Measuring Data Diversity for Instruction Tuning: A Systematic Analysis and A Reliable Metric

Yuming Yang[1*], Yang Nan[1*], Junjie Ye[1], Shihan Dou[1], Xiao Wang[1], Shuo Li[1], Huijie Lv[1], Mingqi Wu[1], Tao Gui[2†], Qi Zhang[1†], Xuanjing Huang[1]

[1] School of Computer Science, Fudan University

[2] Institute of Modern Languages and Linguistics, Fudan University

`yumingyang23@m.fudan.edu.cn` `{qz,tgui}@fudan.edu.cn`

Abstract

Data diversity is crucial for the instruction tuning of large language models. Existing studies have explored various diversity-aware data selection methods to construct high-quality datasets and enhance model performance. However, the fundamental problem of precisely defining and measuring data diversity remains underexplored, limiting clear guidance for data engineering. To address this, we systematically analyze 11 existing diversity measurement methods by evaluating their correlation with model performance through extensive fine-tuning experiments. Our results indicate that a reliable diversity measure should properly account for both inter-sample differences and the information distribution in the sample space. Building on this, we propose *NovelSum*, a new diversity metric based on sample-level "novelty." Experiments on both simulated and real-world data show that *NovelSum* accurately captures diversity variations and achieves a 0.97 correlation with instruction-tuned model performance, highlighting its value in guiding data engineering practices. With *NovelSum* as an optimization objective, we further develop a greedy, diversity-oriented data selection strategy that outperforms existing approaches, validating both the effectiveness and practical significance of our metric.

1 Introduction

Instruction tuning (IT) fine-tunes pretrained large language models (LLMs) with annotated instruction data, enabling them to follow human instructions and perform various tasks effectively (Sanh et al., 2022; Zhang et al., 2023). Recent studies indicate that small-scale, high-quality datasets can outperform larger ones in IT performance (Chen et al., 2023; Zhou et al., 2024), with data diversity playing a crucial role in achieving optimal results (Liu et al., 2023; Bukharin et al., 2024; Zhang et al., 2024; Yang et al., 2025). Consequently, various diversity-aware data selection methods have emerged (Qin et al., 2024; Wang et al., 2024a), driven by different interpretations of data diversity.

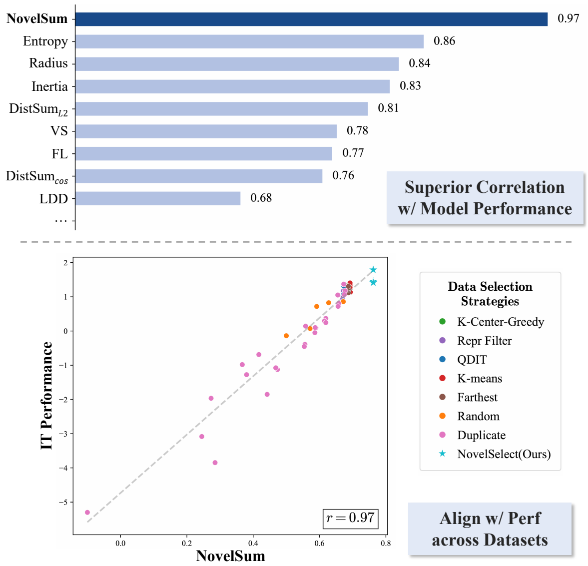


Figure 1: Our diversity metric, *NovelSum*, exhibits superior correlation with model performance compared to existing metrics across IT datasets constructed with various data selection strategies.

However, the fundamental problem of precisely defining and measuring data diversity remains underexplored. This ambiguity has turned data engineering for diversity into a black-box process, leading to data selection methods that often fail to generalize and, at times, perform worse than random selection (Xia et al., 2024; Diddee and Ippolito, 2024). While some diversity metrics have been introduced in IT research (Bukharin et al., 2024; Wang et al., 2024b), a comprehensive evaluation and comparative analysis are still needed to identify a reliable metric that strongly correlates with fine-tuning performance in practice.

To this end, we systematically analyze 11 existing diversity metrics by evaluating their reliability through extensive experiments. Using various

[*]Equal Contribution.
[†]Corresponding Authors.

mainstream diversity-oriented data selection methods, we construct 53 IT datasets and fine-tune models accordingly. We then measure dataset diversity using existing metrics and assess their correlation with model performance. By analyzing the limited correlation of existing metrics, we find that: (1) **A reliable diversity metric must capture differences between samples** to reflect each sample's information uniqueness. Moreover, differences between neighboring samples are more critical for overall diversity but can be overshadowed by variations in distant samples. (2) **Measuring differences between samples should account for both semantic similarity and the uneven distribution of information in space.** In high-density domains like math and code, semantically similar samples can still contain substantial unique information and should therefore be considered more diverse.

Building on these insights, we propose *NovelSum*, a diversity metric that jointly considers inter-sample differences and uneven information density. Specifically, we define dataset diversity as the sum of each sample's unique contribution to overall information, termed "novelty". Just as a research paper's novelty is judged by its distinction from related work based on field-specific standards, we compute a sample's novelty as the proximity-weighted sum of its differences from other samples in the dataset. These differences are measured using density-aware distances, which capture both semantics and local information density.

To validate the effectiveness of *NovelSum*, we conduct both a visualized simulation study and real-world correlation experiments using two different LLMs. The results show that *NovelSum* accurately captures diversity variations and strongly correlates with instruction-tuned model performance, achieving Pearson's $r = 0.98$ and Spearman's $r = 0.95$, outperforming other metrics. This demonstrates *NovelSum*'s potential to effectively guide data engineering practices. Furthermore, we develop *NovelSelect*, a greedy, diversity-oriented data selection strategy that uses *NovelSum* as the optimization objective. Experimental results confirm its superior performance compared to other approaches.

Our main contributions are three-fold:

- We systematically analyze and evaluate the reliability of existing diversity metrics for instruction tuning by computing their correlation with model performance, thereby unveiling pathways to a more reliable metric.
- We propose *NovelSum*, a diversity metric that captures both inter-sample differences and information density, achieving a strong correlation with instruction-tuning performance, substantially exceeding previous metrics.
- We develop *NovelSelect*, a diversity-oriented data selection strategy based on *NovelSum*, which outperforms existing methods and further validates *NovelSum*'s effectiveness and practical value in instruction tuning.

2 Evaluating Existing Diveristy Metrics

We begin by evaluating the correlation between existing diversity metrics and instruction-tuned model performance, identifying limitations to inform the design of a more reliable metric.

Our evaluation follows four steps: (1) Construct multiple IT datasets, each denoted as $\mathcal{X}^{(s)}$, using different data selection strategies from the full data source $\mathcal{X}^{all}$. (2) Measure dataset diversity using existing metrics, denoted as $\mathcal{M}_t(\mathcal{X}^{(s)})$. (3) Fine-tune LLMs on each dataset and evaluate their performance, $\mathcal{P}^{(s)}$, using IT benchmarks. (4) Analyze the correlation between each diversity metric and model performance, denoted as $r_{\mathcal{M}_t,\, \mathcal{P}}$.

2.1 Existing Diversity Metrics

We use 11 existing diveristy metrics for the analysis, categoried into three main types:

Lexical Diversity A classical way to measure textual diversity is by analyzing vocabulary usage, where a higher proportion of unique words indicates greater diversity. Two widely used metrics are the **Type-Token Ratio** (TTR) (Richards, 1987) and **vocd-D** (Malvern et al., 2004), with details in the Appendix A.2.

Distance-based Semantic Diversity Recent studies primarily measure dataset diversity based on the semantics of individual samples, often represented as embeddings $emb(\cdot)$ from language models like BERT. A common approach quantifies diversity by computing distances between samples using their embeddings, encouraging heterogeneity. For example, a straightforward metric sums the pairwise distances among all samples in a dataset:

$$\mathcal{M}_{DistSum}(\mathcal{X}) = \sum_{x_i, x_j \in \mathcal{X}, i \neq j} \Delta(x_i, x_j), \quad (1)$$

where $\Delta(\cdot, \cdot)$ denotes the distances between two samples. Specifically, **DistSum**$_{cosine}$ uses cosine distance and **DistSum**$_{L2}$ uses Euclidean distance. Beyond simple summation, more refined metrics are proposed. The **KNN distance** (Stasaski et al., 2020; Stasaski and Hearst, 2022) measures the average distance of each sample to its k-nearest neighbor, ensuring sample uniqueness:

$$\mathcal{M}_{KNN}(\mathcal{X}) = \frac{1}{|\mathcal{X}|}\sum_{i=1}^{|\mathcal{X}|}\Delta(x_i, N_k(x_i)), \tag{2}$$

where $N_k(x_i)$ denotes the k-th closest neighbor of x_i, typically with $k = 1$. We also compute **Cluster Inertia** (Du and Black, 2019), **Vendi Score** (Pasarkar and Dieng, 2023), **Radius** (Lai et al., 2020) and **Log Determinant Distance** (LDD) (Wang et al., 2024b). Further details are provided in the Appendix A.2.

Distribution-based Semantic Diversity Another notable class of metrics measures diversity from a distributional perspective, assessing how well a selected dataset $\mathcal{X}$ represents the overall sample (semantic) space of $\mathcal{X}^{all}$. One example is the **Facility Location** (FL) function (Farahani and Hekmatfar, 2009), which defines a dataset as diverse if each sample in $\mathcal{X}^{all}$ has a close representative in $\mathcal{X}$, ensuring thorough coverage of space:

$$\mathcal{M}_{FL}(\mathcal{X}) = \sum_{x_j \in \mathcal{X}^{all}} \min_{x_i \in \mathcal{X}} \Delta(x_i, x_j) \tag{3}$$

Another feasible metric, **Partition Entropy**, captures how evenly the selected dataset spans the sample space. It partitions $\mathcal{X}^{all}$ into K clusters using K-means and calculates the entropy of the cluster membership distribution of $\mathcal{X}$.

$$\mathcal{M}_{Entropy}(\mathcal{X}) = -\sum_{k=1}^{K} p_k \log p_k, \tag{4}$$

where p_k is the proportion of selected samples in cluster k. Higher entropy indicates greater distributional uncertainty and a more balanced dataset.

2.2 IT Dataset Construction and Benchmark

Following Liu et al., 2023, we use a combined dataset of WizardLM (Xu et al., 2024), ShareGPT (Chiang et al., 2023), and UltraChat (Ding et al., 2023) as our IT data source, denoted as $\mathcal{X}^{all}$.

We then apply several representative diversity-aware data selection strategies to curate IT datasets from the source $\mathcal{X}^{(s)} \subset \mathcal{X}^{all}$. To minimize the influence of factors beyond diversity, we control for sample quality differences across datasets by removing anomalous source samples and excluding any data quality filters during selection. We also fix the dataset size at 10,000 samples. The strategies used are: **K-Center-Greedy** (Sener and Savarese, 2017; Chen et al., 2023; Du et al., 2023; Wu et al., 2023), which iteratively selects the sample farthest from the current coreset; **Repr Filter** (Liu et al., 2023), which improves $\mathcal{M}_{KNN}$ by applying a minimum distance threshold when adding samples into the coreset; **QDIT** (Bukharin et al., 2024), which optimizes diversity by serially selecting the data point that maximizes $\mathcal{M}_{FL}$; **K-means** (Song et al., 2024), which partitions samples into clusters and evenly select samples from each; and baselines, including **Random** selection and **Farthest**, which ranks samples by their total distances to others and selects the most distant ones. Additionally, we construct datasets with varying amounts of **Duplicate** samples to simulate low-diversity datasets. Each strategy is run at least three times to ensure robustness, yielding 53 IT datasets. Details on dataset curation are in Appendices A.1 and A.3.

We fine-tune LLaMA-3-8B (Dubey et al., 2024) on these datasets and evaluate model performance using two popular IT benchmarks: MT-bench (Zheng et al., 2023) and AlpacaEval (Li et al., 2023). Both use GPT-4 (Achiam et al., 2023) for automatic evaluation, with AlpacaEval assessing single-turn dialogue and MT-bench on multi-turn conversations. To jointly consider both benchmarks, we normalize the results into Z-scores and compute the aggregated performance as

$$\mathcal{P}^{(s)} = z^{(s)}_{MT-bench} + z^{(s)}_{AlpacaEval} \tag{5}$$

2.3 Correlation Analysis

Finally, we compute the correlation between each diversity metric $\mathcal{M}_t$ and model performance $\mathcal{P}$ by averaging their Pearson and Spearman coefficients:

$$r_{\mathcal{M}_t,\,\mathcal{P}} = (r^{Pearson}_{\mathcal{M}_t,\,\mathcal{P}} + r^{Spearman}_{\mathcal{M}_t,\mathcal{P}})/2 \tag{6}$$

The results are shown in Figure 2, with additional plots in Appendix D. Since our experiments minimize the influence of other factors, we believe model performance directly reflects the impact of dataset diversity. Thus, the correlation between diversity metrics and model performance indicates a metric's practical reliability. Overall, we find that each metric favors datasets selected by its own criterion, but may not correlate well with performance, as it overlooks other aspects of diversity:

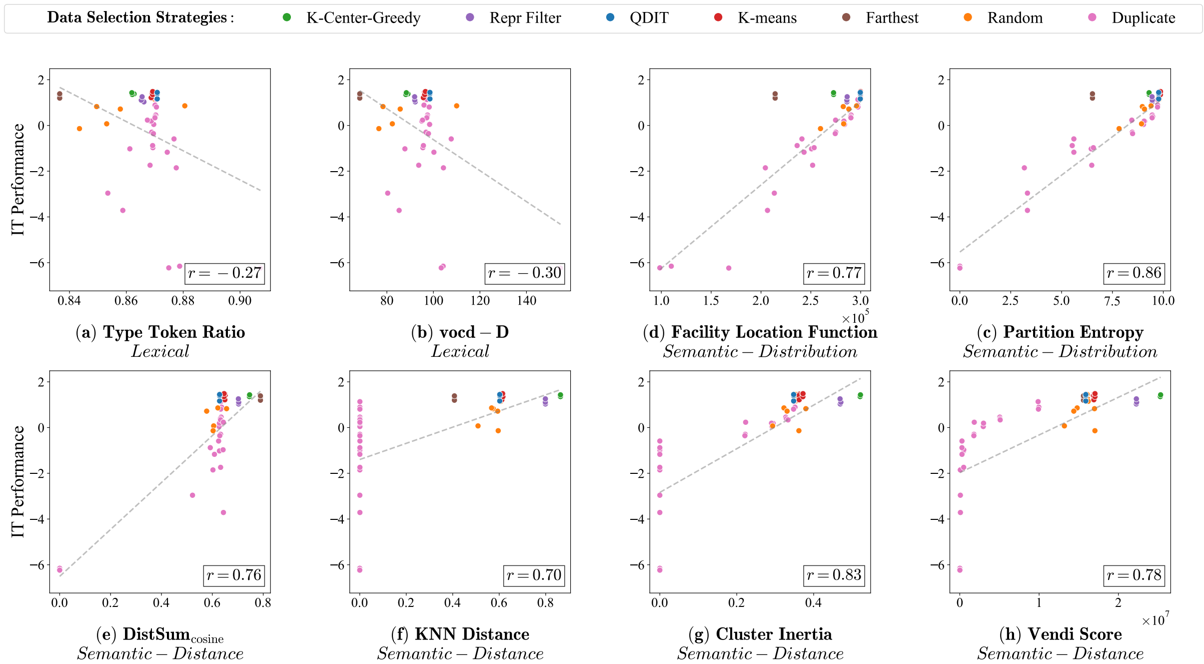


Figure 2: Evaluating existing diversity metrics based on their correlation (Eq. 6) with IT performance (Eq. 5). The X-axis represents diversity measurements. Each point corresponds to a 10k IT dataset constructed using different strategies. Abnormal points highlight the limitations of current metrics and inspire the development of new ones.

Findings 1 *Lexical diversity metrics fail to distinguish between different samples and datasets, showing weak correlation with model performance.*

As shown in Figure 2(a, b), high- and low-performance datasets exhibit similar lexical diversity. This likely results from the widespread use of diverse vocabulary in IT samples, making lexical diversity an ineffective measure for IT datasets.

Findings 2 *Since distribution-based semantic diversity metrics neglect sample uniqueness, they often underestimate the diversity of datasets with large inter-sample distances.*

From Figure 2(c, d), we observe that datasets selected by Farthest and K-Center-Greedy (brown and green points) achieve high IT performance but often receive relatively lower diversity scores from distribution-based diversity metrics, thus weakening their correlation with model performance. This likely occurs because these strategies all prioritize sample uniqueness by selecting samples that are distant from others, a factor not captured by distribution-based metrics. This suggests that overlooking sample uniqueness diminishes the reliability of diversity metrics.

Findings 3 *As distance-based semantic diversity metrics neglect information density in semantic space, they often underestimates datasets taht are close to the overall sample distribution and overestimates datasets with large inter-sample distances.*

From Figure 2(e, f, g, h), we observe common outliers in the fitting line for datasets selected by QDIT and K-means (blue and red points), which receive low diversity scores despite strong performance according to distance-based diversity metrics. In contrast, K-Center-Greedy and Repr Filter (green and purple points) show the opposite trend, weakening the metrics' correlation with the model performance. This is likely because the former two strategies select more samples from dense semantic regions, which better cover the overall sample distribution but conflicts with distance-based diversity calculations. This suggests that ignoring information density in semantic space reduces the reliability of diversity metrics.

Findings 4 *Distance-based metrics often fail to accurately measure diversity in datasets containing redundant samples.*

As shown by the duplicated datasets (pink points) in Figure 2(e, f, g, h), DistSum fails to capture redundancy effectively, as total distances are dominated by variations in distant samples. Meanwhile, other metrics, such as KNN Distance, overly penalize redundant samples by nullifying their contribution to overall diversity.

3 Proposed Metric: *NovelSum*

Extending previous findings, we derive some insights on how to design a more reliable metric: (1) **The uniqueness of individual samples should be a key factor in measuring dataset diversity.** This uniqueness stems from sufficient inter-sample distances, providing diverse information that helps the model learn more generalized patterns. (2) **When quantifying a sample's uniqueness, its distance to nearby and distant samples should be balanced.** Differences with nearby samples define uniqueness and should hold greater importance, with weights assigned smoothly. (3) **When calculating inter-sample distances, both semantic differences and local information density should be considered.** In practical applications of instruction fine-tuning, semantic space varies in information density, with scenarios like math and code having denser data and information. Focusing only on semantics overlooks valuable fine-grained information for the model.

Following these principles, we introduce *NovelSum*, a diversity metric that jointly considers distance and distribution. Specifically, we define dataset diversity as the sum of each sample's uniqueness—its unique contribution to overall information, which we later term "novelty":

$$\mathcal{M}_{NovelSum}(\mathcal{X}) = \sum_{x_i \in \mathcal{X}} v(x_i) \tag{7}$$

Figure 3 and the following paragraphs illustrate how each sample's novelty is computed.

Proximity-Weighted Sum In contrast to *DistSum*, which calculates a sample's uniqueness as a simple sum of distances to other points, we propose a proximity-weighted sum. This method assigns higher weights to closer points, giving them a larger influence on the uniqueness score:

$$v(x_i) = \sum_{x_j \in \mathcal{X},\, x_j \neq x_i} w(x_i, x_j)^{\alpha} \cdot \Delta(x_i, x_j), \tag{8}$$

where the proximity weight is defined as:

$$w(x_i, x_j) = \phi(\pi_i(j))$$

Here, $\pi_i(j)$ is the rank of x_j in the sorted list of distances from x_i to all other points in $\mathcal{X}$, with $\pi_i(j) = 1$ indicating that x_j is the nearest neighbor of x_i. The function $\phi(\cdot)$ is monotonically decreasing, smoothing the weights according to the proximity, for example, we set $\phi(\pi_i(j)) = \frac{1}{\pi_i(j)}$. The hyperparameter α controls the degree to which proximity impacts the uniqueness score.

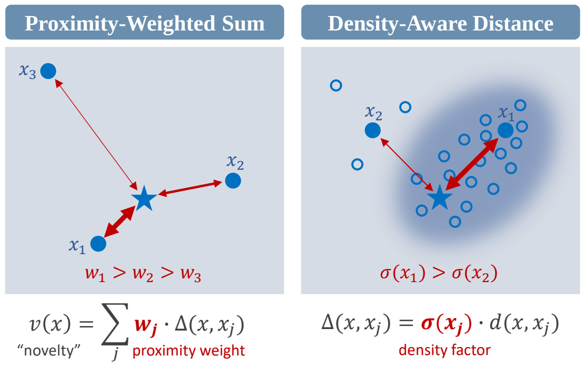


Figure 3: *NovelSum* computes each sample's novelty as a proximity-weighted sum of its density-aware distances to other samples, where closer points have greater influence and high-density regions produce larger distances.

Density-Aware Distance To account for the local information density when calculating $\Delta(x_i, x_j)$, we introduce a density-aware distance that multiplies the original semantic distance by a density factor $\sigma(x_j)$:

$$\Delta(x_i, x_j) = \sigma(x_j)^{\beta} \cdot d(x_i, x_j) \tag{9}$$

Since the probablistic density of the overall sample distribution is intractable, we approximate the density factor by the inverse of the average distance to the K-nearest neighbors of x_j in $\mathcal{X}^{all}$:

$$\sigma(x_j) = \frac{1}{\sum_{k=1}^{K} d(x_j, N_k(x_j))}$$

Here, $d(\cdot, \cdot)$ represents the distance between the embeddings of two samples (e.g., cosine distance), and $N_k(x)$ denotes the k-th nearest neighbor of x. The hyperparameter β controls the extent to which density influences the distance. The computational complexity is discussed in Appendix B.

This approach mirrors how novelty is assessed in academic papers: a paper's novelty depends on its difference from closely related work, and this difference should be considered within the context of its field for a more accurate measure. Therefore, we consider each sample's quantified uniqueness as "novelty" and name our approach "NovelSum."

4 Simulation Study

To validate whether the proposed metric aligns with our design principles and accurately captures dataset diversity, we create a visualizable simulation environment. We generate 150 points in 2D space as the data source and select 20 samples to form a dataset, simulating the data selection process for instruction tuning. As shown in Figure 4,

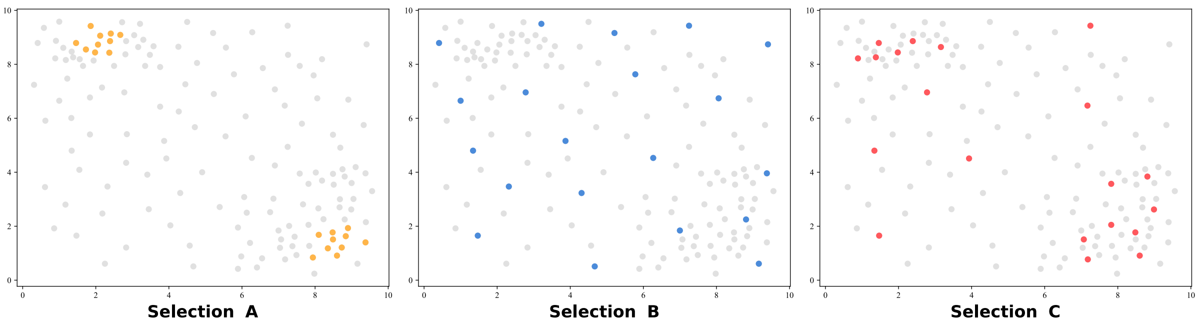


Figure 4: Simulating data selection in a 2D sample space: Selection A represents datasets with redundancy, Selection B optimizes inter-sample distances, and Selection C accounts for both distances and density, which prior analysis suggests yields the highest diversity.

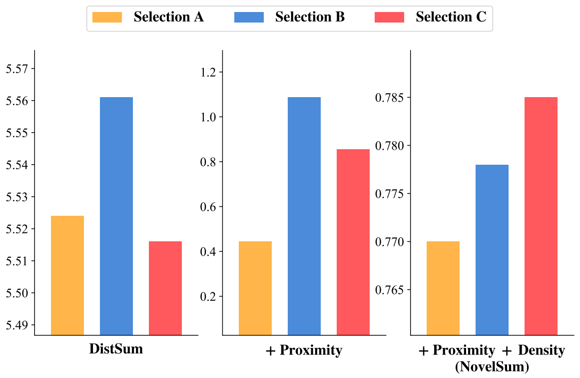


Figure 5: Measuring the diversity of simulated selection A/B/C with various metrics. *NovelSum* accurately captures dataset diversity, exhibiting expected behaviors.

we analyze three data selection scenarios to examine the behavior of our diversity metric. "Selection A" contains samples from two clusters, with most points close to each other, simulating datasets with redundancy. "Selection B", constructed using K-Center-Greedy, consists of samples far apart, simulating datasets optimized for inter-sample semantic distances. "Selection C" considers both inter-sample distances and information density, simulating datasets that best represent the sample space with unique points. Based on prior analysis, the dataset diversity of the three selections should follow $A < B < C$ order intuitively.

Figure 5 presents the diversity measurement results using DistSum, a proximity-weighted version of DistSum, and *NovelSum*. From left to right, we see that DistSum counterintuitively considers $\mathcal{M}(A) \simeq \mathcal{M}(C)$, failing to reflect sample uniqueness. Incorporating the proximity-weighted sum improves uniqueness capture but still exhibits $\mathcal{M}(B) > \mathcal{M}(C)$, overlooking information density. ***NovelSum* resolves these issues, accurately capturing diversity variations in alignment with design principles**, yielding $\mathcal{M}(A) < \mathcal{M}(B) < \mathcal{M}(C)$. This study further validates the necessity of the proximity-weighted sum and density-aware distance for precise diversity measurement.

5 Experiments

Following the settings in Section 2, we evaluate *NovelSum*'s correlation with the fine-tuned model performance across 53 IT datasets and compare it with previous diversity metrics. Additionally, we conduct a correlation analysis using Qwen-2.5-7B (Yang et al., 2024) as the backbone model, alongside previous LLaMA-3-8B experiments, to further demonstrate the metric's effectiveness across different scenarios. Qwen is used for both instruction tuning and deriving semantic embeddings. Due to resource constraints, we run each strategy on Qwen for two rounds, resulting in 25 datasets.

5.1 Main Results

***NovelSum* consistently achieves state-of-the-art correlation with model performance across various data selection strategies, backbone LLMs, and correlation measures.** Table 1 presents diversity measurement results on datasets constructed by mainstream data selection methods (based on $\mathcal{X}^{all}$), random selection from various sources, and duplicated samples (with only $m = 100$ unique samples). Results from multiple runs are averaged for each strategy. Although these strategies yield varying performance rankings across base models, *NovelSum* consistently tracks changes in IT performance by accurately measuring dataset diversity. For instance, K-means achieves the best performance on LLaMA with the highest NovelSum score, while K-Center-Greedy excels on Qwen, also correlating with the highest NovelSum. Table 2 shows the correlation coefficients between

Diversity Metrics	Data Selection Strategies									
	K-means	K-Center -Greedy	QDIT	Repr Filter	Random					Duplicate
					$\mathcal{X}^{all}$	ShareGPT	WizardLM	Alpaca	Dolly	
LLaMA-3-8B										
Facility Loc. $_{\times 10^5}$	2.99	2.73	2.99	2.86	2.99	2.83	2.88	2.83	2.59	2.52
DistSum$_{cosine}$	0.648	0.746	0.629	0.703	0.634	0.656	0.578	0.605	0.603	0.634
Vendi Score $_{\times 10^7}$	1.70	2.53	1.59	2.23	1.61	1.70	1.44	1.32	1.44	0.05
NovelSum (Ours)	0.693	0.687	0.673	0.671	0.675	0.628	0.591	0.572	0.50	0.461
Model Performance	1.32	1.31	1.25	1.05	1.20	0.83	0.72	0.07	-0.14	-1.35
Qwen-2.5-7B										
Facility Loc. $_{\times 10^5}$	3.54	3.42	3.54	3.46	3.54	3.51	3.50	3.50	3.46	3.48
DistSum$_{cosine}$	0.260	0.440	0.223	0.421	0.230	0.285	0.211	0.189	0.221	0.243
Vendi Score $_{\times 10^6}$	1.60	3.09	2.60	7.15	1.41	3.36	2.65	1.89	3.04	0.20
NovelSum (Ours)	0.440	0.505	0.403	0.495	0.408	0.392	0.349	0.336	0.320	0.309
Model Performance	1.06	1.45	1.23	1.35	0.87	0.07	-0.08	-0.38	-0.49	-0.43

Table 1: Measuring the diversity of datasets selected by different strategies using *NovelSum* and baseline metrics. Fine-tuned model performances (Eq. 5), based on MT-bench and AlpacaEval, are also included for cross reference. Darker blue shades indicate higher values for each metric, while darker orange shades indicate lower values. While data selection strategies vary in performance on LLaMA-3-8B and Qwen-2.5-7B, *NovelSum* consistently shows a stronger correlation with model performance than other metrics. More results are provided in Appendix D.

Diversity Metrics	LLaMA			Qwen
	Pearson	Spearman	Avg.	Avg.
TTR	-0.38	-0.16	-0.27	-0.30
vocd-D	-0.43	-0.17	-0.30	-0.31
Facility Loc.	0.86	0.69	0.77	0.08
Entropy	0.93	0.80	0.86	0.63
LDD	0.61	0.75	0.68	0.60
KNN Distance	0.59	0.80	0.70	0.67
DistSum$_{cosine}$	0.85	0.67	0.76	0.51
Vendi Score	0.70	0.85	0.78	0.60
DistSum$_{L2}$	0.86	0.76	0.81	0.51
Cluster Inertia	0.81	0.85	0.83	0.76
Radius	0.87	0.81	0.84	0.48
NovelSum	**0.98**	**0.95**	**0.97**	**0.90**

Table 2: Correlations between different metrics and model performance on LLaMA-3-8B and Qwen-2.5-7B. "Avg." denotes the average correlation (Eq. 6).

various metrics and model performance for both LLaMA and Qwen experiments, where *NovelSum* achieves state-of-the-art correlation across different models and measures.

***NovelSum* can provide valuable guidance for data engineering practices.** As a reliable indicator of data diversity, *NovelSum* can assess diversity at both the dataset and sample levels, directly guiding data selection and construction decisions. For example, Table 1 shows that the combined data source $\mathcal{X}^{all}$ is a better choice for sampling diverse IT data than other sources. Moreover, *NovelSum* can offer insights through comparative analyses, such as: (1) ShareGPT, which collects data from real internet users, exhibits greater diversity than Dolly, which relies on company employees, suggesting that IT samples from diverse sources enhance dataset diversity (Wang et al., 2024b); (2) In LLaMA experiments, random selection can outperform some mainstream strategies, aligning with prior work (Xia et al., 2024; Diddee and Ippolito, 2024), highlighting gaps in current data selection methods for optimizing diversity.

5.2 Ablation Study

NovelSum involves several flexible hyperparameters and variations. In our main experiments, *NovelSum* uses cosine distance to compute $d(x_i, x_j)$ in Eq. 9. We set $\alpha = 1$, $\beta = 0.5$, and $K = 10$ nearest neighbors in Eq. 8 and 9. Here, we conduct an ablation study to investigate the impact of these settings based on LLaMA-3-8B.

Variants	Pearson	Spearman	Avg.
NovelSum	0.98	0.96	0.97
- Use $L2$ distance	0.97	0.83	$0.90_{\downarrow 0.08}$
- $K = 20$	0.98	0.96	$0.97_{\downarrow 0.00}$
- $\alpha = 0$ (w/o proximity)	0.79	0.31	$0.55_{\downarrow 0.42}$
- $\alpha = 2$	0.73	0.88	$0.81_{\downarrow 0.16}$
- $\beta = 0$ (w/o density)	0.92	0.89	$0.91_{\downarrow 0.07}$
- $\beta = 1$	0.90	0.62	$0.76_{\downarrow 0.21}$

Table 3: Ablation Study for *NovelSum*.

In Table 3, $\alpha = 0$ removes the proximity weights, and $\beta = 0$ eliminates the density multiplier. We observe that both $\alpha = 0$ and $\beta = 0$ significantly weaken the correlation, validating the benefits of the proximity-weighted sum and density-aware distance. Additionally, improper values for α and β greatly reduce the metric's reliability, highlighting that *NovelSum* strikes a delicate balance between distances and distribution. Replacing cosine distance with Euclidean distance and using more neighbors for density approximation have minimal impact, particularly on Pearson's correlation, demonstrating *NovelSum*'s robustness to different distance measures.

6 Data Selection Strategy

Introducing *NovelSelect* Given *NovelSum*'s accurate diversity measurement and strong correlation with model performance, we investigate its potential as an optimization objective for selecting samples and generating a diverse dataset:

$$\mathcal{X} = \arg\max_{\mathcal{X} \subset \mathcal{X}^{all}} \mathcal{M}_{NovelSum}(\mathcal{X}), \tag{10}$$

where $\mathcal{M}_{NovelSum}(\mathcal{X})$ is defined in Eq. 7. Since directly solving Eq. 10 is NP-hard (Cook et al., 1994), we propose a greedy approach that iteratively selects the most "novel" sample. The "novelty" of a new sample $v(x)$ relative to an existing set $\mathcal{X}$ is defined as:

$$v(x) = \sum_{x_j \in \mathcal{X}} w(x, x_j)^{\alpha} \cdot \sigma(x_j)^{\beta} \cdot d(x, x_j), \tag{11}$$

where $w(x, x_j)$ and $\sigma(x_j)$ are the proximity weight and density factor from Eq. 8 and 9. The sample with the maximum novelty is then selected: $x^{new} = \arg\max_x v(x), \quad \mathcal{X} \leftarrow x^{new} \cup \mathcal{X}$. This process is repeated from $\mathcal{X} = \emptyset$ until the data budget is reached, resulting in the selected dataset. We refer to this approach as *NovelSelect*.

Algorithm 1 *NovelSelect*

Input: Data pool $\mathcal{X}^{all}$, data budget n
Initialize an empty dataset, $\mathcal{X} \leftarrow \emptyset$
while $|\mathcal{X}| < n$ **do**
    $x^{new} \leftarrow \arg\max_{x \in \mathcal{X}^{all}} v(x)$
    $\mathcal{X} \leftarrow \mathcal{X} \cup \{x^{new}\}$
    $\mathcal{X}^{all} \leftarrow \mathcal{X}^{all} \setminus \{x^{new}\}$
end while
return $\mathcal{X}$

Algorithm 1 outlines the overall process. Notably, by incorporating quality scores into $v(x)$, *NovelSelect* can also seamlessly integrate with quality-based data selection methods.

Strategies	MT-bench	AlpacaEval	Aggregated $\mathcal{P}$
Random	6.18	75.47	1.20
Repr Filter	6.17	72.57	1.05
QDIT	6.21	75.91	1.25
K-Center-Greedy	6.33	75.30	1.31
K-means	6.33	75.46	1.32
NovelSelect	**6.47**	**78.07**	**1.55**

Table 4: Comparisons of different diversity-oriented data selection strategies on IT performance. $\mathcal{P}$ aggregates the performance based on Z-scores (Eq. 5).

Data Selection Experiments We conduct additional data selection experiments on LLaMA-3-8B to evaluate *NovelSelect*'s performance. Following prior settings, we use *NovelSelect* to select 10k samples from $\mathcal{X}^{all}$ and assess the fine-tuned model's performance on MT-bench and AlpacaEval. Results are averaged over three runs.

From Table 4, *NovelSelect* outperforms existing diversity-oriented data selection strategies on both benchmarks, demonstrating superior IT performance. This further validates *NovelSum*'s effectiveness and practical value in IT data engineering.

7 Related Work

Measuring Dataset Diversity Dataset diversity is essential for training generalizable machine learning models, drawing significant research interest (Sun et al., 2024; Zhao et al., 2024; Qin et al., 2024). In NLP, numerous lexical diversity metrics have been proposed to measure text diversity through vocabulary usage (Richards, 1987; Malvern et al., 2004). Recently, semantic embeddings have enabled more flexible diversity measurement from distance (Stasaski and Hearst, 2022; Du and Black, 2019; Dang and Verma, 2024) or distribution perspectives (Shao et al., 2024). Focusing on instruction tuning, while some studies have explored the assessment of IT data diversity (Wang et al., 2024b; Bukharin et al., 2024), the proposed metrics lack sufficient validation of their correlation with IT performance; thus, reliable metrics for guiding data engineering remain underexplored.

Data Selection for Instruction Tuning Instruction tuning trains LLMs to follow human instructions using instruction-response pairs (Zhang et al., 2023). While earlier work focused on large-scale IT datasets (Longpre et al., 2023; Chiang et al., 2023), recent studies show that small, high-quality data sets can reduce costs and improve performance (Chen et al., 2023; Zhou et al., 2024; Dou et al., 2024; Ye et al., 2024). This has led to the development of data selection strategies to identify subsets that boost IT performance (Liu et al., 2023; Du et al., 2023; Wu et al., 2023; Song et al., 2024; Yang et al., 2025). However, the lack of clear definitions and reliable diversity metrics for IT datasets hinders effective optimization. Consequently, some selection methods fail to generalize or perform worse than random selection (Xia et al., 2024; Diddee and Ippolito, 2024). Our work seeks to provide a more reliable diversity metric, based on comprehensive analysis, that accurately reflects the diversity of IT datasets and their instruction tuning performance.

8 Conclusion

In this paper, we investigate the fundamental problem of precisely measuring dataset diversity for instruction tuning and propose *NovelSum*, a reliable diversity metric that correlates well with model performance. Inspired by our systematic analysis of existing diversity metrics, *NovelSum* jointly considers inter-sample distances and information density to effectively capture dataset diversity, achieving superior correlations with model performance compared to previous metrics. Based on *NovelSum*, We further develop a data selection strategy, *NovelSelect*, whose remarkable performance validates the practical significance of *NovelSum*.

Limitations

Although our work systematically analyzes existing and proposed metrics through extensive fine-tuning experiments, we focus solely on the Qwen-2.5-7B and LLaMA-3-8B models as the backbone LLMs. We exclude larger models and other model series due to resource constraints, though these may exhibit different characteristics in terms of data diversity. Additionally, our study focuses on the general instruction-tuning task evaluated by the MT-bench and AlpacaEval benchmarks. Experiments on downstream tasks, such as information extraction and creative writing, are not considered. Their data diversity measurements may differ from those of general instruction-tuning tasks and thus warrant further investigation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. 2024. Data diversity matters for robust instruction tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 3411–3425. Association for Computational Linguistics.

Hao Chen, Yiming Zhang, Qi Zhang, Hantao Yang, Xiaomeng Hu, Xuetao Ma, Yifan Yanggong, and Junbo Zhao. 2023. Maybe only 0.5% data is needed: A preliminary exploration of low training data instruction tuning. *arXiv preprint arXiv:2305.09246*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

William J Cook, William H Cunningham, William R Pulleyblank, and Alexander Schrijver. 1994. Combinatorial optimization. *Unpublished manuscript*, 10:75–93.

Vu Minh Hoang Dang and Rakesh M Verma. 2024. Data quality in nlp: Metrics and a comprehensive taxonomy. In *International Symposium on Intelligent Data Analysis*, pages 217–229. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Harshita Diddee and Daphne Ippolito. 2024. Chasing random: Instruction selection strategies fail to generalize. *arXiv preprint arXiv:2410.15225*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *arXiv preprint arXiv:2305.14233*.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, et al. 2024. Loramoe: Alleviating world knowledge forgetting in large language models via moe-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Wenchao Du and Alan W Black. 2019. Boosting dialog response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Reza Zanjirani Farahani and Masoud Hekmatfar. 2009. *Facility location: concepts, models, algorithms and case studies*. Springer Science & Business Media.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Yi-An Lai, Xuan Zhu, Yi Zhang, and Mona Diab. 2020. Diversity, density, and homogeneity: Quantitative characteristic metrics for text collections. *arXiv preprint arXiv:2003.08529*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pages 22631–22648. PMLR.

David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical diversity and language development*. Springer.

Amey P Pasarkar and Adji Bousso Dieng. 2023. Cousins of the vendi score: A family of similarity-based diversity metrics for science and machine learning. *arXiv preprint arXiv:2310.12952*.

Yulei Qin, Yuncheng Yang, Pengcheng Guo, Gang Li, Hang Shao, Yuchen Shi, Zihan Xu, Yun Gu, Ke Li, and Xing Sun. 2024. Unleashing the power of data tsunami: A comprehensive survey on data assessment and selection for instruction tuning of language models. *arXiv preprint arXiv:2408.02085*.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Ozan Sener and Silvio Savarese. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Yunfan Shao, Linyang Li, Zhaoye Fei, Hang Yan, Dahua Lin, and Xipeng Qiu. 2024. Balanced data sampling for language model training with clustering. *arXiv preprint arXiv:2402.14526*.

Jielin Song, Siyu Liu, Bin Zhu, and Yanghui Rao. 2024. Iterselecttune: An iterative training framework for efficient instruction-tuning data selection. *arXiv preprint arXiv:2410.13464*.

Katherine Stasaski and Marti A Hearst. 2022. Semantic diversity in dialogue with natural language inference. *arXiv preprint arXiv:2205.01497*.

Katherine Stasaski, Grace Hui Yang, and Marti A Hearst. 2020. More diverse dialogue datasets via diversity-informed data collection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4958–4968.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. 2024. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9390–9399.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Jiahao Wang, Bolin Zhang, Qianlong Du, Jiajun Zhang, and Dianhui Chu. 2024a. A survey on data selection for llm instruction tuning. *arXiv preprint arXiv:2402.05123*.

Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar Panda. 2024b. Diversity measurement and subset selection for instruction tuning datasets. *arXiv preprint arXiv:2402.02318*.

Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. 2023. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*.

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

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yuming Yang, Wantong Zhao, Caishuang Huang, Junjie Ye, Xiao Wang, Huiyuan Zheng, Yang Nan, Yuran Wang, Xueying Xu, Kaixin Huang, Yunke Zhang, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025. Beyond boundaries: Learning a universal entity taxonomy across datasets and languages for open named entity recognition. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 10902–10923. Association for Computational Linguistics.

Junjie Ye, Yuming Yang, Qi Zhang, Tao Gui, Xuanjing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. 2024. Empirical insights on fine-tuning large language models for question-answering. *CoRR*, abs/2409.15825.

Jerrold H Zar. 2005. Spearman rank correlation. *Encyclopedia of biostatistics*, 7.

Dylan Zhang, Justin Wang, and Francois Charton. 2024. Instruction diversity drives generalization to unseen tasks. *arXiv preprint arXiv:2402.10891*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Dorothy Zhao, Jerone TA Andrews, AI Sony, Tokyo Orestis Papakyriakopoulos, and Alice Xiang. 2024. Measuring diversity in datasets. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Details of Correlation Evaluation

A.1 Data Processing and Semantic Embeddings

We apply basic preprocessing to remove anomalous samples from the data sources, ensuring more stable results while preserving generality. In the early stage of our work, we observe that short samples often exhibit low quality and tend to be outliers in the semantic space, potentially distorting experimental results. To address this, we filter out samples shorter than 256 tokens using the BERT (Devlin et al., 2019) tokenizer, ensuring consistency for experiments across different LLMs. Furthermore, to ensure the dataset's relevance for English-language tasks and math problems, we exclude samples with a non-English-or-number ratio exceeding 0.8.

When computing sample embeddings, we set the maximum sequence length to 256 to mitigate the impact of varying text lengths. This applies only to embedding computation; fine-tuning uses a much larger maximum length. We extract the last hidden layer of the language model and apply mean pooling, excluding padding tokens, to generate robust sample-level embeddings. For experiments on the LLaMA-3-8B model, we utilize LLaMA-3-8B to compute embeddings for data selection. Similarly, for experiments with the Qwen-2.5-7B model, we employ Qwen-2.5-7B to compute embeddings.

A.2 Details of Existing Diversity Metrics

For lexical diversity, the **Type-Token Ratio** (TTR) quantifies the lexical diversity of a text sequence x_i as the ratio of distinct tokens to the total number of tokens. The overall lexical diversity of a dataset $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ is computed as the average TTR across all samples:

$$\mathcal{M}_{TTR}(\mathcal{X}) = \frac{1}{N}\sum_{i=1}^{N}\frac{|Unique(x_i)|}{|x_i|}. \tag{12}$$

To mitigate the influence of text length on TTR, we randomly sample 30 tokens from each data point to compute the TTR.

To address the sensitivity of TTR to text length, **vocd-D** extends this measure by computing TTR_i^k over sampled sub-sequences of varying lengths k and fitting the following curve:

$$T\hat{T}R_i^k = \frac{D}{k}\left((1+2\frac{k}{D})^{\frac{1}{2}}-1\right), \tag{13}$$

where D is the estimated parameter representing lexical diversity. The vocd-D metric is defined as $\mathcal{M}_{vocd-D} = D_{\text{best fit}}$, with larger values indicating greater lexical diversity. In our experiments, we compute TTR_i^k for $k = 10, 20, 30, 40, 50$ and take the average of the resulting values as the final lexical diversity score.

For distance-based semantic diversity, **Cluster Inertia** (Du and Black, 2019) quantifies diversity by partitioning the dataset into K clusters using K-means and summing the squared distances between each sample and its cluster centroid:

$$\mathcal{M}_{Inertia}(\mathcal{X}) = \sum_{j=1}^{K}\sum_{x_i\in C_j}\|emb(x_i)-\mu_j\|^2, \tag{14}$$

where μ_j is the centroid of cluster C_j. A higher inertia value suggests a greater spread of samples. Additionally, **Vendi Score** (VS) (Pasarkar and Dieng, 2023) measures diversity based on the eigenvalues of the similarity kernel matrix. The generalized VS metric is defined as:

$$\mathcal{M}_{VS}(\mathcal{X}) = \exp\left(\frac{1}{1-\alpha}\log_2\sum_{i=1}^{|\mathcal{X}|}\bar{\lambda}_{i|\theta}^{\alpha}\right), \tag{15}$$

where $\bar{\lambda}_{i|\theta}$ represents the normalized eigenvalues. We set $\alpha = 0.5$ to enhance measurement under severe class imbalance. **Radius** (Lai et al., 2020) characterizes the dispersion of the sample space by approximating embeddings as a multi-variate Gaussian distribution. It computes the geometric mean of the standard deviations along each dimension:

$$\mathcal{M}_{Radius}(\mathcal{X}) = \sqrt[H]{\prod_{j=1}^{H}\sigma_j}, \tag{16}$$

where H is the embedding dimension, and σ_j denotes the radius of the ellipsoid along the j-th axis. Larger values indicate a greater spread of samples in the embedding space. **Log Determinant Distance** (Wang et al., 2024b) utilizes the determinant of the similarity matrix as a measure of dataset diversity. In our work, we employ the cosine similarity function to compute the similarity matrix.

Note that for **DistSum**$_{cosine}$, we use cosine distance $\Delta(x_i, x_j) = 1 - \cos(emb(x_i), emb(x_j))$. For **DistSum**$_{L2}$, we use Euclidean distance $\Delta(x_i, x_j) = \|emb(x_i) - emb(x_j)\|_2^2$.

For **Partition Entropy**, we cluster $\mathcal{X}^{all}$ into 1,000 clusters, while for **Cluster Inertia** (Du and Black, 2019), we cluster $\mathcal{X}^{s}$ into 200 clusters for subsequent computations.

A.3 Details of Data Selection Strategies

All IT datasets in our experiments are selected from $\mathcal{X}^{all}$ and sampled over three rounds (two for Qwen) per strategy variant, unless stated otherwise. We assume these datasets have similar average sample quality, as they come from the same source without any quality filters. Additionally, the dataset size is standardized to 10,000 samples. Thus, our experiments can more accurately reflect the correlation between dataset diversity and model performance, without introducing significant confounders.

K-Center-Greedy (Sener and Savarese, 2017; Chen et al., 2023; Du et al., 2023; Wu et al., 2023) This strategy begins by randomly selecting a data point from the dataset $\mathcal{X}^{all}$ as the initial point of the subset $\mathcal{X}^{(s)}$. Subsequently, it iteratively computes the closest distance between the remaining points in $\mathcal{X}^{all} \setminus \mathcal{X}^{(s)}$ and selected samples in $\mathcal{X}^{(s)}$. The point with the maximum minimum distance (i.e., the farthest point) is added to $\mathcal{X}^{(s)}$. This process continues until the desired subset size is achieved.

Repr Filter (Liu et al., 2023) Unlike the K-Center-Greedy strategy, which selects the farthest point from the remaining data pool, the Repr Filter randomly selects a data point whose similarity with all embeddings in $\mathcal{X}^{(s)}$ is below a predefined threshold. Due to the unique distribution of embeddings across different models, it is necessary to set distinct thresholds for each similarity function and model embedding. To ensure diversity across different experimental rounds, we employ cosine similarity and set the threshold to 0.3 for LLaMA-3-8B and 0.1 for Qwen-2.5-7B.

QDIT (Bukharin et al., 2024) QDIT sampling combines diversity and quality scores for data selection; however, in our work, we focus exclusively on its diversity score. This method computes the sum of similarities between each sample in $\mathcal{X}^{all} \setminus \mathcal{X}^{(s)}$ and its closest data point in $\mathcal{X}^{(s)}$. For each candidate data point, we calculate the similarity sum as if it were added to $\mathcal{X}^{(s)}$, defining its Facility Location (FL) score. The algorithm then iteratively selects the data point with the highest FL score. For the initial selection, it chooses the data point that exhibits the highest overall similarity to all other embeddings. In our experiments, we employ cosine similarity for computing these scores. Since the Facility Location function yields a fixed subset $\mathcal{X}^{(s)}$ for a given $\mathcal{X}^{all}$, and to maintain consistency with other strategies, we utilize the same subset of data but vary the training random seeds across three rounds of experiments.

K-means Clustering (Song et al., 2024) For this strategy, we apply the K-means clustering algorithm to partition all sample embeddings in $\mathcal{X}^{all}$ into K clusters. Subsequently, given a target data budget n, we randomly sample $\frac{n}{K}$ data points from each cluster. For our experiments, we use both 1000 and 100 clusters for LLaMA-3-8B, and 100 clusters for Qwen-2.5-7B.

Random Selection In this baseline strategy, we randomly sample 10,000 data points from $\mathcal{X}^{all}$. To explore the impact of data sources, we also sample from individual datasets, including Alpaca (Taori et al., 2023), Dolly (Conover et al., 2023), WizardLM, UltraChat, and ShareGPT, with similar preprocessing. Although we assume that the average sample quality of these sources does not significantly differ from that of $\mathcal{X}^{all}$, we use only a single round of results from each source in the overall correlation analysis as supplementary data to avoid potential quality differences affecting the outcome.

Duplicate Selection To address the challenge of defining low-diversity datasets, which is crucial for our study, we construct datasets with redundant samples. Given a target data budget n, the dataset is constructed by selecting m unique data points, each duplicated $\frac{n}{m}$ times. We set m to 1, 10, 50, 100, 500, 1000, 2000, and 5000. This approach allows us to systematically control and analyze the impact of diversity on model performance.

A.4 Details of Correlation Measures

We compute the correlation between each diversity metric and model performance using both Pearson (Cohen et al., 2009) and Spearman (Zar, 2005) correlation measures. For example, Pearson's r for a metric $\mathcal{M}_t$ is computed as:

$$r^{Pearson}_{\mathcal{M}_t,\,\mathcal{P}} = \frac{\sum_s (\mathcal{M}_t^{(s)} - \bar{\mathcal{M}}_t)(\mathcal{P}^{(s)} - \bar{\mathcal{P}})}{\sigma_{\mathcal{M}_t}\sigma_{\mathcal{P}}} \quad (17)$$

A.5 Details of Model Fine-Tuning

In our experiments, we leverage four or eight NVIDIA H100 GPUs for training the LLaMA-3-8B and Qwen-2.5-7B models. To enable efficient parallel training, we implement DeepSpeed Zero-Stage 2. Across all experiments conducted in this study, the training parameters are configured as follows: a maximum input length of 4096 tokens,

Data Pool	Dataset Source	Sample Size
$\mathcal{X}^{all}$	ShareGPT	103 K
	UltraChat	207 K
	WizardLM	196 K
$\mathcal{X}^{other}$	Alpaca	52 K
	Dolly	15 K

Table 5: Statistics of Data Pools $\mathcal{X}^{all}$ and $\mathcal{X}^{other}$. The column "Dataset Source" indicates the origin of the data used for sampling, while "Sample Size" denotes the number of samples in each dataset. This table provides an overview of the data used in our experiments.

a batch size of 128, 3 training epochs, a learning rate of 2e-5, and a warm-up ratio of 0.1 utilizing cosine warm-up. We use the official chat templates of LLaMA-3 and Qwen-2.5, respectively, to fine-tune each model. All models are trained with BF16 precision to optimize computational efficiency and memory usage. A single run of fine-tuning on a 10k dataset typically takes about one hour.

B Computational Complexity

NovelSum has a time complexity of $O(n^2 \log n)$ for a dataset of n samples. Computing *NovelSum* on a 10k-sample dataset using precomputed LLaMA-3-8B embeddings takes only 10 seconds on a single H20 GPU, which is negligible compared to model fine-tuning or evaluation.

For our data selection strategy, *NovelSelect*, selecting n samples from a total pool of N has a time complexity of $O(N \cdot n^2 \log n)$, significantly lower than QDIT's $O(N^3)$ since $n \ll N$ in most cases. In practice, *NovelSelect* takes about one hour to select 10k samples from our 396K sample pool using precomputed LLaMA-3-8B embeddings on a single node with eight H100 GPUs. This cost is trivial compared to fine-tuning on all 396K samples, which takes approximately 40 hours under the same setup.

As a key preprocessing step, deriving embeddings for all 396K samples in $\mathcal{X}^{all}$ using LLaMA-3-8B takes under 2 hours on a single node with eight H100 GPUs, leveraging vLLM (Kwon et al., 2023). Since this is a one-time cost for all subsequent processes, it remains practical.

C Data Statistics

Our data sources are detailed in Table 5. After filtering out short data and non-English data, approximately 396K samples remain in $\mathcal{X}^{all}$ for use in our experiments. Note that we use the latest versions of these datasets, which may have a larger size than the initial versions. These datasets encompass samples from a wide range of domains.

D More Results

Additional scatter plots for the analysis in Section 2 are provided in Figure 6, Figure 7 and Figure 8 , illustrating the correlation for DistSum$_{L2}$, Radius, and Log Determinant Distance, respectively.

The full results of the correlation experiments on LLaMA-3-8B and Qwen-2.5-7B are presented in Table 6 and Table 7, respectively. These tables provide a comprehensive comparison of diversity metrics across different experimental configurations.

E Others

E.1 License for Artifacts and Data Consent

In this paper, the artifacts used are all available for academic research work, including ShareGPT, WizardLM, UltraChat, Alpaca and Dolly. The methods compared in this paper can all be used for academic research. All data originates from the original authors' open-source releases and can be used for academic research and publication.

E.2 Data Statement

The training datasets may contain offensive content; however, they do not include any personal information. Furthermore, our training approach is designed to align the model with human preferences without producing harmful content.

E.3 AI Assistant Usage Statement

We solely utilize ChatGPT for writing refinement; however, AI assistants are not employed for research innovation or coding.

E.4 Budgets

We spend approximately one hour of training on a single node with eight H100-80G GPUs for each IT model. Additionally, we spend around $1,000 on the GPT API to evaluate our models using MT-bench and AlpacaEval.

Data Selection Strategy	Runs	NovelSum	DistSum$_{cosine}$	DistSum$_{L2}$	KNN	Inertia	Radius	VS	Entropy	FL	LDD	TTR	vocd-D
Random - alpaca	1	5.72×10^{-1}	6.05×10^{-1}	1.09	5.09×10^{-1}	2.93×10^{-1}	1.10×10^{-2}	1.32×10^{7}	8.93	2.83×10^{5}	-1.70×10^{4}	8.53×10^{-1}	8.23×10^{1}
Random - dolly	1	5.00×10^{-1}	6.03×10^{-1}	1.09	5.96×10^{-1}	3.62×10^{-1}	1.10×10^{-2}	1.70×10^{7}	7.83	2.59×10^{5}	-1.31×10^{4}	8.44×10^{-1}	7.66×10^{1}
Random - sharegpt	1	6.28×10^{-1}	6.56×10^{-1}	1.14	5.74×10^{-1}	3.80×10^{-1}	1.20×10^{-2}	1.70×10^{7}	8.97	2.83×10^{5}	-1.23×10^{4}	8.50×10^{-1}	7.83×10^{1}
Random - ultrachat	1	6.72×10^{-1}	6.22×10^{-1}	1.11	5.67×10^{-1}	3.23×10^{-1}	1.10×10^{-2}	1.48×10^{7}	9.40	2.96×10^{5}	-1.52×10^{4}	8.81×10^{-1}	1.10×10^{2}
Random - wizardlm	1	5.91×10^{-1}	5.78×10^{-1}	1.07	5.94×10^{-1}	3.31×10^{-1}	1.10×10^{-2}	1.44×10^{7}	9.08	2.88×10^{5}	-1.52×10^{4}	8.58×10^{-1}	8.57×10^{1}
Random - $\mathcal{X}^{all}$	3	6.75×10^{-1}	6.34×10^{-1}	1.12	6.06×10^{-1}	3.53×10^{-1}	1.20×10^{-2}	1.61×10^{7}	9.80	2.99×10^{5}	-1.38×10^{4}	8.70×10^{-1}	9.72×10^{1}
Farthest	3	6.87×10^{-1}	7.89×10^{-1}	1.25	4.07×10^{-1}	3.50×10^{-1}	1.20×10^{-2}	1.56×10^{7}	6.52	2.14×10^{5}	-1.25×10^{4}	8.37×10^{-1}	6.83×10^{1}
Duplicate m=1	3	0.00	0.00	0.00	0.00	0.00	0.00	1.08×10^{4}	0.00	1.25×10^{5}	-inf	8.87×10^{-1}	1.21×10^{2}
Duplicate m=10	3	2.68×10^{-1}	5.89×10^{-1}	1.02	0.00	0.00	1.10×10^{-2}	7.16×10^{4}	3.27	2.08×10^{5}	-inf	8.63×10^{-1}	8.99×10^{1}
Duplicate m=50	3	3.88×10^{-1}	6.08×10^{-1}	1.09	1.00×10^{-3}	0.00	1.10×10^{-2}	2.74×10^{5}	5.58	2.40×10^{5}	-inf	8.73×10^{-1}	1.01×10^{2}
Duplicate m=100	3	4.61×10^{-1}	6.34×10^{-1}	1.12	1.00×10^{-3}	0.00	1.10×10^{-2}	4.95×10^{5}	6.50	2.52×10^{5}	-inf	8.66×10^{-1}	9.23×10^{1}
Duplicate m=500	3	5.56×10^{-1}	6.35×10^{-1}	1.12	1.00×10^{-3}	2.22×10^{-1}	1.10×10^{-2}	1.79×10^{6}	8.47	2.75×10^{5}	-inf	8.69×10^{-1}	9.67×10^{1}
Duplicate m=1000	3	5.87×10^{-1}	6.30×10^{-1}	1.12	1.00×10^{-3}	2.92×10^{-1}	1.10×10^{-2}	2.99×10^{6}	9.06	2.83×10^{5}	-inf	8.69×10^{-1}	9.61×10^{1}
Duplicate m=2000	3	6.18×10^{-1}	6.33×10^{-1}	1.12	1.00×10^{-3}	3.30×10^{-1}	1.10×10^{-2}	5.07×10^{6}	9.46	2.90×10^{5}	-inf	8.70×10^{-1}	9.73×10^{1}
Duplicate m=5000	3	6.56×10^{-1}	6.34×10^{-1}	1.12	1.00×10^{-3}	3.49×10^{-1}	1.20×10^{-2}	9.92×10^{6}	9.72	2.97×10^{5}	-inf	8.71×10^{-1}	9.71×10^{1}
K-Center-Greedy	3	6.87×10^{-1}	7.46×10^{-1}	1.22	8.64×10^{-1}	5.22×10^{-1}	1.20×10^{-2}	2.53×10^{7}	9.30	2.73×10^{5}	-7.44×10^{3}	8.62×10^{-1}	8.85×10^{1}
Kmeans Clustering$_{1000}$	3	6.92×10^{-1}	6.46×10^{-1}	1.13	6.15×10^{-1}	3.72×10^{-1}	1.20×10^{-2}	1.70×10^{7}	9.87	2.99×10^{5}	-1.32×10^{4}	8.69×10^{-1}	9.64×10^{1}
Kmeans Cluster$_{100}$	3	6.93×10^{-1}	6.50×10^{-1}	1.13	6.10×10^{-1}	3.62×10^{-1}	1.20×10^{-2}	1.69×10^{7}	9.78	2.99×10^{5}	-1.33×10^{4}	8.69×10^{-1}	9.61×10^{1}
QDIT	3	6.73×10^{-1}	6.29×10^{-1}	1.12	6.02×10^{-1}	3.48×10^{-1}	1.10×10^{-2}	1.59×10^{7}	9.77	2.99×10^{5}	-1.41×10^{4}	8.71×10^{-1}	9.85×10^{1}
Repr Filter	3	6.71×10^{-1}	7.03×10^{-1}	1.18	7.99×10^{-1}	4.70×10^{-1}	1.20×10^{-2}	2.23×10^{7}	9.45	2.86×10^{5}	-9.12×10^{3}	8.66×10^{-1}	9.20×10^{1}
NoveSelect	3	7.62×10^{-1}	8.21×10^{-1}	1.28	7.04×10^{-1}	5.34×10^{-1}	1.30×10^{-2}	2.55×10^{7}	9.23	2.73×10^{5}	-6.27×10^{3}	8.62×10^{-1}	8.79×10^{1}

Table 6: Comprehensive experimental results on LLaMA-3-8B. Each data selection strategy variant is evaluated over three independent runs (except for random selection) to ensure the robustness and reliability of the findings. The results from multiple runs are averaged. Note that *NovelSelect* results are only included in Section 6 and are not part of the correlation calculations. Details of the data selection strategies are provided in Appendix A.3.

Data Selection Strategy	Runs	NovelSum	DistSum$_{cosine}$	DistSum$_{L2}$	KNN	Inertia	Radius	VS	Entropy	FL	LDD	TTR	vocd-D
Random - alpaca	1	3.36×10^{-1}	1.89×10^{-1}	5.96×10^{-1}	2.23×10^{-1}	6.63×10^{-2}	4.00×10^{-3}	1.89×10^{6}	8.66	3.50×10^{5}	-4.40×10^{4}	8.53×10^{-1}	8.24×10^{1}
Random - dolly	1	3.20×10^{-1}	2.21×10^{-1}	6.51×10^{-1}	2.93×10^{-1}	9.81×10^{-2}	5.00×10^{-3}	3.04×10^{6}	7.92	3.46×10^{5}	-3.62×10^{4}	8.44×10^{-1}	7.66×10^{1}
Random - sharegpt	1	3.92×10^{-1}	2.85×10^{-1}	7.31×10^{-1}	2.89×10^{-1}	1.10×10^{-1}	5.00×10^{-3}	3.36×10^{6}	8.87	3.51×10^{5}	-3.34×10^{4}	8.50×10^{-1}	7.83×10^{1}
Random - ultrachat	1	3.89×10^{-1}	2.00×10^{-1}	6.20×10^{-1}	2.52×10^{-1}	7.42×10^{-2}	4.00×10^{-3}	2.09×10^{6}	9.30	3.52×10^{5}	-4.17×10^{4}	8.81×10^{-1}	1.10×10^{2}
Random - wizardlm	1	3.49×10^{-1}	2.11×10^{-1}	6.31×10^{-1}	2.96×10^{-1}	9.29×10^{-2}	4.00×10^{-3}	2.65×10^{6}	9.03	3.50×10^{5}	-3.84×10^{4}	8.58×10^{-1}	8.57×10^{1}
Random $\mathcal{X}^{all}$	2	4.08×10^{-1}	2.30×10^{-1}	6.61×10^{-1}	2.86×10^{-1}	9.16×10^{-2}	4.00×10^{-3}	1.41×10^{6}	9.77	3.54×10^{5}	-3.81×10^{4}	8.69×10^{-1}	9.70×10^{1}
Duplicate m=50	2	2.52×10^{-1}	2.15×10^{-1}	6.38×10^{-1}	1.00×10^{-3}	0.00	4.00×10^{-3}	1.31×10^{5}	5.64	3.44×10^{5}	-inf	8.71×10^{-1}	9.81×10^{1}
Duplicate m=100	2	3.09×10^{-1}	2.43×10^{-1}	6.64×10^{-1}	1.00×10^{-3}	0.00	4.00×10^{-3}	2.03×10^{5}	6.54	3.48×10^{5}	-inf	8.70×10^{-1}	9.77×10^{1}
Duplicate m=500	2	3.57×10^{-1}	2.40×10^{-1}	6.72×10^{-1}	1.00×10^{-3}	5.73×10^{-2}	4.00×10^{-3}	5.10×10^{5}	8.50	3.54×10^{5}	-inf	8.68×10^{-1}	9.49×10^{1}
Duplicate m=1000	2	3.64×10^{-1}	2.29×10^{-1}	6.58×10^{-1}	1.00×10^{-3}	7.60×10^{-2}	4.00×10^{-3}	7.41×10^{5}	9.05	3.56×10^{5}	-inf	8.69×10^{-1}	9.70×10^{1}
Duplicate m=5000	2	3.99×10^{-1}	2.30×10^{-1}	6.61×10^{-1}	1.00×10^{-3}	9.05×10^{-2}	4.00×10^{-3}	1.27×10^{6}	9.68	3.57×10^{5}	-inf	8.70×10^{-1}	9.78×10^{1}
K-Center-Greedy	2	5.05×10^{-1}	4.40×10^{-1}	9.23×10^{-1}	5.01×10^{-1}	2.14×10^{-1}	6.00×10^{-3}	3.09×10^{6}	8.50	3.42×10^{5}	-2.29×10^{4}	8.37×10^{-1}	6.86×10^{1}
K-means Clustering	2	4.40×10^{-1}	2.60×10^{-1}	6.98×10^{-1}	3.01×10^{-1}	1.06×10^{-1}	5.00×10^{-3}	1.60×10^{6}	9.86	3.54×10^{5}	-3.63×10^{4}	8.68×10^{-1}	9.49×10^{1}
QDIT	2	4.03×10^{-1}	2.23×10^{-1}	6.50×10^{-1}	2.83×10^{-1}	9.05×10^{-2}	4.00×10^{-3}	2.60×10^{6}	9.74	3.54×10^{5}	-3.87×10^{4}	8.71×10^{-1}	9.91×10^{1}
Repr Filter	2	4.95×10^{-1}	4.21×10^{-1}	9.01×10^{-1}	4.76×10^{-1}	1.99×10^{-1}	6.00×10^{-3}	7.15×10^{6}	8.59	3.46×10^{5}	-2.42×10^{4}	8.39×10^{-1}	6.98×10^{1}

Table 7: Comprehensive experimental results on Qwen-2.5-7B. Each data selection strategy variant is evaluated over two independent runs (except for random selection).

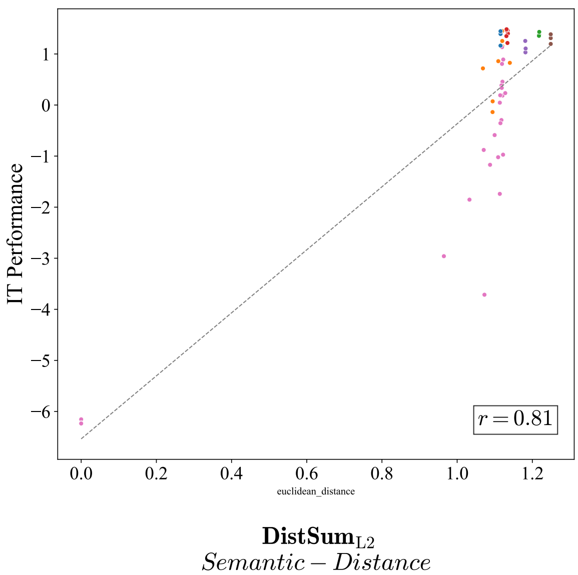


Figure 6: Evaluation of DistSum$_{L2}$ metric by their correlation with IT performance.

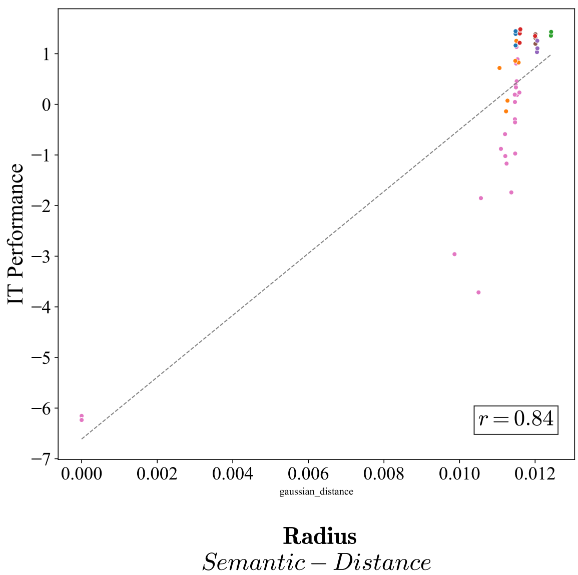
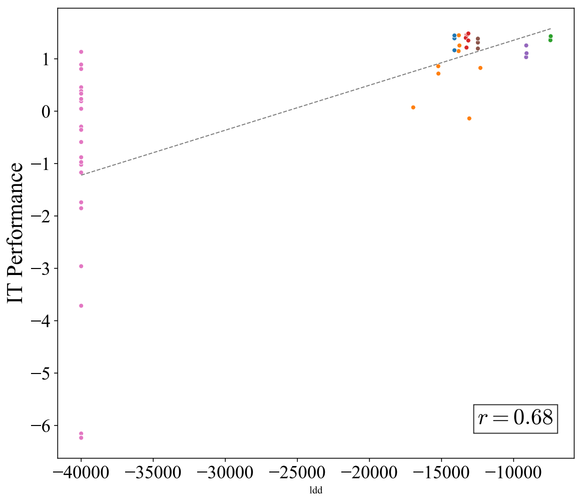


Figure 7: Evaluation of Radius metric by their correlation with IT performance.



Log Determinant Distance
Semantic − Distance

Figure 8: Evaluation of Log Determinant Distance metric by their correlation with IT performance.