller models, we evaluate LLaMA-3.1-8B as both encoder and base model (Table 5). With $\beta = 1024$, CRDS-W improves GSM8K, MBPP, GPQA, and BBH by 2.99, 2.46, 1.86, and 0.38 percentage points, respectively, over the original 4096-dimensional representations. BBH achieves its best result at $\beta = 2048$, with a gain of 0.60 percentage points. These results demonstrate that whitening-based dimensionality reduction can improve data selection across multiple tasks.

7 Conclusion

In this work, we systematically investigate redundancy in semantic representations for instruction-tuning data selection. We propose a distributed computing strategy that accommodates the asymmetric scales of candidate and query data, and develop an effective framework for selecting instruction-tuning data. After identifying substantial redundancy in existing representations, we introduce two methods, CRDS-R and CRDS-W, based on Rademacher projection–concatenation and Whitening, respectively, to optimize semantic representations for similarity computation. Our approach almost always outperforms state-of-the-art data selection baselines, and in the Ling-mini-2.0 setting even surpasses the full-data baseline. It also performs competitively on smaller LLMs, almost always outperforming the compared baselines. These results show that semantic representations remain an effective tool for instruction tuning data selection, complementing methods based on gradient similarity. More broadly, by revealing fundamental redundancy in state-of-the-art encoder representations, our work suggests that representation optimization may also benefit other embedding-based applications, such as clustering, recommendation, and search. These applications represent promising directions for extending this work.

REFERENCES

Richard J. Antonello, Nicole Beckage, Javier Turek, and Alexander Huth. Selecting informative contexts improves language model fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pp. 1072–1085. Association for Computational Linguistics, 2021.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosiute, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: harmlessness from AI feedback. *CoRR*, abs/2212.08073, 2022.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pp. 3411–3425. Association for Computational Linguistics, 2024.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Yicheng Chen, Yining Li, Kai Hu, Zerun Ma, Haochen Ye, and Kai Chen. MIG: automatic data selection for instruction tuning by maximizing information gain in semantic space. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 9902–9915. Association for Computational Linguistics, 2025.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James R. Glass. Diffcse: Difference-based contrastive learning for sentence embeddings. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pp. 4207–4218. Association for Computational Linguistics, 2022.

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. *CoRR*, abs/2110.14168, 2021.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 55–65. Association for Computational Linguistics, 2019.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 6894–6910. Association for Computational Linguistics, 2021.

Yuan Ge, Yilun Liu, Chi Hu, Weibin Meng, Shimin Tao, Xiaofeng Zhao, Mahong Xia, Zhang Li, Boxing Chen, Hao Yang, Bei Li, Tong Xiao, and JingBo Zhu. Clustering and ranking: Diversity-preserved instruction selection through expert-aligned quality estimation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 464–478. Association for Computational Linguistics, 2024.

Yexiao He, Ziyao Wang, Zheyu Shen, Guoheng Sun, Yucong Dai, Yongkai Wu, Hongyi Wang, and Ang Li. SHED: shapley-based automated dataset refinement for instruction fine-tuning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Hamish Ivison, Muru Zhang, Faeze Brahman, Pang Wei Koh, and Pradeep Dasigi. Large-scale data selection for instruction tuning. *CoRR*, abs/2503.01807, 2025.

Tingyu Jiang, Shen Li, Yiyao Song, Lan Zhang, Hualei Zhu, Yuan Zhao, Xiaohang Xu, Kenjiro Taura, and Hao Henry Wang. Importance-aware data selection for efficient LLM instruction tuning. *CoRR*, abs/2511.07074, 2025.

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, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *CoRR*, abs/2411.15124, 2024.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 9119–9130. Association for Computational Linguistics, 2020.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 14255–14273. Association for Computational Linguistics, 2024a.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 7602–7635. Association for Computational Linguistics, 2024b.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Ling-Hao Chen, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. One-shot learning as instruction data prospector for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pp. 4586–4601. Association for Computational Linguistics, 2024c.

Huawei Lin, Jikai Long, Zhaozhuo Xu, and Weijie Zhao. Token-wise influential training data retrieval for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 841–860. Association for Computational Linguistics, 2024.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a.

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 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. TSDS: data selection for task-specific model finetuning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024c.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Baharan Mirzasoleiman, Jeff A. Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6950–6960. PMLR, 2020.

Manish Nagaraj, Sakshi Choudhary, Utkarsh Saxena, Deepak Ravikumar, and Kaushik Roy. TRIM: token-wise attention-derived saliency for data-efficient instruction tuning. *CoRR*, abs/2510.07118, 2025.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. Text and code embeddings by contrastive pre-training. *CoRR*, abs/2201.10005, 2022.

Xinzhe Ni, Yeyun Gong, Zhibin Gou, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Exploring the mystery of influential data for mathematical reasoning. *CoRR*, abs/2404.01067, 2024.

Mahdi Nikdan, Vincent Cohen-Addad, Dan Alistarh, and Vahab Mirrokni. Efficient data selection at scale via influence distillation. *CoRR*, abs/2505.19051, 2025.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Xingyuan Pan, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, and Shanbo Cheng. G-DIG: towards gradient-based diverse and high-quality instruction data selection for machine translation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 15395–15406. Association for Computational Linguistics, 2024.

Jinlong Pang, Jiaheng Wei, Ankit Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. Improving data efficiency via curating llm-driven rating systems. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Wenjie Qu, Wengrui Zheng, Tianyang Tao, Dong Yin, Yanze Jiang, Zhihua Tian, Wei Zou, Jinyuan Jia, and Jiaheng Zhang. Provably robust multi-bit watermarking for ai-generated text. In Lujo Bauer and Giancarlo Pellegrino (eds.), *34th USENIX Security Symposium, USENIX Security 2025, Seattle, WA, USA, August 13-15, 2025*, pp. 201–220. USENIX Association, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Shrihari Sridharan, Deepak Ravikumar, Anand Raghunathan, and Kaushik Roy. Gradientspace: Unsupervised data clustering for improved instruction tuning. *CoRR*, abs/2512.06678, 2025.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *CoRR*, abs/2103.15316, 2021.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023.

Ling Team and Inclusion AI. Every activation boosted: Scaling general reasoner to 1 trillion open language foundation. *CoRR*, abs/2510.22115, 2025.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

Jingtan Wang, Xiaoqiang Lin, Rui Qiao, Pang Wei Koh, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. NICE data selection for instruction tuning in llms with non-differentiable evaluation metric. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025a.

Pingjie Wang, Hongcheng Liu, Yusheng Liao, Ziqing Fan, Yaxin Du, Shuo Tang, Yanfeng Wang, and Yu Wang. Selecting auxiliary data via neural tangent kernels for low-resource domains. *CoRR*, abs/2511.07380, 2025b.

Minghao Wu, Thuy-Trang Vu, Lizhen Qu, and Gholamreza Haffari. The best of both worlds: Bridging quality and diversity in data selection with bipartite graph. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024a.

Tingyu Xia, Bowen Yu, Kai Dang, An Yang, Yuan Wu, Yuan Tian, Yi Chang, and Junyang Lin. Rethinking data selection at scale: Random selection is almost all you need. *CoRR*, abs/2410.09335, 2024b.

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, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024a.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025a.

Xinyu Yang, Junlin Han, Rishi Bommasani, Jinqi Luo, Wenjie Qu, Wangchunshu Zhou, Adel Bibi, Xiyao Wang, Jaehong Yoon, Elias Stengel-Eskin, Shengbang Tong, Lingfeng Shen, Rafael Rafailov, Runjia Li, Zhaoyang Wang, Yiyang Zhou, Chenhang Cui, Yu Wang, Wenhao Zheng, Huichi Zhou, Jindong Gu, Zhaorun Chen, Peng Xia, Tony Lee, Thomas P. Zollo, Vikash Sehwag, Jixuan Leng, Jiuhai Chen, Yuxin Wen, Huan Zhang, Zhun Deng, Linjun Zhang, Pavel Izmailov, Pang Wei Koh, Yulia Tsvetkov, Andrew Gordon Wilson, Jiaheng Zhang, James Zou, Cihang Xie, Hao Wang, Philip Torr, Julian J. McAuley, David Alvarez-Melis, Florian Tramèr, Kaidi Xu, Suman Jana, Chris Callison-Burch, René Vidal, Filippos Kokkinos, Mohit Bansal, Beidi Chen, and Huaxiu Yao. Reliable and responsible foundation models. *Trans. Mach. Learn. Res.*, 2025, 2025b.

Yixin Yang, Qingxiu Dong, Linli Yao, Fangwei Zhu, and Zhifang Sui. Rico: Refined in-context contribution for automatic instruction-tuning data selection. *CoRR*, abs/2505.05327, 2025c.

Yu Yang, Siddhartha Mishra, Jeffrey N. Chiang, and Baharan Mirzasoleiman. Smalltolarge (S2L): scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. LIMO: less is more for reasoning. *CoRR*, abs/2502.03387, 2025.

Junjie Oscar Yin and Alexander M. Rush. Compute-constrained data selection. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

Jia Zhang, Chen-Xi Zhang, Yao Liu, Yi-Xuan Jin, Xiao-Wen Yang, Bo Zheng, Yi Liu, and Lan-Zhe Guo. D3: diversity, difficulty, and dependability-aware data selection for sample-efficient LLM instruction tuning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pp. 8348–8356. ijcai.org, 2025a.

Jipeng Zhang, Yaxuan Qin, Renjie Pi, Weizhong Zhang, Rui Pan, and Tong Zhang. TAGCOS: task-agnostic gradient clustered coreset selection for instruction tuning data. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 4671–4686. Association for Computational Linguistics, 2025b.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Hao Zhao, Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. Long is more for alignment: A simple but tough-to-beat baseline for instruction fine-tuning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *CoRR*, abs/2403.13372, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: less is more for alignment. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.