Add-One-In: Incremental Sample Selection for Large Language Models via a Choice-Based Greedy Paradigm

Zhuo Li[1,2,3], **Yuhao Du**[1,2,3], **Xiaoqi Jiao**[4], **Yiwen Guo**[5], **Yuege Feng**[6], **Xiang Wan**[1,2], **Anningzhe Gao**[1,2], **Jinpeng Hu**[7]*,

[1] Shenzhen International Center for Industrial and Applied Mathematics,
[2] Shenzhen Research Institute of Big Data,
[3] The Chinese University of Hong Kong, Shenzhen,
[4] LIGHTSPEED STUDIOS, [5] Independent Researcher,
[6] Birmingham City University, [7] Hefei University of Technology

Abstract

Selecting high-quality and diverse training samples from extensive datasets plays a crucial role in reducing training overhead and enhancing the performance of Large Language Models (LLMs). However, existing studies fall short in assessing the overall value of selected data, focusing primarily on individual quality, and struggle to strike an effective balance between ensuring diversity and minimizing data point traversals. Therefore, this paper introduces a novel choice-based sample selection framework that shifts the focus from evaluating individual sample quality to comparing the contribution value of different samples when incorporated into the subset. Thanks to the advanced language understanding capabilities of LLMs, we utilize LLMs to evaluate the value of each option during the selection process. Furthermore, we design a greedy sampling process where samples are incrementally added to the subset, thereby improving efficiency by eliminating the need for exhaustive traversal of the entire dataset with the limited budget. Extensive experiments demonstrate that selected data from our method not only surpasses the performance of the full dataset but also achieves competitive results with recent powerful studies, while requiring fewer selections. Moreover, we validate our approach on a larger medical dataset, highlighting its practical applicability in real-world applications. Our code and data are available at https://github.com/BIRlz/comperative_sample_selection.

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2024) have demonstrated exceptional success in AI (Chiang et al., 2023; Hu et al., 2025b; Li et al., 2025; Dai et al., 2025; Hu et al., 2023, 2022, 2025a). Following the knowledge-based pre-training stage, human-oriented supervised fine-tuning (SFT) (Wei et al., 2022; Du et al., 2025) significantly improves LLMs with the most increased performance. However, a huge number of parameters and high complexity of these models also lead to substantial computational and financial demands during SFT, especially when faced with extensive training data.

Recently, some studies have indicated that not all instruction data equally contribute to fine-tuning, with a most representative subset often sufficient to match or surpass the performance of LLMs tuned on full datasets (Zhou et al., 2023). Moreover, many work have shown that the diversity and quality of SFT data are crucial for fully unleashing the potential of LLMs (Ouyang et al., 2022; Xu et al., 2023). Therefore, there is a growing interest in developing sample selectors to identify optimal subsets for more efficient SFT. The construction of data selectors relies on the design of selection criteria, considering both the sources of quality labels and approaches to obtain them, which form the fundamental difference for judging data quality. As highlighted by Liu et al. (2024b), current approaches broadly adopt two strategies. The first one leverages sample internal information and solely relies on its intrinsic characteristics (Li et al., 2024c). The second strategy incorporates quality labels derived from external sources such as LLM preference judgments (Chen et al., 2024; Liu et al., 2024a) or continuous influence scores quantifying a sample's impact on model behavior (Xia et al., 2024; Cao et al., 2024).

Although these models have achieved considerable improvement, they usually suffer from several issues. Scoring results often lack sufficient differentiation, with many instances receiving identical scores like AlpaGasus (Chen et al., 2024) and DEITA (Liu et al., 2024a). On the other hand, Li et al. (2024a) has highlighted that effective pointwise scoring is inherently more challenging for LLMs than pairwise or listwise evaluation. Moreover, independent scoring for individual sample

*Corresponding author, 135858hjp@gmail.com

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 5322–5341
November 4-9, 2025

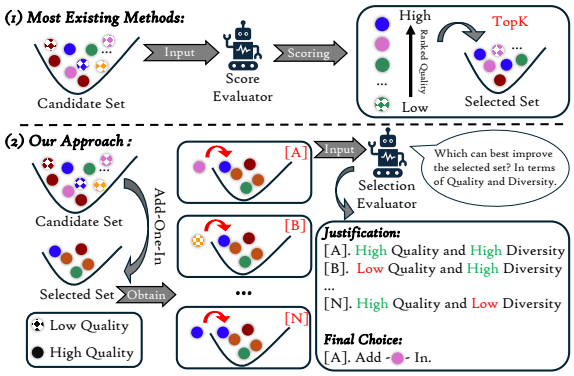

Figure 1: Colors represent data categories, while solid / dapple circles respectively stand for high- / low-quality data. (1) Most existing methods adopt a pointwise approach to produce a subset with top-K representative samples but ignore relationships among them. (2) Our method considers the quality and diversity contribution of each sample to the selected subset. For example, although both ● and ● exhibit high-quality in candidate set, incorporating ● into the selected subset is essential for enhancing its diversity, as the current selected subset already contains ●.

may overlook the internal relationships within the selected subset, which play a crucial role in ensuring the diversity of the data selection. Although some studies (Liu et al., 2024a,c) also propose to improve diversity with the help of explicit diversity measurement like cosine distance and optimal transport (Cuturi, 2013), they typically rely on traversing the entire dataset first, leading to substantial selection demands and time consumption.

Motivated by the Shapley value (Shapley, 1951), which measures the contribution of each individual in a cooperative setting, we propose a novel choice-based greedy selection framework, as shown in Fig. 1. Our framework shifts from traditional individual scoring to an option selection strategy. Specifically, we generate multiple candidate options by combining samples from the current selected subset with those from the remaining candidate set. These options are then evaluated and compared to identify the most suitable one. Unlike conventional methods that rely on explicit utility functions for evaluation, we leverage the language understanding capabilities of LLMs through a carefully designed prompt to assess each option's value in terms of both quality and diversity. To reduce computational overhead, we employ a sampling process to construct these options, where we sample two lists with a fixed window size: one from the selected subset and the other from the remaining dataset, approximating their respective ensembles. We then apply a greedy strategy to iteratively select the optimal option until the desired subset size is achieved. This approach accounts for the interdependencies among samples within the selected subset through an option-based selection mechanism. It also alleviates the need to access and traverse the entire dataset, thereby efficiently identifying a higher-quality and more diverse subset within a limited data budget.

Extensive experiments prove the effectiveness of our method. On the Alpaca instruction tuning dataset (Taori et al., 2023), our method shows that less than 10% of the data can outperform the model trained on the full dataset, and it also exhibits higher effectiveness compared to SOTA methods. Furthermore, we validate the effectiveness of our approach on a larger medical dataset, showcasing its practical applicability in real applications.

2 Related Work

The field of instruction sample selection for LLMs has seen significant advancements in recent years, driven by the need to improve model performance (i.e., safety (Du et al., 2024)) while reducing training costs (Liu et al., 2024b). Early efforts focus on general data selection methods, often relying on human-designed features or simple heuristics to identify high-quality and more diverse samples. For instance, Instruction-Mining (Cao et al., 2024) and InstructionGPT-4 (Wei et al., 2023b) utilize linguistic indicators and GPT-4 scores to guide the selection process, aiming to capture the quality of data through surface-level features. However, these methods often fell short in capturing the nuanced interactions between data and model performance. A more targeted approach emerged with the introduction of model- and data-centric selection criteria, where methods like IFD (Li et al., 2024c) and SuperFiltering (Li et al., 2024b) leverage internal information from the candidate dataset and the target model itself to select data, such as instruction following difficulty and perplexity of each sample. However, these methods are highly model-dependent and rely on the pre-experience training of the target LLM.

Recent advancements have further refined data selection by incorporating external information and producing quality labels. For example, AlpaGasus (Chen et al., 2024) and DEITA (Liu et al., 2024a) introduce discrete quality labels derived from LLM preferences, while methods like

LESS (Xia et al., 2024) use continuous quality labels based on sample influence. However, the formers produce identical quality scores, leading to insufficient differentiation among samples, and the latter can only handle specific-task datasets. Besides, although Liu et al. (2024a) and Liu et al. (2024c) also encourage diversity during the selection, similar to the methods of generating scores, they still rely on traversing the complete dataset first and require to explicitly design a complex diversity measurement function. Unlike these works, we eliminate pre-computed labels and explicit metrics through LLM-powered dynamic combinatorial evaluation, achieving fine-grained differentiation, task-agnosticity, and more time efficiency without full traversal.

In addition, there are also various selection methods that target on facilitating LLMs' pre-training, like D4 (Tirumala et al., 2023), DSIR (Xie et al., 2023), and QuRating (Wettig et al., 2024). Besides, Agrawal et al. (2022) and Nguyen and Wong (2023) expand sample selection into in-context learning and machine translation, while we focus on SFT.

3 Method

Let $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ denote an instruction tuning dataset with size N. Usually each sample x_i is a triplet $\{\text{Instruction}, [\text{Input}], \text{Answer}\}$, where $[\text{Input}]$ is an optional part associated with the instruction. Given a trainable LLM parameterized by $\theta \in \mathbb{R}^d$, we denote $\theta_{\mathcal{D}}$ as the instruction fine-tuned LLM θ on dataset $\mathcal{D}$. Our objective is to effectively find a subset $\mathcal{A} \subseteq \mathcal{D}$ ($|\mathcal{A}| = K \ll |\mathcal{D}|$) with K data budget, such that each selected sample in $\mathcal{A}$ satisfies specific criteria defined by a selection function $F(\cdot)$. Moreover, $\theta_{\mathcal{A}}$ can achieve comparable performance than $\theta_{\mathcal{D}}$ on various downstream tasks.

3.1 Warming-Up

As mentioned above, the diversity of $\mathcal{A}$ and the quality of each sample $x_i \in \mathcal{A}$ affect the performance of the fine-tuned model $\theta_{\mathcal{A}}$. In order to obtain a desirable subset, we begin with the definition of the diversity contribution of a sample x_i to a subset $\mathcal{A}$, where we leverage the following key motivation: *the diversity of a set should depend on how varied or different its elements are.*

Specifically, given an initial set $\mathcal{A}$ and a candidate set $\mathcal{B}$ (initially $\mathcal{B} = \mathcal{D}$) as a start point, for a sample $x_i \in \mathcal{B}$, we define its diversity contribution to the subset $\mathcal{A}$ through a marginal gain $\Delta F(x_i|\mathcal{A})$, where larger $\Delta F(x_i|\mathcal{A})$ indicates that sample x_i can increase the diversity of the set $\mathcal{A}$. We aim at finding the most valuable samples in $\mathcal{B}$ by evaluating $\Delta F(x_i|\mathcal{A})$ for each candidate sample x_i and select the one that maximizes this marginal gain with the help of a carefully designed selection function $F(\cdot)$ by:

$$\begin{aligned} x^* &= \arg\max_{x_i \in \mathcal{B}} \Delta F(x_i|\mathcal{A}) \\ &= \arg\max_{x_i \in \mathcal{B}} \left[F(\mathcal{A} \cup \{x_i\}) - F(\mathcal{A})\right]. \end{aligned} \tag{1}$$

This selection process continues until the subset $\mathcal{A}$ reaches the desired size K and at each iteration, we add the optimal sample x^* into $\mathcal{A}$ and remove it from $\mathcal{B}$. Besides, a desirable selection function $F(\cdot)$ is also expected to have the ability to judge the quality of the input sample x_i. By finding such an effective F, we can take the quality of samples into account during the process of finding a diversity subset using Eq. 1, and ultimately obtain such a desirable subset for LLM tuning. However, it's important to note that finding the exact optimal subset $\mathcal{A}^*$ by Eq. 1 should be NP-hard due to the difficulty of the combinatorial nature of the subset selection task and the computational complexity associated with evaluating all possible subsets (Welch, 1982).

3.2 A LLM-Driven Greedy Method For Contribution-Oriented Subset Selection

To efficiently find a desirable subset $\mathcal{A}$ without incurring the computational overhead of evaluating each candidate individually, we propose a LLM-driven greedy method that leverages the expressive capabilities of LLMs in understanding of the input and capable of handling the long-context to select the most valuable sample in a single computation at each iteration. Specifically, in each selection iteration, we firstly fill a selection prompt with $\mathcal{A}$ and $\mathcal{B}$ simultaneously, then query a LLM to make a choice about which element in $\mathcal{B}$ exhibits higher quality and can contribute the most diversity to $\mathcal{A}$.

As shown in the Fig. 2, we reformulate the selection obj. 1 by serving a LLM as the implementation of F by such a carefully designed selection prompt. In order to prevent the LLM from failing to understand this complex selection prompt well, resulting in mis-following and ultimately being unable to make effective decisions, we adopt an ICL strategy (Dong et al., 2022) to assist the LLM in better accomplishing our selection task and finally obtaining an effective optimal element. Finally, we can

Prompt Template for Our Selection Method:

You are a useful Assistant to help select the optimal element. You will receive two sets: Set A and Candidate Set B. Each element in both sets is a triplet {instruction, input, response}. Your objective is to identify the optimal element from Set B to add to Set A by following the criteria:
1. **Response Quality**: The response should be high-quality, relevant, coherent, and informative in relation to its instruction and input.
2. **Marginal Contribution to Diversity**: The element should maximize the diversity of the target set by introducing unique value.

Example:
Set A: [Element_1, Element_2, ..., Element_N]
Candidate Set B: [[A]-Element, [B]-Element, ..., [N]-Element]

Steps:
1. ***Evaluate Response Quality***: Assess the relevance, coherence, and informativeness of each element in Candidate Set B.
2. ***Add to Set A***: Add each element from Candidate Set B to Set A to form new sets like:
- Adding [[A]-Element] to Set A: [Element_1, Element_2, ..., Element_N, [A]-Element]
- Adding [[B]-Element] to Set A: [Element_1, Element_2, ..., Element_N, [B]-Element]
- ...
- Adding [[N]-Element] to Set A: [Element_1, Element_2, ..., Element_N, [N]-Element]
3. ***Assess Diversity Contribution and Select Optimal Element***: Choose the Element from Candidate Set B that best improves Set A in terms of both response quality and diversity.

Here is the Input:
Set A: [{Set_A}]
Candidate Set B: [{Set_B}]

Finally, ***ONLY*** return the index of selected element from Set B by strictly following the format: [A] for the first one, [B] for the second one, etc. ***AND*** in the subsequent line, please provide a comprehensive justification of your evaluation, avoiding any potential bias.

Your Decision:

Figure 2: The prompt employed in our method to select the optimal element from the candidate set $\mathcal{B}$ with the help of LLM, considering response quality and diversity contribution to the set $\mathcal{A}$.

obtain an optimal element x^* from the candidate set $\mathcal{B}$ by prompting F_{LLM}.

However, given the input length limitation and computational complexity associated with LLMs when processing long contexts, it is impractical to input the complete $\mathcal{A}$ and $\mathcal{B}$ into the LLM all at once. Therefore, during each iteration of selection, we randomly select $\mathcal{A}'$ from the current selected subset $\mathcal{A}$ and $\mathcal{B}'$ from candidate set $\mathcal{B}$, in order to avoid limitations and alleviate the traversal overhead. The selection process can be modified as:

$$x^* = F_{\text{LLM}}(\mathcal{A}', \mathcal{B}', \textit{prompt template}), \quad (2)$$

where sizes of $\mathcal{A}'$ and $\mathcal{B}'$ are denoted as L_A and L_B, respectively. Our LLM-driven greedy selection process are summarized in App. 1, where our method begins with a random initialization of $\mathcal{A}$ and without specific statement, we set $L_A = L_B = 20$.

Algorithm 1: Algorithm Process of Our Method.

Input: $\mathcal{D}$: Full dataset, a LLM and hyper-parameters: $\{K$: Desired subset size, L_A and L_B: window size.$\}$
Output: $\mathcal{A}$: Selected subset

Initialize subset $\mathcal{A} \leftarrow \text{RandomSample}(\mathcal{D}, L_A)$;
Initialize candidate $\mathcal{B} \leftarrow \mathcal{D} \setminus \mathcal{A}$;
while $|\mathcal{A}| < K$ **do**
  Initialize local $B' \leftarrow \text{RandomSample}(\mathcal{B}, L_B)$ and $A' \leftarrow \text{RandomSample}(\mathcal{A}, L_A)$;
  Obtain the optimal sample index by $x^* = F_{\text{LLM}}(\mathcal{A}', \mathcal{B}', \textit{prompt template})$;
  update $\mathcal{A} \leftarrow \mathcal{A} \cup \{x^*\}$;
  update $\mathcal{B} \leftarrow \mathcal{B} \setminus \{x^*\}$;
return $\mathcal{A}$

3.3 Discussion

Our greedy selection method relies on the LLM's ability to comprehend and accurately follow the provided selection instructions, where prior work has shown that LLMs are proficient in understanding and executing detailed prompts (Ouyang et al., 2022; Wei et al., 2023a). Besides, our method can be likened to listwise evaluation approaches, where prior research (Zhuang et al., 2024; Hou et al., 2024) have demonstrated that LLMs are capable of performing listwise selection and ranking tasks effectively. These studies have shown that LLMs can assess and rank a list of items based on certain criteria, which aligns with our approach of selecting items that maximize a set function considering both quality and diversity.

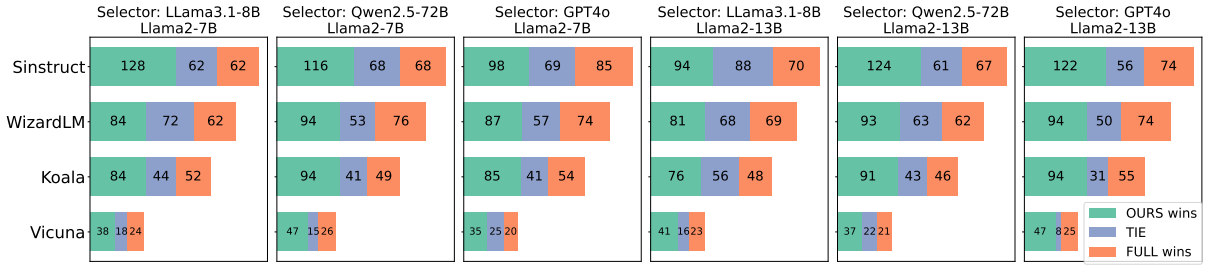

Figure 3: Comparing models fine-tuned on our method (9K) and full data (52K) on Llama2-7B and Llama2-13B with different LLMs as selector.

Moreover, our method parallels the classical greedy algorithm used in submodular optimization. when a LLM can correctly follow the selection prompt, our greedy algorithm could achieve at least a $(1 - e^{-\gamma})$ approximation of the optimal solution (Chen et al., 2018), where the parameter $\gamma \in [0, 1)$ is a submodularity ratio related to the LLM. This means that our method not only efficiently builds a diverse and high-quality subset but also provides theoretical assurance on the solution's near-optimality. In terms of computational cost, each iteration involves querying the LLM to select the next sample, resulting in a total time complexity of $\mathcal{O}(K \cdot T_{\text{LLM}})$, where K is the desired subset size and T_{LLM} is the inference time of the LLM. This linear complexity with respect to K makes our method scalable to large datasets and requires fewer accesses to the entire dataset.

4 Experiment

4.1 Instruction Fine-tuning Dataset

4.1.1 Settings

Datasets We mainly select an effective training subset from the Alpaca dataset that encompasses 52K instruction-following samples (Taori et al., 2023). In order to achieve less biased assessment, we evaluate the performance of our method on four popular testsets: Vicuna (Chiang et al., 2023), Koala (Zhang et al., 2024), WizardLM (Xu et al., 2023) and Self-instruct (Wang et al., 2022), which totally contains 730 human generated instructions from different tasks and sources to ensure the comprehensive convergence of task types.

Selector In order to comprehensively evaluate the flexibility of our method, we employ three types of LLM as the selector: 1) Small-size: Llama3.1-8B-Instruct (Touvron et al., 2023); 2) Medium-size: Qwen2.5-72B-Instruct (Yang et al., 2024); 3) Closed-source: GPT3.5 / GPT4o (OpenAI, 2024).

Baseline We compare the effectiveness of our method with the following methods: 1) Full data; 2) Random; 3) Two popular SOTA methods - AlpaGasus (Chen et al., 2024) and IFD (Li et al., 2024c), both of which require to firstly traverse the entire dataset, then rank, and finally select the top-K highest samples. We mainly consider these two SOTA methods due to the similar scope and experimental settings. For more concise expression, we use the names of these methods to represent the models that are fine-tuned on the corresponding datasets. For instance, "Full" is used to denote θ_{Full}.

Implementation Details Refer to App. A.1.

4.1.2 Evaluation Metrics

Pairwise Comparison Following the common practice of LLM-as-a-judge, we utilize GPT4o and adopt the evaluation template from MT-Bench (Zheng et al., 2023). To alleviate potential positional bias, we present the responses of two models to the judge in two different orders and compare their scores. A model is considered to win only if it does not lose in both orderings. Specifically, we define the outcomes as follows: **Wins**: Outperforms in both orderings or wins in one and ties in the other. **Tie**: Ties in both orderings or wins in one and loses in the other. **Loses**: Lags in both orderings or ties in one and loses in the other.

Benchmark Evaluation To fully understand how our selected samples influence the fine-tuning when compared with baselines, we also evaluate performance on several popular benchmarks, i.e., MMLU (Hendrycks et al., 2021), BBH (Suzgun et al., 2022) and Hellaswag (Zellers et al., 2019).

4.1.3 Results

Pairwise Comparison with Full (52K) By following AlpaGasus, which selects the 9K highest-quality samples scored by GPT3.5, we first demonstrate the effectiveness of our method using the

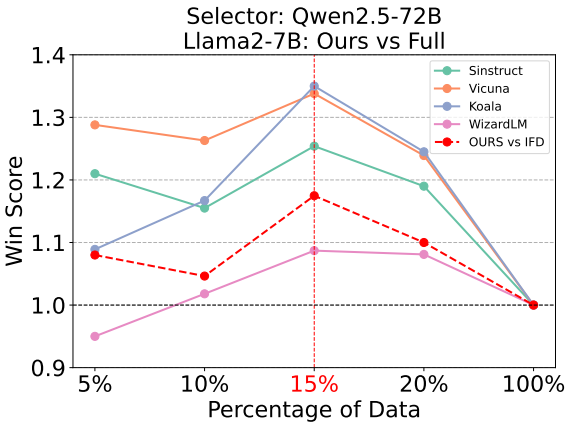

Figure 4: The win score changes with the increasing of data scale by comparing ours with the Full and IFD.

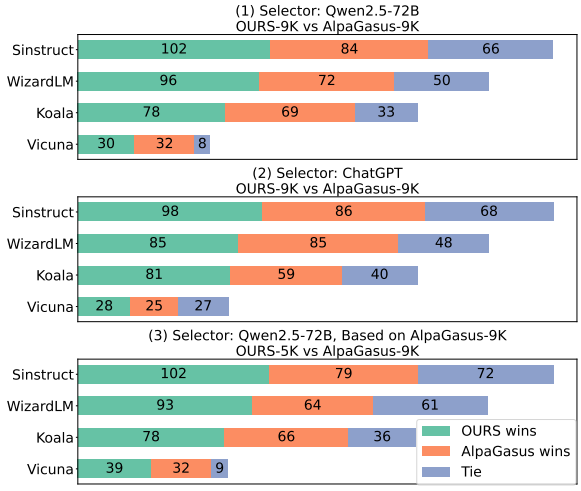

Figure 5: Comparing our method with AlpaGasus under 9K data on Llama2-7B.

same scale of selected samples. As shown in Fig. 3, our model trained with 9K samples significantly outperforms the model trained on the full dataset under various settings. We employ a diverse range of LLMs as selectors and as models to be trained (Llama2-7B-hf and Llama2-13B-hf). These results illustrate that our method effectively identifies valuable and diverse samples for instruction fine-tuning, leveraging the powerful language understanding capability of LLMs, regardless of the specific size or family of models used for selection or training.

Furthermore, we conduct experiments by selecting subsets corresponding to different percentages of the training dataset, ranging from 5% to 20%, and compare the average win score changes relative to the full dataset as the amount of data increases. WS can be calculated as $\frac{\#\text{Win}-\#\text{Lose}}{\#\text{All}} + 1$, where a score greater than 1.0 indicates that the model outperforms the one that is compared against. As shown in Fig. 4, with just 5% of the data it selects, our model can outperform those trained on the full dataset on three of the four test datasets. As the amount of data increases, our method consistently exceeds the performance of the Full across all four datasets. In particular, with 15% data usage, our model achieves the best performance, with an average win rate of 1.257. These results demonstrate that appropriately selecting a subset of the full data is sufficient to enhance the LLM's instruction-following ability, and our incremental method effectively identifies a high-quality and more diverse subset for improved LLM tuning.

Pairwise Comparison with SOTA methods Beyond our comparison with the full data, we also evaluate our performance against the SOTA methods - AlpaGasus and IFD. For comparison with AlpaGasus, we select 9K samples for a fair evaluation at the same data scale, due to the nature of AlpaGasus's discrete quality labels. As shown in (1) and (2) of Fig. 5, ours significantly outperforms AlpaGasus across four datasets, indicating our method's potential to select effective training samples without the need for pointwise traversal of the entire dataset. Furthermore, when we use the 9K high-quality samples selected by AlpaGasus as the candidate pool and apply our method for further selection, we find that the 5K samples chosen by our method can surpass the performance of AlpaGasus, as shown in (3) of Fig. 5. This in-depth exploration reveals that our method does not conflict with pointwise methods; instead, it can effectively enhance existing methods by enabling the selection of more representative samples through a hierarchical and multi-round screening process, thereby facilitating the SFT of more effective LLMs.

In our comparison with IFD, our method demonstrates competitive performance, as illustrated by the *red dotted* line in Fig.4. We consistently outperform IFD across various percentages ranging from 5% to 20%, achieving an optimal win score of 1.17 at 15%, which clearly highlights the strength and practical utility of our selection approach. Using the powerful understanding capabilities of LLMs, we can incrementally identify a diverse subset without having to examine the entire dataset at once, thereby greatly improving efficiency by reducing the number of selection times. Additionally, we provide comparisons with the Random baseline in Fig.6, where our method consistently outperforms the Random baseline across various settings by significant margins, demonstrating the effectiveness

Benchmark	Llama2-7B-hf				
	Full	Random	Alpagasus	IFD	Ours
MMLU-0-Shot	22.09	22.51	23.25	23.12	**23.82**
MMLU-5-Shot	45.45	46.38	46.74	46.10	**47.44**
BBH	**32.10**	31.38	31.32	31.08	30.97
Hellaswag	69.97	70.99	**71.07**	70.55	**71.07**
Average	42.15	42.82	43.10	42.71	**43.33**
Benchmark	Llama2-13B-hf				
	Full	Random	Alpagasus	IFD	Ours
MMLU-0-Shot	28.07	27.21	28.38	26.79	**29.01**
MMLU-5-Shot	53.53	52.34	**54.38**	54.28	54.28
BBH	46.40	44.58	46.21	46.25	**47.28**
Hellaswag	80.55	81.45	81.36	81.36	**81.73**
Average	51.38	51.89	52.58	52.17	**53.08**

Table 1: The benchmark results of models fine-tuned on different subsets selected by corresponding methods. More performance and analysis with varying training samples can be found in Tab. 6.

of the proposed method.

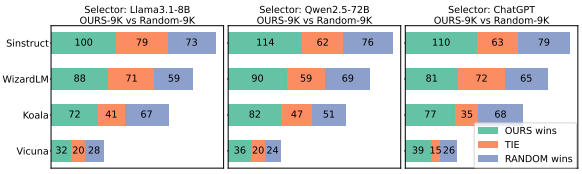

Figure 6: Comparing our method with the Random Baseline with the 9K data on Llama2-7B.

Performance on Benchmarks Following the practice of evaluating LLMs on benchmarks, we also compare our method with Full, Random, AlpaGasus, and IFD, where we set the sample size to 9K for a fair comparison, except for Full. For MMLU, BBH, and Hellaswag, we adopt 0/5-Shot, 3-Shot, and 0-Shot settings, respectively. Tab. 1 suggests that our method provides a promising approach to diverse data selection by achieving competitive performance to baselines with fewer selection times, optimizing both performance and selection cost. More detailed performance with varying scales of training data can be found in App. A.2.

4.2 Analysis

In this section, we conduct an in-depth analysis of our method from the perspective of hyperparameter sensitivity, efficiency analysis, characteristics of the selected samples, and case studies.

Ablation Study As mentioned in Sec. 3.2, in each iteration of selection, we randomly select subsets $\mathcal{A}'$ from the current selected set $\mathcal{A}$ with L_A samples and $\mathcal{B}'$ from the candidate set $\mathcal{B}$ with L_B samples to comply with the LLM input limitation. To empirically examine the impact of L_A and L_B, we conduct a detailed ablation study by varying the length of set $\mathcal{A}'$ and $\mathcal{B}'$ independently, observing its effect on the diversity and quality of the selected subsets. Additionally, we explore a multi-round selection strategy: initially selecting a larger subset, and then performing further selection on this subset in subsequent rounds until we obtain our desired subset size. This strategy aims to enhance diversity by iteratively refining the subset.

Fig. 7 compares different configurations in terms of their ability to select better samples and report pairwise comparison results, indicating superiority in diversity and quality of the selected subsets. Our key findings are as follows: *1) As shown in Fig. 7(1), (2), and (3), increased L_A or L_B leads to better performance consistently across different datasets and model selectors.* A larger L provides the LLMs with more context and a broader range of candidates to choose from, which in turn enhances the diversity and quality of the selected samples. *2) Fig. 7(4) illustrates that our method effectively integrates with a multi-round selection strategy.* By adopting a hierarchical, coarse-to-fine approach, we iteratively refine the subset through multiple selection rounds, gradually narrowing the focus to increasingly promising candidates. The data selected in each round serves as the foundation for constructing new options in the subsequent round, enabling a step-by-step process to identify higher-quality data over iterations. Additional results based on GPT3.5-as-the-selector are in App. A.3.

Efficiency Comparison We investigate whether our method achieves an acceptable trade-off between improved performance and the computational costs associated with the selection, from the perspective of: 1) Time Complexity; 2) Number of Iterations; 3) Practical Time Consumption. Using the Alpaca Dataset (52K), we present an efficiency comparison when selecting the top 10% of the data. The results in Tab. 2 demonstrate that our method significantly reduces selection costs compared to IFD and AlpaGasus. Specifically, operating with a time complexity proportional to K rather than N, our method requires significantly fewer iterations and less practical time, due to the greedy selection process that incorporates the powerful capability of LLMs. This not only accelerates the selection procedure, but also maintains high performance, achieving better efficiency without compromising

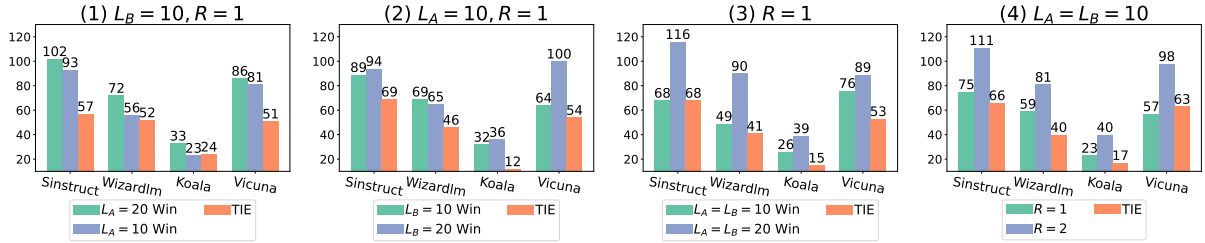

Figure 7: Ablation study on the influence of the lengths of sets $\mathcal{A}'$ and $\mathcal{B}'$ (1, 2, 3), and the number of selection rounds (4). We adopt a pairwise comparison and employ Qwen2.5-72B-Instruct as the selector.

Method	TC	NoI	PTC (Minutes)
AlpaGasus	$\mathcal{O}(N \times T_{\text{LLM}})$	52,002	2466.80
IFD			176.80
OURS	$\mathcal{O}(K \times T_{\text{LLM}})$	**5,200**	**118.83**

Table 2: Efficiency comparison of different methods on the Alpaca Dataset when selecting the top 10% data. T_{LLM} is the time for a single LLM inference. TC, NoI, and PTC indicate the time complexity, number of iterations, and practical time consumption, respectively.

Method	Diversity			Quality		#Tokens
	TTR ↑	MTLD ↑	SDI ↓	DEITA ↑	Helpful ↑	
Full (52K)	95.47	7.9352	0.1067	2.765	1.314	11.35
Random	95.46	7.9421	0.1068	2.764	1.397	11.33
AlpaGasus	96.07	8.0508	0.1086	2.969	2.025	10.93
IFD	96.05	8.0444	0.1091	3.127	2.456	10.80
OURS	**96.24**	**8.4449**	**0.1035**	**3.210**	**2.703**	11.29

Table 3: Performance of different methods on various text diversity and quality metrics. The OURS method shows the best performance across multiple metrics.

the quality of the selected data. In summary, our method is only proportional to the desired subset size K, embodying the principle of "selecting only what is necessary". This characteristic will ensure that our method can guarantee higher efficiency and lower selection costs when faced with extremely large-scale datasets. We provide implementation details in App. A.5.

Data Characteristics We evaluate whether our method identifies a representative subset with higher quality and greater diversity. Tab. 3 presents a comparison using various diversity metrics, such as Type-Token Ratio (TTR) (Richards, 1987), Measure of Textual Lexical Diversity (MTLD) (McCarthy and Jarvis, 2010), and Simpson Diversity Index (SDI) (Simpson, 1997), along with quality metrics including DEITA quality score (Liu et al., 2024a) and helpful reward value (Dong et al., 2023). Higher values generally indicate better diversity or quality, except for the SDI. We also report instruction token counts to assess structural similarity to the full dataset. Baseline methods are evaluated on 9K samples versus 52K with average scores reported. We provide implementation details and explanations to these metrics in App. A.4.

All three sample selection methods outperform the Full and Random baselines across most metrics, indicating that the full dataset contains redundancies and that strategic sample selection efficiently captures the most informative elements. Compared to AlpaGasus and IFD, our method significantly improves diversity and quality by emphasizing both aspects. Higher TTR and MTLD values in our subset reflect richer vocabulary and lexical structure, while a lower SDI suggests a more richer and balanced dataset. Increased DEITA quality scores and reward values confirm our improvements in both diversity and quality. Additionally, our subset's average token count is similar to that of the full dataset, implying that we maintain the original data's structural characteristics without favoring longer/shorter samples and demonstrating that our method effectively captures the semantic richness of the data. In conclusion, our method selects subsets that are both diverse and representative of the full dataset, offering promising implications for more effective LLM tuning.

Case Study We highlight three selection cases in App. B.

4.3 Scaling-Up with Larger Datasets

We conduct experiments in the medical domain to demonstrate the practical effectiveness of our method. Specifically, based on the HuatuoGPT-sft-data-v1 dataset for SFT samples, we evaluate performance on the Chinese Medical Benchmark (CMB) (Wang et al., 2024), utilizing the Baichuan2-7B-Chat model (Baichuan, 2023). Given that AlpaGasus requires querying GPT3.5 for sample scoring, we primarily compare our method with the Base model, Full, and IFD. We provide detailed training settings in App. A.6.

Tab. 4 shows that our method consistently improves performance in various data percentages,

Method	%	Chinese Medical Benchmark (%)						
		Physician	Nurse	Pharmacist	Technician	Disciplines	Exam	Avg.
Base	-	44.20	51.69	46.06	43.83	41.56	36.56	44.29
Full	100	46.35	57.69	50.81	43.83	**49.81**	52.00	49.38
OURS	10	46.35	54.06	48.72	40.58	42.88	43.81	46.65
IFD		42.55	50.19	44.75	**47.75**	41.19	37.19	45.90
OURS	20	**50.65**	**60.44**	51.94	43.00	47.25	**52.19**	**51.33**
IFD		37.50	51.56	46.47	41.67	40.19	44.75	46.94
OURS	30	49.70	58.63	52.09	41.58	46.06	50.00	50.31
IFD		48.00	56.19	49.47	39.67	43.56	44.31	47.54
OURS	40	49.20	60.38	**53.66**	42.58	**48.13**	51.44	51.03
IFD		47.90	58.75	51.31	43.08	45.38	49.63	49.79

Table 4: Performance comparison on the Chinese Medical Benchmark (CMB) using different methods and varying percentages of the dataset for fine-tuning. The percentages (%) indicate the proportion of the dataset utilized. The results are reported across various medical professions.

significantly outperforming both the Full and IFD. For example, with only 20% of the samples selected by our method, we achieve an improvement of 7.04% over the Base model and 4.36% over IFD. These findings highlight potential drawbacks of fine-tuning with the entire dataset and underscore the advantages of our approach in both performance enhancement and time efficiency, suggesting its potential for real-world applications.

5 Conclusion

We propose a novel choice-based sample selection paradigm that shifts the focus from individual scoring to comparing the contribution of each sample when incorporated into a subset. By leveraging the powerful understanding capabilities of LLMs, we are able to simultaneously consider both quality and diversity during the sample selection process. Moreover, we design a greedy process that incrementally builds the subset, which not only eliminates the need to traverse the entire dataset but also significantly reduces selection overhead. Extensive experiments on Alpaca dataset and medical application demonstrate that our method selects more representative subsets with improved selection efficiency compared with SOTA methods, showing as a promising direction for efficient sample selection.

6 Acknowledgments

This work was supported by the Guangxi Key R&D Project (No. AB24010167), the Project (No. 20232ABC03A25), and the Futian Healthcare Research Project (No.FTWS002). This work was also supported by the National Natural Science Foundation of China (NSFC) under Grant 62402158.

7 Limitations and Future Work

While our proposed choice-based sample selection paradigm demonstrates promising results, it is important to acknowledge several limitations and potential areas for future improvement. A significant limitation of our approach is its reliance on LLMs whose capabilities and biases directly impact the sample selection process. The rationality and effectiveness of selecting samples are contingent upon the LLM's ability to accurately assess and compare data points. However, LLMs may have inherent biases inherited from their training data, which can inadvertently influence the selection process (Wang et al., 2023; Ko et al., 2020). Another limitation lies in the need to manually adjust hyperparameters such as the sizes of sets $\mathcal{A}'$ and $\mathcal{B}'$, as well as the design of prompts used to elicit evaluations from the LLM. Selecting the sizes of $\mathcal{A}'$ and $\mathcal{B}'$ impacts the granularity and breadth of the selection process, while the choice of prompts influences the quality of the LLM's assessments.

In the future, enhancing the greedy sampling process by incorporating more sophisticated heuristic methods instead of random selection offers a promising direction. Besides, how to extend my approach to online and sequential data scenarios remains an open challenge. For example, developing adaptive sampling strategies would significantly enhance the practicality of our method. This could involve designing algorithms capable of making immediate selection decisions in real-time applications, ensuring sustained model performance in environments where data arrives continuously.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *Preprint*, arXiv:2212.02437.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. 2024. Instruction mining: Instruction data selection for tuning large language models. *Preprint*, arXiv:2307.06290.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024. Alpagasus: Training a better alpaca with fewer data. *Preprint*, arXiv:2307.08701.

Lin Chen, Moran Feldman, and Amin Karbasi. 2018. Weakly submodular maximization beyond cardinality constraints: Does randomization help greedy? In *International Conference on Machine Learning*, pages 804–813. PMLR.

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

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Chongyuan Dai, Jinpeng Hu, Hongchang Shi, Zhuo Li, Xun Yang, and Meng Wang. 2025. Psyche-r1: Towards reliable psychological llms through unified empathy, expertise, and reasoning. *arXiv preprint arXiv:2508.10848*.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *Preprint*, arXiv:2304.06767.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Yuhao Du, Zhuo Li, Pengyu Cheng, Zhihong Chen, Yuejiao Xie, Xiang Wan, and Anningzhe Gao. 2025. Simplify rlhf as reward-weighted sft: A variational method. *Preprint*, arXiv:2502.11026.

Yuhao Du, Zhuo Li, Pengyu Cheng, Xiang Wan, and Anningzhe Gao. 2024. Detecting ai flaws: Target-driven attacks on internal faults in language models. *Preprint*, arXiv:2408.14853.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. *Preprint*, arXiv:2305.08845.

Jinpeng Hu, DanDan Guo, Yang Liu, Zhuo Li, Zhihong Chen, Xiang Wan, and Tsung-Hui Chang. 2023. A simple yet effective subsequence-enhanced approach for cross-domain ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12890–12898.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022. Graph enhanced contrastive learning for radiology findings summarization. *Preprint*, arXiv:2204.00203.

Jinpeng Hu, Hongchang Shi, Chongyuan Dai, Zhuo Li, Peipei Song, and Meng Wang. 2025a. Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark. *arXiv preprint arXiv:2508.16859*.

Jinpeng Hu, Ao Wang, Qianqian Xie, Hui Ma, Zhuo Li, and Dan Guo. 2025b. Agentmental: An interactive multi-agent framework for explainable and adaptive mental health assessment. *Preprint*, arXiv:2508.11567.

Miyoung Ko, Jinhyuk Lee, Hyunjae Kim, Gangwoo Kim, and Jaewoo Kang. 2020. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1109–1121, Online. Association for Computational Linguistics.

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

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024b. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *Preprint*, arXiv:2402.00530.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024c. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *Preprint*, arXiv:2308.12032.

Zhuo Li, Yuhao Du, Jinpeng Hu, Xiang Wan, and Anningzhe Gao. 2025. Self-instructed derived prompt generation meets in-context learning: Unlocking new potential of black-box LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1857, Vienna, Austria. Association for Computational Linguistics.

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

Ziche Liu, Rui Ke, Feng Jiang, and Haizhou Li. 2024b. Take the essence and discard the dross: A rethinking on data selection for fine-tuning large language models. *Preprint*, arXiv:2406.14115.

Zifan Liu, Amin Karbasi, and Theodoros Rekatsinas. 2024c. Tsds: Data selection for task-specific model finetuning. *Preprint*, arXiv:2410.11303.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *Preprint*, arXiv:2302.11042.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Brian Richards. 1987. Type/token ratios: what do they really tell us? *Journal of child language*, 14:201–9.

Lloyd S Shapley. 1951. Notes on the n-person game—ii: The value of an n-person game.

E. H. Simpson. 1997. Measurement of diversity. *Journal of Cardiothoracic and Vascular Anesthesia*, 11(6):812–812.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *Preprint*, arXiv:2210.09261.

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

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *Preprint*, arXiv:2308.12284.

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.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. Cmb: A comprehensive medical benchmark in chinese. *Preprint*, arXiv:2308.08833.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023a. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Lai Wei, Zihao Jiang, Weiran Huang, and Lichao Sun. 2023b. Instructiongpt-4: A 200-instruction paradigm for fine-tuning minigpt-4. *Preprint*, arXiv:2308.12067.

William J. Welch. 1982. Algorithmic complexity: Three np-hard problems in computational statistics. *Journal of Statistical Computation and Simulation*, 15(1):17–25.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. *Preprint*, arXiv:2402.09739.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *Preprint*, arXiv:2402.04333.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. *Preprint*, arXiv:2302.03169.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *Preprint*, arXiv:2304.12244.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

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

Kaiqi Zhang, Jing Zhao, and Rui Chen. 2024. Koala: Enhancing speculative decoding for llm via multi-layer draft heads with adversarial learning. *Preprint*, arXiv:2408.08146.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. 2024. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 38–47. ACM.

A Experiment

A.1 Implementation Details and Cost

Implementation Details We mainly evaluate the performance of the selection methods on Llama2-7B-hf and Llama2-13B-hf, where we adopt the same training configuration as the original Alpaca using the Stanford codebase[1]. During inference, we employ vLLM (Kwon et al., 2023) to help speed up the generation, where we set the sampling temperature = 0.0 and topK_p = 1 to avoid randomness. Detailed training hyper-parameters and cost can be found in Tab. 5.

Model Size	Data Size	# GPUs	Epoch	LR	Batch Size	Max Length	Training Time (Minutes)
7B	9K	8	3	2e-5	128	512	13.33
7B	52K	8	3	2e-5	128	512	161.40
13B	9K	8	5	1e-5	128	512	43.16
13B	52K	8	5	1e-5	128	512	219.48

Table 5: Detailed hyper-parameter settings and costs.

A.2 Performance of Benchmark on Varying Training Samples

In this section, we conduct an ablation study to analyze the impact of varying training data proportions on the performance of LLMs across different benchmarks. Instead of ablating components of a model architecture, we systematically evaluate models trained on increasing subsets of the full training data: 3K, 6K, and 9K examples. This allows us to isolate the effect of training data size on the model's ability to generalize to unseen tasks, as measured by MMLU, BBH, and Hellaswag. By comparing these results to the performance of a model trained on the complete dataset (52K), we aim to quantify the performance gains achieved with larger training sets and identify potential saturation points or diminishing returns. This analysis provides insights into the data efficiency of LLMs and the importance of training data scale for achieving optimal benchmark performance.

Table 6 presents a comprehensive performance comparison across several few-shot learning methods: Full (utilizing a larger 52K dataset), Random (subsampling to 3K, 6K, and 9K), Alpagasus (3K, 6K, and 9K), IFD (3K, 6K, and 9K), and our proposed approach, "Ours" (evaluated on 3K, 6K, and 9K subsets). The evaluation spans diverse and challenging benchmarks, including MMLU (assessed in both 0-shot and 5-shot settings), BBH, and Hellaswag. The results highlight the performance trends as the size of the training data increases from 3K to 9K, providing insights into the scalability and effectiveness of each method on different tasks. Notably, the final column for each data size (9K) emphasizes the performance of "OURS" often demonstrating competitive or superior results in these higher data regimes. The average performance across all tasks is also provided at the bottom, offering a holistic view of each method's overall effectiveness.

Table 6: Performance Comparison

Benchmarks	Full (52K)	Random (3K)	Alpagasus (3K)	IFD (3K)	OURS (3K)	Random (6K)	Alpagasus (6K)	IFD (6K)	OURS (6K)	Random (9K)	Alpagasus (9K)	IFD (9K)	OURS (9K)
MMLU-0-Shot	22.09	21.10	22.00	21.72	**22.30**	21.80	22.40	22.10	**22.70**	22.51	23.25	23.12	**23.82**
MMLU-5-Shot	45.45	44.00	44.70	44.80	**45.60**	45.00	45.70	45.30	**46.10**	46.38	46.74	46.10	**47.44**
BBH	**32.10**	30.50	30.80	31.40	31.20	30.90	31.10	31.60	31.50	31.38	31.32	31.08	30.97
Hellaswag	69.97	69.00	69.80	69.30	**70.10**	69.50	**71.20**	69.65	70.50	70.99	**71.07**	70.55	**71.07**
Average	42.15	41.15	41.83	41.53	**42.30**	41.80	42.60	41.97	**42.70**	42.82	43.10	42.71	**43.33**

Table 7: The benchmark results of models fine-tuned on different subsets selected by corresponding methods.

[1] https://github.com/tatsu-lab/stanford_alpaca/tree/main

Dataset	Selector	$L_B = 10, R = 1$			$L_A = 10, R = 1$			$R = 1$			$L_A = L_B = 10$		
		$L_A = 20$ Win	$L_A = 10$ Win	TIE	$L_B = 10$ Win	$L_B = 20$ Win	TIE	$L_A = L_B = 10$ Win	$L_A = L_B = 10$ Win	TIE	$R = 1$ Win	$R = 2$ Win	TIE
Sinstruct	Qwen2.5-72B-Instruct	**102**	93	57	89	**94**	69	68	**116**	68	75	**111**	66
Vicuna		**33**	23	24	32	**36**	12	26	**39**	15	23	**40**	17
Koala		**72**	56	52	**69**	65	46	49	**90**	41	59	**81**	40
Wizardlm		**86**	81	51	64	**100**	54	76	**89**	53	57	**98**	63
Sinstruct	GPT3.5	**105**	82	57	88	**94**	70	74	**122**	56	78	**109**	65
Vicuna		**41**	27	12	**32**	**32**	16	25	**47**	56	19	**50**	11
Koala		**73**	65	42	63	**67**	50	55	**94**	31	53	**80**	47
Wizardlm		**96**	71	51	68	**85**	65	74	**94**	50	58	**101**	59

Table 8: Ablation study on the influence of the lengths of sets $\mathcal{A}'$ and $\mathcal{B}'$, and the number of selection rounds. The numbers indicate how many times a configuration wins over another in pairwise comparisons.

A.3 Ablation Study

Ablation Study Intuitively, a larger L_A and L_B results in $\mathcal{A}'$ and $\mathcal{B}'$ containing more elements, which increases the number of candidates and simultaneously elevates the difficulty of assessing diversity among them. This expansion in the candidate subsets provides the LLM with more options to consider, potentially leading to the selection of samples with greater diversity. Our experimental results are summarized in Tab. 8, where we compare different configurations in terms of their ability to select better samples. We report the number of times one configuration wins over another, indicating superiority in diversity and quality of the selected subsets. We conduct evaluations using various LLMs as selectors (Qwen2.5-72B-Instruct and GPT3.5) on multiple datasets (e.g., Sinstruct, Vicuna, Koala, WizardLM). From the results in Tab. 8, we observe several key trends:

1. **Effect of the initialization of $\mathcal{A}'$**: In the preceding experiment, the $\mathcal{A}'$ was initialized through random sampling. It is hypothesized that a more guided initialization, such as employing K-means clustering, could lead to a more advantageous final subset selection and consequently improved model performance. In this section, we experimentally investigate the impact of various initialization strategies for $\mathcal{A}'$ on the characteristics of the ultimately chosen subset. The results presented in Table 9 indicate that utilizing K-means and KNN for the initialization of $\mathcal{A}'$ does indeed result in slightly enhanced Diversity Measurements, as evidenced by metrics like TTR and MTLD. However, when considering Quality Measurement, the Random Initialization approach achieves comparable performance to both K-means and KNN, even outperforming them in terms of Helpful Score. These findings suggest that our method exhibits a degree of robustness with respect to different initializations of $\mathcal{A}'$. We posit that the underlying reason for this resilience stems from our experimental constraint of setting the size of $\mathcal{A}'$ to a maximum of 20, as a consequence, the initial selection within a relatively small pool might not drastically alter the final refined subset obtained through our subsequent selection process.

2. **Effect of Length of Set $\mathcal{A}'$**: Increasing the length of the selected subset $\mathcal{A}'$ from 10 to 20 generally leads to better performance. For instance, on the Sinstruct dataset using Qwen2.5-72B-Instruct as the selector, $\mathcal{A}'$ of length 20 wins 102 times versus 93 times for length 10, with 57 ties. This suggests that providing more context from the current selected set helps the LLM make more informed decisions, enhancing diversity.

3. **Effect of Length of Candidate Set $\mathcal{B}'$**: Similarly, a larger candidate set $\mathcal{B}'$ (length 20) improves performance compared to a smaller candidate set (length 10). For example, on the WizardLM dataset with GPT3.5 as the selector, $\mathcal{B}'$ of length 20 wins 85 times versus 68 wins for length 10. A larger candidate pool offers more options for the LLM to select diverse and high-quality samples.

4. **Combined Effect of Lengths of $\mathcal{A}'$ and $\mathcal{B}'$**: When both $\mathcal{A}'$ and $\mathcal{B}'$ are increased to length 20, we observe a cumulative positive effect. The configuration with both sets at length 20 often wins more comparisons against the configuration where both are at length 10. This indicates that simultaneously increasing both sets' lengths amplifies the benefits in diversity and quality.

5. **Multi-Round Selection Strategy**: Implementing multiple rounds of selection shows a consistent advantage. The "Round 2" configuration, where a second round of selection is performed on an

initially larger subset, often outperforms the "Round 1" configuration. For instance, on the Sinstruct dataset with Qwen2.5-72B-Instruct, Round 2 wins 111 times versus 75 wins for Round 1. This demonstrates that iterative refinement through multiple selection rounds effectively enhances the final subset's diversity and quality.

6. **Consistency Across Models and Datasets**: These trends are observed across different LLM selectors (Qwen2.5-72B-Instruct and GPT3.5) and datasets (Sinstruct, WizardLM, Vicuna, Koala). This consistency suggests that the benefits of larger set sizes and multi-round selection are generalizable. Besides, our method is not limited to specific models or datasets.

Method	Diversity			Quality		#Tokens
	TTR ↑	MTLD ↑	SDI ↓	DEITA ↑	Helpful ↑	
Full (52K)	95.47	7.9352	0.1067	2.765	1.314	11.35
Random	95.46	7.9421	0.1068	2.764	1.397	11.33
AlpaGasus	96.07	8.0508	0.1086	2.969	2.025	10.93
IFD	96.05	8.0444	0.1091	3.127	2.456	10.80
OURS + Random	96.24	8.4449	**0.1035**	3.210	**2.703**	11.29
OURS + Kmeans	**96.58**	**9.0125**	0.1057	3.092	2.613	11.57
OURS + KNN	96.39	8.8721	0.1048	**3.278**	2.581	11.98

Table 9: Performance of different methods on various text diversity and quality metrics. The OURS method shows the best performance across multiple metrics.

In summary, the ablation study confirms that increasing the lengths of sets $\mathcal{A}'$ and $\mathcal{B}'$ allows the LLM to consider more context and a wider range of candidates, leading to the selection of samples with greater diversity and quality. Additionally, employing a multi-round selection strategy further refines the subset by allowing the LLM to iteratively focus on the most promising candidates. However, it is important to balance these benefits with computational considerations, as larger sets and additional rounds may increase inference time. Selecting appropriate lengths for $\mathcal{A}'$ and $\mathcal{B}'$, as well as an optimal number of selection rounds, is crucial for maximizing performance while maintaining efficiency.

A.4 Explanation To Characteristics Metrics

Type-Token Ratio The Type-Token Ratio (Richards, 1987) (TTR) represents the relationship between the quantity of distinct words (types) emerging in a text and their occurrence frequencies. The count of unique words within a text is conventionally termed the number of types. It's worth noting that some of these types recur. The value of the TTR spans from 0 to 100. A larger number of types relative to the total number of tokens (resulting in a higher ratio value) indicates a richer vocabulary. In other words, the text exhibits greater lexical diversity. The Type-Token Ratio is computed as follows:

$$\text{TTR} = \frac{\text{Number of unique types}}{\text{Number of tokens}} * 100 \tag{3}$$

Measure of Textual Lexical Diversity The Measure of Textual Lexical Diversity (McCarthy and Jarvis, 2010) (MTLD) is a metric designed to assess the lexical diversity of a text while mitigating the influence of text length, a limitation often encountered with traditional measures like the Type-Token Ratio (TTR). Unlike TTR, which can vary significantly with the length of the text, MTLD remains relatively stable, providing a more consistent evaluation of lexical diversity across texts of varying lengths. The MTLD value is determined by dividing the total number of words by the total number of factors, effectively representing the average length of word strings that maintain the desired TTR threshold. The MTLD is computed as follows:

$$\text{MTLD} = \frac{\text{Total number of words in the text}}{\text{Number of factors}} \tag{4}$$

A higher MTLD value indicates greater lexical diversity, as it implies longer sequences of words are required before the TTR falls below the threshold, reflecting richer and more varied vocabulary. By

accounting for both the range and distribution of vocabulary, MTLD provides a robust assessment of lexical diversity that is less sensitive to text length compared to other metrics.

Simpson's Diversity Index The Simpson Diversity Index (Simpson, 1997) (SDI), originally developed in ecology to measure biodiversity, can be effectively applied to textual analysis to assess lexical diversity. In textual diversity analysis, unique words are treated as species, and their frequencies as species abundances. The SDI quantifies the probability that two randomly selected words from a text are different and is calculated using the formula:

$$D = \sum_{i=1}^{N} p_i^2, \tag{5}$$

where p_i represents the proportion of the i-th word type in the text, calculated as $p_i = \frac{n_i}{N_t}$. n_i is the frequency of the i-th word and N_t the total number of words. Values close to 1 indicate higher homogeneity, thus lower diversity and values close to 0 indicate higher variability, thus higher diversity. This measure is particularly useful for understanding the complexity and style of a text, as it accounts for both the number of unique words and their frequency distribution.

Helpful Reward Score Reward scores usually help the model learn to maximize helpfulness by adjusting its parameters based on these scores, ensuring that the training data selected is optimally beneficial for improving performance. We adopt a reward model that is trained on the Anthropic Helpful Harmless dataset and achieves a test accuracy of over 75% (Dong et al., 2023), where a higher score indicates the better response quality regarding to its input instruction.

A.5 In-Depth Analysis of Efficiency

Time Complexity (TC) and Number of Iterations (NoI) SOTA methods that rely on LLMs for selection (e.g., IFD, Alpagasus, and DeITA) typically employ pointwise scoring approaches. These methods have a time complexity of $\mathcal{O}(N \times T_{\text{LLM}})$, where N denotes the total number of samples in the full training dataset and T_{LLM} represents the time required for the LLM to perform a single inference. This implies that to select a subset of size K, these methods need to process all N samples, resulting in N data accesses. In contrast, as discussed in Section 3.3, our method operates with a time complexity of $\mathcal{O}(K \times T_{\text{LLM}})$. This means we perform selections proportional only to the desired subset size K, embodying the principle of "selecting only what is necessary". When $K \ll N$, our method demonstrates significantly enhanced efficiency compared to pointwise scoring methods. This efficiency gain largely aligns with the goal of sample selection: to train effective models using less data and computational resources.

Practical Time Consumption (PTC) Beyond theoretical analysis, we provide empirical comparisons of the actual time consumed during the selection process. Based on the Alpaca dataset and under identical hardware conditions, we reproduce IFD and Alpagasus for fair comparison. For both IFD and our method, we utilize Qwen2.5-7B-Instruct and set $L_A = L_B = 20$ in our method. For Alpagasus, we use the official OpenAI API to query GPT3.5. To conserve time and computational resources, we measure the time required to perform selection (for our method) or scoring (for IFD and Alpagasus) on 1,000 samples. We report the total time taken for the process.

Additionally, we also explore how varying the value of L in our method affects the overall practical time consumption (PTC), providing insights into the trade-offs between selection granularity and computational efficiency. Specifically, under the same hardware environment, we record the time consumption for 100 selections under different values of L_A and L_B, when employing Llama3.1-8B-Instruct as the selector. The results in Tab. 10 show that our method does not experience a sharp increase in time consumption as L_A and L_B rise. Although the practical time consumption generally grows with larger L_A and L_B values, the rate of increase is relatively moderate. This indicates that within a certain range, we can adjust L_A and L_B to balance the selection and computational cost without being overly burdened by excessive time expenditure.

$L_A = L_B =$	5	10	15	20	30
PTC (Second)	25.42	40.84	69.95	103.62	108.91

Table 10: Detailed hyper-parameter settings and costs.

In summary, results of these metrics prove that our method strikes an acceptable balance between enhanced performance and computational costs. The results indicate that our approach is more efficient, particularly when the desired subset size K is much smaller than the full dataset size N, fulfilling the objective of achieving better models with less data and reduced time investment.

A.6 Scaling-up with Larger Dataset

Datasets We adopt HuatuoGPT-sft-data-v1[2] as SFT samples that includes 226K Chinese medical QA pairs as SFT samples and evaluate the performance on Chinese Medical Benchmark (CMB) (Wang et al., 2024) that is a comprehensive and all-encompassing Chinese medical quiz assessment benchmark, which contains 12K human-annotated five-option multi-choice questions and cover six aspects in benchmarking a medical LLM. The Tab. 11 highlights the training hyper-parameters.

Model Size	# GPUs	Epoch	LR	Batch Size	Max Length
7B	2*8	3	2.5e-5	128	2048

Table 11: Detailed hyper-parameter settings and costs.

B Case Analysis

In this section, we provide three case selection analyses when employing GPT3.5 as the selector. As shown in Fig. 8, Fig. 9, and Fig. 10, we find LLMs can effectively make a reasonable selection choice given the selection prompt that includes the selected set A and the candidate set B. For example, as shown in Fig. 8, the selector recommends *[A]-Element* for the optimal element to add to the set A, by providing a highly informative response on the Quicksort algorithm and introducing a distinct topic to Set A. The other elements in Candidate Set B also demonstrate high response quality but do not introduce as unique of a concept as the Quicksort algorithm in [A]-Element.

[2] https://huggingface.co/datasets/FreedomIntelligence/HuatuoGPT-sft-data-v1

{"role": "system", "content": "You are a useful Assistant to help select the optimal element. You will receive two sets: Set A and Candidate Set B. Each element in both sets is a triplet [instruction, input, response]'. Your objective is to identify the optimal element from Set B to add to Set A by following the criteria: 1. Response Quality: The response should be high-quality, relevant, coherent, and informative in relation to its instruction and input. 2. Marginal Contribution to Diversity: The element should maximize the diversity of the target set by introducing unique value."},
{"role": "user", "content": "### Example: \n Set A: [Element_1, Element_2, \u2026, Element_N] \n Candidate Set B: [[A]-Element, [B]-Element, \u2026, [N]-Element]\n ###Steps: \n 1. Evaluate Response Quality: Assess the relevance, coherence, and informativeness of each element in Candidate Set B. \n 2. Add to Set A: Add each element from Candidate Set B to Set A to form new sets like: \n - Adding [[A]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [A]-Element] \n - Adding [[B]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [B]-Element] \n - \u2026 \n - Adding [[N]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [N]-Element] \n 3. Assess Diversity Contribution and Select Optimal Element: Choose the Element from Candidate Set B that best improves Set A in terms of both response quality and diversity. \n
Here is the Input:
Set A: [1. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Provide a brief overview of the French Revolution.### Response:The French Revolution began in 1789 and lasted for slightly over a decade. It was a major event in the history of France and Europe, and was driven by the desire to achieve political liberty, social justice, and economic freedom. During this period, property and unprivileged rights were redistributed, and the French monarchy was abolished. The French Revolution had a profound impact on the entire European political system.

2. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Arrange the following numbers from least to greatest: 9, -3, 15, 0.### Response:The numbers from least to greatest are -3, 0, 9, 15.

3. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Create a script by using a classical story structure### Response:Act 1: Introduction - Introduce the characters, setting and the struggle. Act 2: Conflict - Introduce the source of struggle and the climax. Act 3: Resolution - Introduce the resolution and how the characters learned from the struggle. Act 4: Conclusion - Introduce the resolution, where the characters and the setting stand post-struggle.

4. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Write a query to return all rows from a MySQL table named \"Items\" where the column \"category\" is \"clothing\" and the column \"color\" is either \"red\" or \"white\".### Response:SELECT * FROM Items WHERE category=\"clothing\" AND (color=\"red\" OR color=\"white\");

5. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Describe what the sky looks like in a sentence.### Response:The sky is a lush blue, dotted with white fluffy clouds that drift lazily across the expanse.]

Candidate Set B: [
[[A]]: Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Give an example of an algorithm and explain what it does.### Response:An example of an algorithm is Quicksort. Quicksort is a sorting algorithm that uses a form of divide and conquer to quickly sort an array of data. It works by Divide by selecting a pivot element from the array and partitioning the other elements into two sub-arrays, according to whether they are less than or greater than the pivot element. It then recursively sorts the sub-arrays.

[[B]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Summarize the steps needed to solve this problem.### Input:Given a string of 10 characters, print the character at index 3.### Response:To solve this problem, one needs to take a string of 10 characters, determine the index at which the desired character is located, and then print out the character located at that index.

[[C]]: Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Suggest a suitable activity for family members to do together.### Response:A great activity for family members to do together is to have a movie night. Plan to watch a movie that all members of the family can enjoy, get some snacks, and enjoy a fun bonding time watching the film together.

[[D]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Generate a PDF report of information about a company.### Input:ACME Co. is a successful software company that has been on the market for over 10 years. The company specializes in developing solutions with innovative and user-friendly interfaces, while focusing on customer satisfaction. The company employs 150+ people in 5 different countries and offers a wide range of services.### Response:Using Python, one could generate a PDF report of the information about ACME Co. by converting the information into a beautiful PDF template. The template could include details about the company such as the number of years ACME Co. has been in business, the number of employees, the countries where they are based, and the services they offer. The PDF report could also be supplemented with additional information such as the company mission statement, vision and core values, case studies, customer testimonials, and information about the technology and tools they use.

[[E]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Sort these animals by their habitat.### Input:owl, leopard, whale### Response:Owl (land), Leopard (land), Whale (water)]

Finally, ONLY return the index of selected element from Set B by strictly following the format: [A] for the first one, [B] for the second one, etc. And in the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias..\n ### Your Decision:"}],

"response": "After evaluating the response quality and potential contribution to diversity of each element in Candidate Set B, my recommendation for the optimal element to add to Set A is [A]-Element. \n\nThe response to the instruction is highly informative and demonstrates a clear understanding of the topic. The Quicksort algorithm is an important and widely used sorting algorithm, and the response provides a clear explanation of how it works. Additionally, this element introduces a new topic to Set A that is distinct from the existing elements, thus contributing to the diversity of the set. \n\nWhile the other elements in Candidate Set B also demonstrate high response quality and potential contribution to diversity, they do not provide as distinct of a topic as the Quicksort algorithm does. For example, [C]-Element suggests watching a movie together as a family, which may be a common activity for many families. [D]-Element and [E]-Element are more technical in nature, but may not introduce as unique of a concept as the Quicksort algorithm. [B]-Element is also technical in nature but introduces a more specific problem, which may limit its contribution to diversity. \n\nIn summary, [A]-Element is the optimal choice due to its high response quality and unique contribution to the diversity of Set A. \n\nTherefore, my decision would be: \nThe selected element from Set B is [A]. \n\nExplanation: \n- [A]-Element provides a highly informative response on the Quicksort algorithm and introduces a distinct topic to Set A. \n- The other elements in Candidate Set B also demonstrate high response quality but do not introduce as unique of a concept as the Quicksort algorithm. \n- Therefore, [A]-Element is the optimal choice for improving the response quality and diversity of Set A."}

Figure 8: Selection case analysis.

{"role": "system", "content": "You are a useful Assistant to help select the optimal element. You will receive two sets: Set A and Candidate Set B. Each element in both sets is a triplet [instruction, input, response]'. Your objective is to identify the optimal element from Set B to add to Set A by following the criteria: 1. Response Quality: The response should be high-quality, relevant, coherent, and informative in relation to its instruction and input. 2. Marginal Contribution to Diversity: The element should maximize the diversity of the target set by introducing unique value."},
{"role": "user", "content": "### Example: \n Set A: [Element_1, Element_2, \u2026, Element_N] \n Candidate Set B: [[A]-Element, [B]-Element, \u2026, [N]-Element]\n ###Steps: \n 1. Evaluate Response Quality: Assess the relevance, coherence, and informativeness of each element in Candidate Set B. \n 2. Add to Set A: Add each element from Candidate Set B to Set A to form new sets like: \n - Adding [[A]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [A]-Element] \n - Adding [[B]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [B]-Element] \n - \u2026 \n - Adding [[N]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [N]-Element] \n 3. Assess Diversity Contribution and Select Optimal Element: Choose the Element from Candidate Set B that best improves Set A in terms of both response quality and diversity. \n
Here is the Input:
Set A: [1. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Summarize the essay \"The Value of Games and Sport\"### Response:The essay \"The Value of Games and Sport\" explores the benefits that come from engaging in physical activities such as sports and games. It argues that these activities have the potential to foster physical health, cognitive development, and social skills. Moreover, engaging in these activities can be a great source of fun, relaxation, and proper exercise. It is concluded that playing games and participating in sports has the ability to positively influence the mental and physical well-being of individuals.

2. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Brainstorm five other methods that can be used to measure air quality.### Response:Five other methods to measure air quality include:- monitoring of ambient air concentrations and emissions- remote sensing techniques (e.g. satellite imaging)- establishing regional monitoring networks- collecting and analyzing in-situ samples - measurement of deposition and aerosols.

3. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Brainstorm 5 ideas for creating a better world.### Response:1. Invest in renewable energy sources, such as solar and wind energy. 2. Increase access to education for everyone, regardless of location or socio-economic status. 3. Plant more trees to help fight climate change and purify air quality. 4. Implement sustainable farming practices to preserve land and protect native species. 5. Establish public policies that value and protect human rights.

4. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:If a customer calls and wants to find out their order status, what steps should the customer service representative take?### Response:When a customer calls to find out their order status, the customer service representative should first ask the customer for their order number. Then they should check the customer's order history and provide the customer with an update on their order status. This can include information such as order processing, shipping, and estimated delivery times. Finally, the customer service representative should provide the customer with a confirmation number for their call.

5. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Compare and contrast capitalism and socialism.### Response:Capitalism and socialism are two opposite economic systems. In capitalism, the means of production and distribution are privately owned, and the market determines what goods and services should be produced. In socialism, the means of production and distribution are collectively owned and managed by the people, and the government is responsible for managing the economy. Capitalism emphasizes individual achievement and profitability, while socialism emphasizes collective ownership and meeting the needs of all people equally.]

Candidate Set B: [[[A]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Construct a sentence where all words start with the same letter### Input:NA### Response:Superstorm Sandy smashed several structures severely.

[[B]]: Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Create a recipe for making a simple but tasty burger dip.### Response:Ingredients:- 8 ounces cream cheese, softened- 1 can (10 oz) diced tomatoes with green chilies, drained- 1 cup shredded cheddar cheese- 1/2 cup sliced scallions- 1/2 teaspoon garlic powder- 1/4 teaspoon salt- 1/4 teaspoon ground pepper- 1/4 cup sliced pickled jalape\u00f1os (optional)Instructions:1. Preheat oven to 350\u00b0F.2. In a medium bowl, mix the cream cheese, tomatoes, cheddar, scallions, garlic powder, salt, and pepper until combined.3. Spread the mixture evenly in a 9-inch pie plate. Top with the pickled jalape\u00f1os, if using.4. Bake for 25 minutes until the cheese is melted and bubbly.5. Let cool for 5 minutes before serving. Serve with chips, crackers, or veggies.

[[C]]: Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Explain why socialism is considered an economic system of equality.### Response:Socialism is considered an economic system of equality because it focuses on distributing resources and wealth between different individuals in a fair and equitable way. Socialism seeks to reduce the gap between the wealthy and the poor and provide everyone with access to basic goods and services. It emphasizes social cooperation, collective decision-making, and public ownership of the means of production and distribtion. This system of economic and social organization provides a level of equality which is not present in other economic systems.

[[D]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Rewrite given statement so that it is in the third person.### Input:I am excited to go shopping.### Response:She is excited to go shopping.

[[E]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Tell me if the following statement is true or false.### Input:Two bars of a piano have the same frequency.### Response:False] \n Finally, ONLY return the index of selected element from Set B by strictly following the format: [A] for the first one, [B] for the second one, etc. And in the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias..\n ### Your Decision:"}],

"response": "After evaluating the response quality and assessing the diversity contribution, I would suggest adding element [C]-Element to Set A. \n\nThis element is a well-written response that provides a clear and concise explanation of why socialism is considered an economic system of equality. It fulfills the criteria of being informative, relevant, and coherent. Additionally, it introduces a unique perspective to Set A, which is focused on task completion and customer service. The other candidates either repeat information already present in Set A or do not add significant value to the existing set.\n\nTherefore, I recommend adding [C]-Element to Set A, as it maximizes diversity and quality of responses in the set."}

Figure 9: Selection case analysis.

{"role": "system", "content": "You are a useful Assistant to help select the optimal element. You will receive two sets: Set A and Candidate Set B. Each element in both sets is a triplet [instruction, input, response]'. Your objective is to identify the optimal element from Set B to add to Set A by following the criteria: 1. Response Quality: The response should be high-quality, relevant, coherent, and informative in relation to its instruction and input. 2. Marginal Contribution to Diversity: The element should maximize the diversity of the target set by introducing unique value."},
{"role": "user", "content": "### Example: \n Set A: [Element_1, Element_2, \u2026, Element_N] \n Candidate Set B: [[A]-Element, [B]-Element, \u2026, [N]-Element]\n ###Steps: \n 1. Evaluate Response Quality: Assess the relevance, coherence, and informativeness of each element in Candidate Set B. \n 2. Add to Set A: Add each element from Candidate Set B to Set A to form new sets like: \n - Adding [[A]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [A]-Element] \n - Adding [[B]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [B]-Element] \n - \u2026 \n - Adding [[N]-Element] to Set A: [Element_1, Element_2, \u2026, Element_N, [N]-Element] \n 3. Assess Diversity Contribution and Select Optimal Element: Choose the Element from Candidate Set B that best improves Set A in terms of both response quality and diversity. \n
Here is the Input:
Set A: [1. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Give an example of an algorithm and explain what it does.### Response:An example of an algorithm is Quicksort. Quicksort is a sorting algorithm that uses a form of divide and conquer to quickly sort an array of data. It works by Divide by selecting a pivot element from the array and partitioning the other elements into two sub-arrays, according to whether they are less than or greater than the pivot element. It then recursively sorts the sub-arrays.

2. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Summarize the essay \"The Value of Games and Sport\"### Response:The essay \"The Value of Games and Sport\" explores the benefits that come from engaging in physical activities such as sports and games. It argues that these activities have the potential to foster physical health, cognitive development, and social skills. Moreover, engaging in these activities can be a great source of fun, relaxation, and proper exercise. It is concluded that playing games and participating in sports has the ability to positively influence the mental and physical well-being of individuals.

3. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Produce a list of 5 strategies to conserve water.### Response:1. Install water-efficient fixtures and appliances in your home.2. Collect and reuse household water, such as from showering, laundry and dishwashing.3. Minimize outdoor irrigation and use water-efficient landscaping practices.4. Utilize rainwater harvesting and water recycling systems.5. Actively reduce water waste and spread awareness of water conservation.

4. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Give an example of the use of robotics in the medical field.### Response:An example of the use of robotics in the medical field is with robotic-assisted surgery. In this type of surgery, a surgeon uses a robotic system to perform precise, delicate operations through smaller incisions and with greater accuracy. This can also help reduce patient recovery times.

5. Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Generate an article title for a news article discussing the side-effects of the coronavirus mandated mask-wearing.### Response:The Unexpected and Unseen Consequences of Wearing a Mask During the Coronavirus Pandemic.]

Candidate Set B: [[[A]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Create a \u201cWhat would you do?\u201d scenario based on the following situation.### Input:Your car has run out of gas in the middle of nowhere### Response:What would you do if your car ran out of gas in the middle of nowhere? You would have to figure out a creative way to figure out your location and find the nearest gas station. You could try to find help from local businesses, flag down passing cars, or even try to fix the car yourself. Ultimately, the best solution would depend on your resources and the amount of time available.

[[B]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Following the input, list two ways companies can use facial recognition technology to enhance customer experience.### Input:Facial Recognition### Response:Companies can use facial recognition technology to enhance customer experience in several ways. First, facial recognition enables customers to quickly verify their identity, eliminating long lines and wait times. Second, facial recognition can be used to detect emotions and reactions to capture customer feedback in real-time.

[[C]]: Below is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Summarize the given article and keep the summary to less than 50 words.### Input:https://www.sciencemag.org/news/2019/04/biased-algorithms-are-killing-people-prominent-ai-researchers-warn### Response:AI algorithms have experienced moral hazard due to imprecise decision-making processes, and AI researchers are warning that if the issue isn\u2019t addressed, algorithms will result in wrongful death and other adverse outcomes. AI researchers are issuing warnings that biased algorithms have the potential to cause serious harm.

[[D]]: Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Convert the RGB color (255,153,0) to a hex color code.### Response:#FF9900

[[E]]: Below is an instruction that describes a task. Write a response that appropriately completes the request.### Instruction:Summarize what we know about the coronavirus.### Response:The coronavirus, also known as COVID-19, is a contagious virus that affects the respiratory system and is believed to have originated in China. It has been spreading rapidly around the world and is now considered a pandemic by the World Health Organization. Symptoms include fever, cough, shortness of breath, and difficulty breathing. There is no known cure for the virus, but treatments are available to help reduce symptoms and prevent complications. Prevention measures include frequent hand-washing, wearing a face mask, avoiding close contact with people who are sick, and staying up-to-date on the latest travel advisories.] \n Finally, ONLY return the index of selected element from Set B by strictly following the format: [A] for the first one, [B] for the second one, etc. And in the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias..\n ### Your Decision:"}],

"response": "The optimal element to add to Set A is [B]-Element: \n\nBelow is an instruction that describes a task, paired with an input that provides further context. Write a response that appropriately completes the request.### Instruction:Following the input, list two ways companies can use facial recognition technology to enhance customer experience.### Input:Facial Recognition### Response:Companies can use facial recognition technology to enhance customer experience in several ways. First, facial recognition enables customers to quickly verify their identity, eliminating long lines and wait times. Second, facial recognition can be used to detect emotions and reactions to capture customer feedback in real-time. \n\nEvaluation: \n\nResponse Quality: \nThe response provides two clear and specific ways that companies can use facial recognition technology to enhance customer experience. The language used is easy to understand and free from errors or ambiguity. Hence, the response quality is high. \n\nMarginal Contribution to Diversity: \nThis response adds diversity to Set A because it addresses a unique topic (facial recognition technology for customer experience) that is not covered by any of the existing elements in Set A. It is also more practical and business-oriented compared to the other responses, which are more academic. Hence, this response has a significant contribution to the diversity of Set A. \n\nTherefore, based on the high response quality and significant contribution to diversity, [B]-Element is the optimal element to add to Set A."

Figure 10: Selection case analysis.