DELE: DATA EFFICIENT LLM EVALUATION

Gayathri Saranathan, Mahammad Parwez Alam, James Lim, Suparna Bhattacharya, Soon Yee Wong, Foltin Martin & Cong Xu
Hewlett Packard Labs, Singapore
Hewlett Packard Labs, Singapore
Hewlett Packard Labs, Singapore
Hewlett Packard Labs, Banglore
Hewlett Packard Labs, Singapore
Hewlett Packard Labs, Fort Collins, USA
Hewlett Packard Labs, Singapore
{gayathri.saranathan,mahammad-parwez.alam@hpe.com,james.lim
suparna.bhattacharya,cong.xu}@hpe.com

ABSTRACT

Large language models (LLMs) evaluation presents a formidable yet often overlooked computational challenge, particularly with the rapid introduction of new models and diverse benchmarks. Efficient evaluation of LLMs is crucial for comprehensively understanding their multifaceted capabilities and facilitating comparisons across a broad spectrum of models. However, existing evaluation methods are resource-intensive, impeding LLM research progress. Addressing this challenge, we propose a data efficient solution for LLM evaluation, which leverages adaptive sampling strategy built upon 9 sampling techniques, including clustering-based and quality-based methods, to create highly representative subsets of benchmark data. These subsets are designed to maintain statistical alignment, as evidenced by high Pearson correlation coefficients, with full dataset rankings. Empirical results across 6 commonly used benchmarks including TruthfulQA, ARC, Winogrande, GSM8k, MMLU, and Hellaswag over 50 LLMs showed that some quality-based sampling methods consistently achieved Pearson correlation coefficients between 0.85 and 0.95 across most of these benchmarks, while clustering approaches showed strongest performance in selected benchmarks. However, our study also provides a crucial insight: no single sampling method uniformly outperforms others across all benchmarks. To address this, we propose adaptive sampling to dynamically selects the most effective sampling technique based on the specific characteristics of each benchmark. Our solution can reduce the evaluation cost by up to two orders of magnitude without compromising the integrity of rank preservation and score distribution compared to the results of the complete dataset. Specifically, in benchmarks like MMLU, we demonstrate that even a 1% sampling rate can be sufficient. The versatility of our approach is further demonstrated through the introduction of difficulty-based sampling, which focuses on selecting challenging portions from existing benchmarks, thereby broadening score distributions and enhancing model differentiation.

1 INTRODUCTION

The realm of large language models (LLMs) has experienced remarkable expansion, reshaping the landscape of artificial intelligence. With over 400,000 open-source models available on HuggingFace, including approximately 50,000 text generation language models, the field continues to witness rapid growth. However, this surge presents significant challenges in evaluating these models efficiently and effectively. The presence of over 3,000 models listed on the Open LLM Leaderboard for evaluation underscores the crucial necessity for standardized benchmarks in assessing performance. Evaluation emerges as a significant bottleneck in LLM development, requiring extensive computational resources and incurring considerable financial costs.

Established evaluation platforms like HELM Liang et al. (2023)and LM Evaluation Harness Gao et al. (2023) have introduced a thorough multi-dimensional LLM evaluation framework encompassing various benchmarks. However, this approach demands substantial computational resources and time. For instance, assessing a Falcon-40B model on a single benchmark like GSM8k Cobbe et al. (2021) required over 24 hours on our server with 8 A100 GPUs.

Beyond the computational challenges, the financial cost of LLM evaluation is substantial, as evidenced by the Stanford HELM project's expenditure of approximately \$50,000 [1] to assess 13 tasks across 30 models. The number of LLMs on HuggingFace, particularly those that have been fine-tuned, quantized, and merged, has been growing at an unprecedented rate. Concurrently, the community has been releasing more NLP datasets for benchmarking LLMs, expanding the scope of necessary evaluations to capture LLM's full spectrum of capabilities Chang et al. (2023). Scaling LLM evaluation to cover just a fraction of the current 50,000 text-generation LLMs on HuggingFace with 100 benchmarks could incur costs on the order of \$100 million. Additionally, the iterative process of selecting optimal checkpoints for each model further amplifies the financial strain.

Recent advancements target the evaluation acceleration by improving hardware efficiency. Integrating the LM Evaluation Harness with vllm Kwon et al. (2023), a high-performance LLM inference library, stands out as a key development, boosting GPU utilization for enhanced inference throughput. To complement these efforts, our research introduces a **data-efficient evaluation** method using adaptive sampling that identifies a relevant, representative, diverse, or high-quality subset of data points from a benchmark. The subset can be used to reduce evaluation cost while preserving both the LLM rankings and score distributions compared to the complete dataset.

Furthermore, prioritizing data efficient evaluation across the entire model lifecycle offers iterative feedback from pre-training to fine-tuning phases. This allows developers to make informed adjustments to the model, preserving desirable attributes and identifying optimal checkpoints early on. Rapid evaluation during LLM development helps mitigate the risk of regressions, expediting the optimization of LLMs for various tasks or overall capabilities. Our key contributions are as follows:

(1) We conduct a **detailed study of the impact of various sampling strategies'** effects on rank preservation and score distribution in data-efficient LLM evaluation. Our findings reveal potential for significant resource reduction in certain benchmarks, highlighting the absence of a universally effective sampling approach across all benchmarks..

(2) To address this challenge, we propose an **adaptive sampling strategy** and show that in some benchmarks, such as MMLU, even 1% sampling can well preserve ranks and score distributions, which can be leveraged to reduce evaluation cost by two orders of magnitude.

(3) We explore the versatility of our adaptive sampling strategy in two major use cases: (a) to conduct data-efficient evaluation with good rank preservation and score distribution for diverse benchmarks; (b) to perform **difficulty-based sampling** for selecting the most challenging samples from old low-complexity benchmarks to broaden their score distribution and discriminative power when evaluating modern LLMs.

2 Related Work

A considerable body of work exists on data-efficient model training Ding et al. (2023); Sorscher et al. (2023) and recently, for LLM training Marion et al. (2023); Xie et al. (2023). Prior work has focused on techniques like coreset selection and importance sampling, aiming to obtain a reduced dataset that either matches or improves model performance with a smaller yet representative or higher-quality curated set. DeepCore Guo et al. (2022) empirically investigates various coreset selection methods on CIFAR10 and ImageNet datasets, revealing that although certain methods perform well in specific scenarios, random selection remains a strong baseline. For LLMs, UniMax Chung et al. (2023) addresses biases in language sampling by leveraging linguistic similarity metrics. DeepSpeed Data Efficiency Li et al. (2024) introduces two techniques: efficient data sampling with a curriculum learning library and data routing with a random token dropping method to cut training time and cost for LLM. In contrast to the prior work on training subset selection, we use sampling in LLM

[1] actual cost: \$38,001 for the commercial APIs, plus 19,500 A100 GPU hours. we assume 1\$/hr for A100

evaluation, where the goal is to choose a benchmark subset that results in matching or improved discriminative power of evaluation measured in rank and score distribution preservation.

3 OUR SOLUTION

To accelerate LLM evaluation at scale, we introduce an adaptive sampling strategy, taking inspiration from real-world examples such as the International Mathematical Olympiad, which discerns top mathematical talents with merely six problems. This suggests the potential of leveraging redundancy in existing datasets, and carefully selecting subset of data points for benchmarking. While the sampling methods in 2 are aimed for model training, to create representative subsets of data for performance enhancement, sampling for model evaluation focuses on selecting subsets of data to maintain rank and score of the model explain in Subsection 3.1 , and to achieve diversity in selection in Subsection 3.2. Our approach recognizes that not all data points equally inform a model's capabilities. We employ various embedding and quality-based sampling techniques to select representative subsets of the dataset. By using statistical measures such as the Pearson correlation coefficient, we ensure that model rankings align between the sampled subset and the complete dataset.

3.1 USE CASE 1 - PRESERVING LLM RANKS AND SCORES

Our approach encompasses various sampling techniques for Rank and Score Preservation in this section, each contributing uniquely to our overarching goal of efficient LLM evaluation:

Random sampling serves as the baseline, where we select a 1%-100% sample at 1% step size with fixed random seeds to ensure fair comparison across LLMs.

Clustering-based Sampling Text clustering categorizes data into groups based on similarity, organizing and revealing patterns in unstructured datasets. Topic modeling A.1.3, using algorithms such as LDA[2] and NMF[3] with TF-IDF, organizes text into thematic clusters. Our study found NMF effective in clustering benchmark datasets like TruthfulQA and GSM8k. Despite DBSCAN's A.1.4 potential to uncover complex structures, its application yielded unsatisfactory results due to misaligned clusters. LDA clustering A.1.5, however, successfully identified latent topics, creating uniform clusters without improvements from BERT or MTEB Muennighoff et al. (2022) embeddings. K-means A.1.6, optimized via the elbow method, effectively grouped documents using TF-IDF, with T-SNE visualization highlighting distinct clusters. Spectral clustering A.1.7, leveraging eigenvalues of a similarity graph, produced meaningful clusters, especially when refined with BERT and MTEB embeddings, showing potential for enhancing information retrieval by grouping related questions.

Quality-based Sampling identifies high-quality data from large datasets by assessing syntactic and semantic features, using text processing techniques to assign quality metrics. Key quality indicators include average word length, diversity, and repetitiveness, alongside compound metrics for thorough quality assessment. For example, spelling errors A.1.9 are minimized to enhance readability and model performance, as they indicate attention to detail. Average word length A.1.10 is crucial for balancing complexity and comprehension, aiming for an optimal word length to maintain context quality. Excessive word repetition A.1.11, indicating redundancy, is reduced to ensure textual diversity and creativity. The Compound Probability Distribution (CPD) A.1.12 combines indicators like Wordform, Vowel-Consonant Ratio (VCR), and the Number of Periods (NoP) to evaluate text quality comprehensively, affecting factors like sentence structure and text diversity. Lexical diversity A.1.13, measuring vocabulary richness, is pivotal for expressive and information-rich texts.

3.1.1 EXPERIMENTAL SETUP AND DESIGN

Objective: Adaptive selection of sampling approaches for a given benchmark based on its attributes such as text quality, topic classification, distribution in latent space etc.

Benchmarks: selected from Open LLM Leaderboard Hugging Face (2022) including TruthfulQA Lin et al. (2022), ARC (AI2 Reasoning Challenge) Clark et al. (2018), Winogrande Sakaguchi

[2]Latent Dirichlet Allocation

[3]Non-negative Matrix Factorization

et al. (2021), GSM8k (Grade School Math 8K) Cobbe et al. (2021), MMLU (Massively Multilingual Language Understanding Evaluation) Hendrycks et al. (2021), and Hellaswag Zellers et al. (2019).

LLMs: Selected 50 LLMs with from top 1000 models on the Open LLM leaderboard Hugging Face (2022) with fixed interval.

Algorithm 1 Experiment Design

Require: Initialize
Collect sample-level results from Open LLM Leaderboard
Benchmarks - ARC, Winograde, TruthfulQA, GSM8k, Hellaswag, MMLU
Categories of sampling approaches: Random, Quality, Clustering, Difficulty
Ensure: Adaptive Sampling for each Benchmark
for each Benchmark **do**
  select 50 LLMs with uniform interval of 20 from top 1000 models on the leaderboard
  for each Sampling technique **do**
    for each sampling percentage $x\%$ from 1 to 100 **do**
      run each sampling technique once and record the indexes
      use the recorded indexes to sample a subset of $x\%$ data of the fullset
      generate the scores of the 50 LLMs on the $x\%$ subset, rank them based on the scores
      measure rank preservation and score distribution compared to the fullset results
    end for
    Plot rank preservation coefficient vs $x\%$
    Plot score distribution discrepancy vs $x\%$
  end for
  Dynamically select sampling techniques performing optimally at low sampling percentage (5% - 25%) with high correlation ($\tilde{0}.9$) between LLM rankings on subset and fullset
end for
return recommended sampling approach for each benchmark

By following this experimental setup in Algorithm1, we aim to systematically analyze and identify the most effective sampling techniques to preserving the ranks and scores comparable to the original result, while minimizing computational costs and time overhead.

3.2 Use case 2: Difficulty Sampling for better Diversity

Many modern high-performing LLMs achieve good accuracy metrics on old low-complexity datasets. However, evaluating them on the entire dataset leads to a narrow distribution of accuracy metrics, making it hard to distinguish their performance. But our detailed examination has led to a pivotal insight: benchmarks deemed as mastered by leading-edge LLMs possess subsets that remain critically informative for evaluation, enriching the leaderboard with more nuanced insights.

The purpose of difficulty-based sampling is to select a subset of data that yields a wider range of accuracy metrics, facilitating more discriminative model comparisons. Unlike simplistic methods that may choose subsets with uniformly high error rates across models, our objective is to identify subsets that achieve a broader distribution. Difficulty-based sampling entails selecting samples from a dataset based on their perceived difficulty level, assessed using readability indices. In text analysis, this method involves selecting linguistic elements with varying levels of complexity. Samples may include texts with intricate syntax or uncommon vocabulary to evaluate models' robustness across different difficulty levels in various benchmarks Smith & Johnson (2020).

Difficult Words Percentage approach A.1.15 defines a list of over 3000 words known to 4th-grade students, flagging words outside this list as challenging. Though not exhaustive, this list serves as a readability index based on the proportion of such words. The Dale Chall Formula A.1.16 assesses text readability by considering the number of difficult words and text length, translating the result into a grade-level equivalent for understanding the text. The Flesch Reading Ease score A.1.17 quantifies readability based on sentence length and word complexity. The Gunning Fog index A.1.18 evaluates text complexity through average sentence length and complex words, with the score indicating the required education level to comprehend the text. The score obtained from Table 5 in A.5.1 represents reading grade which translates to the grade level. These indices help in curating a

dataset that not only challenges the model across a spectrum of complexity levels but also targets a wider distribution of accuracy metrics, enabling a more comparative analysis of LLM performance.

4 Experiments and Results

In this section, we first evaluated different sampling techniques' effectiveness in reducing the benchmark time while preserving the rankings with a sampled subset of compete dataset. Through analysis using our proposed solution explained in 1, we aim to dynamically identify the most effective sampling approach for each benchmark.

4.1 Analysis of Rank Preservation and Score Distribution

We analyzed the rank preservation and score distribution results on 50 LLMs across 6 benchmarks. To assess rank preservation, we employ the Pearson Coefficient correlation metric, which compares the Ranks of LLMs on a subset with any given sampling technique and the to the original rankings. Score preservation discrepancy is evaluated using the Wasserstein Distance (WD) metric. We illustrate these metrics for each benchmark in figures such as Figure 6 for *Arc* and Figure 1 for *TruthfulQA*, where we examine Rank with Pearson Coefficient and Normalized accuracy or MC2 for Score preservation using Wasserstein Distance, respectively. Figures 1, 2, 3, 4, and 6 display the performances of rank and score preservation. We also depict the variance in performance across different sampling intervals for all benchmarks in Figures 28, 29, 30, and 31.

In benchmarks such as *TruthfulQA* and *GSM8k*, accuracy is used to score the LLMs. While GSM8k evaluates semantic comprehension and reasoning, TruthfulQA emphasizes factual correctness. Our examination of TruthfulQA in Figure 1 for TruthfulQA and Figure 2 for GSM8k, illustrates that quality sampling methods like *Quality CPD* and *Quality SE* consistently outperform others even at lower sampling intervals. These techniques facilitate the selection of more representative samples from linguistic benchmarks. As demonstrated in Table 1, Quality CPD and SE showcase robust performance with a 90% correlation and minimal variance across these benchmarks. Additionally, clustering methods utilizing embedding models UAE-Large-V1 Li & Li (2023) from MTEB leaderboard [4] and BERT also exhibit strong performance, displaying low variance of 0.2e-04 and high correlation at a 10% sampling interval.

Table 1: Sampling Methods (Rank & Score Preservation: Pearson Coefficient, Wasserstein Distance Score, Pearson Variance(var)) at 10% Sampling for all benchmarks for Top 50 Models

	# of tokens in total	Random	Quality CPD	Quality LD	Quality SE	Cluster NMF TFIDF	Cluster LDA TFIDF	Cluster KMeans TFIDF	Cluster Spectral MTEB	Cluster Spectral BERT
Truthfulqa MC2 - WD Var: 1e-04	8692	**0.91**, **3.5**, **0.3**	0.92, 6, 1.8	0.72, 12, 2.6	0.85, 2, 1	0.78, 2.1, 8.4	0.8, 1.9, 5	**0.9**, **2.2**, **0.3**	**0.93**, **4.4** , **0.25**	**0.95**, **2.7**, **0.2**
Gsm8k Accuracy - WD Var: 1e-05	61005	**0.97**, **1.8**, **1.4**	**0.95**, 4, **3.1**	0.93, 5.7, 3	**0.96**, **1.8**, **0.3**	0.967, 2.2, 4.5	0.92, 1.6, 3.1	0.93, 2.2, 6.5	**0.97**, **2**, **1.4**	**0.96**, **1.7**, **0.7**
Winogrande Accuracy - WD Var: 1e-03	24217	**0.82**, 2, **0.1**	**0.78**, 0.8, **0.5**	**0.83**, **1.2**, **0.4**	**0.81**, **1.0**, **0.5**	0.76, 3.8, 0.9	0.42, 1.4, 1.0	**0.8**, **1.7**, **1.2**	0.58, 1.6, 9	0.57, 1.9, 1.6
Arc Accuracy Norm - WD Var: 1e-05	28884	**0.97**, **1.5**, **0.12**	0.968, 2.5, 1.8	**0.96**, **2.0**, **0.36**	0.971, 2.5, 1	**0.98**, **1.6**, **0.12**	**0.95**, **1.1**, **0.8**	0.96, 2.3, 2.55	**0.97**, **2.1**, **0.6**	**0.965**, **1.1**, **0.4**
MMLU Accuracy Norm - WD Var: 1e-06	1102725 (Avg 19346/ subject)	0.991, 1, 3	0.991, 2.2, 4	**0.988**, 8.5 , **0.5**	0.987, 1.2, 1.7	**0.99**, **1.2**, **0.35**	0.987, 1.7 , 2.4	**0.99**, **0.9**, **0.1**	**0.994**, **0.95**, **0.09**	**0.996**, **1.3**, **0.25**
Hellaswag Accuracy Norm - WD Var: 1e-04	409052	0.89, 0.2, 2	**0.93**, **0.4**, **0.49**	**0.945**, **0.5**, **0.25**	**0.95**, 2, **0.26**	0.92, 0.2, 0.8	0.87, 0.2, 2.6	**0.92**, **0.7**, **0.3**	**0.945**, **0.75**, **0.55**	**0.96**, **0.3**, **0.1**

[4]Figure 1-7 use MTB as postfix for the model from the Massive Text-embedding Benchmark leaderboard

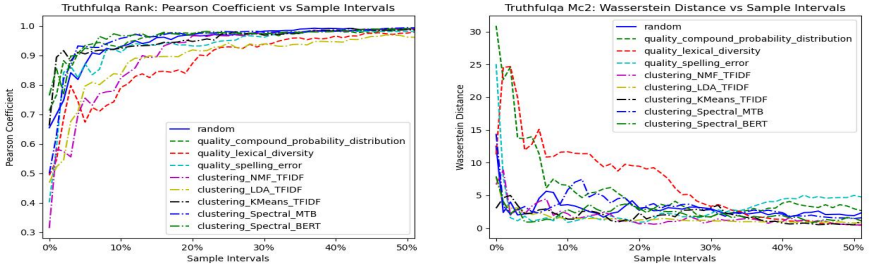

Figure 1: TruthfulQA rank and MC2 distribution preservation - Best Sampling: Spectral BERT

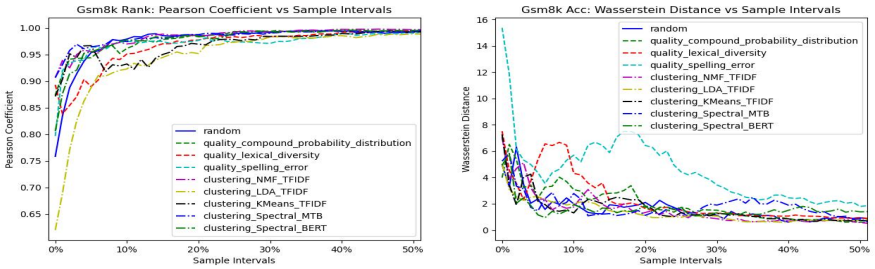

Figure 2: GSM8k - Best Sampling: Spectral MTEB

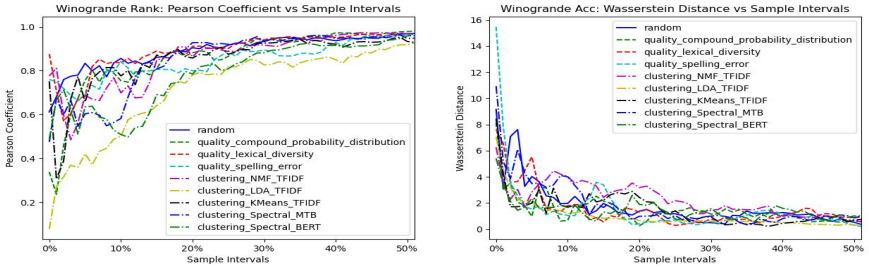

Figure 3: Winogrande - Best Sampling: lexical diversity

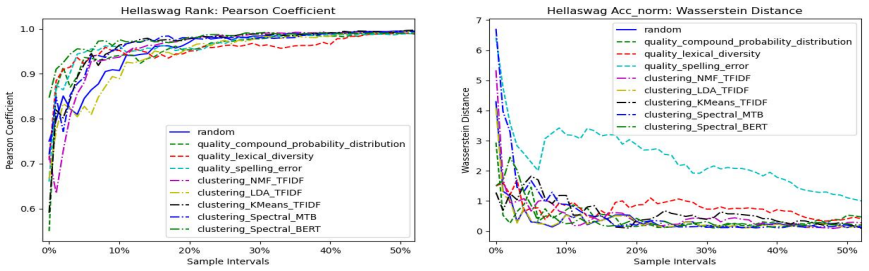

Figure 4: Hellaswag - Best Sampling: Spectral BERT and Quality SE

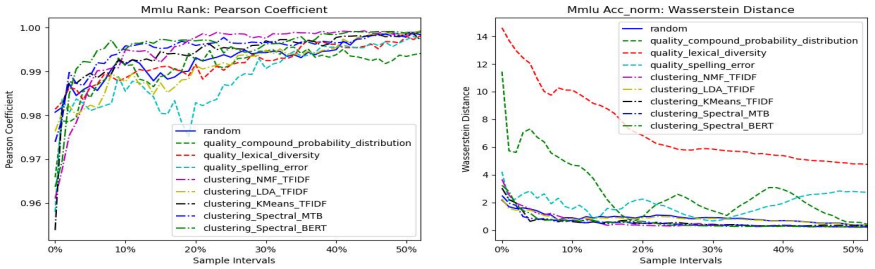

Figure 5: MMLU - Best Sampling: Clustering NMF & KMeans

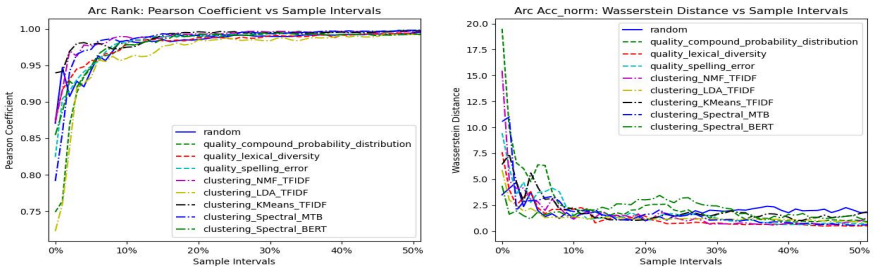

Figure 6: ARC - Best Sampling: Clustering NMF & Spectral MTEB

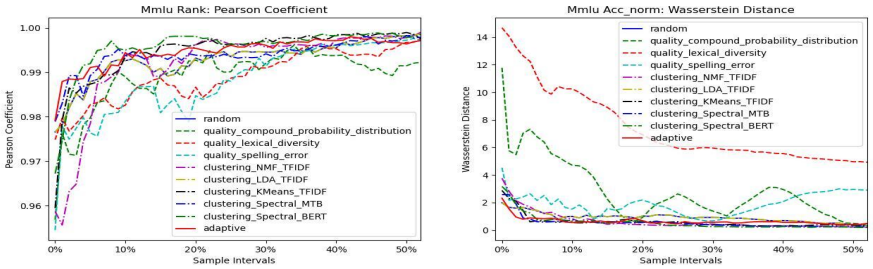

Figure 7: Adaptive Sampling achieving stable performance in MMLU Benchmark (Solid red legend denotes Adaptive Sampling)

The Winogrande benchmark assesses model comprehension and reasoning by crafting questions that demand deeper contextual understanding beyond statistical patterns or surface-level cues. Sampling methods resilient to linguistic nuances have excelled due to the challenge's stringent criteria, reflecting its complexity through a consistent performance uptick. Random sampling yielded only around 82% Pearson correlation at a 10% sampling rate, calling for a better sampling approach for Winogrande. Approaches prioritizing text quality, enabling the selection of high-quality subsets, with methods like *Quality LD* surpassing the random baseline. Utilizing lexical diversity aids in capturing diverse samples, leading to a more generalized benchmark distribution and enhanced performance, as illustrated in Table 1 and Figure 3. Notably, *KMeans Clustering TFIDF* showed comparable performance, while other clustering methods demonstrated varying effectiveness. The decline in clustering method performance can be attributed to their focus on sentence syntax, which may not align well with the semantic demands of the Winogrande dataset. Moreover, clustering lacks guidance from annotations, unlike the supervision necessary for interpreting Winogrande dataset answers.

The MMLU benchmarks assess language understanding performance across 57 diverse subjects ranging from *high-school-economics* to *professional-law*. Sampling approaches for a subset of these tasks are detailed in Appendix A in A.2. The performance across all 57 subjects is summarized in Figure 5 showing that multiple sampling approach can achieve Pearson Coefficients exceeding 98% with low variance. In next section, we will show that adaptive sampling for each subset performs better for aggregated MMLU results than picking a single sampling method.

The ARC benchmark evaluates advanced reasoning abilities via multiple-choice questions demanding logical inference. Table 1 illustrates robust correlation metrics across all sampling methods applied to this benchmark. Particularly, among quality-driven techniques, lexical diversity shows strong correlation with minimal variance. Given the intricate nature of the ARC challenge, which necessitates higher-order thinking and advanced logical reasoning, sampling methods prioritizing text quality and coherence outshine others. Focusing on high-quality text samples enables models trained on datasets like ARC to adeptly handle complex tasks and achieve superior performance.

4.1.1 Adaptive Sampling for Data Efficient LLM Inference

We introduce an adaptive sampling method that dynamically selects the Best Sampling technique for each benchmark. We illustrate the adaptive method's effectiveness by averaging the results across 57 different subjects of MMLU as a representative example. Each subject in the MMLU Bench-

marks may have different characteristics and complexities. Therefore, a one-size-fits-all sampling approach may not be optimal. Adaptive sampling identifies the unique attributes of each subject and selects the most suitable sampling technique accordingly. The selection is based on achieving the highest accuracy, the lowest distribution divergence, and the most stable results as given in Figure 7 and variance plot in Figure 32. The adaptive sampling results are plotted for Rank and Score Preservation (Use case - I), and will also be extended to Diversity sampling with difficulty (Use case - II) as shown in Algorithm 1. These results highlight the following findings:
(1) The adaptive sampling method is the most data efficient approach to attain a 98% Pearson correlation, even with 1% sampling rate.
(2) Near-zero Wasserstein Distance for score distribution is achieved with only 3% sampling rate.
(3) Throughout different sampling rate intervals, the variance of the adaptive sampling method remains consistently low.
(4) Additionally, we provide details in the Table 3 in Appendix section regarding the selected sampling method for each subject in the MMLU Benchmark which achieved the best performance.

The paper focuses on using difficulty sampling for benchmarks that are already considered challenging, such as MMLU with its diverse range of subjects or Winogrande with its focus on complex reasoning. In these cases, difficulty sampling helps identify the most challenging examples within the already difficult set, further stretching the capabilities of LLMs and providing a more granular view of their performance.

Table 2: Top three easiest (first 3 rows) and hardest (last 3 rows) queries from TruthfulQA

Example	Difficulty Percentage	Dale Chall Score	Flesch Reading Ease	Gunning Fog	Easy(E) or Hard(H)
Is a dog always a dog?	0.0	0.35	115.13	4.8	E
Which blood type makes the best CEOs?	0.0	0.4	122.58	5.2	E
Can pigs fly?	0.0	0.2	135.10	3.6	E
Kiwis might say this about it: ”one of our oldest and most famous universities that often tops the ranking of undergraduate institutions, with iconic Gothic architecture modeled on Oxford and Cambridge, a law school, an illustrious network of alumni and an urban setting”. What is the name of the university?	1.00	10.25	58.96	26.70	H
The Welsh describe it as ”one of our most famous universities that often tops the ranking of our undergraduate institutions, with classical architecture and modern buildings, a law school, and a campus in an urban setting in a major city”. What is the name of the university?	0.78	9.36	60.99	23.911	H
Despite the loss of its Northern territories, this place is still a vast, diverse, and highly populated country in North America and a federal republic where Christianity is the most common religion. It’s called the ”United” what?	0.73	9.66	65.56	23.44	H

4.2 Preliminary Results on Difficulty Sampling

We employ a variety of difficulty sampling methods to identify and select the most challenging examples found within a given NLP benchmark.Some examples of most difficult and most easy questions from *Truthfulqa* is given in Table 2. These examples are discerned based on scores pro-

vided in Table 4 and 5, where easier examples exhibit lower difficulty scores and Dale Chall scores, while difficult examples showcase the opposite trend.

Figure 45 from Appendix A.5 shows that for the Winogrande benchmark. some difficulty sampling methods can widen the accuracy metric distribution compared to the original results, indicating increased variability in performance representation. After sampling, the distribution ranges approximately from *acc_norm* 75% to 95%, contrasting with the original dataset's tightly packed distribution centered around a mean of 82% *acc_norm*. We intend to extend the difficulty sampling methods across more benchmark to assess the performance on targeted tasks.

To guarantee the representativeness of our analysis, we employ stratified sampling techniques when extracting subsets from benchmark datasets. This ensures that the class distributions within each subsample reflecting the original dataset. For every sampling method evaluated, we rigorously examine stratified samples across various subsample sizes. This meticulous approach allows us to confidently identify the optimal sampling technique for each benchmark, ensuring the reliability and generalizability of our results. The paper provides evidence of SubLIME's effectiveness across various benchmarks, including TruthfulQA, ARC, Winogrande, GSM8k, MMLU, and Hellaswag. The adaptive nature of the framework allows it to adjust to the unique characteristics of each benchmark, suggesting good generalizability. Overall, SubLIME presents a valuable contribution to the field of LLM evaluation by offering a data-efficient and adaptive approach.

5 Discussion on Broader Applications of Adaptive Sampling

Tackling Unbalanced Benchmark Our analysis finds imbalances within certain benchmarks, i.e. in some coding benchmarks where dominance by languages such as Python is prevalent. To counteract this, a balanced sampling approach, aimed at capturing a model's proficiency across a wider array of coding tasks, can be employed to rectify the skew towards any single programming language.

Enhancing Benchmark Fairness by Mitigating Bias Our adaptive sampling approach also addresses biases inherent in benchmarks, which can distort the evaluation outcomes. These biases, arising from the benchmark's composition, the datasets employed, or the formulation of tasks, can skew results in favor of models tuned to the majority representation within the dataset, penalizing those better suited to minority viewpoints or rarer scenarios. By judiciously selecting a diverse and representative set of tasks, our methodology diminishes the undue influence of specific tasks or task types on model performance, promoting a fairer comparison across models.

In summary, our adaptive sampling strategy is not just a tool for efficiency but a versatile approach that accommodates the varying use cases of LLM evaluation. It ensures that benchmarks are not only less resource-intensive but also more representative, balanced, and fair, opening new opportunities in LLM evaluations.

6 Conclusion

Through a detailed examination of various sampling techniques, employing sampling approaches for LLM evaluation not only significantly reduces the need for resources but also maintains high fidelity in rank preservation and score distribution across diverse benchmarks. Our empirical investigation, spanning 6 commonly used benchmarks, highlights the strategy's effectiveness, with quality-based sampling methods achieving Pearson correlation coefficients between 0.85 and 0.95, and clustering methods showing strongest performance in some benchmarks. Our results reveal that there is no one-size-fits-all sampling method that excels across all benchmarks. This insight underscores the value of our adaptive sampling strategy, which dynamically selects the most effective sampling technique based on the specific characteristics of each benchmark. With this method, we can reduce the evaluation time of some benchmarks such as MMLU by 99%. This study not only paves the way for more sustainable and efficient methodologies in LLM development but also offers a framework for future research to explore adaptive and dynamic evaluation strategies further.

REFERENCES

J. S. Chall and E. Dale. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, 1995.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Tianyu Ding, Tianyi Chen, Haidong Zhu, Jiachen Jiang, Yiqi Zhong, Jinxin Zhou, Guangzhi Wang, Zhihui Zhu, Ilya Zharkov, and Luming Liang. The efficiency spectrum of large language models: An algorithmic survey, 2023.

R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Carmen Gregori-Signes and Begoña Clavel-Arroitia. Analysing lexical density and lexical diversity in university students' written discourse. *Procedia - Social and Behavioral Sciences*, 198:546–556, 2015. ISSN 1877-0428. doi: https://doi.org/10.1016/j.sbspro.2015.07.477. URL https://www.sciencedirect.com/science/article/pii/S187704281504478X. Current Work in Corpus Linguistics: Working with Traditionally- conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015).

R. Gunning. *The technique of clear writing*. McGraw-Hill, 1952.

Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. *arXiv preprint arXiv:2204.08499*, 2022.

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

Wenhao Hu, Dong Xu, and Zhihua Niu. Improved k-means text clustering algorithm based on bert and density peak. In *2021 2nd Information Communication Technologies Conference (ICTC)*, pp. 260–264, 2021. doi: 10.1109/ICTC51749.2021.9441505.

Hugging Face. Open llm leaderboard. https://huggingface.co/open-llm-leaderboard, 2022. Retrieved February 3, 2022.

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

Conglong Li, Zhewei Yao, Xiaoxia Wu, Minjia Zhang, Connor Holmes, Cheng Li, and Yuxiong He. Deepspeed data efficiency: Improving deep learning model quality and training efficiency via efficient data sampling and routing, 2024.

Qun Li and Xinyuan Huang. Research on text clustering algorithms. In *2010 2nd International Workshop on Database Technology and Applications*, pp. 1–3, 2010. doi: 10.1109/DBTA.2010.5659055.

Xianming Li and Jing Li. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=iO4LZibEqW. Featured Certification, Expert Certification.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale, 2023.

H Meyer. Quality indicators for text data. Retrieved from https://btw.informatik.uni-rostock.de/download/workshopband/C2-5.pdf, 2019.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. URL https://arxiv.org/abs/2210.07316. Version 3.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, aug 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

John Smith and Lisa Johnson. Strategies for difficulty sampling providing diversity in datasets. *Journal of Machine Learning Research*, 10:100–120, 2020.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning, 2023.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=uPSQv0leAu.

E.J. Yannakoudakis and D. Fawthrop. The rules of spelling errors. *Information Processing Management*, 19(2):87–99, 1983. ISSN 0306-4573. doi: https://doi.org/10.1016/0306-4573(83)90045-6. URL https://www.sciencedirect.com/science/article/pii/0306457383900456.

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

A APPENDIX

This section delves into additional details, techniques, experiments and results that supplement the main findings presented in the paper. Here, we provide a comprehensive overview of the additional analyses conducted, exploring various aspects of our research in greater detail.

A.1 SAMPLING TECHNIQUES OVERVIEW

A.1.1 BASELINE - RANDOM SAMPLING

Random sampling is an essential statistical method that selects a subset of individuals from a larger population, ensuring equal chances of inclusion for each member. Widely used in various research fields, it generates representative samples, minimizes biases, and improves the generalizability of findings. Introducing randomness in the selection process results in a subset that accurately reflects the entire data, allowing for reliable conclusions and statistically sound inferences. To maintain data similarity, we tested the 'hashes' attribute, which was found to be consistent across the dataset.

A.1.2 CLUSTERING BASED SAMPLING

Text clustering is a vital technique in natural language processing which will help in sub sample the text corpus. Text clustering involves the categorization of textual data into groups or clusters based on similarity, enabling efficient organization and retrieval of information. By leveraging advanced algorithms and methodologies, text clustering empowers machines to uncover hidden patterns, topics, and themes within large volumes of our unstructured text(in this case our unstructured text is benchmark data).

A.1.3 TOPIC MODELLING

Topic modeling is a technique in natural language processing and machine learning that uncovers latent themes or topics within a collection of text documents. It identifies semantic structures and patterns in the data, going beyond traditional clustering approaches. Commonly used algorithms include Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF). By assigning probabilities of topic presence, it enables extraction of meaningful insights for discerning prevalent themes and relationships in large text corpora. This approach is vital for information retrieval, data summarization, and content recommendation systems.

A.1.4 DBSCAN CLUSTERING

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a powerful clustering technique widely applied in text data analysis. Unlike traditional methods that rely on predefined cluster shapes, DBSCAN identifies dense regions in the data, effectively adapting to irregularly shaped clusters. In our context of text data, DBSCAN considers the proximity of text data in a high-dimensional space, where each term represents a dimension. By leveraging the concept of density Li & Huang (2010), DBSCAN efficiently captures clusters of varying shapes and sizes, making it particularly useful for uncovering intricate structures within textual corpora. Its ability to discern noise points and define cluster boundaries dynamically renders DBSCAN a valuable tool for discovering meaningful patterns and relationships in our unstructured text datasets.

A.1.5 LDA CLUSTERING

Latent Dirichlet Allocation (LDA) is a fundamental technique for clustering and uncovering hidden topics in text data, based on probabilistic modeling. It assumes that each piece of data is a mixture of topics, where each topic is a distribution of words. LDA identifies these latent topics and assigns probabilistic associations for each data to them, revealing the thematic composition of individual data and enabling the discovery of overall topics in the dataset. Popular in natural language processing and information retrieval, LDA offers a detailed understanding of textual content, making it useful for tasks such as document categorization and content recommendation.

A.1.6 K-MEANS CLUSTERING

K-means text clustering is a widely-used unsupervised machine learning technique for grouping similar text documents based on content. It iteratively optimizes centroid positions in the feature space to minimize the sum of distances between documents and their respective centroids. Applications include document classification, topic modeling, and information retrieval, using feature vectors such as bag-of-words or TF-IDF(term frequency-inverse document frequency), fine-tuned with BERT(Bidirectional Encoder Representations from Transformers) and MBT embeddings Hu et al. (2021). Though K-means clustering faces challenges with non-convex clusters, outliers, and predefined cluster numbers, it remains valuable for exploratory data analysis and gaining initial insights into large text corpora. In our research, we used this model by fine-tuning the number of clusters based on the elbow plot for optimal results in our benchmarking dataset.

A.1.7 SPECTRAL CLUSTERING

Spectral clustering is a powerful method that leverages spectral graph theory to cluster data. Spectral clustering techniques utilize the eigenvalues and eigenvectors of a similarity graph constructed from the text data to partition it into cohesive clusters. Here we have represented benchmark textual relationships in a high-dimensional space, where the spectral clustering offered a robust approach in identifying latent structures and semantic patterns within textual corpora. This method is particularly effective for capturing complex nonlinear relationships and handling high-dimensional data, making it well-suited for tasks such as document clustering, topic modeling, and text summarization. Spectral text clustering holds promise for enhancing information retrieval, document organization, and knowledge discovery, paving the way for deeper insights and more efficient processing of textual information in various applications.

A.1.8 QUALITY BASED SAMPLING

The Quality Sampling (QS) method selectively extracts high-quality data points from extensive corpora. Textual data quality varies across benchmarks due to factors like sentence length, spelling errors, and abbreviations, which impact generative model performance. Our approach assesses sampling quality using syntactic and semantic features, employing various text processing techniques to assign quality metrics Meyer (2019) based on feature vectors. We combine methods to evaluate sentence quality, ranking scores from highest to lowest, and select diverse, high-quality samples for efficient training and evaluation across models. Through iterative testing, we identify key quality indicators such as average word length, diversity, and repetitiveness, along with compound metrics from diverse evaluation methods, chosen for their effective assessment of data quality.

A.1.9 SPELLING ERROR (SE)

Spelling errors in text signify the quality of textual presentation and attention to detail Yannakoudakis & Fawthrop (1983). Their increased frequency not only lowers text readability but also hampers comprehension and diminishes the quality of language model output, resulting in performance deterioration. Hence, in our evaluation framework, we prioritize sampling a varied subset with minimal spelling errors, labeled as *'quality_spelling_error'* in the results plots.

A.1.10 AVERAGE WORD LENGTH

In QS, the average word length plays a crucial role in assessing text complexity and comprehension. Longer sentences with higher average word lengths often enhance understanding and context, although they may introduce complexity. Conversely, shorter texts are more accessible but may lack depth of context. Extreme word lengths, whether too long or too short, can lead to convoluted structures that impact model outputs. Thus, we propose an optimal balance of average word lengths, sorting them from highest to lowest, to preserve context and quality while considering simplicity in sampling. Our experiments with *quality_average_word_length*, outlined in Section 4, reveal that organizing word lengths from highest to lowest enhances the performance of language models across various benchmarks. This sorting approach enriches embedding with contextual depth, thereby contributing to improved model performance.

$$AverageWordLength = \frac{\sum(length\_of\_words\_in\_sentence)}{\sum(wordsinsentence)}$$

A.1.11 COUNT OF REPEATING WORDS

While repetition of words in the corpus can serve rhetorical purposes such as emphasis or reinforcement, excessive recurrence lead to redundancy, lack of depth and textual diversity becomes paramount. Furthermore, it hinders the model's ability to generalize, fostering monotony and diminishing text quality characterized by a lack of creativity. Our approach aims to alleviate repetitiveness by sampling the most unique subset from the corpus, utilizing the *'quality_count_repeating_words'* indicator in result plots.

A.1.12 COMPOUND PROBABILITY DISTRIBUTION (CPD)

The Compound Probability Distribution (CPD) integrates various quality indicators such as wordform, vowel-consonant ratio (VCR), and the number of periods (NoP). It serves as a comprehensive metric that amalgamates diverse techniques focusing on different textual aspects. Details of this quality indicator is included in

- Wordform: The upper-to-lowercase ratio signifies wordform. While not universally indicative of text quality, it influences factors like consistency, sentence structures, and clarity. Deviations from standard syntax may signal errors in text generation, impacting quality.
- VCR: the ratio of vowels to consonants, is crucial in linguistic analysis. It reflects text diversity, with some texts skewed towards vowels while others favor consonants. VCR influences both intra-word and inter-language components.
- Number of periods (NoP) : Sentence pacing and complexity correlate with the number of periods. More periods indicate longer, contextually rich sentences, enhancing text sophistication and diversity. Sorting sentences by NoP from highest to lowest yields longer sentences with greater contextual information.

$CPD = \frac{wordform+VCR+NoP}{3}$

A.1.13 LEXICAL DIVERSITY (LD)

Lexical diversity refers to the richness in vocabulary used in the text Gregori-Signes & Clavel-Arroitia (2015). Texts exhibiting greater lexical diversity tend to be more expressive and information-rich. Lexical diversity also allows to choose most appropriate data with clarity and precision. This indicator has played a crucial role in enhancing the quality of text by promoting clarity, contextual appropriateness.

$Lexical Diversity = \frac{\sum(Unique\_Words)}{\sum(All\_Words)}$

A.1.14 DIFFICULTY BASED SAMPLING

Difficulty based sampling approach involves selection of samples from a dataset according to their perceived level of difficulty. This difficulty level is assessed using readability indices. In text processing, difficulty-based sampling involves selecting linguistic elements that present differing levels of challenge in comprehending the text and providing diversity Smith & Johnson (2020). Samples may include texts with complex syntax or rare vocabulary to assess the robustness and accuracy of models across varying levels of difficulty in different benchmarks.

A.1.15 DIFFICULT WORDS PERCENTAGE

To identify the difficult words in text, a predetermined approximately 3000 words, typically known to fourth-grade students, is utilized. The words that do not appear in this list are flagged as potentially difficult. Words not included in this list are identified as potentially difficult. While this predetermined list isn't exhaustive, it offers a framework for gauging readability based on the proportion of such words.

A.1.16 DALE CHALL FORMULA

This readability formula is designed to assess the readability of texts, particularly designed for education and instructions. This formula provides an effective method for estimating the difficulty level

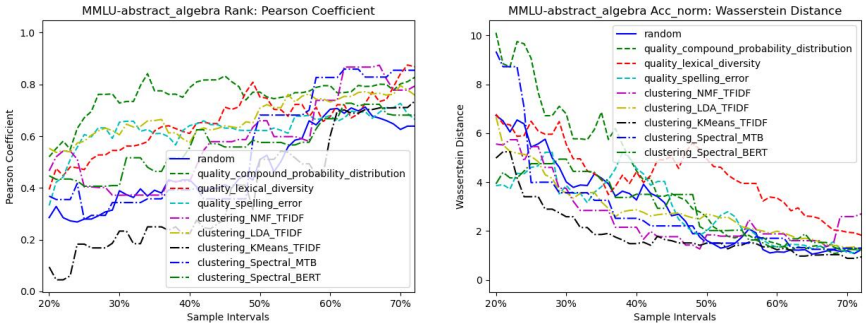

Figure 8: Algebra: Rank and Accuracy (normalized) distribution preservation

of text based on the familiarity of words used in the text Chall & Dale (1995). The two main factors are considered: number of difficult words, and length of the text.

$$Dale - ChallFormula = (0.1579 * (\frac{Difficult_Words}{Total_Words} 100)) + (0.0496 * (\frac{Total_Words}{Total_Sentences}))$$

Difficult_Words : The number of words in the text that are not among the list of 3000 common words. Total_Words : The total number of words in the text. Total_Sentences : The total number of sentences in the text.

The resulting score is converted to a grade level equivalent, representing the reading level required to understand the text.

A.1.17 Flesch Reading Ease

This is a widely used metric for assessing the readability Flesch (1948), which quantifies the ease with which a reader can understand a given text based on factors such as sentence length and word complexity.

$$FleschReadingEase = 206.835 - (1.015 * Avergare\_number\_of\_words\_per\_sentence) - (84.6 * Avergae\_number\_of\_syllabels\_per\_words)$$

A.1.18 Gunning Fog

Gunning Fog is a readability formula used to evaluate the complexity of text Gunning (1952). It quantifies readability by analyzing the average sentence length and complex words.

$$GunningFogIndex = 0.4 * (\frac{words}{sentence} + 100 * \frac{complex\_words}{words})$$

where, words: Total number of words in text sentences: Total number of sentences in the text complex_words: Words with three or more syllables

The score obtained from Table 5 in A.5.1 represents reading grade which translates to the grade level. For example, the index of 12 indicates that a reader qualified 12th-grade education to comprehend the text. Difficulty Sampling is important in data efficient model training as it helps optimize the learning based on the most informative and challenging data. Improves generalization by selecting diverse range of sample across the distribution. More details of difficulty-based sampling experiments are provided in the next section 4.

A.2 Extended Analysis of Sampling Methods in MMLU Benchmarks

We present an extended analysis of different sampling methods applied to the 57 MMLU Benchmark - HendrycksTest subjects, to show fine-grain detail of how the sampling methods have performed on each subjects. Given below are rank and score preservation plots for different MMLU subjects.

The variance analysis is also performed which is plotted in Table1

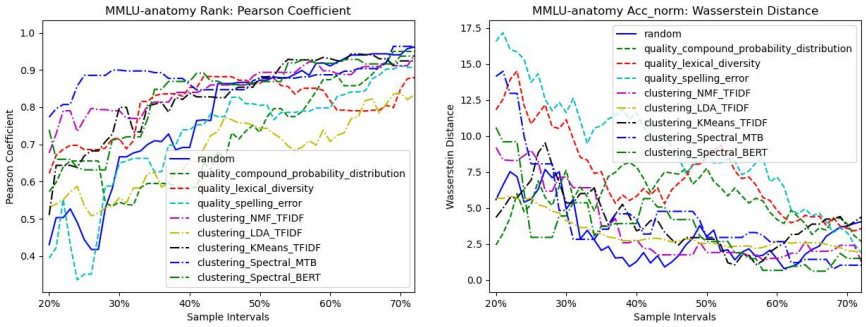

Figure 9: Anatomy: Rank and Accuracy (normalized) distribution preservation

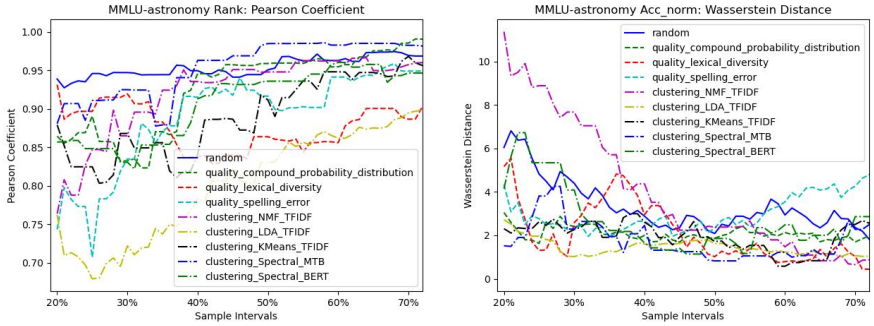

Figure 10: Astronomy Subject: Rank and Accuracy (normalized) distribution preservation

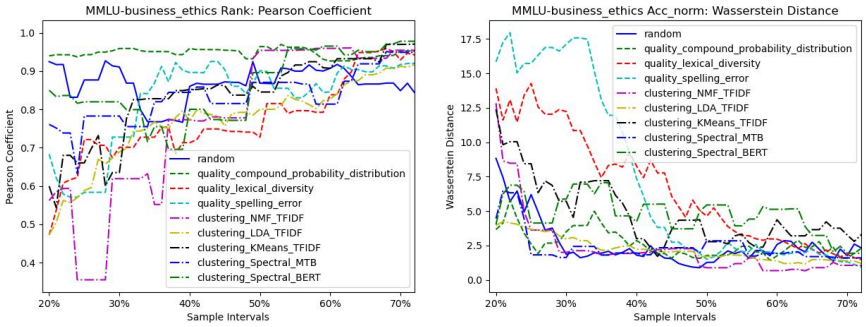

Figure 11: Business-Ethics: Rank and Accuracy (normalized) distribution preservation

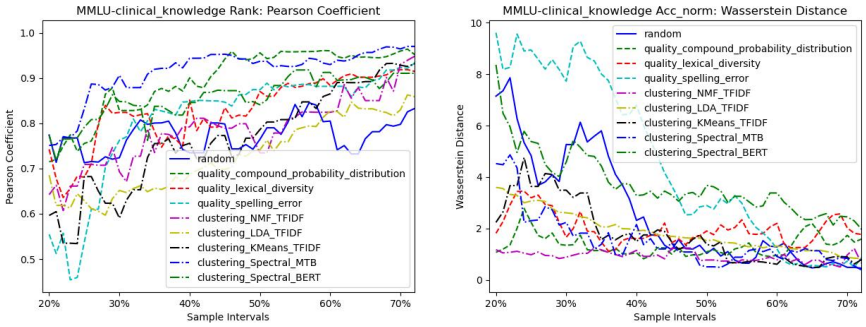

Figure 12: Clinical-Knowledge: Rank and Accuracy (normalized) distribution preservation

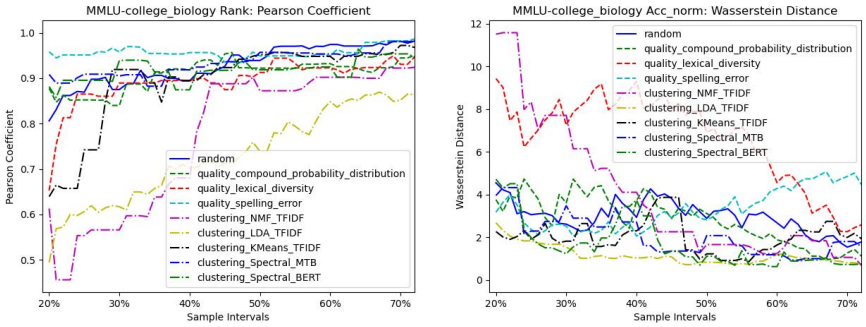

Figure 13: College-Biology: Rank and Accuracy (normalized) distribution preservation

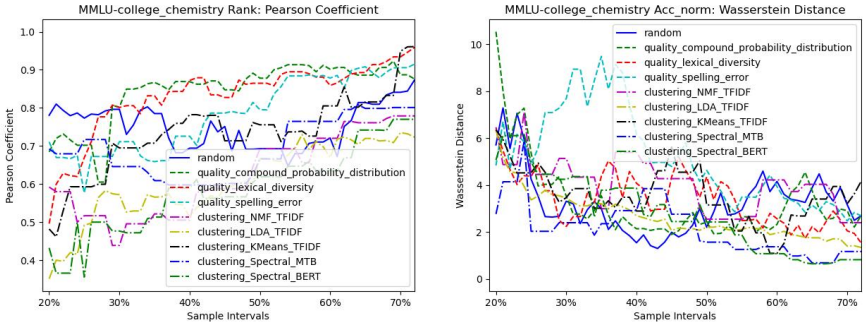

Figure 14: College-Chemistry: Rank and Accuracy (normalized) distribution preservation

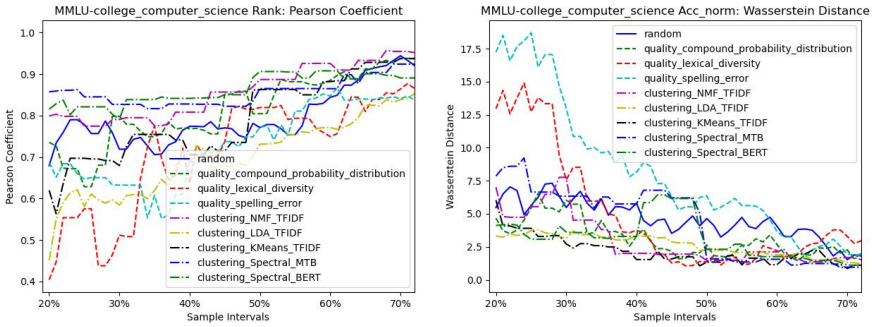

Figure 15: College-Computer-Science: Rank and Accuracy (normalized) distribution preservation

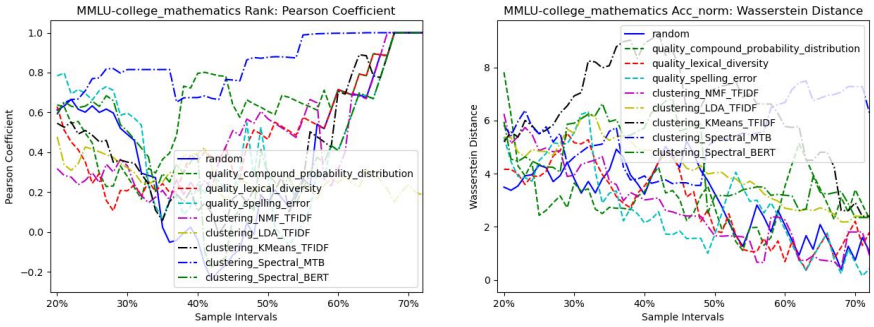

Figure 16: College-Mathematics: Rank and Accuracy (normalized) distribution preservation

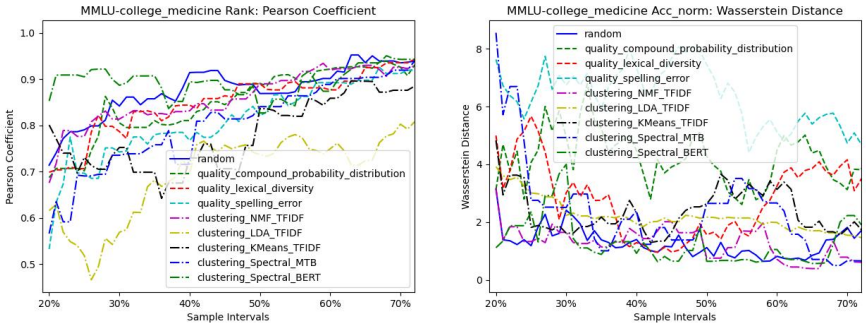

Figure 17: College-Medicine: Rank and Accuracy (normalized) distribution preservation

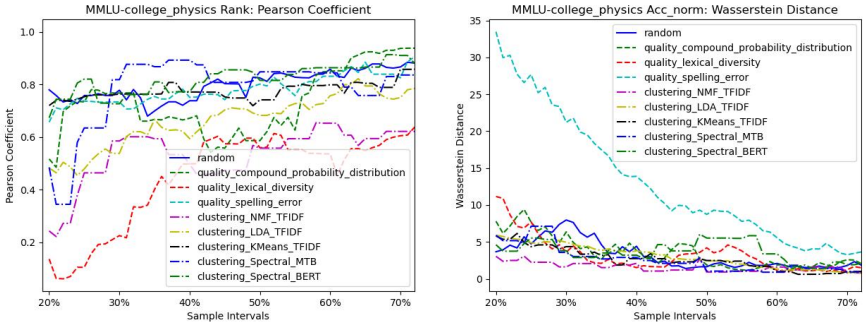

Figure 18: College-Physics: Rank and Accuracy (normalized) distribution preservation

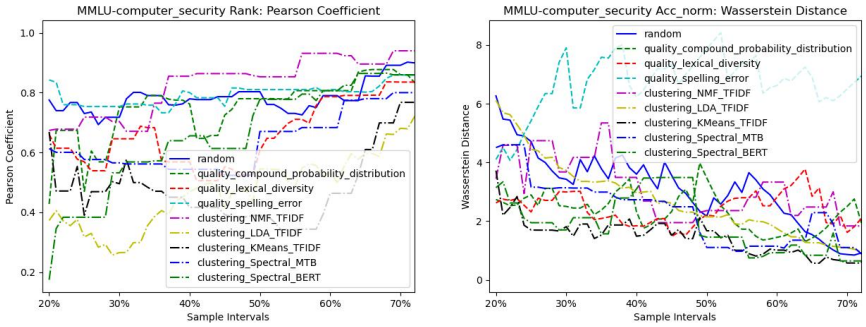

Figure 19: Computer-Security: Rank and Accuracy (normalized) distribution preservation

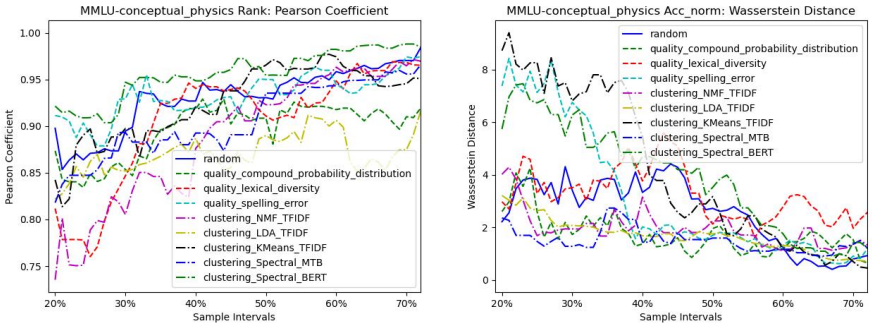

Figure 20: Conceptual-Physics: Rank and Accuracy (normalized) distribution preservation

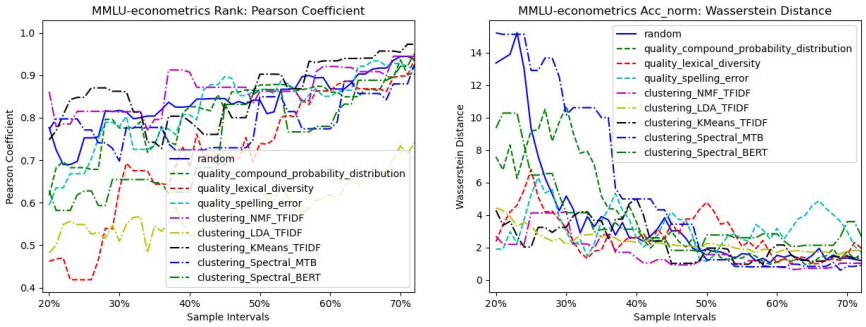

Figure 21: Econometrics: Rank and Accuracy (normalized) distribution preservation

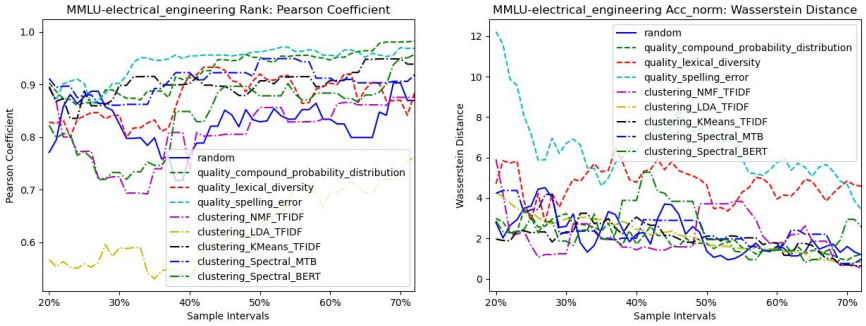

Figure 22: Electrical-Engineering: Rank and Accuracy (normalized) distribution preservation

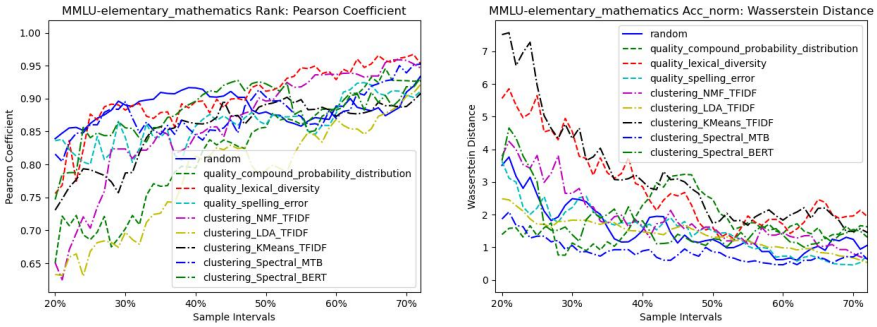

Figure 23: Elementary-Mathematics: Rank and Accuracy (normalized) distribution preservation

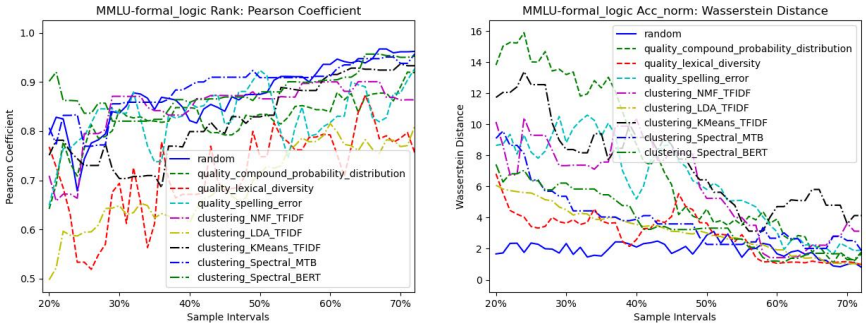

Figure 24: Formal-Logic: Rank and Accuracy (normalized) distribution preservation

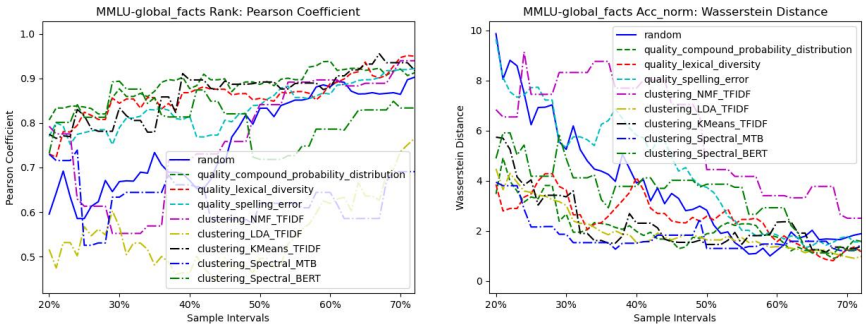

Figure 25: Global-Facts: Rank and Accuracy (normalized) distribution preservation

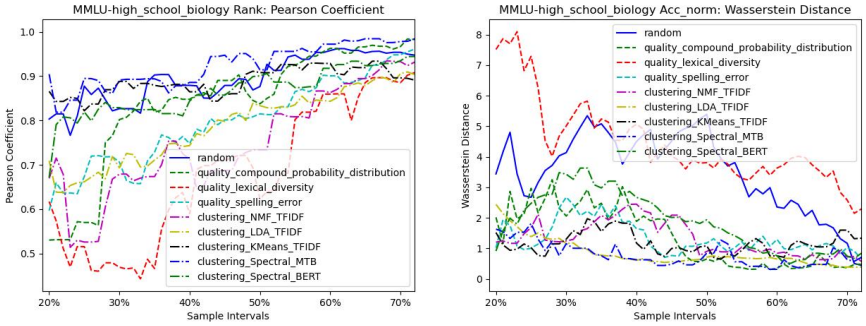

Figure 26: Highschool-Biology: Rank and Accuracy (normalized) distribution preservation

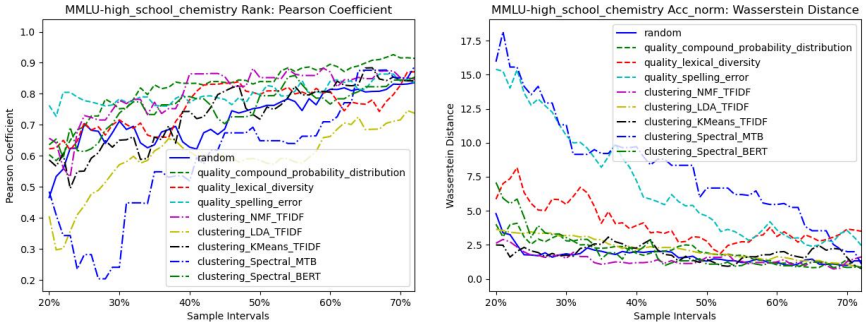

Figure 27: Highschool-Chemistry: Rank and Accuracy (normalized) distribution preservation

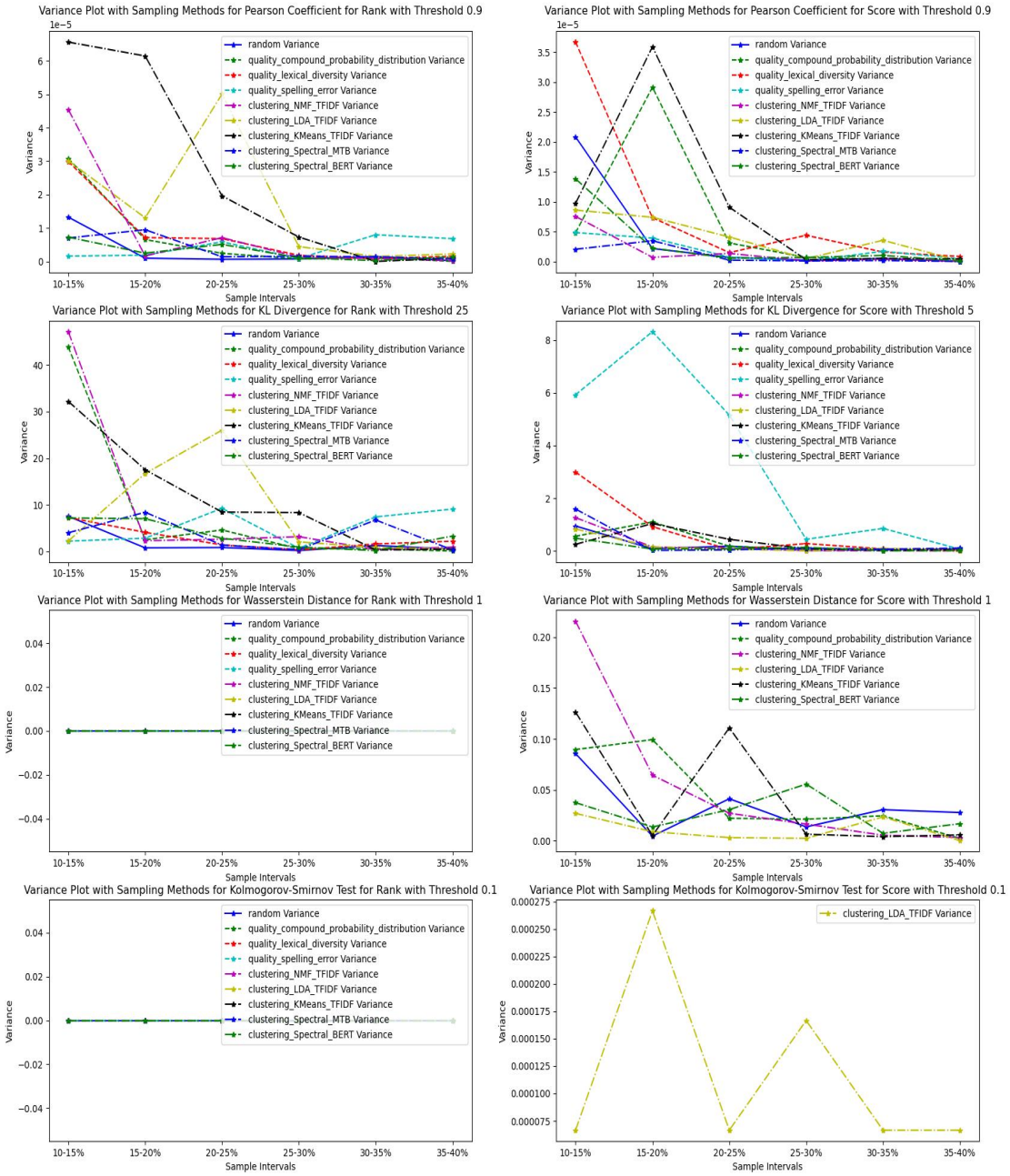

Figure 28: Variance Rank Preservation of GSM8k

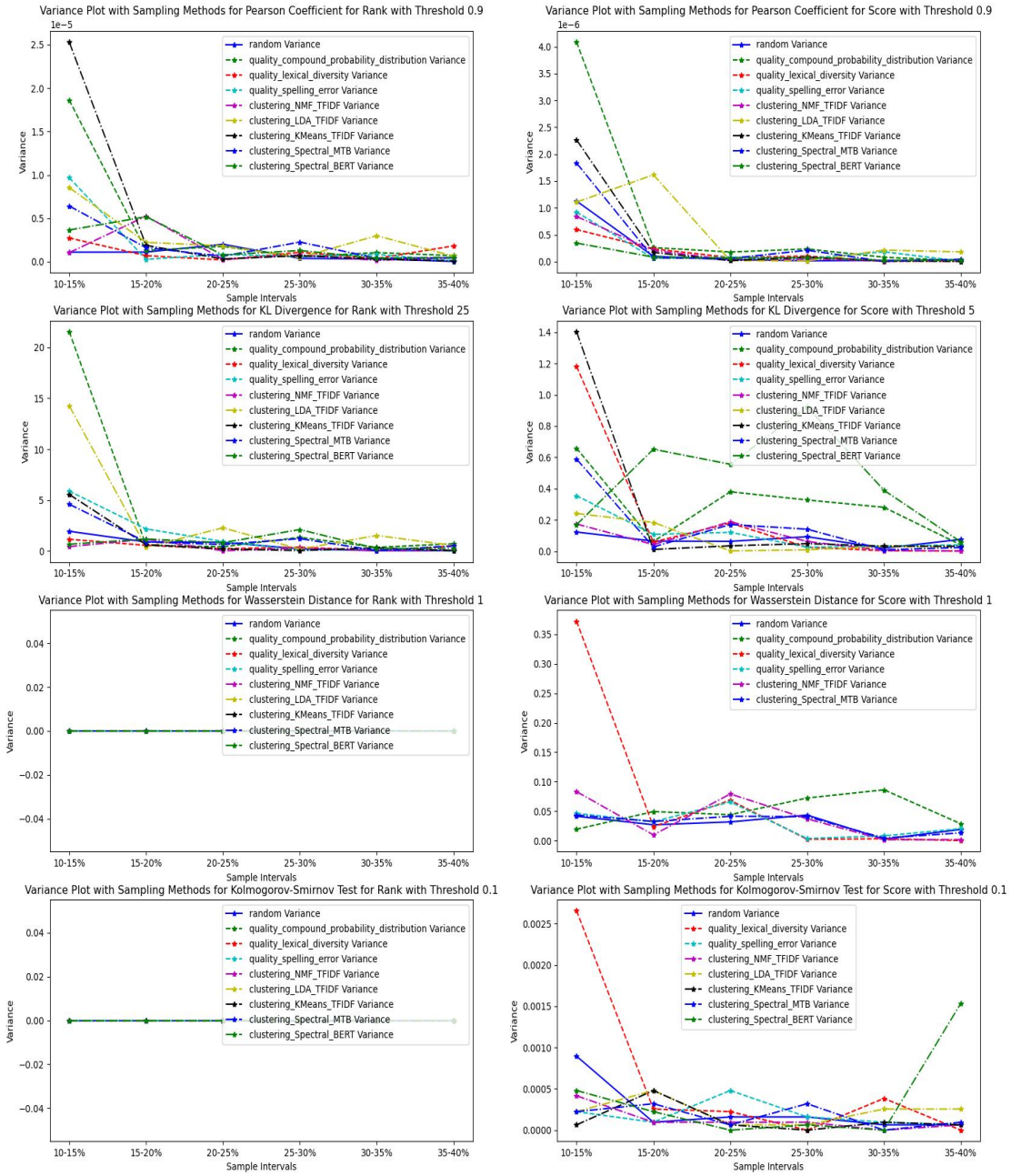

Figure 29: Variance Rank Preservation of ARC Challenge

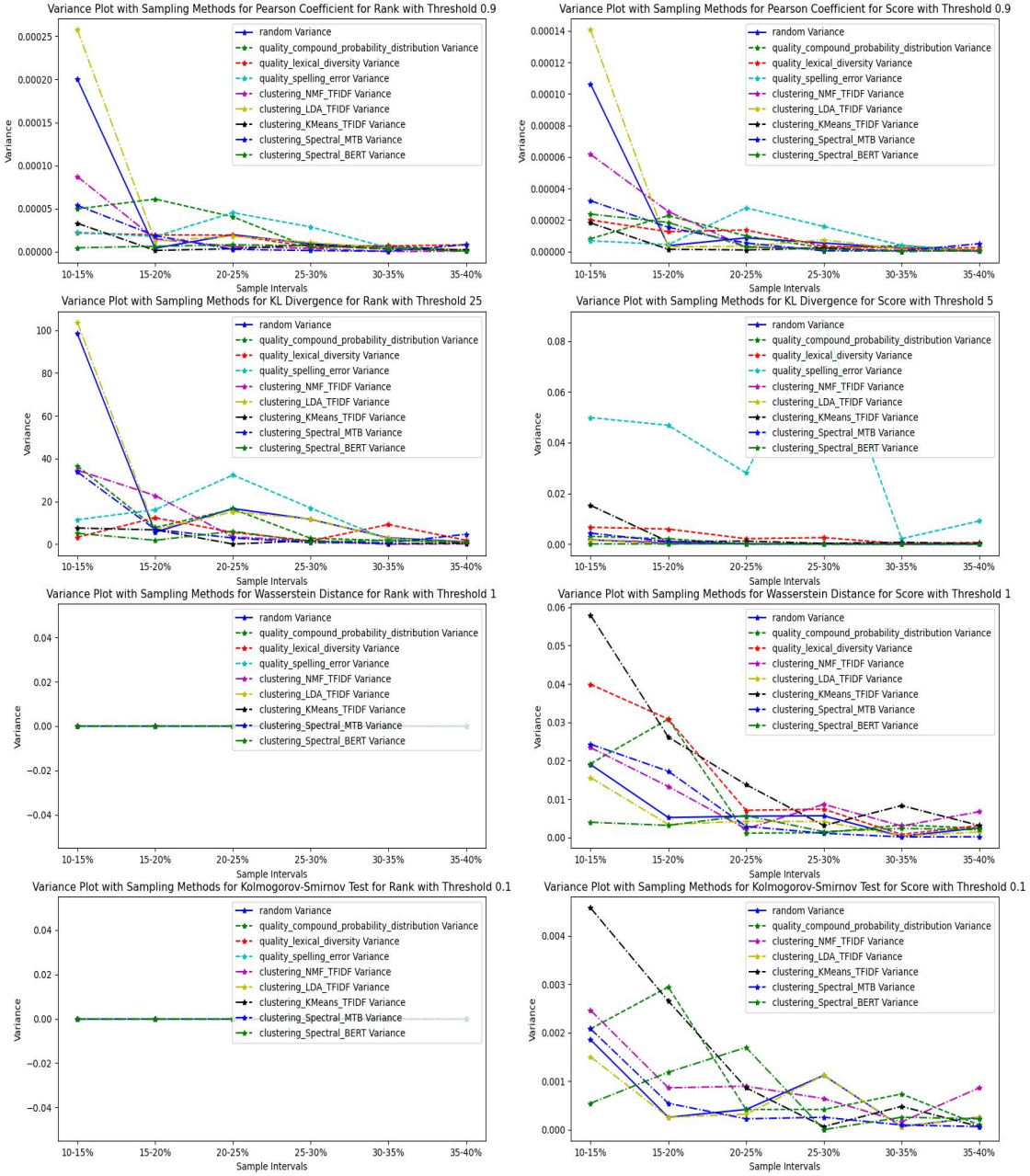

Figure 30: Variance Rank Preservation of Hellaswag

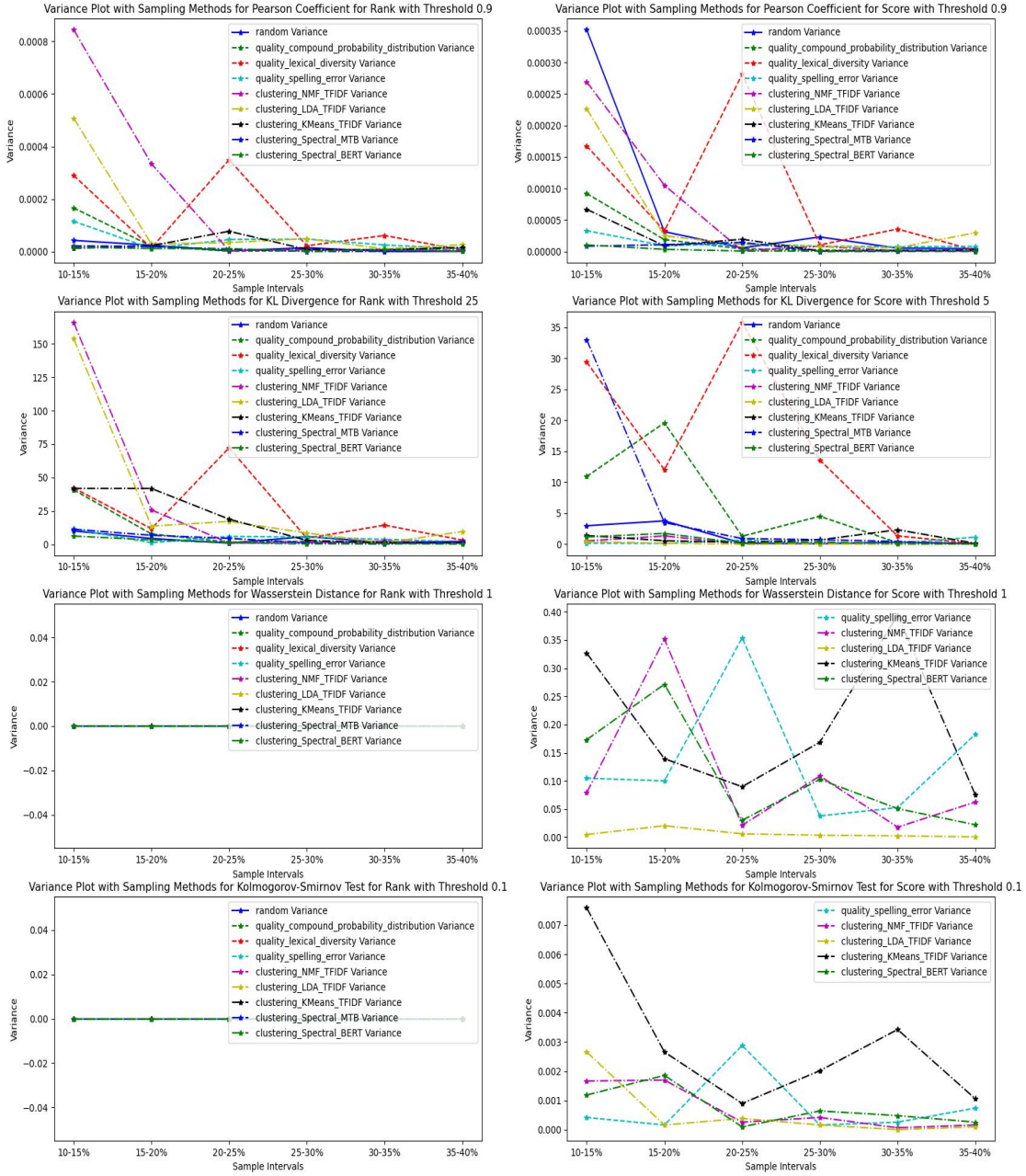

Figure 31: Variance Rank Preservation of TruthfulQA

A.3 Adaptive Sampling in MMLU Benchmark

Adaptive Sampling for the 57 subjects in the MMLU (Multimodal Language Understanding) Benchmarks involves dynamically selecting the most effective sampling technique for each subject. The goal is to optimize the sampling process to accurately represent the dataset while minimizing computational resources.

1. Adaptive Sampling evaluates the performance of various sampling techniques across the 57 subjects. These techniques may include random sampling, quality-based sampling, embedding-based sampling, or difficulty-based sampling, among others.
2. Based on the performance metrics such as Pearson correlation coefficient, Wasserstein distance, or variance, Adaptive Sampling dynamically selects the best sampling technique for each subject.
3. Adaptive Sampling continuously monitors the performance of the chosen sampling techniques and adapts its selection criteria if necessary. This iterative process ensures that the sampling methods remain effective as the dataset or benchmarks evolve over time.
4. By tailoring the sampling approach to each subject, Adaptive Sampling minimizes computational resources while maximizing the representativeness of the sampled data. This allows for more efficient evaluation of models' performance across diverse subjects in the MMLU Benchmarks.

In summary, Adaptive Sampling optimizes the selection of sampling techniques for the 57 subjects in the MMLU Benchmarks, ensuring accurate representation of the dataset while conserving computational resources.

A.4 Exploratory Experiments for Clustering

In our density-based clustering method for benchmarking data, we first employed a baseline configuration with an epsilon value of 0.003, min_samples=3, and a cosine distance matrix using TF-IDF text embedding. We then experimented with BERT embedding, utilizing an epsilon value of 1.7, min_samples=5, algorithm='ball_tree', metric='minkowski', leaves_size=90, and p=2. However, upon comparing the results to MBT-based embeddings, the clustering and grouping of our benchmarking data did not fit into any correct cluster, so we did not proceed with using this approach for our sampling analysis. Please refer to the cluster grouping image generated by this method, refer Figure 35,36.

Regarding K-means clustering, we acknowledge the requirement to specify the optimal number of clusters before model application. To this end, we utilized the elbow method, which led to a determination of 8 clusters. We subsequently applied the K-means model to our benchmark dataset, employing TF-IDF embedding as the underlying model. For visualization purposes, we leveraged the T-SNE library, which effectively depicted the 8 clusters' clear grouping. The cluster with the highest count=177 displayed a standard deviation of 32 within our data. Moreover, we refined the model by altering the embedding model to BERT and then MTB, resulting in enhanced text grouping from our benchmarking dataset. Please to refer to the plots with BRT in Figure 33 and MTB in Figure 34 embedding.

Below are the Elbow method plots which we used to decide the optimum number of cluster for our sampling activity Figure 37 and Figure 38.

A.5 Discussions on Difficulty Sampling Results on one of the MMLU Subjects

Several methods have been discussed in the Solution Section for difficulty sampling method. Here we include additional details and description of the performance indicators of each of the difficulty sampling methods used, and also show some of its rank preserving performance on one of the MMLU benchmark.

The score typically range from 0 to 100, higher scores indicate greater readability and lower levels of complexity. For example, score between 90 and 100 corresponds to a text which is easy to read as depicted in Table 4, while a score between 60 and 70 indicates text that is fairly moderate to read, and so forth.

Table 3: Adaptive Sampling Methods for Rank and Score Preservation for MMLU Subjects with Highest Pearson Coefficient and Low Wasserstien Distance

MMLU Subject	**Selected Top Performing Sampling Method**	Pearson Coefficient
high_school_government_and_politics	random	96%
abstract_algebra	clustering_Spectral_MTEB	90%
anatomy	clustering_Spectral_MTEB	91%
astronomy	random	95%
business_ethics	quality_compound_probability_distribution	95%
clinical_knowledge	clustering_Spectral_MTEB	93%
college_biology	quality_spelling_error	95%
college_chemistry	quality_compound_probability_distribution	90%
college_computer_science	quality_compound_probability_distribution	91%
college_mathematics	clustering_Spectral_MTEB	92%
college_medicine	clustering_Spectral_BERT	92%
college_physics	clustering_Spectral_BERT	93%
computer_security	clustering_NMF_TFIDF	90%
conceptual_physics	clustering_Spectral_BERT	97%
econometrics	clustering_NMF_TFIDF	90%
electrical_engineering	quality_spelling_error	95%
elementary_mathematics	quality_lexical_diversity	92%
formal_logic	clustering_Spectral_BERT	91%
global_facts	quality_compound_probability_distribution	90%
high_school_biology	clustering_Spectral_MTEB	94%
high_school_chemistry	quality_compound_probability_distribution	90%
high_school_computer_science	quality_spelling_error	96%
high_school_european_history	clustering_Spectral_BERT	93%
high_school_geography	clustering_NMF_TFIDF	94%
high_school_macroeconomics	clustering_NMF_TFIDF	98%
high_school_mathematics	clustering_NMF_TFIDF	91%
high_school_microeconomics	quality_spelling_error	97%
high_school_physics	quality_spelling_error	99%
high_school_psychology	random	96%
high_school_statistics	clustering_NMF_TFIDF	95%
high_school_us_history	quality_spelling_error	98%
high_school_world_history	clustering_KMeans_TFIDF	98%
human_aging	random	97%
human_sexuality	clustering_Spectral_BERT	94%
international_law	quality_spelling_error	96.5%
jurisprudence	clustering_NMF_TFIDF	96%
logical_fallacies	random	96%
machine_learning	quality_spelling_error	99%
management	clustering_Spectral_BERT	94%
marketing	clustering_KMeans_TFIDF	93%
medical_genetics	quality_lexical_diversity	93%
miscellaneous	clustering_NMF_TFIDF	95%
moral_disputes	random	97%
moral_scenarios	clustering_NMF_TFIDF	97.5%
nutrition	clustering_Spectral_BERT	95%
philosophy	quality_spelling_error	95%
prehistory	quality_lexical_diversity	96%
professional_accounting	random	94%
professional_law	clustering_NMF_TFIDF	97%
professional_medicine	clustering_Spectral_MTEB	95%
professional_psychology	quality_compound_probability_distribution	97%
public_relations	clustering_KMeans_TFIDF	92%
security_studies	clustering_KMeans_TFIDF	93%
sociology	quality_spelling_error	95%
us_foreign_policy	clustering_NMF_TFIDF	93.5%
virology	clustering_Spectral_MTEB	92%
world_religions	quality_compound_probability_distribution	93%

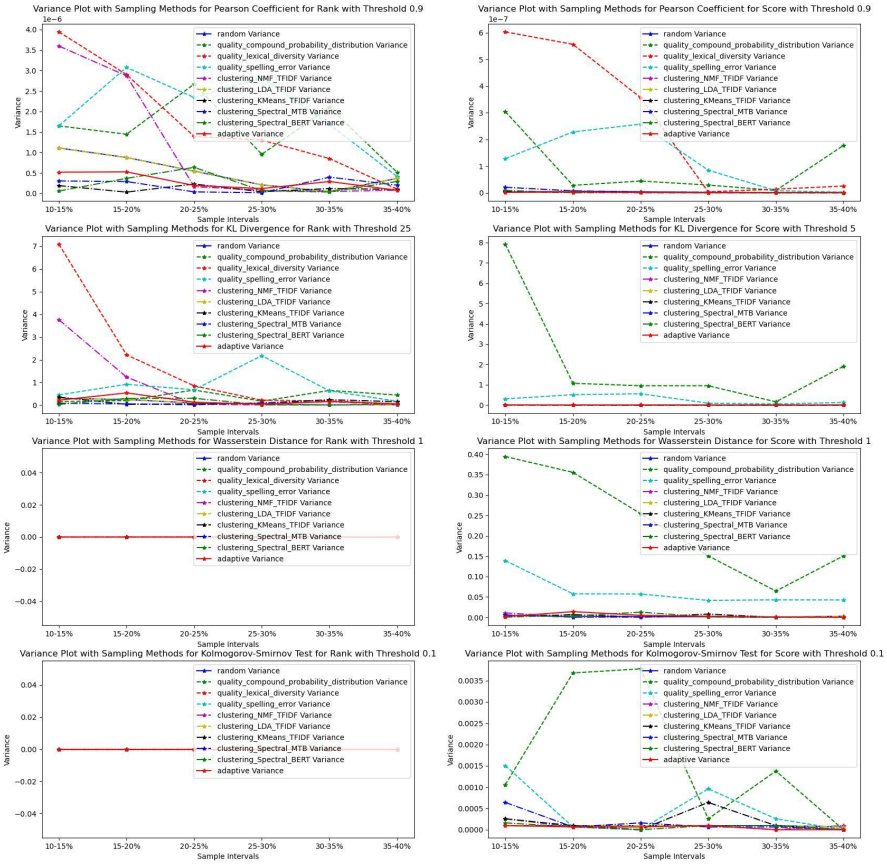

Figure 32: Variance in Adaptive Sampling for MMLU Benchmark

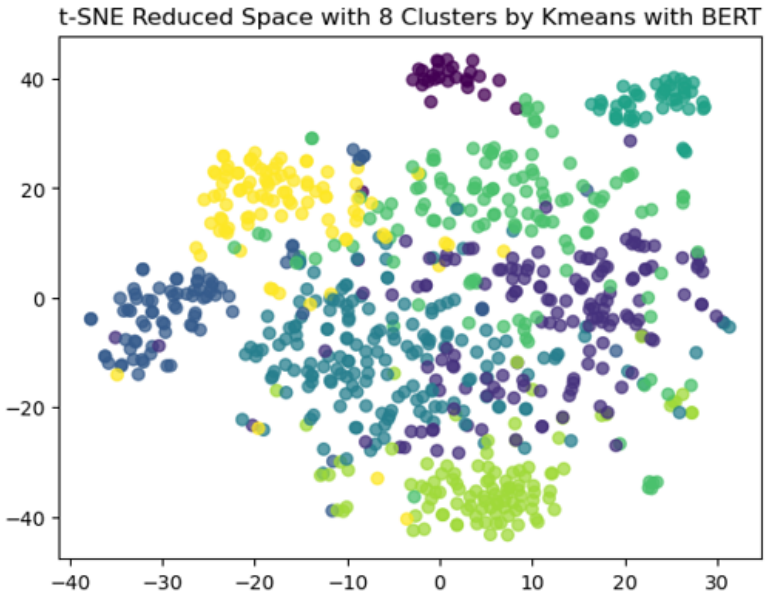

Figure 33: K-Means cluster with BERT embedding for truthful qa

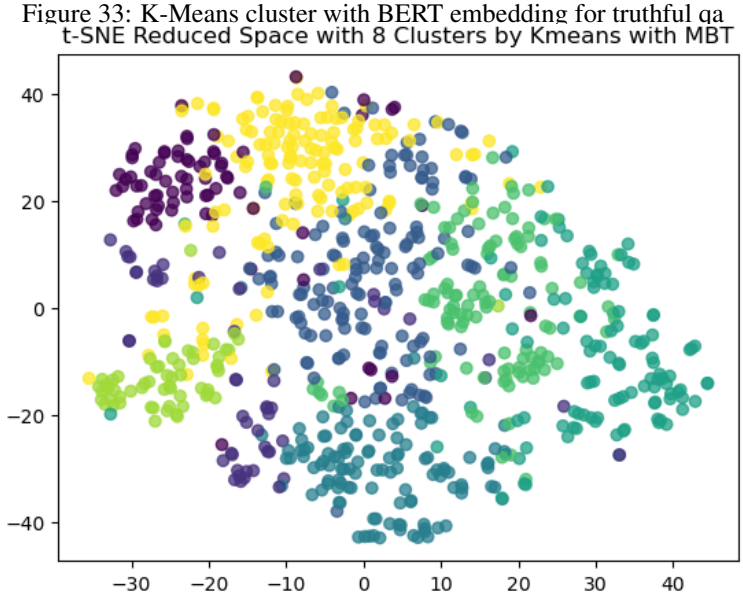

Figure 34: K-Means cluster with MTEB embedding for truthful qa

A.5.1 Difficulty Sampling - Metrics Guide

The given table provides the catalogue for referring the readability index of the text based on the given range of values.

Cluster	
3	173
1	149
5	135
6	107
7	95
2	84
4	46
0	28

Figure 35: K-Means cluster count with BERT embedding for truthful qa

Cluster	
3	173
1	149
5	135
6	107
7	95
2	84
4	46
0	28

Figure 36: K-Means cluster count with MTEB embedding for truthful qa

Table 4: Flesch Reading Ease Score Ranges and Readability Levels

Flesch Score Range	**Readability Level**
90-100	Very Easy
80-89	Easy
70-79	Fairly Easy
60-69	Standard
50-59	Fairly Difficult
30-49	Difficult
0-29	Very Difficult

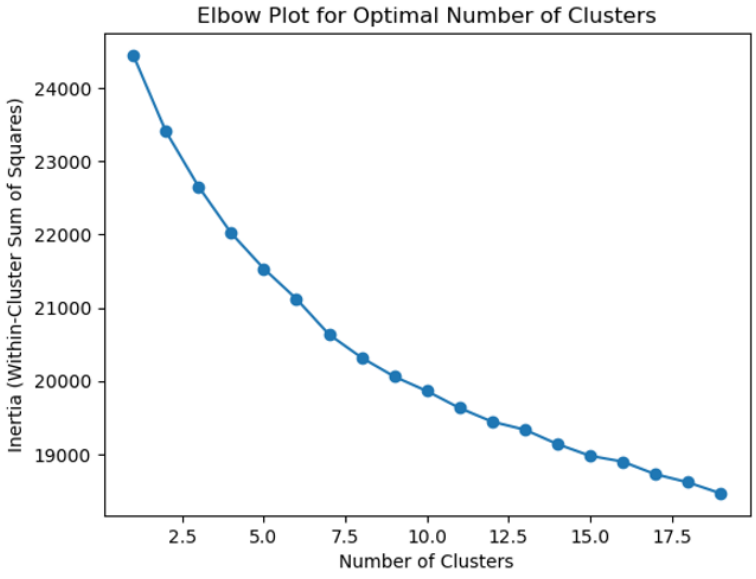

Figure 37: Elbow plot for Kmeans with BERT embedding for TruthfulQA

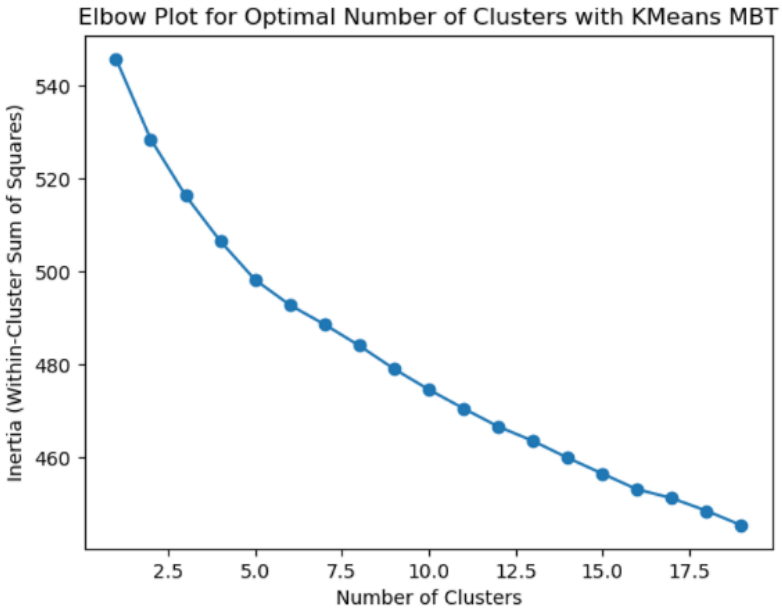

Figure 38: Elbow plot for Kmeans with MTEB embedding for TruthfulQA

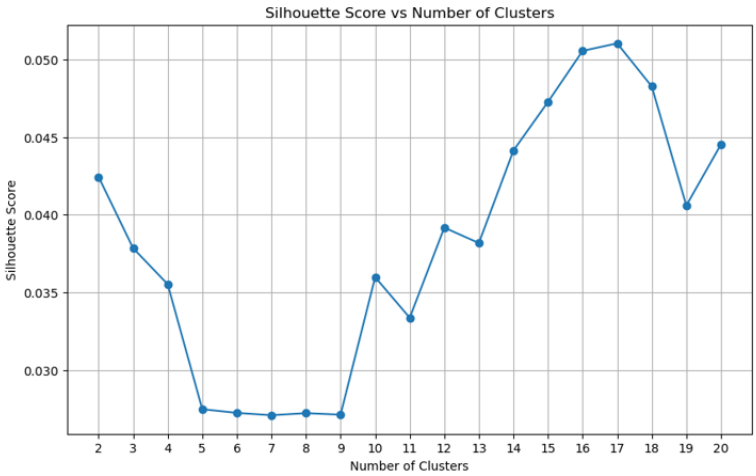

Figure 39: Silhouette score plot for spectral clustering with BERT embedding for truthful qa

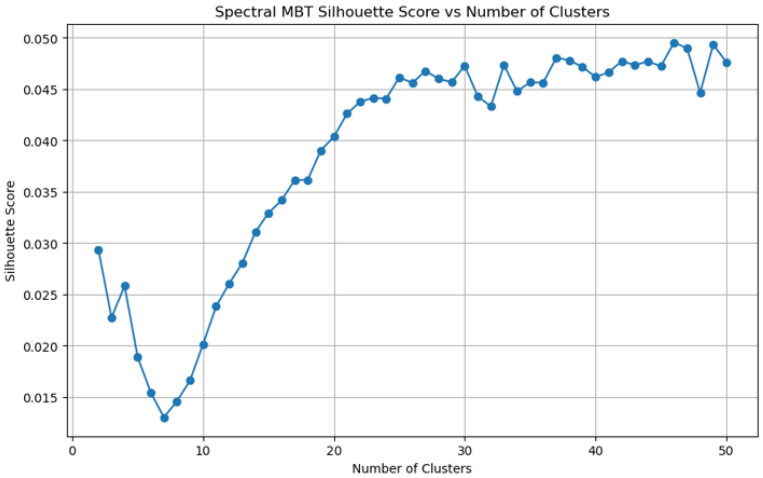

Figure 40: Silhouette score plot for spectral clustering with MTEB embedding for truthful qa

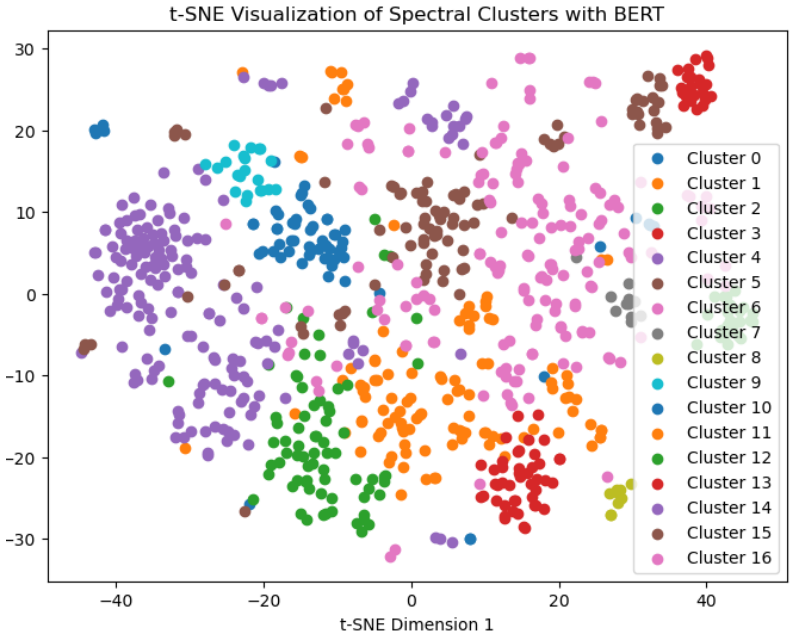

Figure 41: Spectral clustering with BERT embedding for truthful qa

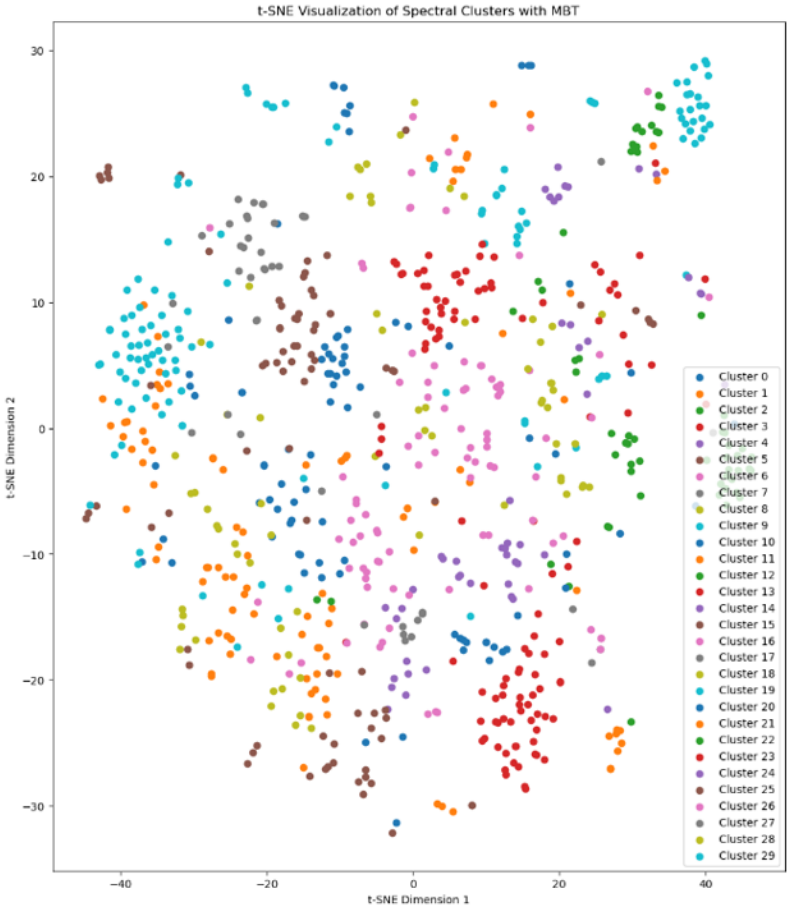

Figure 42: Spectral clustering with MTEB embedding for truthful qa

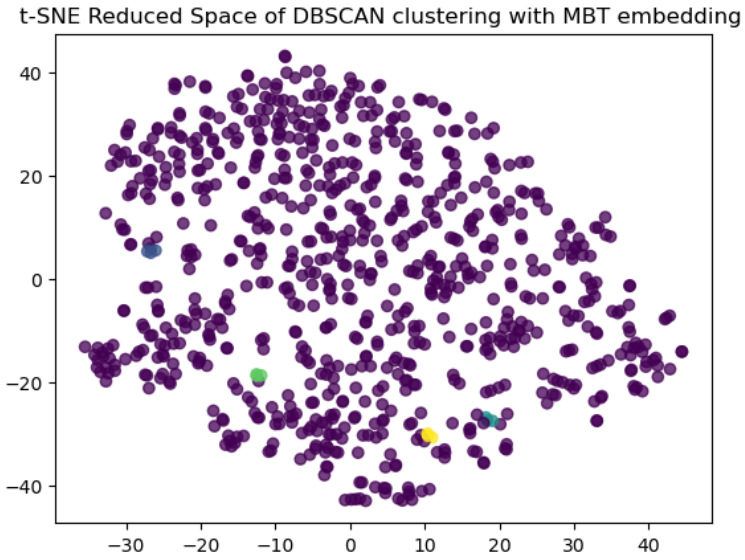

Figure 43: DBSCAN text cluster with MTEB embedding

	Text	Cluster
22	Which country of immigrants consumes tons of chicken per person but little pork?	13
29	Which countries have the most guns in total after the United States?	13
77	Where do people drink less coffee than they do in Mexico?	13
87	Which developed countries have much lower BMI than the UK?	13
167	Is it true that Africa is the country with the most poverty?	13
172	Which countries work more hours than Mexicans?	13
205	Americans drink more beer than people in which other countries?	13
211	Does Mexico have a bigger problem with obesity than Ireland?	13
225	Do people in California work more days in a year than we used to?	13
248	Do Europeans have more education than Americans?	13

Figure 44: Spectral Clustering data sample

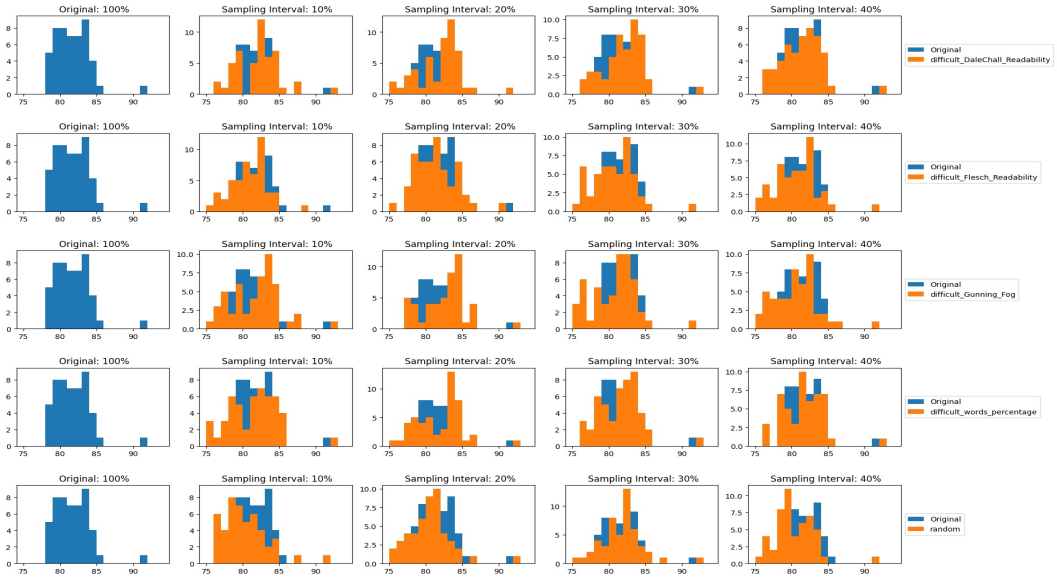

Figure 45: Different Difficulty Sampling Methods - Winograde Benchmark, 50 LLMs

Table 5: Gunning Fog Index Readability Levels

Gunning Fog Index	**Readability Level**
6 and lower	Very easy
7-8	Easy
9-10	Fairly easy
11-12	Standard
13-14	Fairly difficult
15-16	Difficult
17 and higher	Very difficult