

Learning from the Best, Differently: A Diversity-Driven Rethinking on Data Selection

Hongyi He^{1*} Xiao Liu^{2†} Zhenghao Lin² Mingni Tang^{3*} Yi Cheng^{3*}

Jintao Wang^{1,4†} Wenjie Li³ Peng Cheng² Yeyun Gong^{2†}

¹ Tsinghua University, ² Microsoft Research, ³ The Hong Kong Polytechnic University

⁴ Beijing National Research Center for Information Science and Technology (BNRist)

{hehy22@mails, wangjintao@}.tsinghua.edu.cn,

{xiaoliu2, zhenghaolin, pengc, yegong}@microsoft.com,

{minnie17.tang, alyssa.cheng}@connect.polyu.hk,

cswjli@comp.polyu.edu.hk

Abstract

High-quality pre-training data is crucial for large language models (LLMs), yet existing score-based selection approaches often fail to balance quality and diversity. Empirical studies have revealed a non-monotonic relationship between dataset scores and downstream performance: simply selecting the top-scored data often leads to performance degradation due to the collapse of correlated evaluation dimensions, which sacrifices diversity. To address this, we propose **Orthogonal Diversity-Aware Selection (ODiS)**, a novel framework to preserve both quality and diversity by decomposing correlated evaluation metrics into orthogonal feature dimensions. First, ODiS evaluates data from multiple dimensions, covering language quality, knowledge quality, and comprehension difficulty. The resulting multi-dimensional scores are then decorrelated via Principal Component Analysis (PCA), yielding orthogonal evaluation dimensions. For each dimension, a RoBERTa-based scorer is trained to regress the data onto PCA-projected scores, enabling scalable inference on large corpora. Finally, ODiS constructs the training dataset by selecting the top-scored data from each orthogonal dimension, thereby ensuring both quality and diversity. Empirical results show that ODiS-selected data exhibit less than 2% inter-dimension overlap, confirming the orthogonality between dimensions. More importantly, models trained on ODiS-selected data significantly outperform baselines across multiple downstream benchmarks, underscoring the necessity of orthogonal, diversity-aware data selection for LLMs.

1 Introduction

Pretraining is the primary stage for models to acquire fundamental abilities, such as language understanding, text generation, and information extraction (Brown et al., 2020; Chowdhery et al., 2023;

Roberts et al., 2020). These capabilities are largely determined by data quality and diversity. Quality captures authenticity, reliability, and semantic integrity, ensuring that models learn accurate and well-structured knowledge. Diversity, on the other hand, emphasizes coverage and comprehensiveness, enabling models to generalize across domains and tasks. With increasing model and corpus sizes, designing efficient data selection methods that jointly account for these two aspects has become critical for improving model performance.

Existing works have proposed diverse methods for selecting data based on quality and diversity. Quality-based methods utilize rule-based heuristics (Laurençon et al., 2022; Weber et al., 2024; Penedo et al., 2023; Lee et al., 2022), such as document length constraint and content de-duplication, or score-based techniques (Touvron et al., 2023; Wettig et al., 2024; Penedo et al., 2024; Su et al., 2025), where classifiers or perplexity models assign a single quality score to filter noisy or irrelevant data. Diversity-based methods (He et al., 2024; Zhang et al., 2025; Yang et al., 2025) instead focus on broadening coverage by domain mixing or embedding-space clustering to reduce redundancy and expand the distribution. Recent methods further combine quality and diversity via multi-dimensional scoring or domain-level mixing (Zhuang et al., 2025; Liu et al., 2025; Bai et al., 2025). However, the intrinsic correlation between dimensions makes weight tuning challenging, and naive aggregation often results in redundant, overlapping signals. In summary, current work encounters three key challenges: (1) The top-scoring data are not always optimal, and sampling from a broader range is necessary to achieve satisfactory performance, but the cause of this selection bias remains unexplored. (2) Score-based methods combine multiple aspects into a one-dimensional signal, resulting in dimension collapse and failing to capture both quality and diversity. (3) Delicate

*Work done during internship at Microsoft Research Asia.

†Corresponding author.

hyper-parameter tuning is required to balance the influence from different dimensions, undermining generality and deployment of the methods.

To address these challenges, we first revisit the data selection process through the lens of bias and correlation, identifying the neglect of diversity as the fundamental cause for performance degradation. Guided by this insight, we propose **Orthogonal Diversity-Aware Selection (ODiS)**, a framework for effectively constructing a dataset that balances quality and diversity. Specifically, we annotate a reference dataset from 11 dimensions, covering four main categories: language quality, knowledge quality, comprehension difficulty, and information quality. After analyzing the correlation across the dimensions, we reveal strong entanglement among them, which introduces redundancy and biases the score-based selection. To decouple these features, we apply Principal Component Analysis (PCA) to scores and derive orthogonal evaluation dimensions. To accelerate and scale, RoBERTa-based scorers are trained to predict scores for each orthogonal dimension on the target dataset. Finally, ODiS constructs the training set by selecting the top samples along each orthogonal dimension, ensuring both quality and diversity of the dataset. Unlike conventional methods that treat data quality as a monolithic or aggregated metric, ODiS selects data from orthogonal dimensions, ensuring that the best data are captured across non-overlapping features.

Empirical results show that ODiS outperforms DSIR, PPL, and Nemotron-CC across downstream benchmarks. We analyze the source of performance gains by comparing top-scored data with samples drawn from broader ranges of each dimension, and find that top-scored subsets consistently underperform, whereas combining data across dimensions substantially improves performance. Data analysis further confirms strong pre-PCA correlations and demonstrates that orthogonal dimensions capture distinct data aspects, thereby validating quality and diversity in the selected data. Finally, ablations further reveal an efficiency-performance trade-off in the number of dimensions.

The main contributions of this work are as follows: (1) We provide the first analysis of selection bias and performance degradation in top-scored data, identifying inter-dimensional correlation and neglected diversity as the underlying cause. (2) We propose ODiS, a score-based framework that explicitly ensures quality and diversity via dimension decomposition. (3) Extensive experiments

show that ODiS outperforms baselines and benefits from dimension decomposition, inspiring future data selection methods to reduce inter-dimensional correlations and enhance data diversity.

2 Method

2.1 Selection Bias in Score-based Methods

We begin by analyzing the causes of selection bias and the role of data sampling by examining model performance with a fixed amount of training data drawn from candidate data pools of varying sizes. Following (Wettig et al., 2024; Zhuang et al., 2025), we establish multiple metrics targeting different semantic features, details of which are provided in Section 2.3. Then, we utilize a score-based method to construct candidate pools of varying sizes from the Nemotron-CC dataset (Su et al., 2025), which selects data based on the average scores from the above metrics. Specifically, we select the data with the top-k scores at different thresholds, yielding data pools of 100B to 900B tokens. After that, we sample 100B tokens from each candidate pool to construct training sets, and train a 1.5B-parameter model from scratch on each of these datasets. Finally, model performance is evaluated across multiple benchmarks, and the results are provided in Figure 1. More training details and benchmark selection are provided in Section 3.1.

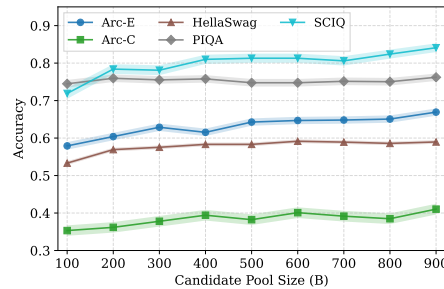
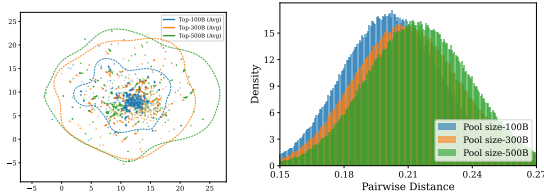


Figure 1: Performance of training subsets drawn across candidate pools. The x-axis denotes the size of the top-k selected pool, filtered by average quality scores. The subset is constructed by randomly sampling 100B tokens from each candidate pool, and a 1.5B model is then trained and evaluated. The leftmost point (100B) represents training on the highest-scored data, while moving right indicates including lower-scored data. The standard error is shown with the shaded region.

Figure 1 demonstrates that the data with the highest score (e.g., the candidate pool size equals 100B) performs the worst, while including data with lower scores (e.g., the candidate pool size larger than 100B) leads to improvements. This reveals a non-monotonic relationship between the data score and

model performance, indicating selection bias in score-based methods and complicating data selection. Since the top-scored data has the highest quality, we further investigate its diversity through an embedding-based visualization to determine the reason for the selection bias.

The diversity is examined through both qualitative and quantitative analysis. We sample the same amount of data from the above candidate pools. Qualitatively, Figure 2a illustrates the UMAP projection of text embeddings. As can be observed, the purple points (Top-100B) tend to cluster in a limited region, while the yellow points (Top-500B) span a broader space. This demonstrates that directly selecting the data with the highest score results in more homogeneous features, whereas broader selection achieves comprehensive semantic coverage. Quantitatively, Figure 2b demonstrates that as the selection scope increases (e.g., from 100B to 500B), the pairwise distance distribution shifts rightward, indicating an increase in data distinction and semantic diversity. Therefore, we can conclude: *Score-based methods with aggregated scores suffer selection bias, which sacrifices data diversity and homogenizes the selected data.*



(a) UMAP projection. (b) Pairwise distance distribution.

Figure 2: Visualization of data embeddings. The selection scope indicates the number of top-ranked tokens based on average scores. The embeddings are generated with the m3e-base model (Wang et al., 2023) on 1000 randomly sampled texts. The left figure shows embeddings reduced to 2 dimensions with UMAP and the estimated boundary with kernel density estimation (KDE), and the right figure shows pairwise distances computed using cosine distance.

2.2 Problem Formulation

Motivated by the analysis in Section 2.1, we focus on selecting high-quality data that maintains diversity. Our objective is to select the most valuable subset $\mathcal{D}_s$ from a massive target corpus $\mathcal{D}_t$. We define an evaluation function $E : \mathcal{X} \rightarrow \mathbb{R}^m$ that annotates an m -dimensional quality vector expressed as $\alpha^{(i)} = (\alpha_1^{(i)}, \dots, \alpha_m^{(i)})$ to data x_i . Due to selection bias, relying on raw evaluations leads to suboptimal selection. Therefore, we will learn

a mapping function $\Phi : \mathbb{R}^m \rightarrow \mathbb{R}^K$, which transforms the raw score vector $\alpha^{(i)}$ into a latent feature representation $\beta^{(i)} = \Phi(\alpha^{(i)}) \in \mathbb{R}^K$. This latent representation $\beta^{(i)}$ aims to decouple the underlying semantic features and correct the bias inherent in raw evaluation. Since directly learning the above functions on a large-scale target dataset $\mathcal{D}_t$ is computationally costly, we introduce a reference dataset $\mathcal{D}_r = \{x_i\}_{i=1}^N$ with $N \ll |\mathcal{D}_t|$, learn a parameterized proxy scorer $S_\theta : \mathcal{X} \rightarrow \mathbb{R}^K$ approximating the composite mapping $\Phi \circ E$ on $\mathcal{D}_r$ and then infer $\beta^{(i)}$ on $\mathcal{D}_t$. Based on representation $\beta^{(i)}$, we adopt a selection method $\text{Sel}(\cdot)$ that constructs subset $\mathcal{D}_s$.

We formulate this task as the following optimization problem, where we aim to maximize the downstream performance $G(\cdot)$ under token budget B by designing $E(\cdot)$, $\Phi(\cdot)$, $S_\theta(\cdot)$, and $\text{Sel}(\cdot)$:

$$\mathcal{D}_s^* = \arg \max_{\mathcal{D}_s \subset \mathcal{D}_t} G(\mathcal{M}(\mathcal{D}_s; \omega)), \quad (1a)$$

$$\text{s.t. } \mathcal{D}_s = \text{Sel}(\mathcal{D}_t, S_{\theta^*}(\mathcal{X})), \sum_{x \in \mathcal{D}_s} l^{\text{tok}}(x) \leq B, \quad (1b)$$

$$\theta^* = \arg \min_{\theta} \sum_{x \in \mathcal{D}_r} \mathcal{L}(S_\theta(x), \Phi(E(x))), \quad (1c)$$

where $\mathcal{M}(\cdot; \omega)$ denotes LLM pre-training on the selected data, $G(\cdot)$ represents the evaluation on downstream benchmarks, $l^{\text{tok}}(\cdot)$ denotes the number of tokens, $\mathcal{L}$ is a regression loss function, and B is the data budget.

2.3 Multi-Dimension Data Evaluation

To avoid directly optimizing towards task-specific signals, we evaluate data by setting $m = 11$ dimensions across four general aspects: Language quality, Knowledge quality, Comprehension difficulty, and Information Quality (Zhuang et al., 2025). For generality, we still use m to denote the dimension number in the framework. We briefly describe the above dimensions, with details in Appendix C:

- **Language quality.** We prefer data that is (i) coherent in structure, (ii) concise without redundancy, and (iii) correct in spelling/grammar and word choice (Penedo et al., 2023).
- **Knowledge quality.** We value content with (i) sufficient coverage and (ii) depth, (iii) useful reasoning signals, (iv) clear educational value, and (v) practical value (Gunasekar et al., 2023; Guo et al., 2025).
- **Comprehension difficulty.** We assess the difficulty level, i.e., conceptual complexity and domain professionalism, as greater difficulty

improves generalization (Agrawal and Singh, 2023).

- **Information quality.** We require (i) factual accuracy and (ii) sufficient completeness so models can learn reliable, fully specified facts (Chang et al., 2024).

Each document in the reference dataset $\mathcal{D}_r$ is assigned a score of 0 to 5 for each dimension using the OpenAI GPT API, as widely adopted in previous works related to LLM-as-a-Judge (Zhuang et al., 2025; Su et al., 2025; Penedo et al., 2024) and LLM training (Grattafiori et al., 2024). Detailed metric definitions and prompts are provided in Appendix O. The resulting scores constitute a

$$\text{matrix } \mathbf{X} = \begin{bmatrix} \alpha_1^{(1)} & \cdots & \alpha_m^{(1)} \\ \vdots & \ddots & \vdots \\ \alpha_1^{(N)} & \cdots & \alpha_m^{(N)} \end{bmatrix} \in \mathbb{R}^{N \times m}.$$

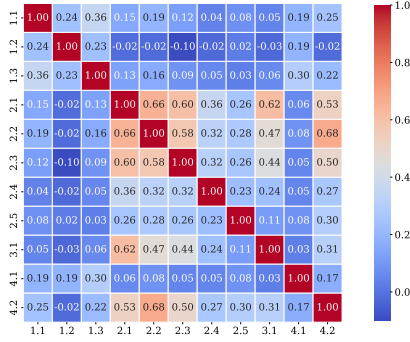


Figure 3: Correlation across dimensions. The index denotes each dimension, with mapping in Appendix C.

We calculate the correlation coefficient across the 11 dimensions above, and the results are visualized in Figure 3. We observe moderate correlations across all dimensions, such as knowledge depth vs. knowledge richness and completeness vs. knowledge richness, indicating partial overlap in their coverage, aligning with the results found in (Zhuang et al., 2025). Such redundancy implies that directly aggregating these metrics would amplify common signals at the expense of unique features. Combined with the analysis in Section 2.1, these results indicate that: *While metrics are conceptually distinct in design, they exhibit significant intrinsic correlations in practice. This redundancy biases selection and leads to diversity loss, motivating dimension decomposition to isolate independent quality signals.*

2.4 Orthogonal Diversity-Aware Selection

Dimension decomposition. To decouple the correlated metrics shown in Figure 3, our framework seeks a transformation Φ that maps scores from

the raw evaluation space into a latent feature space where the dimensions are uncorrelated. While various decorrelation techniques exist (e.g., ICA, Autoencoders), we adopt Principal Component Analysis (PCA) as the implementation of Φ due to its parameter-free robustness and interpretability in variance ranking.

We first standardize the scores by centering each dimension and scaling it to unit variance. For the j -th dimension, we compute $\mu_j = \frac{1}{N} \sum_{i=1}^N \alpha_{ij}$, $\sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^N (\alpha_{ij} - \mu_j)^2}$, and obtain the standardized score matrix $\tilde{\mathbf{X}}$ as $\tilde{\alpha}_{ij} = \frac{\alpha_{ij} - \mu_j}{\sigma_j}$.

After that, we compute the covariance matrix Σ and adopt eigen decomposition to Σ :

$$\Sigma = \frac{1}{N-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T, \quad (2)$$

where $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_m]$ are eigenvectors and $\mathbf{\Lambda} = \text{diag}(\lambda_1 \geq \dots \geq \lambda_m \geq 0)$ are eigenvalues. Each eigenvector represents an orthogonal combination of original metrics, with the eigenvalue quantifying the variance contribution, i.e., the proportion of feature representation that the eigenvector accounts for. A higher eigenvalue indicates that the data spreads more along this direction, and the eigenvector contains a more representative feature.

Score projection to the latent space. We select the first K principal components (PC) to satisfy the explained-variance ratio: $\frac{\sum_{k=1}^K \lambda_k}{\sum_{k=1}^m \lambda_k} \geq \tau$, where τ is the threshold. Then, we construct the projection matrix: $\mathbf{W}_K = [\mathbf{v}_1, \dots, \mathbf{v}_K] \in \mathbb{R}^{m \times K}$. Through projection, the compressed score vector for x_i is calculated as: $\beta^{(i)} = \mathbf{W}_K^T (\alpha^{(i)})^T \in \mathbb{R}^K$. Each element of $\beta^{(i)}$ represents data x_i 's score along an orthogonal dimension, ensuring the score in one dimension does not imply redundancy in another.

Learning the latent mapping via model-based scorer. To scale this evaluation to the large-scale target dataset $\mathcal{D}_t$, we train RoBERTa-based scorers $r_k(\cdot)$ to approximate the projection on the k -th PC dimension $\Phi_k(E(x))$. This directly solves the optimization objective defined in (1c). The scorer minimizes the regression loss between the k -th projected score and the raw text. All the scorers $r_k(\cdot)$, $k \in \{1, \dots, K\}$ constitute the proxy scorer S_θ . We then label the target dataset $\mathcal{D}_t$ with scorer $r_k(\cdot)$ to obtain the scores $\hat{\beta}_k^{(i)}$, $k \in 1, \dots, K, i \in \{1, \dots, |\mathcal{D}_t|\}$. The details of the parameter setting and the training of RoBERTa-based scorers are in Appendix E. The overall cost of data selection, including training RoBERTa-based scorers and the

annotation process, is in Appendix F.

Constructing dataset with the best data in each dimension. We allocate a data budget B_k to each PC dimension considering its contribution and the total data budget: $B = \sum_{k=1}^K B_k$. For each dimension, a score threshold t_k is set such that the selected data satisfy the budget constraint B_k , and the corresponding subset is denoted as $\mathcal{D}_s^k$. Finally, the dataset is constructed as: $\mathcal{D}_s = \{x_i \in \mathcal{D}_t \mid \exists k \in 1, \dots, K, \hat{\beta}_k^{(i)} > t_k\}$. Alternatively, this can be viewed as the union of top-scored subsets across all dimensions, i.e., $\mathcal{D}_s = \cup_{k=1}^K \mathcal{D}_s^k$. This construction ensures that each dimension contributes the highest-quality data, while the orthogonality of the PC dimensions enhances diversity in the resulting dataset. The overall procedure is summarized in Algorithm 1 in Appendix H.

3 Experiments

3.1 Experiment Setup

Dataset. We use the Nemotron-CC dataset (Su et al., 2025) as the target dataset for selection, which is a large-scale dataset for pretraining large language models. The dataset comprises both real-world and synthetic data, covering major domains, such as web knowledge and question answering. The reference dataset $\mathcal{D}_r$ is constructed by 460k samples from the FineWeb-Edu dataset (Penedo et al., 2024). For tokenization, we use the LLaMA-3-8B tokenizer and set the maximum sequence length to 4096.

Evaluation. We evaluate model performance with lm-eval-harness framework (Gao et al., 2024). We first monitor task-level performance fluctuations across training steps, with results presented in Appendix M. Based on the above result, we use the “fine task” metric (Kydlicek et al., 2024) to select tasks with obvious performance variation as training progresses. We select five tasks covering main categories: **General Knowledge** (including Arc-Easy/Challenge (Clark et al., 2018)), **Commonsense Reasoning** (including Hellaswag (Zellers et al., 2019), SCIQ (Welbl et al., 2017)), and **Physical Reasoning** (PIQA (Bisk et al., 2020)).

Training. Unless specified in the context, each experiment is conducted under a training data budget of 100B tokens and a 1.5B parameter model, and ODiS selects data from the first 4 PC dimensions and allocates data budget evenly to each dimension. The *top* selection represents directly selecting a 100B token dataset for training, while *sample* se-

lection represents selecting a 700B-token dataset and randomly sampling training data from it. More training details can be found in Appendix D.

3.2 Main Results

The main results are summarized in Table 1. We compare ODiS with DSIR (Xie et al., 2023), PPL (Ankner et al., 2025), and Nemotron-CC (Su et al., 2025), and methods adopting a similar separate-then-select paradigm, including Web-Organizer (Wettig et al., 2025), Semdedup (Abbas et al., 2023), and selection from the original 11 dimensions. The results show that the model trained on data selected by ODiS achieves the best average performance and surpasses all the methods across task categories, highlighting its versatility in addressing a wide range of downstream tasks. The gains are typically evident in Arc-E and Arc-C, indicating improvements in overall knowledge and content diversity. In contrast, PPL and DSIR focus solely on data quality, limiting diversity and hampering overall performance. Moreover, compared with other methods that employ a similar separate-then-select paradigm, ODiS benefits from effective orthogonal dimension construction, which avoids the interference and correlation from distinct dimensions. Furthermore, the superiority over the data selected with the original 11 dimensions confirms that the performance gain of ODiS primarily comes from the PCA dimension decomposition, rather than the multi-dimensional data evaluation. These results demonstrate the importance of maintaining diversity during data selection and that decomposition is promising for improving diversity.

3.3 Analysis Result

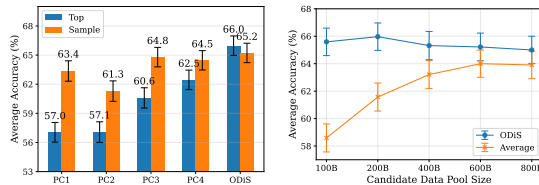
3.3.1 Inspection of Data Bias

ODiS mitigates data selection bias. Figure 4a shows that models trained with top-scored data within a single dimension consistently underperform, while sampling from a broader score range yields significant performance gains. This observation highlights the inherent bias in relying exclusively on single-dimension scoring, where overemphasis on one metric neglects complementary aspects of data quality and diversity. Furthermore, Figure 4b illustrates that averaging scores from multiple dimensions only partially alleviates the data bias issue, as correlation across dimensions discussed in Section 3.3.2 continues to bias the data selection. In contrast, ODiS decomposes the dimensions into orthogonal components and selects

Table 1: Performance across data selection methods. The results are reported as percentage accuracy, with $\pm$ standard error shown in a smaller font size. The CI indicates the 95% confidence interval of accuracy difference compared with ODiS to demonstrate the performance gain is statistically meaningful.

Method	Arc-C	Arc-E	Hellaswag	SCIQ	PIQA	Average	95% CI
Random	35.0 $\pm$ 1.4	62.7 $\pm$ 1.0	58.3 $\pm$ 0.5	85.5 $\pm$ 1.1	74.5 $\pm$ 1.0	63.2 $\pm$ 1.0	[2.0, 3.5]
Nemotron-HQ	37.3 $\pm$ 1.4	64.6 $\pm$ 1.0	57.7 $\pm$ 0.5	83.9 $\pm$ 1.1	73.6 $\pm$ 1.0	63.4 $\pm$ 1.0	[1.8, 3.3]
PPL- <i>Top</i>	37.9 $\pm$ 1.4	62.8 $\pm$ 1.0	54.7 $\pm$ 0.5	83.4 $\pm$ 1.2	74.7 $\pm$ 1.0	62.7 $\pm$ 1.0	[2.5, 4.1]
PPL- <i>Sample</i>	36.1 $\pm$ 1.4	64.3 $\pm$ 1.0	58.4 $\pm$ 0.5	85.8 $\pm$ 1.1	74.8 $\pm$ 1.0	63.9 $\pm$ 1.0	[1.3, 2.8]
DSIR	27.8 $\pm$ 1.3	48.5 $\pm$ 1.0	54.6 $\pm$ 0.5	78.5 $\pm$ 1.3	71.0 $\pm$ 1.1	56.1 $\pm$ 1.0	[9.0, 10.8]
PC Avg- <i>Top</i>	35.3 $\pm$ 1.4	57.9 $\pm$ 1.0	53.3 $\pm$ 0.5	71.9 $\pm$ 1.4	74.5 $\pm$ 1.0	58.6 $\pm$ 1.1	[6.5, 8.2]
PC Avg- <i>Sample</i>	39.2 $\pm$ 1.4	64.8 $\pm$ 1.0	58.9 $\pm$ 0.5	80.6 $\pm$ 1.3	75.1 $\pm$ 1.0	63.7 $\pm$ 1.0	[1.5, 3.0]
Web-Org <i>Format</i>	36.4 $\pm$ 1.4	64.5 $\pm$ 1.0	57.5 $\pm$ 0.5	84.6 $\pm$ 1.1	73.7 $\pm$ 1.0	63.3 $\pm$ 1.0	[1.9, 3.4]
Web-Org <i>Topic</i>	36.2 $\pm$ 1.4	62.1 $\pm$ 1.0	60.3 $\pm$ 0.5	83.7 $\pm$ 1.2	75.8 $\pm$ 1.0	63.6 $\pm$ 1.0	[1.6, 3.1]
Ori-11-Dim	41.0 $\pm$ 1.4	66.5 $\pm$ 1.0	54.8 $\pm$ 0.5	86.5 $\pm$ 1.1	73.4 $\pm$ 1.0	64.5 $\pm$ 1.0	[0.8, 2.2]
Semdedup	39.6 $\pm$ 1.4	66.4 $\pm$ 1.0	57.8 $\pm$ 0.5	85.0 $\pm$ 1.1	74.4 $\pm$ 1.0	64.6 $\pm$ 1.0	[0.6, 2.1]
ODiS	41.6 $\pm$ 1.4	66.9 $\pm$ 1.0	58.4 $\pm$ 0.5	85.6 $\pm$ 1.0	77.4 $\pm$ 1.0	66.0 $\pm$ 1.0	

high-quality data from each dimension, effectively mitigating data bias and enhancing diversity. Notably, ODiS achieves superior performance with the smallest subset of data, avoiding both excessive sampling ranges and unnecessary data waste.



(a) Performance with data selected from PC dimensions (b) Performance across data pools with different sizes

Figure 4: Model performance across selection methods and scopes. “Average” selects the data with the average scores from PC dimensions. Each data point represents a model trained with 100B tokens drawn from the candidate pool, with error bar indicating standard error.

ODiS effectively enhances data diversity. To validate the diversity gain from ODiS, we compare the embedding visualization of ODiS-selected data against score-based baselines. Figure 5 shows that the data selected by ODiS has a significantly larger average pairwise distance, even compared with a larger selected data pool, indicating enhanced data diversity. Similarly, UMAP visualization reveals that the ODiS-selected data spans a wider region in the compressed space, while the data selected with baseline methods tend to cluster around a narrower region. These results suggest that ODiS mitigates the selection bias and captures a wider range of semantic features during data selection.

3.3.2 Correlation between Dimensions

Figure 6 illustrates the correlations between original dimensions and PC dimensions after PCA, indicating that PCs retain meaningful data features.

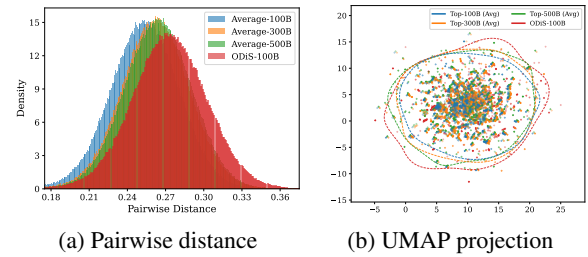


Figure 5: Diversity visualization across selection methods. The same number of samples is drawn from the selected data and then an embedding visualization is performed with similar setting in Figure 2.

For example, PC1 aligns strongly with knowledge quality and comprehension difficulty, highlighting its central role in characterizing overall data quality. By contrast, language-related dimensions exhibit weaker direct alignment with the leading PCs. This observation indicates that linguistic factors may already be embedded in other dimensions, e.g., knowledge quality, and thus do not emerge as dominant signals. More generally, the orthogonal principal components are linear combinations of the original dimensions. They should not be interpreted as single semantic dimensions, but rather as abstract features that combine multiple correlated attributes and capture salient aspects of the dataset from the model’s perspective.

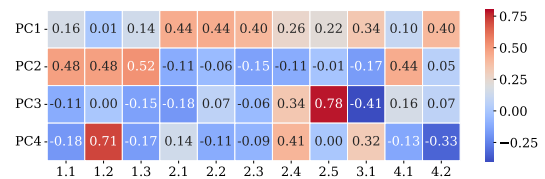
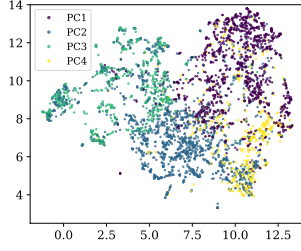
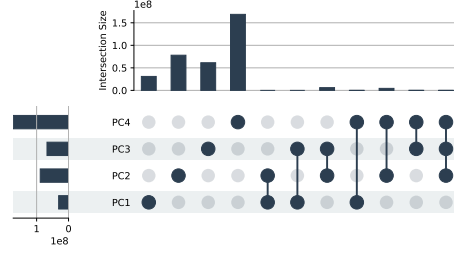


Figure 6: Correlation analysis of original evaluation dimensions and the PC dimensions. The name of the dimension index can be found in Appendix C.



(a) UMAP for embeddings of randomly sampled data from PC dimensions (i.e., $\mathcal{D}_s^k$).



(b) UpSet plot of top-scored subsets (i.e., $\mathcal{D}_s^k$), with intersections indicating data that simultaneously belong to multiple top-scored subsets.

Figure 7: Data orthogonality across PC dimensions, with top-scored 100B tokens from each PC dimension.

3.3.3 Orthogonality Between Dimensions

To assess the distinctiveness across PC dimensions, we visualize the embeddings of top-scored data from each dimension using UMAP projection and an Upset plot, as shown in Figure 7. The UMAP visualization in Figure 7a reveals that samples from different dimensions occupy separable regions, validating the orthogonality of the dimensions and suggesting that combining data across dimensions can yield complementary data subsets. Figure 7b illustrates the marginal intersection of the data from different PC dimensions, with an overlapping ratio of 2% after tokenization, further confirming the effectiveness of the dimension decomposition. However, the data selected from the 11 original dimensions suffers from about 50% duplication.

3.3.4 Scaling with More PC Dimensions

To examine whether increasing the number of PCs improves performance, we conduct an ablation study with varying PC dimensions, as summarized in Table 2. The results show that performance improves steadily up to four dimensions, after which the gain saturates. Beyond four PCs, additional dimensions contribute marginally to the model performance, likely because lower-variance PCs contain less information or noisy signals. Additionally, selecting more PCs will increase computational cost, as labeling a large corpus is both time- and resource-intensive. These findings suggest that a small set of carefully selected orthogonal dimensions is sufficient for robust data selection, balancing efficiency and effectiveness.

3.3.5 Scaling with Different Model Sizes and Data Budgets

Table 3 summarizes results with model sizes of 400M, 1.5B, and 8B, under token budgets of 100B, 200B, and 25B, respectively, to demonstrate the scalability and generalization of the proposed ODiS. The results are summarized in Table 3. We can observe that ODiS achieves consistently the

Table 2: Performance across data from PC subsets. “PC1- k ” denotes the dataset constructed by the union of top-scored subsets from the first k PC dimensions.

Method	Arc-C	Arc-E	H-Swag	SCIQ	PIQA	Avg
PC1	36.4	60.4	43.1	81.1	64.4	57.1
PC1-2	35.8	61.1	53.7	83.0	75.1	61.7
PC1-3	36.4	60.9	59.4	79.1	77.5	62.6
PC1-4	41.6	66.9	58.4	85.6	77.4	66.0
PC1-5	39.4	66.9	56.9	88.0	76.7	65.6
PC1-6	42.5	71.0	56.1	87.2	76.0	66.5

best performance across these settings. The performance gain is more pronounced on PIQA, indicating an enhanced ability in the physical common-sense reasoning. Compared with Nemotron-HQ, the 400M model exhibits a larger margin on Hellaswag, while the 8B model outperforms on Arc-C and Arc-E. These results demonstrate that the proposed ODiS scales effectively across different models and data budgets, enabling more efficient training in diverse settings.

4 Related Works

Existing data selection methods have been proposed to enhance data quality and diversity through three main directions: non-classifier-based methods, single-classifier-based methods, and multi-classifier-based methods.

Non-classifier-based methods. Various works have relied on rule-based filtering with explicit heuristics or deterministic criteria (Laurençon et al., 2022; Weber et al., 2024; Penedo et al., 2023; Raffel et al., 2020; Lee et al., 2022), including language identification, URL blocks, content deduplication, and document length thresholds. Moreover, these approaches are combined into multi-stage pipelines that sequentially perform cleaning, deduplication, quality, and safety filtering (Nguyen et al., 2024). While rule-based methods effectively improve data quality and reduce noisy data, they fail to inspect semantic-level information and will introduce distribution bias.

Table 3: Performance across methods with model sizes 400M, 1.5B, and 8B and data budgets 100B, 200B, 25B tokens, respectively. The results are reported as percentage accuracy with standard error.

Model	Method	Arc-C	Arc-E	Hellaswag	SCIQ	PIQA	Average
400M	Nemotron-HQ	29.6±1.3	54.3±1.0	40.7±0.5	75.4±1.4	67.9±1.1	53.6±1.1
	PPL-sample	28.2±1.3	50.7±1.0	41.7±0.5	76.3 ±1.4	69.9±1.1	53.4±1.1
	DSIR	24.6±1.3	40.7±1.0	36.5±0.5	66.6±1.5	65.0±1.1	46.7±1.1
	PC-Avg-Top	29.2±1.3	48.7±1.0	41.6±0.5	69.9±1.5	70.5±1.1	52.0±1.1
	PC-Avg-Sample	30.1 ±1.3	53.2±1.0	43.2 ±0.5	75.0±1.4	70.1±1.1	54.3±1.1
	ODiS	30.0±1.3	54.3 ±1.0	42.5±0.5	74.5±1.4	72.5 ±1.0	54.8 ±1.1
1.5B	Nemotron-HQ	40.3±1.4	67.9±1.0	59.9±0.5	74.5±1.0	87.1±1.1	65.9±1.0
	PPL-Sample	38.2±1.4	67.5±1.0	60.5±0.5	75.6±1.0	88.8 ±1.0	66.1±1.0
	DSIR	30.2±1.3	50.8±1.0	58.1±0.5	72.2±1.1	79.4±1.3	58.1±1.0
	PC Avg-Top	36.6±1.4	59.1±1.0	54.8±0.5	74.9±1.0	72.8±1.4	59.6±1.1
	PC Avg-Sample	40.0±1.4	66.3±1.0	61.4 ±0.5	76.4±0.5	83.4±1.2	65.5±0.9
	ODiS	40.9 ±1.4	68.4 ±1.0	59.1±0.5	78.4 ±1.0	88.2±1.0	67.0 ±1.0
8B	Nemotron-HQ	36.2±1.4	62.1±1.0	58.4±0.5	84.3±1.2	74.9±1.0	63.2±1.0
	PPL-sample	36.2±1.4	63.1±1.0	57.9±0.5	84.9±1.1	73.3±1.0	63.1±1.0
	DSIR	30.9±1.4	49.3±1.0	55.8±0.5	77.8±1.3	71.7±1.1	57.1±1.0
	PC-Avg-Top	37.1±1.4	59.1±1.0	54.2±0.5	72.4±1.4	75.2±1.0	59.6±1.1
	PC-Avg-Sample	41.1 ±1.4	65.7±1.0	59.5 ±0.5	85.1±1.1	75.5±1.0	65.4±1.0
	ODiS	40.1±1.4	67.2 ±1.0	57.6±0.5	85.5 ±1.1	77.3 ±1.0	65.5 ±1.0

Single-classifier-based methods. In contrast to the rule-based method, it utilized a learned scoring function or discriminator to label and filter out high-quality data (Wenzek et al., 2020; Touvron et al., 2023; Wettig et al., 2024; Penedo et al., 2024; Su et al., 2025; Wang et al., 2025). Among them, language modeling perplexity has been adopted to identify data with high quality (Wenzek et al., 2020; Touvron et al., 2023). Methods like QuRating (Wettig et al., 2024) and FineWeb-Edu (Penedo et al., 2024) utilize classifiers that focus on specific aspects of LLM capabilities, such as reading comprehension and knowledge acquisition. Moreover, works like Ultra-FineWeb (Wang et al., 2025) and DSIR (Xie et al., 2023) utilize a target dataset to guide the classifier in predicting the quality of the data. Although methods with a single classifier can effectively filter out data with certain desirable features, their reliance on a single evaluation dimension limits data diversity and often leads to imbalanced capabilities in the trained models.

Multi-classifier-based methods. Compared with single-dimensional methods, the combination of multi-dimensional classifiers can provide a more comprehensive evaluation of data. However, it remains a challenge to balance the influence from different dimensions. One line of the research typically assigns weights to each dimension through performance tests on small proxy models (Zhuang et al., 2025; Bai et al., 2025). However, the dimensional correlations are not well-addressed, resulting in bias in the combined scores and reduced data diversity. Another line of the research design sam-

pling ratios for different domains (Liu et al., 2025) to ensure data diversity. However, the inherent overlapping of the domains will still lead to bias in the selected data. As a result, the traditional top- k method is ineffective in the multi-dimensional setting, leaving the question of integrating multi-dimensional evaluations open.

5 Conclusion

In this work, we investigated the underlying cause of bias in score-based data selection and identified that inter-dimensional correlations and subsequent neglected diversity lead to non-monotonicity between the dataset scores and model performance. To address this, we proposed ODiS, a framework that explicitly mitigates feature correlation by selecting the best data from orthogonal dimensions, ensuring data diversity and retaining high-quality data. The experiment results demonstrated that ODiS effectively reduces inter-dimensional correlation, enhances data diversity, and consistently improves model performance across various downstream tasks compared to several baselines. These findings highlight that effective data selection for pre-training models should consider both quality and diversity jointly, and that the correlation across different dimensions must be appropriately addressed. Looking forward, we encourage future data selection work to move beyond aggregated scoring and prioritize orthogonal evaluation to prevent feature and diversity collapse and ensure robust model generalization.

Limitations

Our analysis pipeline relies on GPT-based scoring as the initial source of supervision. Specifically, we first use GPT to assign scores to documents, then perform PCA over the resulting score space, and finally train a RoBERTa-based model to predict the PCA-derived component scores. While this pipeline enables scalable and fine-grained analysis, the resulting components inevitably inherit potential biases from the GPT judge. For example, GPT-based evaluation may favor certain document styles, reasoning formats, levels of verbosity, or surface-level plausibility. As a result, the learned principal components should be interpreted as reflecting the structure of a GPT-induced evaluation space, rather than a fully unbiased or human-grounded representation of document quality.

Although the principal components provide a compact way to decompose and analyze different dimensions of model responses, their human-interpretable meanings remain limited. PCA identifies directions of maximal variance in the score space, but these directions do not necessarily correspond to clearly separable human concepts such as factuality, coherence, reasoning depth, helpfulness, or linguistic quality. Therefore, while we can observe that different components capture distinct patterns in model behavior, assigning precise semantic interpretations to each component remains challenging. More detailed human annotation or controlled probing would be needed to better understand what each component represents from a human perspective.

We do not systematically examine the sensitivity of our findings to the choice of reference dataset and model-based analysis method. Since the PCA space is constructed from scores assigned to a particular set of responses, changing the reference dataset, prompt distribution, model outputs, or scoring model may alter the resulting components. Similarly, the RoBERTa-based predictor may introduce its own modeling bias when approximating PCA-derived scores. Therefore, the conclusions drawn from the current component space may depend on the specific datasets, GPT judge, and fitting model used in our pipeline. Future work should evaluate the robustness of the proposed analysis across multiple reference datasets, alternative judge models, different embedding or fitting architectures, and, ideally, human-annotated validation sets.

Acknowledgement

Hongyi He and Jintao Wang are supported by the National Natural Science Foundation of China under Grant 625B2106.

References

- Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.
- Ameeta Agrawal and Suresh Singh. 2023. Corpus complexity matters in pretraining language models. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 257–263.
- Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max M Marion, Matthew Leavitt, and Mansheej Paul. 2025. [Perplexed by perplexity: Perplexity-based data pruning with small reference models](#). In *International Conference on Learning Representations*, volume 2025, pages 100615–100634.
- Tianyi Bai, Ling Yang, Zhen Hao Wong, Fupeng Sun, Xinlin Zhuang, Jiahui Peng, Chi Zhang, Lijun Wu, Qiu Jiantao, Wentao Zhang, Binhang Yuan, and Conghui He. 2025. [Efficient pretraining data selection for language models via multi-actor collaboration](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9465–9491, Vienna, Austria. Association for Computational Linguistics.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. 2024. [How do large language models acquire factual knowledge during pretraining?](#) In *Advances in Neural Information Processing Systems*, volume 37, pages 60626–60668. Curran Associates, Inc.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul

- Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Nan He, Weichen Xiong, Hanwen Liu, Yi Liao, Lei Ding, Kai Zhang, Guohua Tang, Xiao Han, and Yang Wei. 2024. [SoftDedup: an efficient data reweighting method for speeding up language model pre-training](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4011–4022, Bangkok, Thailand. Association for Computational Linguistics.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Hynek Kydlicek, Guilherme Penedo, Cl  mentine Fourier, Nathan Habib, and Thomas Wolf. 2024. [Finetasks: Finding signal in a haystack of 200+ multilingual tasks](#).
- Hugo Lauren  on, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo Gonz  lez Ponferrada, Huu Nguyen, J  rg Froberg, Mario   a  ko, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, and 35 others. 2022. [The BigScience roots corpus: A 1.6TB composite multilingual dataset](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. [Deduplicating training data makes language models better](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.
- Fengze Liu, Weidong Zhou, Binbin Liu, Zhimiao Yu, Yifan Zhang, Haobin Lin, Yifeng Yu, Bingni Zhang, Xiaohuan Zhou, Taifeng Wang, et al. 2025. QuaDMix: Quality-diversity balanced data selection for efficient LLM pretraining. *arXiv preprint arXiv:2504.16511*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.
- Guilherme Penedo, Hynek Kydli  ek, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data only. *Advances in Neural Information Processing Systems*, 36:79155–79172.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

- Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. [How much knowledge can you pack into the parameters of a language model?](#) In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.
- Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. [Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, Vienna, Austria. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#). *Preprint*, arXiv:2302.13971.
- Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, Xu Han, and Zhiyuan Liu. 2025. [Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data](#). *Preprint*, arXiv:2505.05427.
- Yuxin Wang, Qingxuan Sun, and Sicheng He. 2023. M3e: Moka massive mixed embedding model.
- Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. [RedPajama: an open dataset for training large language models](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 116462–116492. Curran Associates, Inc.
- Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. [Crowdsourcing multiple choice science questions](#). In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.
- Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. [CCNet: Extracting high quality monolingual datasets from web crawl data](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.
- Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. [QuRating: Selecting high-quality data for training language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 52915–52971. PMLR.
- Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. [Organize the web: Constructing domains enhances pre-training data curation](#). *arXiv preprint arXiv:2502.10341*.
- Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023. [Data selection for language models via importance resampling](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34201–34227. Curran Associates, Inc.
- Xianjun Yang, Shaoliang Nie, Lijuan Liu, Suchin Gururangan, Ujjwal Karn, Rui Hou, Madian Khabsa, and Yuning Mao. 2025. [Diversity-driven data selection for language model tuning through sparse autoencoder](#). *arXiv preprint arXiv:2502.14050*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Qiu Jiantao, Lei Cao, Ju Fan, Ye Yuan, Guoren Wang, and Conghui He. 2025. [Harnessing diversity for important data selection in pretraining large language models](#). In *International Conference on Learning Representations*, volume 2025, pages 72980–73003.
- Xinlin Zhuang, Jiahui Peng, Ren Ma, Yinfan Wang, Tianyi Bai, Xingjian Wei, Qiu Jiantao, Chi Zhang, Ying Qian, and Conghui He. 2025. [Meta-rater: A multi-dimensional data selection method for pre-training language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10856–10896, Vienna, Austria. Association for Computational Linguistics.

A Ethics Statement

This work focuses on pre-training data selection for large language models. The experiments use publicly available datasets, corpora, benchmarks, and evaluation tools, including Nemotron-CC, FineWeb-Edu, lm-eval-harness, and standard downstream benchmarks. We cite the original creators of these artifacts and use them only for research purposes, following their licenses and terms of use whenever applicable.

Our work does not involve direct user interaction, human-subject experiments, crowdsourcing, clinical or legal decision-making, or physical-world actions. We do not collect new personal data or intentionally use personally identifiable information. Since the large-scale corpora and benchmarks used in this work are released by prior work, we rely on their original preprocessing, filtering, and terms of use for handling potentially identifying or offensive content. We do not release newly collected raw text, personally identifying information, offensive content, or user-facing model outputs. Therefore, we do not introduce additional anonymization requirements and do not identify any apparent risks of immediate personal, physical, or societal harm beyond the general risks associated with training and deploying large language models.

B Use of Large Language Models

During manuscript preparation, we occasionally utilized large language models (LLMs) to refine language expression, such as improving sentence fluency, and enhancing readability. The model was not involved in generating original research contributions, including research direction formulation, methodologies selection, experimental design, result analysis. All the core intellectual work, such as idea development, experiment execution, and results interpretation, was carried out independently by the authors. Any linguistic suggestions offered by the LLM were carefully reviewed and selectively incorporated, ensuring that accuracy, originality, and scholarly integrity were fully maintained. The authors alone take responsibility for the research content and conclusions, and the LLM is not listed as a contribution or author.

C 11 Evaluation Dimensions

To comprehensively evaluate the quality and usefulness of training data, we design an 11-dimensional evaluation framework. These dimensions aim to

capture complementary aspects of data that jointly determine its contribution to pretraining large language models (LLMs). Specifically, the dimensions are grouped into four general categories: language quality, which reflects the clarity and fluency of expression; knowledge quality, which measures the depth, diversity, and utility of information; comprehension difficulty, which reflects the complexity of content and its potential to improve generalization; and information quality, which ensures factual correctness and completeness. Together, these dimensions provide a multi-perspective evaluation of data, enabling a balanced assessment of both quality and diversity. The index for each dimension is provided in Table 4

Below, we provide the details of each category:

1. **Language quality:** LLMs fundamentally rely on languages to understand the content and interact with humans. High-quality data contributes directly to improving models' comprehension and generation abilities. A high-quality document should present ideas in a logical and rational organization, avoid redundant and irrelevant content, and use accurate spelling and grammar to convey meaning (Penedo et al., 2023). To capture these properties, we evaluate language quality in three dimensions: (i) *coherence*, which reflects whether the text follows a logical and consistent structure; (ii) *conciseness*, which measures whether the information is conveyed efficiently without unnecessary repetition; and (iii) *spelling/grammar accuracy*, which ensures correctness in word usage and sentence construction.
2. **Knowledge quality:** Beyond linguistic clarity, high-quality data must provide comprehensive knowledge to enrich an LLM's understanding of the world. This dimension mainly measures whether a document contains valuable, diverse, and practical information that can improve the model's reasoning ability and enhance the factual knowledge base (Gunasekar et al., 2023). Recent work has also shown that the small models can approach the performance of larger ones if trained with reasoning data (Guo et al., 2025). To comprehensively capture knowledge quality, we define five dimensions: (i) *knowledge depth*, assessing the extent to which a document explores concepts beyond superficial descriptions; (ii)

Table 4: Mapping of dimension indices to feature names used in Figure 3 and Figure 6.

Idx	Feature Name	Idx	Feature Name
1.1	Language Coherence	1.2	Language Conciseness
1.3	Language Spelling Accuracy	2.1	Knowledge Depth
2.2	Knowledge Richness	2.3	Knowledge Reasoning Level
2.4	Educational Value	2.5	Practical Helpfulness
3.1	Comprehension Difficulty	4.1	Factual Accuracy
4.2	Completeness		

knowledge richness, measuring the breadth of covered topics and perspectives; (iii) *reasoning*, capturing the presence of explicit logical inference, argumentation, or step-by-step derivations; (iv) *educational value*, evaluating whether the content provides clear, structured explanations suitable for learning or instruction; and (v) *practical helpfulness*, which assesses the applicability of the knowledge to real-world problems or everyday use.

3. **Comprehension difficulty:** Challenging and complex data can enhance better generalization and adaptability in LLMs. Texts with higher conceptual complexity, specialized domain knowledge, or multi-step expository structures push models to develop deeper comprehension abilities and to generate more professional and domain-specific responses (Agrawal and Singh, 2023). This dimension, therefore, evaluates the difficulty of the data, considering factors such as the abstractness of concepts, the level of technicality or professionalism, and the requirement for multi-stage reasoning.
4. **Information Quality:** For models to learn accurate and reliable representations, it is crucial that the training data provides factual, complete, and unambiguous information. Documents with inaccurate or incomplete facts risk propagating errors and degrading downstream performance (Chang et al., 2024). To evaluate this aspect, we define two evaluation dimensions: (i) *factual accuracy*, which measures the degree to which the document presents correct and verifiable information; and (ii) *completeness*, which reflects whether the information is sufficiently detailed and covers all key aspects of the described topic, thereby reducing the risk of the model learning partial or misleading facts.

D Training Details

The training model is based on LLaMA-3 architecture. We train it from scratch with the selected data. Training uses a global batch size of 512 and the AdamW optimizer (Loshchilov and Hutter, 2019), with a peak learning rate of $3e-4$, cosine decay scheduling, and linear warmup. The hyperparameters for different model sizes are provided in Table 5.

Table 5: Hyperparameters for 400M, 1.5B, and 8B Models.

Hyper-param	400M	1.5B	8B
Hidden Size	1,024	2,048	4,096
Layers	12	22	32
Attention Heads	16	32	32
KV Heads	4	8	8
Intermediate Size	2,816	5,632	14,336

E Model-based Scorer

We selected RoBERTa-base as the foundation for our scorer due to its cost-effectiveness. The maximum context window was set to 512 tokens. Each model was finetuned for 5 epochs with a batch size of 256. The learning rate was dynamically adjusted based on validation performance. We trained the RoBERTa-base scorers to regress the FineWeb-Edu 460k samples to the PCA-transformed scores. A 20% set was used for validation. The optimized learning rate and the resulting performance of the RoBERTa scorers are reported as follows:

F Computational Cost Analysis

We analyze the computational cost of our proposed framework, which consists of four stages: labeling the reference dataset, training the RoBERTa-based scorers, annotating the large-scale dataset, and pre-training the model on the selected data.

Table 6: Performance comparison of different models. We report the learning rate, Accuracy, F1 score, and Spearman correlation coefficient.

Model	Lr	Acc (%)	F1	Spear. Corr
PC1	5e-5	72.1	0.725	0.92
PC2	3e-5	62.7	0.611	0.68
PC3	1e-4	63.3	0.603	0.67
PC4	3e-5	68.9	0.620	0.55

Labeling reference dataset For the Fineweb-Edu dataset with 460k samples, the average input sequence consists of 750 data tokens and 290 prompt tokens, with an average generation length of 50 tokens. When labeling the reference dataset across 11 dimensions, the cumulative consumption is approximately 5.5×10^9 tokens.

Following (Kaplan et al., 2020; Hoffmann et al., 2022), we approximate the FLOPs for training a Transformer-based model using Equation 3:

$$C_{\text{train}} \approx 6 \times P \times D_{\text{train}} \times E, \quad (3)$$

where P denotes size of model parameter, D_{train} denotes the tokens of the training set, and E denotes the number of training epochs. Similarly, the inference FLOPs can be approximated as:

$$C_{\text{infer}} \approx 2 \times P \times D_{\text{infer}}, \quad (4)$$

where D_{infer} denotes the number of tokens to infer on.

Training the scorer We train 4 RoBERTa-based scorers, with 5 epochs each, on the Fineweb-Edu dataset. Therefore, the total cost is about 1.04×10^{18} FLOPs.

Data Annotation. The trained scorers are used to annotate the entire dataset. This requires inference over the full dataset for each of the 4 scorers. The total cost is about 2.2×10^{21} FLOPs.

Target Model Training. Finally, we train the target model (with 1.5B parameters) on the selected dataset with 100B token training budget. The training cost is about 9×10^{20} FLOPs.

While the annotation cost is comparable to the training budget of the 1.5B model under a 100B token budget, we emphasize that this is a one-time fixed investment. Since the annotated data and trained scorers can be reused to filter large training datasets or train significantly larger models (e.g., 7B+), the cost is effectively amortized at scale, making the method highly cost-efficient for real-world pre-training.

G Results with single PC

Table 7 demonstrates the general evaluation results of data selected with the single PC scorer. We can observe that each PC exhibits strength in certain areas: PC1 and PC4 perform better on Arc-C and Arc-E, indicating a better ability at general knowledge, while PC2 and PC3 perform better on Hellaswag and PIQA, indicating a better ability for commonsense and physical reasoning. Moreover, models trained with top-scored data from each PC dimension consistently underperform, while sampling from a larger score range enhances the performance. These results highlight that different PC scorers focus on distinct data features, and using one of them alone can not achieve the best performance.

Table 7: Performance across different PC dimensions. The results are reported as percentage accuracy. ‘‘Hellaswag’’ is abbreviated as ‘‘H-Swag’’, and ‘‘Sample’’ is abbreviated as ‘‘Smp’’.

Method	Arc-C	Arc-E	H-Swag	SCIQ	PIQA	Avg
PC1- <i>top</i>	36.4	60.4	43.1	81.1	64.4	57.1
PC2- <i>top</i>	30.7	51.0	55.7	72.7	75.3	57.1
PC3- <i>top</i>	33.1	55.5	61.8	74.1	78.5	60.6
PC4- <i>top</i>	40.5	70.4	44.8	87.9	68.6	62.5
PC1- <i>Smp</i>	39.5	65.2	54.6	86.3	71.2	63.4
PC2- <i>Smp</i>	36.9	61.0	57.6	76.5	74.5	61.3
PC3- <i>Smp</i>	36.9	65.6	61.1	84.6	75.8	64.8
PC4- <i>Smp</i>	40.9	68.6	53.6	86.1	73.2	64.5

H Orthogonal Diversity-Aware Selection

Dimension decomposition. To decouple the correlated metrics shown in Figure 3, our framework seeks a transformation Φ that maps scores from the raw evaluation space into a latent feature space where the dimensions are uncorrelated. While various decorrelation techniques exist (e.g., ICA, Autoencoders), we adopt Principal Component Analysis (PCA) as the implementation of Φ due to its parameter-free robustness and interpretability in variance ranking.

We first standardize the scores by centering each dimension and scaling it to unit variance. For the j -th dimension, we compute

$$\mu_j = \frac{1}{N} \sum_{i=1}^N \alpha_{ij}, \sigma_j = \sqrt{\frac{1}{N} \sum_{i=1}^N (\alpha_{ij} - \mu_j)^2} \quad (5)$$

and obtain the standardized score matrix $\tilde{\mathbf{X}}$ as

$$\tilde{\alpha}_{ij} = \frac{\alpha_{ij} - \mu_j}{\sigma_j}. \quad (6)$$

After that, we compute the covariance matrix Σ and adopt eigen decomposition to Σ :

$$\Sigma = \frac{1}{N-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T, \quad (7)$$

where $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_m]$ are orthogonal eigenvectors and $\mathbf{\Lambda} = \text{diag}(\lambda_1 \geq \dots \geq \lambda_m \geq 0)$ are eigenvalues. Each eigenvector represents an orthogonal combination of the original metrics, with the eigenvalue quantifying the variance contribution, i.e., the proportion of the feature representation that the eigenvector accounts for. The higher eigenvalue indicates that the data spreads more along this direction, and the eigenvector contains a more representative feature.

Score projection to the latent space. We select the first K principal components (PC) to satisfy the explained-variance ratio: $\frac{\sum_{k=1}^K \lambda_k}{\sum_{k=1}^m \lambda_k} \geq \tau$, where τ is the threshold. Then, we construct the projection matrix: $\mathbf{W}_K = [\mathbf{v}_1, \dots, \mathbf{v}_K] \in \mathbb{R}^{m \times K}$. Through projection, the compressed score vector for x_i is calculated as: $\beta^{(i)} = \mathbf{W}_K^T (\alpha^{(i)})^T \in \mathbb{R}^K$. Each element of $\beta^{(i)}$ represents data x_i 's score along an orthogonal dimension, ensuring the score in one dimension does not imply redundancy in another. The sign ambiguity of PCA components is resolved using a deterministic orientation convention, as detailed in Appendix I.

Learning the latent mapping via model-based scorer. To scale this evaluation to the large-scale target dataset $\mathcal{D}_t$, we train RoBERTa-based scorers $r_k(\cdot)$ to approximate the projection on the k -th PC dimension $\Phi_k(E(x))$. This directly solves the optimization objective defined in (1c). The scorer minimizes the regression loss between the k -th projected score and the raw text. All the scorers $r_k(\cdot), k \in \{1, \dots, K\}$ constitute the proxy scorer S_θ . We then label the target dataset $\mathcal{D}_t$ with scorer $r_k(\cdot)$ to obtain the scores $\hat{\beta}_k^{(i)}, k \in 1, \dots, K, i \in \{1, \dots, |\mathcal{D}_t|\}$. The details of the parameter setting and the training of RoBERTa-based scorers are in Appendix E. The overall cost of data selection, including training RoBERTa-based scorers and the annotation process, is in Appendix F.

Constructing dataset with the best data in each dimension. We allocate a data budget B_k to each PC dimension considering its contribution and the total data budget: $B = \sum_{k=1}^K B_k$. For each dimension, a score threshold t_k is set

Algorithm 1 Orthogonal Diversity-Aware Selection

Require: Reference dataset $\mathcal{D}_r$, target dataset $\mathcal{D}_t$, dimensions for evaluation, data budget B

Ensure: Selected dataset $\mathcal{D}_s$

- 1: For each $x_i \in \mathcal{D}_r$, obtain the m dimensional score vector $\alpha^{(i)} = (\alpha_1^{(i)}, \alpha_2^{(i)}, \dots, \alpha_m^{(i)})$;
 - 2: Compute the mean vector μ and construct the centered matrix $\tilde{\mathbf{X}}$ as (6);
 - 3: Compute the covariance matrix Σ , and perform eigen-decomposition to obtain eigenvalues λ_i and eigenvectors $\mathbf{v}_i$;
 - 4: Determine the number of principal components (PCs) K with the threshold τ ;
 - 5: Construct the project matrix $\mathbf{W}_K$, and project scores into the orthogonal space $\beta^{(i)} = \mathbf{W}_K^T (\alpha^{(i)})^T \in \mathbb{R}^K$;
 - 6: Allocate budget B_k to each PC dimension such that $\sum_{k=1}^K B_k = B$;
 - 7: **for** $k = 1, \dots, K$ **do**
 - 8: Train a RoBERTa-based scorer $r_k(\cdot)$ by regressing text $x_i \in \mathcal{D}_r$ to the PCA-transformed scores $\beta_k^{(i)}$;
 - 9: Apply $r_k(\cdot)$ to obtain predicted scores $\{\hat{\beta}_k^{(i)}\}$ for all $x_i \in \mathcal{D}_t$;
 - 10: Given budget B_k , determine threshold t_k , such that $\sum_{x_i \in \mathcal{D}_s^k} l^{\text{tok}}(x_i) = B_k$;
 - 11: Select $\mathcal{D}_s^k = \{x_i \mid \hat{\beta}_k^{(i)} > t_k, x_i \in \mathcal{D}_t\}$;
 - 12: **end for**
 - 13: Construct the final selected dataset $\mathcal{D}_s = \cup_{k=1}^K \mathcal{D}_s^k$.
-

such that the selected data satisfy the budget constraint B_k , and the corresponding subset is denoted as $\mathcal{D}_s^k$. Finally, the dataset is constructed as: $\mathcal{D}_s = \{x_i \in \mathcal{D}_t \mid \exists k \in 1, \dots, K, \hat{\beta}_k^{(i)} > t_k\}$. Alternatively, this can be viewed as the union of top-scored subsets across all dimensions, i.e., $\mathcal{D}_s = \cup_{k=1}^K \mathcal{D}_s^k$. This construction ensures that each dimension contributes the highest-quality data, while the orthogonality of the PC dimensions enhances diversity in the resulting dataset.

I PCA Sign Convention.

Principal components are sign-ambiguous, i.e., both $\mathbf{v}_k$ and $-\mathbf{v}_k$ define the same PCA direction. Since ODiS selects samples according to high PC scores, we resolve this ambiguity using a deterministic convention. For each principal component $\mathbf{v}_k$, we orient its sign such that the original quality

dimension with the largest absolute loading has a positive coefficient:

$$j^* = \arg \max_j |v_{k,j}|, \mathbf{v}_k \leftarrow \text{sign}(v_{k,j^*}) \mathbf{v}_k. \quad (8)$$

Since all original quality dimensions are positively oriented, where larger scores indicate higher quality, this convention gives each PC a consistent quality-oriented direction according to its dominant loading and makes the subsequent selection deterministic.

J Token Budget Allocation

We include a comparison between equal budget allocation and non-uniform budget allocation, where the budgets are set to be proportional to the explained variance and inversely proportional to the explained variance of the PCs (e.g., the ratios between PC1-PC4 are set as 1:2:4:4 and 4:2:1:1), respectively, to show that the gains are not driven by manual weighting. The results are provided in Table 8. The result demonstrates that the gains of ODiS are not driven by delicate manual weighting. In particular, placing more emphasis on the high-variance PCs does not consistently improve performance, whereas allocating relatively more budget to later PCs can be beneficial, likely because these dimensions contribute more complementary diversity. Moreover, the PC dimension can be understood as a feature direction that captures a distinct part of data, and the performance gain comes from combining multiple orthogonal and complementary dimensions, rather than from tuning a particular weighting rule.

Table 8: Comparison of different token budget allocation strategy. The ratio indicates the ratio of token budget allocated to PC1 to PC4 under a total budget of 100B tokens.

Ratio	Arc-C	Arc-E	H-Swag	SciQ	PiQA	Avg
1:1:1:1	41.6	66.9	58.4	85.6	77.4	66.0
4:2:1:1	39.0	63.8	55.4	85.6	74.9	63.7
1:2:4:4	41.0	67.2	58.3	87.3	77.1	66.2

K Diversity Analysis with more metrics

To validate diversity from more angles, we additionally computed n-gram and self-BLEU metrics on the data pools with results summarized in Table 9. The results show that selecting from a larger range or with the proposed ODiS yields higher

Distinct-1 and lower Self-BLEU, consistent with the higher pairwise embedding distance (as shown in Figure 5a). Therefore, the proposed ODiS enhances data diversity and the model’s downstream performance.

Table 9: Quantitative comparison of data diversity across subsets of different scales and the subset selected by ODiS.

Subset	Distinct-1 ↑	Self-BLEU ↓
Top-100B	0.1048	0.1252
Top-300B	0.1069	0.1223
Top-500B	0.1093	0.1160
ODiS	0.1173	0.1114

L Ablation Studies

Robustness to scoring sources. To test whether ODiS depends on a GPT-specific scoring feature, we also annotated the reference dataset (e.g., FineWeb-Edu 460K) with two open-weight scorers, Qwen3-Next-80B-A3B-Instruct and Llama-3.1-70B-Instruct, using the same 11-dimensional metrics. As shown in Table 10, although the raw scores differ across scorers, they remain strongly aligned with GPT scores, with within- ± 1 score agreement of 0.846 and 0.903, respectively. More importantly, the induced PCA spaces are highly stable, since the PC1 cosine similarity is 0.945/0.996, and the mean direction angle over the top-4 PCs is only $9.1^\circ/6.9^\circ$.

Comparison of decorrelation methods. To test whether nonlinear dependence dominates our score space, we compute distance correlation before and after PCA, which captures both linear and nonlinear correlation. As shown in Table 11, PCA reduces the distance correlation from 0.21 to 0.06, indicating that broader dependence is substantially reduced. We further compare PCA with ICA, kernel PCA, and an autoencoder-based decomposition. Although the autoencoder achieves the lowest distance correlation, it introduces a parameterized nonlinear representation and additional hyperparameters, while ICA and kernel PCA provide no clear practical advantage. In contrast, PCA remains deterministic, parameter-free, and directly targets the aggregation-induced redundancy that motivates ODiS. Since our goal is to provide a general framework for orthogonalizing multi-dimensional quality signals rather than exhaustively optimize the

Table 10: Robustness of the scoring space to different LLM scorers on the FineWeb-Edu reference set. Higher cosine similarity and lower direction angle indicate stronger alignment with the GPT-induced PCA space.

Metric	GPT vs. Qwen	GPT vs. Llama
Score Mean Absolute Error ↓	0.76	0.62
Within-±1 Score Agreement ↑	0.846	0.903
PC1 Alignment (cos) ↑	0.945	0.996
Mean Top-4 PC Direction Angle ↓	9.1°	6.9°

decomposition module, we use PCA as a simple and stable instantiation of ODiS.

Table 11: Comparison of different decorrelation methods. Pearson, Spearman, and distance correlation (dCor) measure the residual dependence among the resulting dimensions; lower values indicate weaker dependence.

Method	Pearson ↓	Spearman ↓	dCor ↓
Raw	0.22	0.22	0.21
PCA	0.00	0.02	0.06
ICA	0.00	0.04	0.06
kernelPCA	0.00	0.02	0.08
Autoencoder	0.06	0.03	0.05

Robustness to the reference corpus. We further test whether the PCA-induced quality space is sensitive to the choice of reference corpus. We annotate C4, RefinedWeb, FineWeb, and DCLM-Baseline, using the same 11-dimensional metrics. We then apply PCA on scores from each corpus and compare the resulting quality spaces with the original FineWeb-Edu PCA space. To make the comparison independent of corpus-specific samples, we also applied each corpus-derived PCA basis to the FineWeb-Edu dataset and compared the induced PC scores. As can be observed in Table 12, the resulting spaces are highly aligned. The leading PC differs by only 3.3°–4.0°, the mean absolute cosine ranges from 0.954 to 0.991, and the subspace alignment R^2 ranges from 0.983 to 0.996. Here, subspace alignment R^2 measures the overall overlap between the PC1 to PC4 subspaces with values close to 1 indicating the same space of two PCA bases.

M Benchmarks Selection

To select appropriate downstream benchmarks that can effectively reflect the model performance, we observe the accuracy fluctuation as the trained data increases, with results reported in Figure 8. Arc-C, Arc-E, HellaSwag, PIQA, and SCIQ have obvious

variation as the training progresses, while the rest of the benchmarks have a smaller performance improvement. Since our model and the trained data budget are relatively small, some benchmarks can not obviously reflect the training outcomes of the model. Therefore, we select Arc-C, Arc-E, HellaSwag, PIQA, and SCIQ as our benchmarks.

N Score Distribution across PC Dimensions

Figure 9 demonstrates the score distribution over different PC dimensions on different domains. The domains are pre-divided by the Nemotron-CC dataset. We can observe that different PC dimensions emphasize distinct aspects, and joint selection across dimensions enhances data diversity.

O Metrics and Prompt for each dimension

To evaluate data quality and diversity across multiple dimensions, we design a set of prompts that guide large language models to score documents on 11 distinct dimensions, as described in Appendix C. Each dimension is assessed on a tailored Likert scale ranging from 0 to 3/4/5 points depending on the property being measured, with detailed criteria provided for each score level. The prompts are designed to generate both a quantitative score and a brief qualitative justification, ensuring transparency in the evaluation process. These scores are then aggregated into the score matrix used in our PCA-based data selection framework (see Section 2.3).

```

1 Below is an extract from a web page.
  Evaluate whether the text demonstrates
  high coherence in terms of language
  quality. Please follow the following
  guideline to assess the language
  quality of the given extract on a 4
  likert scale:
2
3 0 Point: Incomprehensible
4 - The text is grammatically chaotic and
  difficult to understand.
```

Table 12: Robustness of the PCA-induced quality space across different reference corpora. All results are measured relative to the PCA basis learned from FineWeb-Edu.

Reference corpus	PC1 angle ↓	Mean $ \cos $ ↑	Subspace alignment ↑
C4	4.0°	0.987	0.996
RefinedWeb	3.3°	0.991	0.996
FineWeb	3.4°	0.982	0.991
DCLM-Baseline	3.3°	0.954	0.983

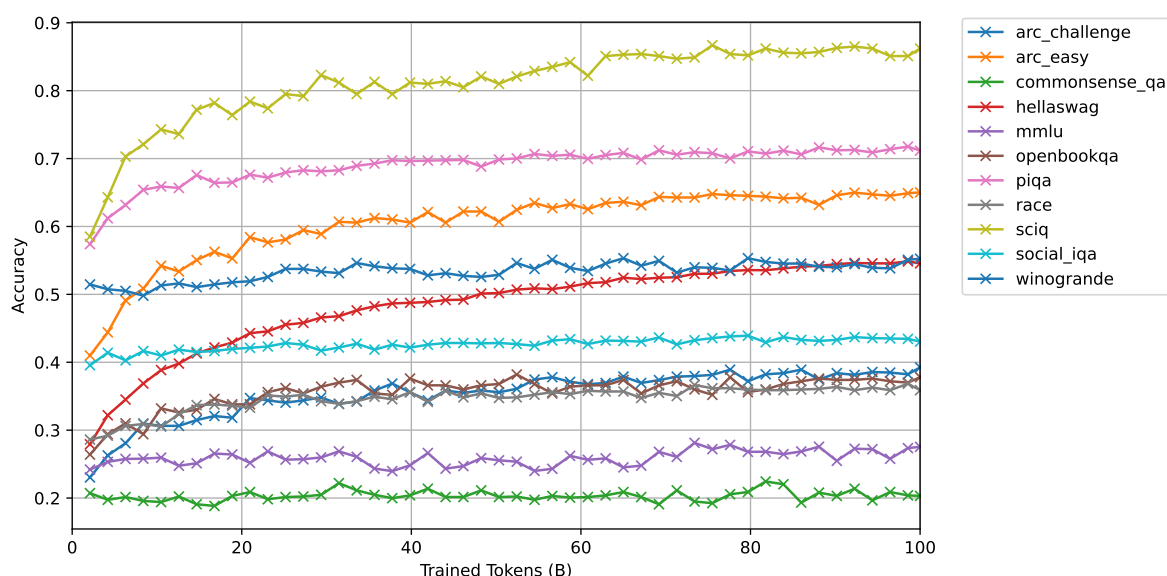


Figure 8: Performance across downstream tasks

5 - Severe errors in structure, agreement, and tense prevent understanding.

6

7 1 Point: Partially Readable

8 - Some sentences are clear, but overall clarity is lacking.

9 - Noticeable grammatical errors and inconsistency disrupt smooth reading.

10

11 2 Points: Moderately Coherent

12 - Occasional language issues but overall understandable.

13 - Logical flow is maintained with some awkward phrasing.

14

15 3 Points: Generally Coherent with Minor Errors

16 - Paragraphs progress logically with minor, infrequent language errors.

17 - Sentences are generally well-formed with consistent tense and clear subject-verb agreement.

18

19 4 Points: Exceptionally Coherent

20 - The text is grammatically flawless, with precise subject-verb agreement and tense usage.

21 - Sentence and paragraph structure is logically ordered and fluid.

22 - Punctuation and syntax enhance the clarity and flow of ideas.

23

24 The extract: {text}

25 After examining the extract:

26 - Briefly justify your total score, up to 50 words.

27 - Conclude begin with the score using the format: "Language Coherence Score: <total points>"

Listing 1: Prompt for Coherence

1 Below is an extract from a web page. Evaluate whether the text demonstrates a high level of conciseness. Please follow the following guideline to assess the conciseness of the given extract on a 4 likert scale:

2

3 0 Point: Excessively Wordy

4 - The extract is filled with redundant, unrelated, or repetitive language.

5 - Nearly every sentence could be significantly shortened or removed without loss of meaning.

6 - Core ideas are obscured or lost in verbosity.

7

8 1 Point: Somewhat Wordy

9 - The text is clear but contains noticeable repetition or unnecessary words.

10 - Some sentences are overly elaborate.

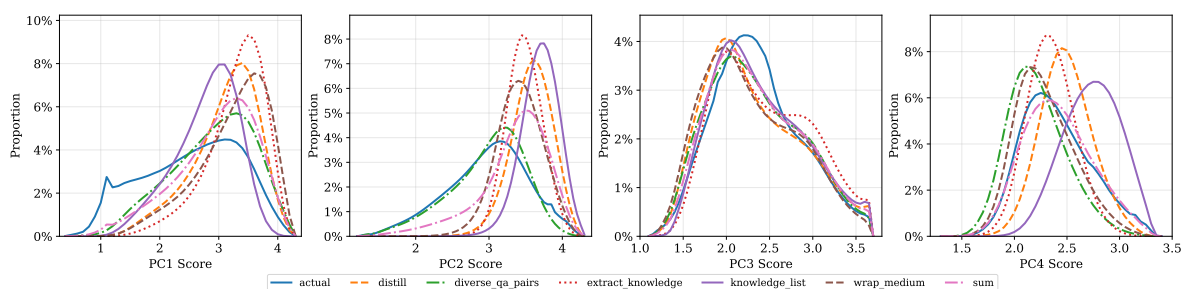


Figure 9: Score distribution over different domains.

```

11 2 Points: Moderately Concise
12 - The extract avoids major redundancy but
13   may include some unnecessary
   elaboration.
14 - Most sentences convey meaning
   efficiently, though small improvements
   in brevity are possible.
15 - The main points are clear and not lost in
   superfluous language.
16
17 3 Points: Concise and Effective
18 - Ideas are expressed clearly and directly,
   with minor redundancy or unnecessary
   details.
19 - Minimal to no repetition or fluff.
20
21 4 Points: Exceptionally Concise
22 - Every word is essential and contributes
   directly to the meaning.
23 - No repetition, filler, or unnecessary
   elaboration.
24 - The writing is focused, impactful, and
   efficient.
25
26
27 - The extract: {text}
28
29 After examining the extract:
30 - Briefly justify your total score, up to
   50 words.
31 - Conclude begin with the score using the
   format: "Language Conciseness Score:
   <total points>"

```

Listing 2: Prompt for Conciseness

```

1 Below is an extract from a web page.
   Evaluate whether the text demonstrates
   high accuracy of word usage, which
   contributes to the as overall language
   quality. Please follow the following
   guideline to assess the accuracy of
   word usage in the given extract on a 4
   likert scale:
2
3 0 Points: Severe Inaccuracy
4 - The extract contains frequent incorrect
   word usages.
5 - Frequent typos, incorrect word forms, or
   misuse of words make the text almost
   unreadable.
6 - Errors severely hinder understanding.
7
8 1 Points: Limited Accuracy

```

```

9 - Spelling mistakes appear regularly but
   are not overwhelming.
10 - Occasional misuse of words or minor typos
   affect clarity.
11 - The overall message is still
   understandable but occasionally unclear.
12
13 2 Points: Moderate Accuracy
14 - Most of the text is correctly spelled,
   with some minor errors or infrequent
   typos.
15 - Occasional confusion between
   similar-sounding words may appear but
   does not significantly affect meaning.
16 - The extract remains mostly readable and
   understandable.
17
18 3 Points: Strong Accuracy
19 - Spelling is generally correct throughout.
20 - Only rare, minor typos or homophone
   errors are present, and they do not
   interfere with comprehension.
21 - The extract demonstrates clear attention
   to written accuracy.
22
23 4 Points: Perfect Accuracy
24 - The extract is free from any spelling
   errors, typos, or homophone confusion.
25 - All words are used appropriately and are
   correctly spelled.
26 - The writing is polished and precise,
   reflecting excellent language control.
27
28 The extract: {text}
29
30 After examining the extract:
31 - Briefly justify your total score, up to
   50 words.
32 - Conclude begin with the score using the
   format: "Language Spelling Accuracy
   Score: <total points>"

```

Listing 3: Prompt for Spelling Accuracy

```

1 Below is an extract from a web page.
   Evaluate whether the text demonstrates
   an appropriate depth of knowledge,
   particularly with regard to the grade
   level it targets. The following
   guideline is used to assess whether a
   text has a high knowledge depth on a 5
   likert scale:
2
3 0 Points: No Knowledge Depth

```

4 - The extract contains no meaningful or accurate knowledge.

5 - It lacks substance entirely and offers no educational value at any grade level.

6

7 1 Point: Shallow and Common Knowledge for Pre-K to Grade 1

8 - The content is understandable even to early primary grades (Pre-K to Grade 1).

9 - Contain simple, basic facts or common knowledge (e.g., basic facts like "grass is green" or "2 + 2 = 4").

10

11 2 Points: Basic Knowledge for Lower Grades (Grades 2-4)

12 - The content is at lower elementary levels.

13 - Introduces simple concepts and provides very short, basic explanations.

14 - Requires understanding of simple definitions and explicit information.

15

16 3 Points: Introductory Knowledge for Middle Grades (Grades 5-7)

17 - Understandable for upper elementary to early middle school.

18 - Explains foundational concepts with some detail and structure.

19 - Some depth is present. It may require understanding of cause-and-effect relationships and ability to follow multi-step explanations.

20

21 4 Points: Substantive Knowledge for Secondary Levels (Grades 8-12)

22 - Content is well-developed and appropriate for high school.

23 - Explores concepts in depth, including underlying principles, reasoning, and potential implications.

24 - Characterized by complex sentence structures, theoretical concepts, evidence or examples to support points; resembles textbook content.

25

26 5 Points: Advanced Knowledge Depth (college-level or graduate-level)

27 - The extract reflects college-level or graduate-level understanding.

28 - The knowledge is usually only known to the professional people in a certain field.

29 - May presents complex information, including detailed analysis, theoretical frameworks, multiple perspectives, and nuanced arguments.

30

31 The extract: {text}

32

33 After examining the extract:

34 - Briefly justify your total score, up to 50 words.

35 - Conclude begin with the score using the format: "Knowledge Depth Score: <total points>"

Listing 4: Prompt for Knowledge Depth

1 Below is an extract from a web page. Evaluate whether the text demonstrates a high degree of knowledge density in

its content. The following curriculum is used to assess whether a text has dense knowledge on a 4 likert scale:

2

3 0 Point: No Meaningful Knowledge

4 - The extract lacks any meaningful or specific content.

5 - No concrete facts, data, or identifiable concepts

6

7 1 Point: Minimal Knowledge Content

8 - Contains only 1-2 disjointed factual statements

9 - No context, sourcing, or explanation

10

11 2 Points: Moderately Knowledge Density

12 - The extract includes several points of useful knowledge.

13 - Support with some details, examples, or explanations.

14

15 3 Points: Substantially Rich in Knowledge

16 - The content provides a well-rounded and informative discussion.

17 - Ideas are explained with clarity and supported by relevant details or examples.

18

19 4 Points: Exceptionally Knowledge-Rich

20 - The extract offers a dense, nuanced, and well-connected presentation of knowledge.

21 - The content shows breadth and depth, encouraging comprehensive understanding.

22

23 The extract: {text}

24

25 After examining the extract:

26 - Briefly justify your total score, up to 50 words.

27 - Conclude begin with the score using the format: "Knowledge Richness Score: <total points>"

Listing 5: Prompt for Knowledge Richness

1 Below is an extract from a web page. Evaluate whether the text demonstrates a high level of reasoning level. The following curriculum is used to assess whether a text has a high reasoning level:

2

3 0 Points: No Reasoning Present

4 - The text lacks any evidence of thinking or reasoning from the writer.

5

6 1 Point: Minimal Reasoning

7 - Some claims are made, but reasoning is largely absent or extremely shallow.

8 - No causal relationships or inferential steps are evident.

9 - Readers are not encouraged to reflect or engage intellectually.

10

11 2 Points: Limited Reasoning

12 The text demonstrates some basic thinking and reasoning, such as:

13 - a straightforward application of a known technique

```

14 - simple analysis of a problem.
15
16 3 Points: Moderate Reasoning
17 The text demonstrates adequate level
18 thinking and reasoning, such as
19 - a consideration of multiple approaches to
20 a problem.
21 - A discussion of the trade-offs between
22 different solutions.
23
24 4 Points: Strong Reasoning
25 The text demonstrates significant thinking
26 and reasoning, such as:
27 - Multi-step reasoning chains to solve a
28 complex problem.
29 - Advanced reasoning patterns often used in
30 specialized science domains
31
32 5 Points: Exceptional Reasoning Quality
33 The text exemplifies exceptional thinking
34 and reasoning, such as:
35 - A highly innovative and creative approach
36 to solving a complex problem in
37 specialized domains.
38 - Combining multiple reasoning and thinking
39 techniques, with novel abstraction of
40 the problem.
41
42 The extract: {text}
43
44 After examining the extract:
45 - Briefly justify your total score, up to
46 50 words.
47 - Conclude begin with the score using the
48 format: "Knowledge Reasoning Score:
49 <total points>"

```

Listing 6: Prompt for Reasoning Level

```

1 Below is an extract from a web page.
2 Evaluate whether the page has a high
3 educational value for teaching from
4 kindergarten to graduate education. The
5 following curriculum is used to assess
6 whether a text has a high educational
7 value on a 3 point scale:
8
9 **0 Point: No Educational Value**
10 - Not even a single bit of information is
11 worth learning.
12 - Note that if there is even a single bit
13 of information that is worth learning,
14 the score should be at least 1 point.
15
16 **1 Point: Minimal Educational Relevance**
17 - The extract provides some useful
18 information pertinent that is worth
19 learning or teaching, but does not
20 align closely with educational
21 standards.
22 - It may include a large amount of
23 non-educational content (e.g.,
24 advertisements, promotional material)
25 that detracts from its usefulness.
26
27 **2 Points: Suitable for Educational Use**
28 - The extract provides a lot of useful
29 information that is worth learning or
30 teaching. The content is fluent and
31 coherent.

```

```

13 - It may include a small amount of
14 non-educational content. It may have
15 limitations, such as incomplete
16 coverage or extraneous information.
17
18 **3 Points: Highly Relevant and Beneficial**
19 - The extract has very high educational
20 value. It contains high density of
21 information that is worth learning or
22 teaching, either for any level of
23 education.
24 - Content is clear, consistent, and
25 focused, with minimal irrelevant
26 information.
27 - May resemble a snippet from a textbook,
28 tutorial, exercises, solutions, or any
29 structured learning materials.
30
31 The extract:
32 {text}
33
34 After examining the extract:
35 - Briefly justify your total score, up to
36 50 words.
37 - Conclude begin with the score using the
38 format: "Educational score: <the
39 assigned score>"

```

Listing 7: Prompt for Educational Value

```

1 Below is an extract from a web page.
2 Evaluate whether the content
3 demonstrates a high degree of practical
4 helpfulness, particularly in terms of
5 offering applicable knowledge for
6 real-world utility. The following
7 curriculum is used to assess whether a
8 text has a high practical helpfulness
9 on a 4 likert scale:
10
11 0 Points: No Practical Helpfulness
12 - The extract contains no useful or
13 applicable knowledge.
14 - May be purely entertainment or
15 advertisement with zero actionable
16 takeaways
17 - May contain misinformation or harmful
18 suggestions
19
20 1 Point: Minor Utility
21 - The text may hint at applicable ideas but
22 lacks clarity, specificity, or guidance.
23 - It is too general or abstract to be put
24 into use.
25
26 2 Points: Moderate Helpfulness
27 - The knowledge can be applicable in some
28 uncommon scenarios (targets <1%
29 audience) that only relate to a small
30 portion of people.
31
32 3 Points: Broadly Helpful
33 - The extract includes practical
34 information that could be applied in
35 common contexts.
36 - Offers validated strategies for common
37 needs
38
39 4 Points: Substantially Helpful

```

```

20 - The extract offers clear, applicable
    knowledge or skills that are useful in
    real-world scenarios that frequently
    occur.
21 - Addresses frequent pain points (>10%
    audience)
22
23 The extract: {text}
24
25 After examining the extract:
26 - Briefly justify your total score, up to
    50 words.
27 - Conclude begin with the score using the
    format: "Knowledge Practical
    Helpfulness Score: <total points>"
28

```

Listing 8: Prompt for Practical Helpfulness

```

1 Here is an extract from a webpage. Please
  evaluate the percentage of the global
  population that is likely to be able to
  comprehend the knowledge text. The
  following scale is used to assess the
  comprehension difficulty, with a
  5-point Likert scale:
2
3 0 Points: No value to understand
4 - The content is incomprehensible due to
  its low language quality.
5 - Contains gibberish, severe grammar
  errors, or formatting problems.
6 - Examples: Advertisement,
  machine-translated nonsense, corrupted
  text
7
8 1 Point: Universal Comprehension
9 - The content is very simple and direct,
  easily understood by the vast majority
  of people.
10 - Requires basic vocabulary (<4th grade
    level), commonsense knowledge, with no
    jargon.
11 - Examples: Weather reports, simple
    recipes, basic safety instructions
12
13 2 Points: Majority Effortless
14 The content is clear and easily
    understandable for almost everyone,
    with only a very small percentage
    finding it difficult.
15 - Requires conversational language level
    and general world knowledge
16 - Examples: social media posts, most new
    articles
17
18 3 Points: Educated Majority
19 - The content is accessible to the majority
    of people, with some difficulty, but
    most people should be able to
    understand and comprehend it after some
    effort.
20 - Requires high school reading level and
    secondary education concepts
21 - Examples: Government pamphlets, workplace
    training manuals, simple financial
    advice.
22
23 4 Points: Specialized Audience

```

```

24 - The content is understood by a small
    portion of people, but it remains
    challenging for the majority.
25 - The content may require some expertise.
26 - Requires undergraduate-level training in
    field
27 - Examples: College textbooks, legal
    contracts, financial advice
28
29 5 Points: Expertise
30 - The content may be very professional or
    academic.
31 - Requires graduate-level expertise.
32 - Examples: Quantum physics proofs, AI
    architecture patents, genomic research
33
34 Extract:
35 {text}
36
37 After reviewing the text:
38 Briefly justify your total score in up to
    50 words.
39 Conclude begin with the score using the
    format: "Comprehension Difficulty
    Score: "

```

Listing 9: Prompt for Comprehension Difficulty

```

1 Here is an extract from a webpage. Evaluate
  whether the content demonstrates a high
  level of factual accuracy as part of
  its overall information quality.
2 Note that:
3 - the text may include some facts that are
  unknown to you. In these cases, you can
  ignore these unknown or uncertain facts
  and only focus on identify those
  obvious factual errors that are known
  to you.
4 - In some special contexts, such as
  fictions, it is allowed to contain some
  imaginary facts.
5
6
7 The following guideline is used to assess
  the factual accuracy, with a 3-point
  Likert scale:
8
9 0 Point: Evidently Inaccurate
10 - The extract is filled with incorrect
    information.
11 - Key claims are demonstrably wrong or
    contradict well-established facts.
12
13 1 Point: Highly Unreliable
14 - The extract contains multiple factual
    inaccuracies or distortions.
15 - Misleading phrasing or vague statements
    obscure the truth.
16 - While not entirely false, it cannot be
    trusted as a reliable source of
    information.
17
18 2 Points: Generally Accurate with Minor
    Issues
19 - <2 minor errors in peripheral details
20 - Occasional imprecise language without
    distorting meaning
21 - Preserves core truth despite technical
    imperfections

```

```

22
23 3 Points: Accurate and Trustworthy
24 - No detectable errors in verifiable claims.
25
26 Extract:
27 {text}
28
29 After reviewing the text:
30 - Briefly justify your total score in up to
31 50 words.
32 - Conclude begin with the score using the
    format: "Information Factual Accuracy
    Score:"

```

Listing 10: Prompt for Factual Accuracy

```

    answers potential reader questions.
28
29 Extract:
30 {text}
31
32 After reviewing the text:
33 Briefly justify your total score in up to
    50 words.
34 Conclude begin with the score using the
    format: "Information Completeness
    Score: "

```

Listing 11: Prompt for Completeness

```

1 Here is an extract from a webpage. Evaluate
  whether the content demonstrates a high
  degree of completeness, specifically in
  terms of how fully the topic is covered
  and whether the information is
  presented with sufficient context. The
  following scale is used to assess the
  information completeness, with a
  4-point Likert scale:
2
3
4 0 Point: Severely Incomplete
5 The extract offers only fragments of
  information or vague references to the
  topic.
6 Key background, definitions, or context are
  missing.
7 The presentation leaves readers with more
  questions than answers.
8
9 1 Point: Limited Completeness
10 The extract touches on parts of the topic
    but leaves significant gaps.
11 It may assume prior knowledge or skip
    necessary context.
12 Information is partial or unevenly
    distributed.
13
14 2 Points: Moderately Complete
15 The extract introduces the main topic and
    provides sufficient context to follow
    the discussion.
16 Some areas may be underdeveloped or
    missing, but overall understanding is
    possible.
17 It resembles a summary or introductory
    passage.
18
19 3 Points: Substantially Complete
20 The extract covers the topic in a
    well-rounded and balanced manner.
21 Most relevant aspects are addressed, with
    clear and sufficient context.
22 There may be minor omissions, but they do
    not disrupt comprehension.
23
24 4 Points: Exceptionally Complete
25 The extract thoroughly explores the topic
    with comprehensive coverage.
26 All necessary context is included, with no
    critical gaps.
27 It reflects a deep and well-structured
    presentation that anticipates and

```