GMDS: Group-Level Multi-Objective Data Selection for Large Language Model Pretraining

Anonymous ACL submission

Abstract

Data selection plays a crucial role in organizing pretraining corpora and has a significant impact on the performance of large language models (LLMs) across downstream tasks. However, existing data selection methods often overlook the potential interference between different knowledge types, leading to improvements in some objectives but performance degradation on others. To address this limitation, we propose a Group-level Multi-objective Data Selection (GMDS) method. This approach extracts semantic features and then applies a clustering algorithm to partition the dataset into different groups. Within each group, samples are quantitatively evaluated along three dimensions: knowledge density, information value, and difficulty. These evaluations are used to establish an initial weight distribution, providing a stable prior for subsequent optimization. Based on this distribution, we employ the NSGA-III algorithm to perform nondominated sorting and iterative search across multiple benchmark tasks, constructing a Pareto front to optimize group-level weights for data selection. Experimental results demonstrate that pretraining data selected by GMDS improves LLM performance across multiple downstream tasks, outperforming existing selection methods and validating its effectiveness in the pretraining stage.

1 Introduction

Pre-training data play a crucial role in determining the capabilities of large language models (LLMs). Prior studies (Brown et al., 2020; Hoffmann et al., 2022; Touvron et al., 2023) show that pre-training on high-quality and diverse corpora can significantly enhance model performance and improve adaptability across multi-task and multi-domain settings. Motivated by this finding, efficiently selecting, deduplicating, and balancing multi-source corpora has become a key focus of current research(Penedo et al., 2024; Li et al., 2024). This trend grows even more pronounced as model scales increase and downstream task complexity rises, making data selection an indispensable component of large-scale pre-training (Wettig et al., 2024).

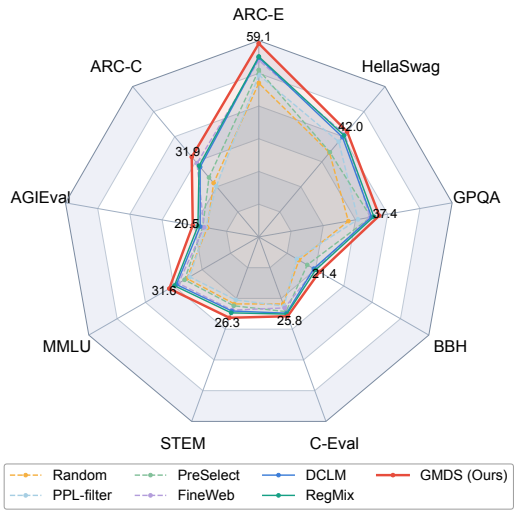

Figure 1: Comparison of different baselines and our method on various benchmarks, demonstrating that models pretrained on data selected by our approach exhibit superior performance on downstream tasks.

Existing data selection methods typically follow two paradigms, consisting of heuristic-based methods and model-driven methods. Heuristic methods (Liu et al., 2024b; Kang et al., 2024; Gu et al., 2025a; Lin et al., 2024) utilize human-designed rules to filter out low-quality noise from large-scale datasets to optimize data quality. Specifically, common manually designed filters include average word length, stop-word ratios, and word repetition rates. In model-driven approaches, trained models are employed to calculate semantic scores for samples to help identify data with high instructional value for downstream tasks. For instance, FineWeb-Edu (Penedo et al., 2024) trained a BERT classifier to prioritize educational content, while DCLM (Li et al., 2024) uses a fastText-based scorer to select

data for supervised fine-tuning. The objective of these methods is to extract high-quality subsets from raw datasets so that LLMs trained on these subsets can achieve similar or even superior performance on downstream tasks compared to those trained on the complete dataset.

However, existing data selection strategies mainly focus on identifying high-quality samples from vast amounts of data, while often overlooking the heterogeneity within high-quality corpora and their differential contributions to the model's various capabilities. These selection approaches fail to adequately recognize the potential conflicts in optimization goals posed by sample groups with different semantic features and complexity levels. For instance, within the same corpus judged to be of high quality, samples focusing on logical reasoning and those emphasizing linguistic expression may compete during parameter updates, leading to a situation where improving a specific skill comes at the cost of performance in other dimensions. This performance trade-off, arising from the lack of fine-grained balance, suggests that simply increasing the data quality cannot automatically resolve knowledge interference in multi-task scenarios.

To address these limitations, we propose a framework termed Group-level Multi-objective Data selection (GMDS). We first partition the dataset into distinct groups using a clustering algorithm based on embedding vectors, thereby mitigating the adverse effects of internal heterogeneity. After the groups are established, we assess the intrinsic value of each group by calculating metrics including knowledge density, information value, and difficulty, which provide structured prior information for the subsequent optimization process. Subsequently, we employ the NSGA-III algorithm to explore the Pareto optimal front across multiple competing objectives, such as mathematical reasoning and linguistic proficiency. This multi-objective search identifies a balanced configuration that enhances overall model performance while preventing the degradation of specific capabilities.

To validate the effectiveness of the GMDS framework, we conduct large-scale pretraining using models of different scales, across two representative heterogeneous datasets, RefinedWeb and AICC. As shown in Figure 1, experiments on nine representative benchmarks spanning general knowledge, domain-specific knowledge, and logical reasoning demonstrate that our approach improves downstream task accuracy while maintaining high training efficiency.

In conclusion, our contributions are as follows:

- We propose a group level data selection framework that partitions corpora into fine grained subsets to reduce knowledge interference in pretraining LLMs.
- We introduce a multi-objective optimization strategy based on the NSGA-III, designed to identify the Pareto front across multiple benchmarks for high-quality data selection.
- Experimental results on multiple downstream benchmarks show that LLMs pre-trained with our selected data outperform baselines, demonstrating the effectiveness of GMDS.

2 Related Work

2.1 Data Selection for LLMs Pretraining

Training data selection is crucial for LLM pretraining (Touvron et al., 2023), and existing methods can be broadly categorized into heuristic and model-driven approaches. Heuristic methods, including text extraction (Barbaresi, 2021), language filtering (Joulin et al., 2017), and deduplication (Tirumala et al., 2023), are efficient but generalize poorly across tasks and domains. Model-driven methods score data with pretrained models at the sample or token level to optimize quality, diversity, and distribution matching(Liu et al., 2024b; Lee et al., 2025; Gu et al., 2025b; Wang et al., 2025; Deng et al., 2026), leveraging reference data(Engstrom et al., 2024; Yu et al., 2024), reference models(Ankner et al., 2024; Wettig et al., 2024), or influence functions(Wang et al., 2024). Group-level methods such as RegMix (Liu et al., 2024a) instead organize data into pools and perform cross-group mixing to mitigate knowledge interference. However, most existing methods operate at only one selection level and do not jointly optimize both group-level and sample-level data selection. In contrast, GMDS integrates group-level data organization with Pareto-based multi-objective scoring, enabling fine-grained and scalable sample selection across diverse tasks.

2.2 Multi-Objective Optimization

Multi-Objective Optimization Problems (MOPs) frequently arise in engineering and computational tasks, where multiple objectives may conflict. Traditional single-objective methods struggle with such trade-offs, motivating the development of Multi-Objective Evolutionary Algorithms

(MOEAs) (Deb, 2015). NSGA-II (Deb et al., 2002) uses non-dominated sorting and crowding distance to maintain diversity while approximating the Pareto front; SPEA-II (Wahid et al., 2015) assigns fitness based on dominance strength and preserves an external archive of solutions. PAES (Knowles and Corne, 2000) relies on a simple archiving strategy combined with mutation to evolve a diverse Pareto set. Decomposition-based methods, such as MOEA/D (Zhang and Li, 2007), split MOPs into scalar subproblems and exploit neighborhood information to reduce computational cost and produce uniformly distributed solutions. NSGA-III (Yuan et al., 2014), a modern extension of NSGA-II, introduces reference points to guide the search toward a well-spread Pareto front, improving scalability and diversity maintenance compared to earlier algorithms. These methods provide a principled foundation for balancing competing objectives in complex tasks. In this work, we adopt NSGA-III to guide multi-objective data allocation across structured groups, providing a scalable and principled approach for balancing informativeness, diversity, and difficulty in LLM pretraining.

3 Methodology

In this section, we propose the GMDS framework designed to identify high-value samples for downstream training tasks. As illustrated in Figure 2, the pipeline consists of three core components. We first obtain dense vector embeddings from raw text data using a pre-trained model and perform clustering to group samples into semantically coherent clusters. Next, an intra-group knowledge evaluation is performed to score each sample across three critical dimensions: knowledge density, information value, and difficulty. This multi-dimensional assessment ensures that the selected data is both diverse and intellectually dense. Finally, we formulate the data selection task as a multi-objective optimization problem and employ NSGA-III to optimize the allocation in different groups.

3.1 Problem Formulation

The goal of data selection is to identify a subset of training samples that maximizes model performance on downstream tasks. Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ denote the full training dataset, where x_i and y_i are the input and corresponding label, respectively. A data selection strategy π selects a subset $S_\pi \subseteq D$, which is then used to train a model f_{θ_π}. The objective is to find an optimal strategy π^* such that the resulting model achieves the best performance on a target test set D': $\pi^* = \arg\max_\pi R(f_{\theta_\pi}; D')$, where θ_π is the model parameters obtained by training on the selected subset S_π, and $R(\cdot; D')$ is a task-specific performance metric evaluated on the test set D'.

3.2 Dataset Grouping Strategies

Treating a dataset as a uniform collection of samples can obscure its internal structure and reduce selection efficiency. Therefore, we propose a grouping strategy that partitions the dataset into several clusters to capture latent semantic patterns. Formally, for each sample $x_i \in \mathcal{D}$, we first extract its feature embedding: $\mathbf{z}_i = \phi(x_i)$, where $\phi(\cdot)$ denotes BGE embedding model (Chen et al., 2024). We then apply a clustering algorithm to partition the set of all embeddings $\{\mathbf{z}_i\}_{i=1}^{N}$ into K disjoint groups:

$$\{G_1, G_2, \ldots, G_K\} = \text{Clustering}(\{\mathbf{z}_i\}_{i=1}^{N}).$$

Each group G_k consists of samples that are close to each other in the feature space. For further analysis, we compute a group-level feature vector $\mathbf{f}_k$ as the mean of the embeddings within each cluster:

$$\mathbf{f}_k = \frac{1}{|G_k|} \sum_{x_i \in G_k} \phi(x_i).$$

3.3 Intra-Group Knowledge Evaluation

Within each group, evaluating the knowledge carried by samples provides a prior distribution for the Pareto-based selection and accelerates convergence in the multi-objective optimization space. We assess each sample along three dimensions, namely knowledge density, information value, and difficulty, motivated by a preliminary feature study detailed in Appendix B.

We first establish an automated scoring mechanism based on LLMs. Specifically, we employ meticulously designed prompts to guide the LLM in quantitatively scoring samples across three dimensions: knowledge density $K(x_i)$, information value $I(x_i)$, and difficulty $Dif(x_i)$. Subsequently, we utilize these text score pairs as supervision data to train a BERT based regression model, effectively distilling the assessment capability into a lightweight and high speed scorer. Using this scorer, each sample x_i is efficiently mapped to a three dimensional feature score vector:

$$\mathbf{s}_i = [K(x_i), I(x_i), Dif(x_i)]^T.$$

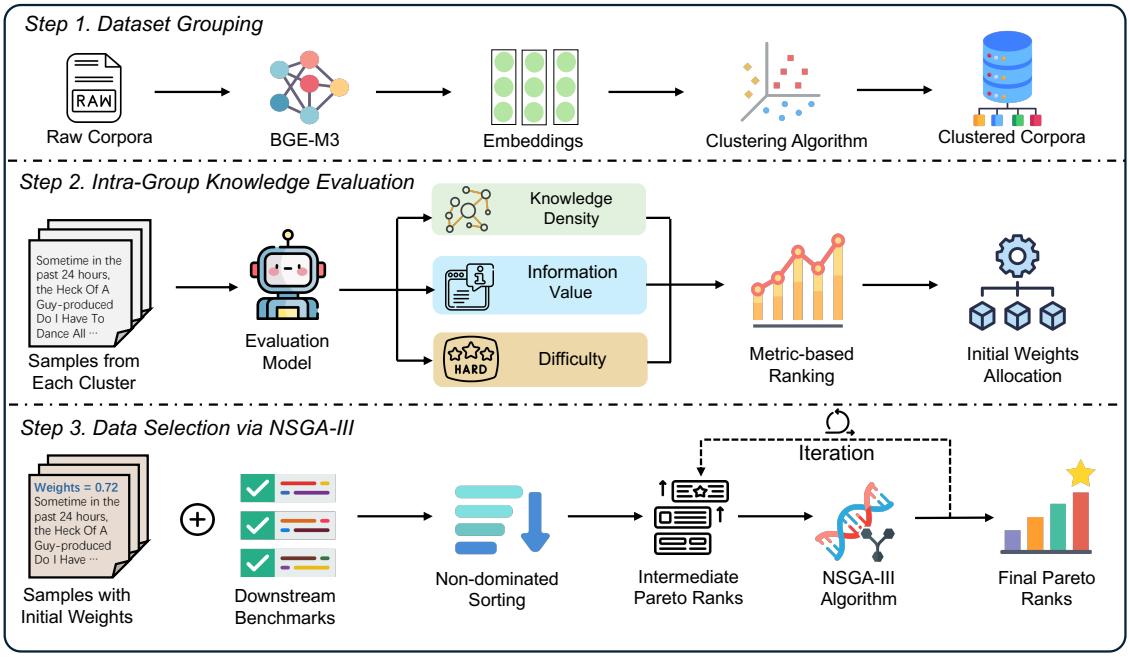

Figure 2: Overview of the GMDS framework, illustrating the data selection process from semantic clustering and multi-dimensional knowledge evaluation to Pareto-based multi-objective optimization.

Once the three dimensional scores are obtained, we combine them into a composite knowledge score S_i using linear hyperparameters λ_k, λ_i, and λ_d:

$$S_i = \lambda_k K(x_i) + \lambda_i I(x_i) + \lambda_d Dif(x_i).$$

To further transform individual scores into optimization signals, we introduce a piecewise function that assigns a corresponding weight v_i to each sample based on the interval in which S_i resides:

$$v_i = \begin{cases} \alpha, & S_i \in [0, \tau_1) \\ \beta, & S_i \in [\tau_1, \tau_2] \\ \gamma, & S_i \in (\tau_2, 1] \end{cases}$$

where τ_1, τ_2 are interval thresholds, and α, β, γ are adjustable weight hyperparameters.

Finally, we calculate the mean weight of all samples within the same cluster to derive the prior proportion $\bar{w}_k$ for that cluster:

$$\bar{w}_k = \frac{1}{|G_k|} \sum_{v_i \in G_k} v_i,$$

which serves as the initial candidate for the subsequent optimization.

3.4 Data Selection via NSGA-III

We optimize the mixing proportions of samples across groups using a multi-objective evolutionary algorithm. Each candidate solution represents a vector of mixing proportions across G groups:

$$x_i = [p_{i,1}, p_{i,2}, \ldots, p_{i,G}], \quad p_{i,j} \geq 0,$$

where $p_{i,j}$ is the proportion of samples selected from group j for individual x_i.

Population Initialization. The initial population $\mathcal{P}^{(0)} = \{x_1, \ldots, x_N\}$ of size N is generated by leveraging the cluster level prior weights $\bar{w}$ derived from the knowledge assessment phase.

Non-dominated Sorting. Each individual $x_i \in \mathcal{P}^{(t)}$ is first evaluated on B benchmark tasks to obtain its fitness vector:

$$f_i = [f_1(x_i), f_2(x_i), \ldots, f_B(x_i)],$$

where $f_b(x_i)$ is the performance of a proxy model trained on samples drawn according to x_i and evaluated on benchmark b. Each objective is then normalized across the population:

$$\tilde{f}_{i,b} = \frac{f_b(x_i) - f_b^{\min}}{f_b^{\max} - f_b^{\min}}, \quad b = 1, \ldots, B.$$

The normalized fitness matrix $\tilde{F} = [\tilde{f}_1; \ldots; \tilde{f}_N]$ is used as input to a non-dominated sorting algorithm:

$$\text{PFs}, r = \text{ND_Sort}(\tilde{F}),$$

where PFs is a set of Pareto fronts and r_i denotes the non-dominated rank of individual x_i.

Iterative Evolution. In each generation, the algorithm produces offspring by iteratively applying crossover and mutation to the selected parents. The evolution proceeds through the following steps:

❶ *Selection.* Individuals are chosen for the mating pool via binary tournament selection based on Pareto rank r_i:

$$mating_pool[i] = \begin{cases} x_{ind1}, & r_{ind1} \leq r_{ind2} \\ x_{ind2}, & r_{ind1} > r_{ind2} \end{cases},$$

where $ind1$ and $ind2$ index two individuals drawn uniformly at random from the current population.

❷ *Crossover.* Parent pairs (x^a, x^b) generate offspring via simulated binary crossover (SBX):

$$x^{\text{child}} = \frac{x^a + x^b}{2} \pm \frac{\beta}{2}(x^a - x^b),$$

with $\beta \in [0, 1]$ sampled according to a distribution controlled by crossover index η_c. Each gene undergoes crossover with probability p_c.

❸ *Mutation.* Polynomial mutation perturbs each gene x_j of the offspring with probability p_m:

$$x'_j = x_j + \delta(u_j - l_j), \quad \delta \sim \text{Poly}(\eta_m),$$

where $[l_j, u_j]$ are the bounds of x_j. The resulting vector is normalized so that $\sum_j x'_j = 1$.

❹ *Group Update.* The parent and offspring populations are combined:

$$\mathcal{P} = \mathcal{P}^{(t)} \cup \mathcal{Q}^{(t)}.$$

Each individual $x_i \in \mathcal{P}$ is assigned a Pareto rank r_i, which is defined as follows:

$$r_i = \begin{cases} 0, & \text{if } D_i = \emptyset, \\ 1 + \min_{x_j \in D_i} r_j, & \text{otherwise}, \end{cases}$$

where $D_i = \{x_j \in \mathcal{P} \mid x_j \succ x_i\}$.

For individuals in the last partially selected front, niche preservation is enforced via reference vectors V (Das and Dennis, 1998):

$$\text{assoc}(x_i) = \arg\min_{v \in V} \left\| \mathbf{f}'(x_i) - \frac{\mathbf{f}'(x_i) \cdot v}{\|v\|^2} v \right\|,$$

selecting individuals associated with the most underrepresented reference directions one at a time until population size is reached.

Pareto Front and Knee-Point Selection. After T generations of iterative evolution, the resulting Pareto front collects every non-dominated solution retained in the final population:

$$PF = \{x_i \in \mathcal{P}^{(T)} \mid r_i = 0\}.$$

The final knee-point solution x^* is selected to balance all objectives:

$$x^* = \arg\max_{x_i \in PF} \frac{d_{\text{anti}}(x_i)}{d_{\text{ideal}}(x_i)}, \tag{1}$$

$$d_{\text{ideal}}(x_i) = \sqrt{\sum_{b=1}^{B} \left(1 - \tilde{f}_b(x_i)\right)^2}, \tag{2}$$

$$d_{\text{anti}}(x_i) = \sqrt{\sum_{b=1}^{B} \left(\tilde{f}_b(x_i)\right)^2}. \tag{3}$$

The selected x^* determines the optimal mixing proportions of samples across groups, effectively balancing multiple benchmark objectives.

4 Experiment

4.1 Experiment Settings

Training Corpora. RefinedWeb (Penedo et al., 2023) is a pretraining dataset consisting purely of web text, entirely based on data crawled from CommonCrawl. It achieves a scale of 5 trillion tokens through MinHash-based fuzzy deduplication and suffix-array-based exact deduplication, and 600 billion tokens have been open-sourced. AICC (Ma et al., 2025) is a multilingual pretraining dataset with a scale of 7.3 trillion tokens, entirely based on data crawled from Common Crawl. Its key feature is the use of the MinerU-HTML parser, built on a 0.6B-parameter language model, which preserves complex structured information in web pages such as code and tables.

Benchmarks. ARC-E and ARC-C (Chollet, 2019) contain 5,197 elementary questions and 2,590 challenging science reasoning questions, respectively. AGIEval (Zhong et al., 2024) consists of approximately 8,062 questions designed to evaluate a model's ability to handle complex real-world tasks. MMLU (Hendrycks et al., 2020) covers 57 subject areas with a total of 15,908 questions, aiming to assess models' depth of multi-task knowledge, with STEM as its core category comprising 18 science and engineering disciplines. C-Eval (Huang et al., 2023) includes 13,948 questions spanning 52 subject areas and serves as a

comprehensive benchmark for evaluating models' multidisciplinary cognitive abilities. BBH (Suzgun et al., 2023) consists of 23 highly challenging tasks with 6,511 evaluation samples, focusing on advanced logical reasoning. GPQA (Rein et al., 2024) contains 448 graduate-level questions. HellaSwag (Zellers et al., 2019) comprises approximately 70,000 samples and is primarily used to evaluate models' understanding of commonsense natural language inference.

Baselines. We compare our method with six representative data selection baselines. **Random** uniformly samples documents from the raw pool. **PPL-filtering** scores each document with a pretrained model and drops those with the highest perplexity. **FineWeb** (Penedo et al., 2024) applies a cascade of heuristic rules to remove low-quality web pages. **PRESELECT** (KaShun et al., 2025) trains a fastText scorer whose target is a model's compression efficiency on each document, and keeps the top-scoring samples. **DCLM** (Li et al., 2024) uses fastText/BERT classifiers trained on instruction-style positives; we apply their released scorers directly to our pool. **RegMix** (Liu et al., 2024a) treats data mixing as a regression problem and searches for the mixture that minimizes the predicted proxy-model loss.

Implementation Details. We use the fine-grained sparse Mixture-of-Experts (MoE) architecture of Ling (Team et al., 2025) in two configurations, MoE-E64 and MoE-E8, with identical expert parameterization across scales. All runs use a global batch size of 2,048, a learning rate of 4.78×10^{-4}, and a single epoch, executed on an NVIDIA H800 GPU cluster with PyTorch (Imambi et al., 2021) and HuggingFace Transformers (Wolf et al., 2019). To prevent contamination, the NSGA-III fitness is computed on a held-out 0.8B proxy suite whose validation splits are disjoint from the test splits used for final reporting. Further details and an end-to-end cost analysis appear in Appendix A and Appendix D.

4.2 Main Results

As shown in Table 1, GMDS achieves the highest average score across every model scale and dataset in our experiment settings.

At the 0.8B parameter scale, GMDS attains an average of 24.51 on RefinedWeb and 26.28 on AICC, exceeding the strongest baselines by 1.10 and 1.28 respectively, and obtains the top score on several individual benchmarks including ARC-E and MMLU. At the 1.4B parameter scale, GMDS further reaches 32.89 on RefinedWeb and 30.00 on AICC, outperforming the strongest baselines by 1.95 and 2.06 respectively, and the gap over RegMix widens consistently as the model grows. Under the same 1.4B RefinedWeb setting, GMDS also attains 20.50 on AGIEval and 26.30 on STEM, both above the strongest baseline, indicating a clear advantage on reasoning and knowledge-intensive tasks. We further report 3B results in Appendix C, where GMDS remains the best performing method on both RefinedWeb and AICC. Gradient-level analysis in Appendix E further verifies that GMDS reduces knowledge interference.

In conclusion, experimental results show that GMDS outperforms existing data selection baselines across different model sizes and data distributions, sustaining higher average scores and achieving the best performance on most benchmarks, and these performance gains further expand as the model size scales from 0.8B to 3B.

5 Analysis

5.1 Ablation Study

In this section, we conduct the ablation study to evaluate the contribution of each GMDS component. The experiments are conducted on a data scale of 100 billion tokens and compare four configurations: full GMDS with semantic clustering, initial weight allocation, and Pareto optimization; removing clustering, using randomly grouped data; without group selection, skipping initial weight allocation and directly applying Pareto optimization; and without Pareto optimization, omitting the NSGA-III algorithm.

The results in Table 2 show that full GMDS achieves the best performance with an average score of 24.51, demonstrating the effectiveness of combining clustering, initial evaluation, and multi-objective optimization. Removing Pareto optimization causes the largest performance drop, reducing the average score by 1.23, indicating that heuristic scoring is insufficient for challenging tasks such as STEM and BBH. In addition, Skipping initial weight allocation decreases the average score by 0.92, confirming that quantitative evaluation of knowledge density and difficulty provides a stable prior distribution and improves the efficiency of subsequent optimization. Furthermore, semantic clustering demonstrates a clear advantage over

Dataset	Method	ARC-E	ARC-C	AGIEval	MMLU	STEM	C-Eval	BBH	GPQA	HellaSwag	Avg.	Gain (Δ)
Results of 1.4B Models												
RefinedWeb	Random	47.50	21.80	16.20	25.80	22.10	22.10	14.50	28.20	33.90	25.79	-
	PPL-filter	50.40	20.50	16.80	24.80	20.80	22.80	13.80	31.40	37.40	26.52	-
	PreSelect	51.20	24.10	17.70	26.20	22.70	24.50	17.40	35.10	33.60	28.06	-
	FineWeb	54.70	30.30	16.30	29.80	24.30	23.40	20.40	35.50	39.80	30.50	-
	DCLM	56.20	28.10	17.60	30.90	24.50	25.60	19.80	35.30	40.50	30.94	-
	RegMix	54.80	28.20	18.70	29.40	24.30	24.80	20.10	35.60	40.30	30.69	-
	GMDS	**59.10**	**31.90**	**20.50**	**31.60**	**26.30**	**25.80**	**21.40**	**37.40**	**42.00**	**32.89**	**+1.95**
AICC	Random	31.00	26.80	17.40	24.50	24.10	23.60	21.40	25.80	36.00	25.62	-
	PPL-filter	30.80	27.20	17.30	24.30	23.10	24.30	22.10	26.40	36.50	25.78	-
	PreSelect	32.40	27.00	18.30	25.20	24.80	24.10	22.10	26.70	37.20	26.42	-
	FineWeb	33.70	29.60	18.50	26.20	25.30	26.30	24.30	27.40	37.70	27.67	-
	DCLM	33.50	30.40	19.20	25.40	25.80	24.80	23.90	28.40	37.10	27.61	-
	RegMix	33.90	30.20	19.00	26.40	25.80	25.80	24.40	28.20	37.80	27.94	-
	GMDS	**35.80**	**31.40**	**22.50**	**28.60**	**28.20**	**27.20**	**26.50**	**29.30**	**40.50**	**30.00**	**+2.06**
Results of 0.8B Models												
RefinedWeb	Random	21.20	18.20	15.10	23.00	16.40	19.50	11.80	19.80	27.50	19.17	-
	PPL-filter	21.50	18.10	16.30	22.00	16.10	24.20	13.80	21.20	29.70	20.32	-
	PreSelect	23.80	18.90	16.80	22.90	16.70	22.80	13.70	24.60	30.80	21.22	-
	FineWeb	24.80	23.50	16.60	24.50	19.30	**26.20**	17.90	25.20	32.70	23.41	-
	DCLM	25.80	**24.40**	16.20	24.10	19.20	22.50	17.50	**26.70**	33.40	23.31	-
	RegMix	25.10	22.50	16.10	24.30	18.80	24.50	17.30	24.40	33.20	22.91	-
	GMDS	**27.50**	23.90	**17.30**	**26.20**	**20.50**	26.00	**19.20**	24.20	**35.80**	**24.51**	**+1.10**
AICC	Random	23.50	21.30	17.20	22.90	19.50	22.20	17.60	21.60	32.50	22.03	-
	PPL-filter	23.30	21.50	17.40	22.00	19.00	22.90	19.20	23.20	33.60	22.46	-
	PreSelect	25.50	21.40	17.80	22.70	19.90	22.60	19.30	24.60	34.80	23.18	-
	FineWeb	27.60	24.90	17.90	24.10	22.00	24.90	21.50	25.00	35.40	24.81	-
	DCLM	28.10	**25.80**	18.10	23.90	22.10	24.10	21.30	**25.70**	35.90	25.00	-
	RegMix	27.90	24.80	17.90	24.40	22.50	24.40	21.70	25.10	35.80	24.94	-
	GMDS	**29.00**	25.40	**19.20**	**26.80**	**23.80**	**26.10**	**23.40**	25.50	**37.30**	**26.28**	**+1.28**

Table 1: We conduct experiments on two datasets, RefinedWeb and AICC, using models with 1.4B and 0.8B parameters. **Bold** indicates the best performance on each benchmark, while underlined values denote the second-best results. The rows highlighted in blue correspond to our method (GMDS).

Method	ARC-E	ARC-C	AGIEval	MMLU	C-Eval	STEM	BBH	GPQA	Hella	Avg.
GMDS (Ours)	27.50	23.90	17.30	26.20	26.00	20.50	19.20	24.20	35.80	**24.51**
- w/o Clustering	27.07 $_{-0.43}$	23.51 $_{-0.39}$	16.94 $_{-0.36}$	25.76 $_{-0.44}$	25.56 $_{-0.44}$	20.03 $_{-0.47}$	18.80 $_{-0.40}$	23.87 $_{-0.33}$	35.35 $_{-0.45}$	24.10 $_{-0.41}$
- w/o Group Selection	26.53 $_{-0.97}$	23.03 $_{-0.87}$	16.50 $_{-0.80}$	25.21 $_{-0.99}$	25.01 $_{-0.99}$	19.45 $_{-1.05}$	18.30 $_{-0.90}$	23.45 $_{-0.75}$	34.79 $_{-1.01}$	23.59 $_{-0.92}$
- w/o Pareto Optimization	26.21 $_{-1.29}$	22.74 $_{-1.16}$	16.23 $_{-1.07}$	24.88 $_{-1.32}$	24.68 $_{-1.32}$	19.09 $_{-1.41}$	18.00 $_{-1.20}$	23.21 $_{-0.99}$	34.46 $_{-1.34}$	23.28 $_{-1.23}$

Table 2: Ablation study on our GMDS method. We pre-train a 0.8B model on the RefinedWeb dataset. Subscripts in black indicate the performance loss relative to the column-wise maximum.

random grouping. This validates that clustering similar data together provides a necessary foundation for the subsequent multi-dimensional scoring and Pareto optimization processes.

5.2 Scaling Law of Data Size

To investigate the scaling law of data size, we analyze the performance of the GMDS framework under varying training data scales. All experiments adopt a unified 1.4B-parameter model architecture and use the RefinedWeb dataset as the raw corpus pool, with FineWeb and DCLM serving as baselines. Model training spans data scales ranging from 0 to 100B tokens.

As shown in Figure 3, GMDS achieves the best performance and reaches the highest average score of 32.89 at 100B tokens, outperforming both DCLM and FineWeb. Moreover, we observe that GMDS attains the performance level of baseline methods trained on 100B tokens using only 60B tokens, reducing pretraining computational cost by approximately 40%. This result further demonstrates the effectiveness of the proposed data selection method.

For training stability, GMDS exhibits a steady performance improvement as the training data scale increases. In contrast, baselines may suffer from performance degradation when the data scale continues to expand. These results indicate that GMDS provides a more stable data selection strategy.

5.3 Distribution of Selected Data

To assess the quality of data selection, we use RefinedWeb as the raw training corpus and adopt

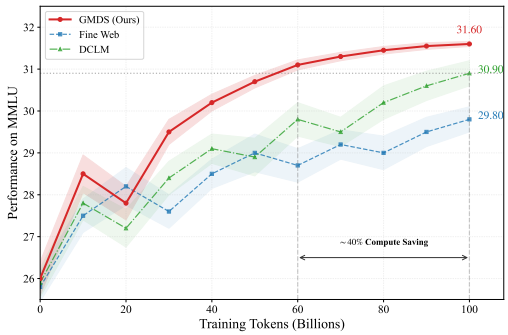

Figure 3: Comparison of scaling performance on data size between GMDS and baseline models.

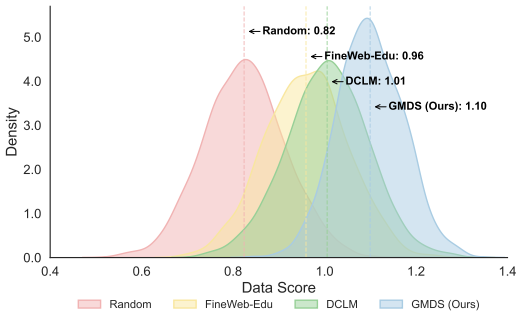

Figure 4: Distribution of data quality scores across different data selection methods.

Random, FineWeb-Edu, and DCLM as baseline methods. From the subset selected by each method, we randomly sample 10,000 data points for analysis. The x-axis (Data Score) denotes the quality score, which is computed by jointly evaluating three dimensions: knowledge density, information value, and difficulty. The y-axis (Density) represents the probability density, illustrating the distribution frequency of selected data across different score ranges.

As shown in Figure 4, the distribution curve of GMDS (Ours) is shifted furthest to the right, achieving the highest average score of 1.10 among all methods. This is followed by DCLM with an average score of 1.01 and FineWeb-Edu with 0.96, while the Random baseline attains the lowest average score of 0.82. These results indicate that GMDS selects a substantially larger proportion of high-scoring data compared to other baselines, with its density peak closest to the high-quality region, demonstrating its superiority in identifying and extracting high-value data points.

5.4 Impact of NSGA-III Iterations

To investigate the impact of different iteration numbers in the NSGA-III algorithm on model performance, we pretrain LLMs using datasets selected by NSGA-III under varying iteration settings. In this experiment, iteration 0 corresponds to data selected based on an initial weight distribution constructed from three dimensions: knowledge density, information value, and difficulty. We then evaluate the resulting pretrained models on three benchmark tasks: ARC-E, MMLU, and STEM.

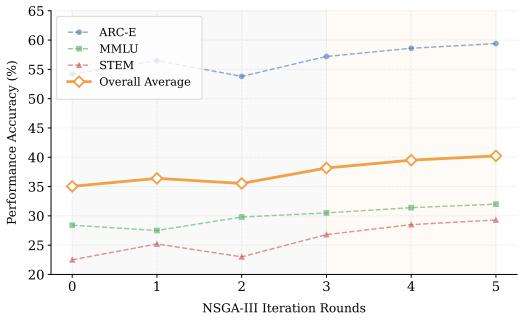

Figure 5: Model performance on downstream benchmarks across NSGA-III iteration rounds.

As shown in Figure 5, as the number of iterations increases, model performance across multiple benchmarks exhibits noticeable fluctuations in the early stages, followed by a stable upward trend. Taking the overall average performance (Overall Average) as an example, the Avg. metric steadily improves after the initial fluctuations, increasing from 38.3% at iteration 3 and reaching a peak performance of 40.2% at iteration 5. Compared with the initial weight-based selection strategy, this result represents an improvement, showing the effectiveness of the iterative process in NSGA-III.

6 Conclusion

In this work, we propose GMDS, a group-level multi-objective data selection method, which aims to alleviate knowledge interference caused by data heterogeneity in the pretraining of LLMs. GMDS first performs fine-grained clustering of the corpus based on semantic representations, then derives prior proportions from a multi-dimensional intra-group knowledge evaluation, and finally refines these proportions through the NSGA-III algorithm to obtain high-quality training data. Experimental results show that models pretrained on data selected by GMDS outperform existing data selection methods across multiple downstream tasks.

In future work, we will extend this method to larger models and broader data domains, including multilingual corpora.

Limitations

The intra-group knowledge scores in GMDS are produced by teacher LLMs and then distilled into a BERT model, so the BERT model has the same preferences with the teacher. In areas where the teacher itself is weak, the predicted scores tend to be lower than the true values of the samples. To remove bias, we combine the scores of multiple teacher LLMs from different model families after adjusting their outputs.

Our experiments focus on RefinedWeb and AICC, both of which are web-based and mostly English, and the largest model trained from start to finish has 1.4B parameters, with a 3B model reported in the appendix. Due to limited computing resources, we have not tested GMDS at the 7B scale or larger, nor on non-English data.

References

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. 2024. Perplexed by perplexity: Perplexity-based pruning with small reference models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

François Chollet. 2019. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*.

Indraneel Das and John E Dennis. 1998. Normal-boundary intersection: A new method for generating the pareto surface in nonlinear multicriteria optimization problems. *SIAM journal on optimization*, 8(3):631–657.

Kalyanmoy Deb. 2015. Multi-objective evolutionary algorithms. In *Springer handbook of computational intelligence*, pages 995–1015. Springer.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE transactions on evolutionary computation*, 6(2):182–197.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. 2026. Less is more: Improving llm alignment via preference data selection. *Advances in Neural Information Processing Systems*, 38:161259–161285.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. 2024. Dsdm: Model-aware dataset selection with datamodels. In *International Conference on Machine Learning*, pages 12491–12526. PMLR.

Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. 2025a. Data selection via optimal control for language models. In *The Thirteenth International Conference on Learning Representations*.

Yuxian Gu, Li Dong, Hongning Wang, Yaru Hao, Qingxiu Dong, Furu Wei, and Minlie Huang. 2025b. Data selection via optimal control for language models. In *International Conference on Learning Representations*, volume 2025, pages 61932–61959.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, and 1 others. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36:62991–63010.

Sagar Imambi, Kolla Bhanu Prakash, and GR Kanagachidambaresan. 2021. Pytorch. In *Programming with TensorFlow: solution for edge computing applications*, pages 87–104. Springer.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomáš Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers*, pages 427–431.

Feiyang Kang, Hoang Anh Just, Yifan Sun, Himanshu Jahagirdar, Yuanzhi Zhang, Rongxing Du, Anit Kumar Sahu, and Ruoxi Jia. 2024. Get more for less: Principled data selection for warming up fine-tuning

in llms. In *The Twelfth International Conference on Learning Representations*.

SHUM KaShun, Yuzhen Huang, Hongjian Zou, YiXuan Liao, Xiaoxin Chen, Qian Liu, Junxian He, and 1 others. 2025. Predictive data selection: The data that predicts is the data that teaches. In *Forty-second International Conference on Machine Learning*.

Joshua D Knowles and David W Corne. 2000. Approximating the nondominated front using the pareto archived evolution strategy. *Evolutionary computation*, 8(2):149–172.

JoonHo Lee, JuYoun Son, Juree Seok, Wooseok Jang, and Yeong-Dae Kwon. 2025. Preference consistency matters: Enhancing preference learning in language models with automated self-curation of training corpora. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12150–12169.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, and 1 others. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, Weizhu Chen, and 1 others. 2024. Not all tokens are what you need for pretraining. *Advances in Neural Information Processing Systems*, 37:29029–29063.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2024a. Regmix: Data mixture as regression for language model pre-training. In *The Thirteenth International Conference on Learning Representations*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2024b. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Ren Ma, Jiantao Qiu, Chao Xu, Pei Chu, Kaiwen Liu, Pengli Ren, Yuan Qu, Jiahui Peng, Linfeng Hou, Mengjie Liu, and 1 others. 2025. Aicc: Parse html finer, make models better–a 7.3 t ai-ready corpus built by a model-based html parser. *arXiv preprint arXiv:2511.16397*.

Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, and Thomas Wolf. 2024. Fineweb: decanting the web for the finest text data at scale. *HuggingFace. Accessed: Jul*, 12.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Ling Team, Ang Li, Ben Liu, Binbin Hu, Bing Li, Bingwei Zeng, Borui Ye, Caizhi Tang, Changxin Tian, Chao Huang, and 1 others. 2025. Every activation boosted: Scaling general reasoner to 1 trillion open language foundation. *arXiv preprint arXiv:2510.22115*.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36:53983–53995.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Abdul Wahid, Xiaoying Gao, and Peter Andreae. 2015. Multi-objective clustering ensemble for high-dimensional data based on strength pareto evolutionary algorithm (spea-ii). In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE.

Jingtan Wang, Xiaoqiang Lin, Rui Qiao, Chuan-Sheng Foo, and Bryan Kian Hsiang Low. 2024. Helpful or harmful data? fine-tuning-free shapley attribution for explaining language model predictions. In *Proceedings of the 41st International Conference on Machine Learning*, pages 50960–50991.

Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, and 1 others. 2025. Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data. *arXiv preprint arXiv:2505.05427*.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. In *Forty-first International Conference on Machine Learning*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Zichun Yu, Spandan Das, and Chenyan Xiong. 2024. Mates: Model-aware data selection for efficient pretraining with data influence models. *Advances in Neural Information Processing Systems*, 37:108735–108759.

Yuan Yuan, Hua Xu, and Bo Wang. 2014. An improved nsga-iii procedure for evolutionary many-objective optimization. In *Proceedings of the 2014 annual conference on genetic and evolutionary computation*, pages 661–668.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Qingfu Zhang and Hui Li. 2007. Moea/d: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation*, 11(6):712–731.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

A Implementation Details

The implementation of the GMDS framework begins with the dataset grouping stage, where we utilize the BGE-M3 embedding model to transform raw text into dense 1024-dimensional vectors. For our experiments, we set the number of disjoint clusters K to 200 to capture a diverse range of latent semantic patterns while maintaining computational efficiency. During the intra-group knowledge evaluation phase, we employ a BERT-base regression model distilled from 50,000 samples scored by LLMs, including GPT-5.4 and Gemini-3.5-Flash. The composite knowledge score S_i for each sample is calculated using linear hyperparameters $\lambda_k = 0.4$, $\lambda_i = 0.3$, and $\lambda_d = 0.3$, and the weight mapping thresholds are defined at $\tau_1 = 0.3$ and $\tau_2 = 0.7$ with assigned weights of $\alpha = 0.1$, $\beta = 0.5$, and $\gamma = 1.0$.

For the multi-objective optimization conducted via the NSGA-III algorithm, we initialize a population of size $N = 100$ and evolve the mixing proportions over $T = 10$ generations. The evolutionary process employs Simulated Binary Crossover (SBX) with a crossover probability $p_c = 0.9$ and a distribution index $\eta_c = 20$. Polynomial mutation is applied with a probability $p_m = 1/G$, where G represents the number of groups, and a mutation index $\eta_m = 20$. All model scales, including the 0.8B and 1.4B parameter variants, are trained with a global batch size of 2,048 and a uniform learning rate of 4.78×10^{-4} for a single epoch. The training is executed on an NVIDIA H800 GPU cluster using the AdamW optimizer with a cosine decay scheduler and a warmup period of 2,000 steps to ensure training stability.

B Justification for the Three Evaluation Dimensions

The three evaluation dimensions used in GMDS, namely knowledge density, information value, and difficulty, are not chosen arbitrarily. They are designed to characterize each data sample from three complementary aspects, corresponding to content quality, information richness, and learning complexity. To verify that these three aspects provide sufficient and non-redundant signals, we conducted a preliminary correlation study over a broader set of candidate features.

Candidate Features Considered. In addition to the three adopted dimensions, we examined two intuitively plausible features as alternatives, namely sentiment polarity and text length.

- **Sentiment Polarity.** A pre-trained sentiment classifier was used to assign polarity scores ranging from -1.0 to 1.0 to a random subset of pretraining data. Sentiment captures affective tone but does not reflect factual or reasoning content.
- **Text Length.** Token count of each sample was used as a simple proxy for information amount, under the intuition that longer texts may carry more knowledge.

Correlation Analysis. For each candidate feature, we trained 0.8B proxy models on subsets selected by that single feature and measured the Pearson correlation between the feature score and downstream reasoning performance on representative benchmarks such as GSM8K accuracy. As reported in Table 3, sentiment polarity yields a near-zero correlation of 0.04, indicating that affective tone is uninformative for pretraining utility. Text length attains a weak correlation of 0.12. Longer documents tend to carry more raw content, but they also introduce greater redundancy and irrelevant noise, which lowers effective knowledge density. In contrast, the three adopted dimensions all exhibit substantially higher correlations with downstream performance, supporting their role as primary objectives.

Status	Feature	r
Discarded	Sentiment Polarity	0.04
	Text Length	0.12
Adopted	Knowledge Density	0.47
	Information Value	0.41
	Difficulty	0.38

Table 3: Pearson correlation r between candidate sample-level features and downstream reasoning performance. The three adopted dimensions clearly dominate the discarded alternatives.

Complementarity of the Three Dimensions. The three adopted dimensions are designed to be complementary rather than redundant. Knowledge density measures how much factual content is packed into a sample, information value reflects the utility and novelty of the content beyond surface form, and difficulty captures the cognitive complexity required to learn from the sample. A high-quality pretraining sample need not maximize all three at the same time. For example, a concise defi-

nition of a scientific concept can be high in knowledge density yet moderate in difficulty, whereas a multi-step reasoning derivation can be moderate in density yet high in difficulty. Treating these aspects as separate objectives allows NSGA-III to explore trade-offs along the Pareto front and select group-level mixtures that are balanced across capability types, rather than collapsing them into a single scalar score that would obscure such trade-offs.

Config	λ_k	λ_i	λ_d	**Avg.**
Random	–	–	–	19.68
High K	0.8	0.1	0.1	23.45
High I	0.1	0.8	0.1	23.10
High D	0.1	0.1	0.8	22.85
GMDS (Ours)	0.4	0.3	0.3	**24.51**

Table 4: Sensitivity of the composite score to the dimension weights $\lambda_k, \lambda_i, \lambda_d$. Single-dimension settings underperform the balanced configuration used in GMDS.

Empirical Sensitivity. To further confirm that the three dimensions jointly contribute to the final performance, we vary the linear weights λ_k, λ_i, and λ_d used in the composite score. As reported in Table 4, single-dimension configurations, in which one weight is set to 0.8 and the others to 0.1, all underperform the balanced setting adopted in GMDS. Knowledge density is the most influential individual dimension, which aligns with our choice of $\lambda_k = 0.4$ in the main configuration. These observations indicate that the three adopted dimensions are both necessary and sufficient for the data selection objective considered in this work.

C Results at 3B Model Scale

We additionally pretrain 3B parameter models on RefinedWeb and AICC, and Table 5 reports the experimental results. As shown in the experiments, our proposed GMDS remains the best-performing method at the 3B model scale.

D Analysis of Computational Cost

In this section, we discuss the computational cost of GMDS.

Embedding Extraction. We use BGE-M3 to encode each document into a 1024-dimensional vector. This is a one-time forward pass and accounts for less than 0.5% of total pretraining FLOPs on the same hardware.

Clustering. We set $K = 200$ clusters and accelerate the clustering process with FAISS. Clustering billion-scale embeddings finishes within hours on a single CPU node, which is negligible compared with the GPU time required for pretraining.

Intra-Group Scoring. We distill a lightweight BERT-base regression model from $50{,}000$ LLM-scored samples. The distilled scorer is orders of magnitude cheaper to run than the labeling LLM, so the scoring pass over the full corpus remains a small fraction of the pretraining cost.

NSGA-III Search. The mixing proportion search relies on a small proxy model and feedback from a held-out validation set, rather than full-scale training runs. The cost is fixed, independent of the final pretraining scale, and dominated by the proxy evaluations.

Net Effect. The overhead of all four GMDS stages combined remains a small fraction of pretraining FLOPs. More importantly, GMDS reaches the average performance of the strongest baseline at 100B tokens using only 60B tokens, which translates to roughly 40% savings in pretraining FLOPs. The selection overhead is therefore amortized many times over by the reduction in training compute, yielding a favorable cost-benefit trade-off in practice.

E Analysis of Knowledge Conflict Alleviation

To directly observe the interference phenomenon, we calculate the gradient cosine similarity $\cos\theta$ between mini-batches randomly sampled from different semantic clusters (generated via BGE-M3 clustering) during the pretraining of the 1.4B RefinedWeb model. Table 6 reports the average similarity and the gradient conflict ratio (fraction of pairs with $\cos\theta < 0$) under three pairing types.

The empirical analysis shows that over 60% of gradient directions between heterogeneous clusters are negatively correlated, confirming that knowledge interference is a primary driver of performance degradation. The GMDS-selected subset preserves positive average alignment and reduces the conflict ratio to 18%, providing direct training-dynamics evidence that the group-level mixture mitigates interference rather than only improving benchmark averages.

Dataset	Method	ARC-E	ARC-C	AGIEval	MMLU	STEM	C-Eval	BBH	GPQA	HellaSwag	Avg.	Gain (Δ)
RefinedWeb	Random	52.83	24.61	17.82	28.57	24.13	25.64	17.05	32.21	41.38	29.36	-
	PPL-filter	55.42	23.79	18.44	27.63	23.21	26.28	17.51	35.18	46.42	30.43	-
	PreSelect	57.36	28.29	19.63	30.42	25.87	27.82	21.41	38.92	41.53	32.36	-
	FineWeb	60.81	33.72	21.34	34.07	28.46	28.23	24.55	40.11	48.07	35.48	-
	DCLM	61.27	33.94	21.43	34.71	28.92	28.51	24.83	40.06	48.52	35.80	-
	RegMix	62.05	33.81	22.27	35.34	29.41	28.97	24.79	39.82	49.37	36.20	-
	GMDS	**65.47**	**38.42**	**24.81**	**38.23**	**32.06**	**30.88**	**27.39**	**43.54**	**52.61**	**39.27**	**+3.07**
AICC	Random	35.42	31.18	20.13	27.06	26.34	26.11	23.82	28.91	39.77	28.75	-
	PPL-filter	35.21	31.63	20.42	27.18	26.07	26.83	24.47	29.32	40.21	29.04	-
	PreSelect	36.78	32.41	21.57	28.82	27.93	27.21	25.36	30.43	41.37	30.21	-
	FineWeb	39.23	34.86	22.91	30.57	29.78	30.04	28.13	32.47	43.62	32.40	-
	DCLM	39.41	36.12	23.42	30.91	30.27	31.13	28.41	33.71	44.02	33.04	-
	RegMix	39.06	35.47	24.08	31.82	30.83	30.96	28.18	33.59	44.71	33.19	-
	GMDS	**42.53**	**38.47**	**27.21**	**35.32**	**34.81**	**33.48**	**32.74**	**36.13**	**48.19**	**36.54**	**+3.35**

Table 5: Additional results at the 3B model scale.

Table 6: Gradient similarity statistics across sample pairing types on the 1.4B RefinedWeb model.

Sample Pairing Type	Avg. $\cos\theta$	Conflict Ratio ($\cos\theta < 0$)
Intra-group	$+0.45$	8%
Inter-group (Heterogeneous)	-0.18	62%
GMDS Selected Subset	$+0.31$	18%

F Sensitivity Analysis of Hyperparameter

K controls the granularity of the grouping stage. A small K merges heterogeneous documents into the same group and reintroduces intra-group gradient conflict, while an excessively large K produces sparse groups whose intra-group scores become unreliable. As reported in Table 7, performance is stable in the range $K \in [100, 300]$ and peaks at the default value $K = 200$ used throughout the paper.

K	50	100	200	300	500
Avg.	23.42	24.18	**24.51**	24.30	23.65

Table 7: Sensitivity to the number of clusters K on 1.4B RefinedWeb.

G Prompts

To ensure the reproducibility of our data selection and mixing process, we provide the specific prompt templates used in our experiments. Specifically, we leverage LLM-as-a-judge to perform intra-group knowledge evaluation: Figure 6 presents the prompt for evaluating the Information Value, focusing on the utility and novelty of the content; Figure 7 details the criteria for Knowledge Density, identifying how much factual information is packed into a given sample; and Figure 8 shows the prompt for assessing Difficulty, which helps the model differentiate between foundational and complex reasoning data. Furthermore, Figure 9 illustrates the prompt utilized for the FineWeb-based quality filtering baseline.

Prompt for Information Value Evaluation

Please rate the information value of the following text on a 5-point scale (1-Very Low, 5-Very High). Consider the following dimensions:
Coverage: Number of topics/subjects covered, diversity of perspectives
Depth: Level of detail on individual topics, presence of in-depth analysis
Density: Amount of meaningful information per unit of text length
Value: Degree of novel insights, new knowledge, or practical guidance provided
Scoring Criteria:
1: Vacuous content with minimal substantive information (e.g., simple greetings, repetitive formulas)
2: Limited information, only basic facts or superficial descriptions
3: Moderate information with clear topic and basic development, but lacking depth
4: Rich information with multi-dimensional development and some depth of analysis
5: Very rich information covering multiple dimensions with deep exploration and high-value insights
Output Format:
Information Density Score: [1-5]
Detailed Rationale: [Bullet-point explanation with at least 3 points]
Improvement Suggestions: [If applicable, specific directions to enhance information density]

Figure 6: Prompt for Assessing Information Value.

Prompt for Knowledge Density

Please evaluate the knowledge density of the following text based on two dimensions: "Novelty of Knowledge" and "Knowledge Concentration."
Evaluation Dimensions:
Novelty of Knowledge (Proportion of new knowledge):
Brand new knowledge: First appearance in the field or latest research findings
New perspective knowledge: Novel interpretations or applications of known knowledge
Integrated knowledge: Innovative combinations of multiple known knowledge pieces
Conventional knowledge: Common sense or basic knowledge in the field
Knowledge Concentration (Amount of knowledge per unit text):
Core concept density: Number of important concepts per 100 words
Relational insight density: Revelation of deep relationships between concepts (causal, comparative, etc.)
Methodological density: Detailed presentation of specific methods, techniques, or procedures
Comprehensive Scoring:
Low knowledge density: Mainly conventional knowledge, low concept density, lacking deep relationships
Medium knowledge density: Some new knowledge or perspectives, moderate concept density
High knowledge density: Considerable new knowledge or innovative integration, high concept density, profound relational insights
Very high knowledge density: Dense frontier knowledge, detailed methodology, breakthrough insights
Output Format:
Knowledge Density Rating: [Low/Medium/High/Very High]
Novelty Analysis: [Proportion of new vs. conventional knowledge with specific examples]
Concentration Analysis: [Performance in concept density, relational insights, methodology]
Knowledge Value Assessment: [Potential value in knowledge dissemination/innovation]

Figure 7: Prompt for Assessing Knowledge Density in Text Samples.

Prompt for Difficulty Evaluation

Please rate the reading difficulty of the following text on a 7-point scale (1-Very Easy, 7-Very Difficult).
Consider these dimensions:
Linguistic Complexity:
Lexical difficulty: Proportion of technical terms, academic vocabulary, rare words
Syntactic difficulty: Sentence length, nested structures, grammatical complexity
Rhetorical difficulty: Use of metaphors, symbolism, complex rhetorical devices
Cognitive Load:
Concept abstraction: Ratio of concrete to abstract concepts
Logical complexity: Length of argument chains, rigor of reasoning
Prerequisite knowledge: Depth and breadth of background knowledge required
Structural Complexity:
Text organization: Linear narrative vs. multi-threaded structure
Information organization: Chronological, logical progression, comparative analysis, etc.
Scoring Criteria:
1-2: Suitable for elementary students, simple vocabulary, short sentences, concrete concepts
3-4: Suitable for middle school students/general adults, everyday vocabulary, straightforward logic
5-6: Suitable for college students/professionals, some specialized vocabulary, moderately complex logic
7: Suitable for domain experts, highly specialized, abstract concepts, rigorous and complex logic
Output Format:
Difficulty Score: [1-7]
Target Audience: [Detailed description of suitable reader groups]
Core Challenges: [List 2-3 main difficulty points]
Simplification Suggestions: [If applicable, specific suggestions to reduce difficulty]

Figure 8: Prompt for Assessing Sample Difficulty.

Prompt for FineWeb

Below is an extract from a web page.
Please evaluate whether the page has a high educational value and could be useful in an educational setting for teaching from primary school to grade school levels using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction of each criterion.
Add 1 point if the extract provides some basic information relevant to educational topics, even if it includes some irrelevant or non-academic content like advertisements and promotional material. Add another point if the extract addresses certain elements pertinent to education but does not align closely with educational standards. It might mix educational content with non-educational material, offering a superficial overview of potentially useful topics, or presenting information in a disorganized manner and incoherent writing style.
Award a third point if the extract is appropriate for educational use and introduces key concepts relevant to school curricula. It is coherent though it may not be comprehensive or could include some extraneous information. It may resemble an introductory section of a textbook or a basic tutorial that is suitable for learning but has notable limitations like treating concepts that are too complex for grade school students.
Grant a fourth point if the extract is highly relevant and beneficial for educational purposes for a level not higher than grade school, exhibiting a clear and consistent writing style. It could be similar to a chapter from a textbook or a tutorial, offering substantial educational content, including exercises and solutions, with minimal irrelevant information, and the concepts aren't too advanced for grade school students. The content is coherent, focused, and valuable for structured learning.
Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for teaching either at primary school or grade school. It follows detailed reasoning, the writing style is easy to follow and offers profound and thorough insights into the subject matter, devoid of any non-educational or complex content.
The extract: <{}>
After examining the extract, briefly justify your total score, up to 100 words Conclude with the score using the format: "Educational score: <total points>".

Figure 9: Prompt for FineWeb-based Data Selection.