SubLIME: Less is More for LLM Evaluation

MahammadParwez Alam*
Gayathri Saranathan*
James Lim†
mahammad-parwez.alam@hpe.com
gayathri.saranathan@hpe.com
jamesl@hpe.com
Hewlett Packard Labs

Suparna Bhattacharya
Foltin Martin
Soon Yee Wong
suparna.bhattacharya@hpe.com
martin.foltin@hpe.com
soon-yee_wong@hpe.com
Hewlett Packard Labs

Cong Xu‡
cong.xu@hpe.com
Hewlett Packard Labs

ABSTRACT

Evaluating large language models (LLMs) presents a substantial computational challenge, a critical aspect often overlooked. Efficient evaluation of LLMs is crucial for comprehensively understanding their multifaceted capabilities and facilitating comparisons across ever growing number of new models and benchmarks. In response, we introduce an innovative data-efficient approach that utilizes adaptive sampling based on various techniques such as clustering and quality-based methods, to generate representative subsets of benchmarks. This strategy ensures statistically aligned LLM ranks compared to the complete dataset, as demonstrated by high Pearson correlation coefficients. Through empirical analysis of six benchmarks, we find that: (1) quality-based sampling consistently yields strong correlations (0.85 to 0.95) with the full datasets at a 10% sampling rate; (2) clustering methods stand out in certain benchmarks; (3) no single method universally outperforms others across all metrics. Our adaptive sampling framework dynamically selects the optimal technique for each benchmark, significantly reducing evaluation costs while maintaining the integrity of ranking and score distribution. Remarkably, a minimal sampling rate of 1% is effective in benchmarks like MMLU. Furthermore, employing difficulty-based sampling to focus on more challenging segments of benchmarks enhances model differentiation, leading to broader score distributions.

KEYWORDS

Large Language Model, Data Efficiency, Benchmarking, Evaluation, Sampling, Open-Source Model

1 INTRODUCTION

Large language models (LLMs) have witnessed remarkable growth, revolutionizing artificial intelligence. With over 400,000 open-source models, including about 56,000 text generation models on HuggingFace, their rapid expansion poses challenges in efficiently evaluating them. Over 3,000 models are present on the Open LLM Leaderboard, emphasizing the necessity for standardized performance benchmarks. Evaluation becomes a significant bottleneck, demanding extensive computational resources and substantial financial costs.

The financial burden of LLM evaluation is considerable, evidenced by the HELM [21] project's spending of about $50,000[1] on just 30 LLMs with 13 tasks. The proliferation of LLMs on HuggingFace, including those fine-tuned, quantized, and merged, is releasing at an unprecedented pace. Simultaneously, the community is releasing more NLP datasets for benchmarking, expanding the evaluation scope to capture the full range of LLM capabilities [3]. Scaling LLM evaluation to cover even a fraction of the current 56,000 text-generation LLMs on HuggingFace with 100 benchmarks could result in costs on the order of $100 million.

Recent advancements in fast evaluation focus on hardware efficiency optimization. The LM Evaluation Harness [10] is integrated with vllm [17], a high-performance LLM inference library, to improve GPU utilization and enhance throughput. To complement these efforts, we introduce a **data-efficient** solution that employs adaptive sampling to identify a relevant, representative, diverse, or high-quality subset of data points from a given benchmark, aiming to reduce evaluation costs while maintaining LLM rankings and score distributions compared to the complete dataset.

1. We **analyze different sampling strategies'** effects on rank preservation and score distribution in data-efficient LLM evaluation. Our findings suggest notable resource reduction potential in certain benchmarks, underlining the absence of a universally effective sampling method across all benchmarks.

2. We demonstrate the effectiveness of **adaptive sampling** in reducing the evaluation time by orders of magnitude for benchmarks like MMLU, where even a 1% sampling rate can well preserve ranks and score distributions.

3. Our sampling strategy was applied in two different scenarios: (a) achieving efficient evaluation with consistent rank preservation and score distribution; (b) employing **difficulty-based sampling** to select challenging samples from older, less complex benchmarks, enhancing their score distribution and discriminative capacity in evaluating LLMs.

2 RELATED WORK

A significant part of literature delves into data-efficient model training [7, 29], extending recently to LLM [24, 31]. Previous studies concentrated on methodologies such as coreset selection and importance sampling, aiming to derive a condensed dataset that either aligns with or enhances model performance by training with a smaller yet representative or higher-quality dataset. DeepCore [12] empirically explores various core-set selection methods on CIFAR10 and ImageNet datasets, demonstrating that while specific methods excel in certain scenarios, random selection remains a robust baseline. For LLMs, UniMax [4] addresses biases in language sampling by leveraging linguistic similarity metrics. DeepSpeed Data Efficiency [18] introduces two techniques: efficient data sampling

*First authors
†Core author
‡Senior lead author
[1]actual cost: $38,001 for commercial APIs, plus 19,500 A100 GPU hours, assuming a rate of $1/hr for A100

with curriculum learning and data routing with a random token dropping method to cut training time for LLM. In contrast to the prior work that focus on training, we use sampling in LLM evaluation, where the goal is to choose a benchmark subset that results in similar LLM rankings and score distributions compared to the complete dataset.

Difficulty sampling selects hard examples to speedup convergence during training[32]. We use diverse difficulty sampling to identify the most challenging yet representative examples in NLP benchmarks, to prioritize LLM evaluation discriminatory power over solely enhancing model performance.

3 OUR SOLUTION

To expedite LLM evaluation at scale, we propose an adaptive sampling strategy inspired by real-world examples like the International Mathematical Olympiad, which identifies top talents with only six problems. This underscores the potential of leveraging existing dataset redundancy and selecting data subsets judiciously for benchmarking while preserving LLM ranking and score distributions. Various sampling techniques were compared to identify the optimal approach given the data characteristics of a benchmark. Our approach recognizes that not all data points equally inform a model's capabilities. We employ different sampling techniques to select best 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

In this section, we present a range of sampling techniques aimed at rank and score preservation. Each method contributes uniquely to our overarching goal of efficient LLM evaluation.

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

Clustering-based Sampling involves categorizing data into similarity-based groups, revealing patterns in unstructured datasets. Topic modeling methods like latent Dirichlet allocation (LDA) [27] and Non-negative matrix factorization (NMF) [19] with TF-IDF [23] organize text into thematic clusters. NMF proved effective in clustering datasets like TruthfulQA and GSM8k. However, DBSCAN failed due to misaligned clusters, while LDA successfully identified latent topics without enhancements from BERT or top embedding models on Massive Text Embedding Benchmark (MTEB) leaderboard [25]. K-means [14] grouped documents effectively using TF-IDF, and spectral clustering [30] produced meaningful clusters, especially when refined with BERT and MTEB embedding models.

Quality-based Sampling identifies high-quality data from large datasets through the evaluation of syntactic and semantic features, utilizing text processing methods to establish quality metrics. Key indicators of quality encompass average word length, diversity, and repetitiveness, along with compound metrics for comprehensive assessment. For instance, subset with minimize spelling errors [15] contributes to improved readability and model performance, signaling attention to detail. Maintaining an optimal average word length [1] is strikes a balance between complexity and comprehension, thereby preserving context quality. Reducing excessive word repetition [9] ensures textual diversity and fosters creativity. The Compound Probability Distribution (CPD) method, integrating metrics like Wordform, Vowel-Consonant Ratio, and Number of Periods, facilitates a comprehensive text quality evaluation, influencing aspects such as sentence structure and text diversity. Additionally, lexical diversity, which gauges vocabulary richness [34], contributes to the text expressiveness and informativeness.

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 [16] including TruthfulQA [22], ARC (AI2 Reasoning Challenge) [5], Winogrande [26], GSM8k (Grade School Math 8k) [6], MMLU (Massive Multitask Language Understanding) [13], and Hellaswag [33].
LLMs: Sellected 50 LLMs with from top 1000 models on the leaderboard [16] with uniform interval.

Algorithm 1 Experiment Design

Require: Initialize
Collect sample-level model 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 interval of 20 from Top 1000 models
  for each Sampling technique **do**
    for sampling rate $x\%$ from 1 to 100 at step size 1 **do**
      Run each sampling once and record the indexes
      Use these indexes to sample subset of $x\%$ from fullset
      Generate scores of the 50 LLMs on $x\%$ subset, rank them based on the scores
      Measure rank and score preservation wrt fullset results
    end for
    plot (a) rank preservation coefficient and (b) 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

3.2 Use case 2: Difficulty Sampling for better Diversity

Modern high-performing LLMs often excel in accuracy metrics when evaluated on older, less complex datasets. However, assessing them across the entire dataset results in a limited distribution of accuracy metrics, complicating performance distinction. Yet, our

analysis reveals a key insight: subsets of benchmarks considered mastered by advanced LLMs retain critical evaluative value, enriching the leaderboard with nuanced insights.

The goal of difficulty-based sampling is to choose a subset of data that provides a broader spectrum of accuracy metrics, enabling more insightful model comparisons. Unlike simplistic approaches that may opt for subsets with consistently high error rates across models, our aim is to pinpoint subsets that achieve a more diverse distribution. Difficulty-based sampling involves selecting samples from a dataset based on their perceived difficulty level, which is evaluated using readability indices. In text analysis, this method entails selecting linguistic elements with varying degrees of complexity. Samples may include texts with intricate syntax or uncommon vocabulary to evaluate models' robustness across different difficulty levels in various benchmarks [28].

The Difficult Words Percentage approach 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 [2] 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 [8] quantifies readability based on sentence length and word complexity. The Gunning Fog index [11] evaluates text complexity through average sentence length and complex words, with the score indicating the required education level to comprehend the text. 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 assessed various sampling techniques' effectiveness in reducing the benchmark time while maintaining rankings using a subset of the complete dataset. Using our proposed method outlined in 1, we aim to dynamically pinpoint the best sampling approach for each benchmark.

4.1 Analysis of Rank Preservation and Score Distribution

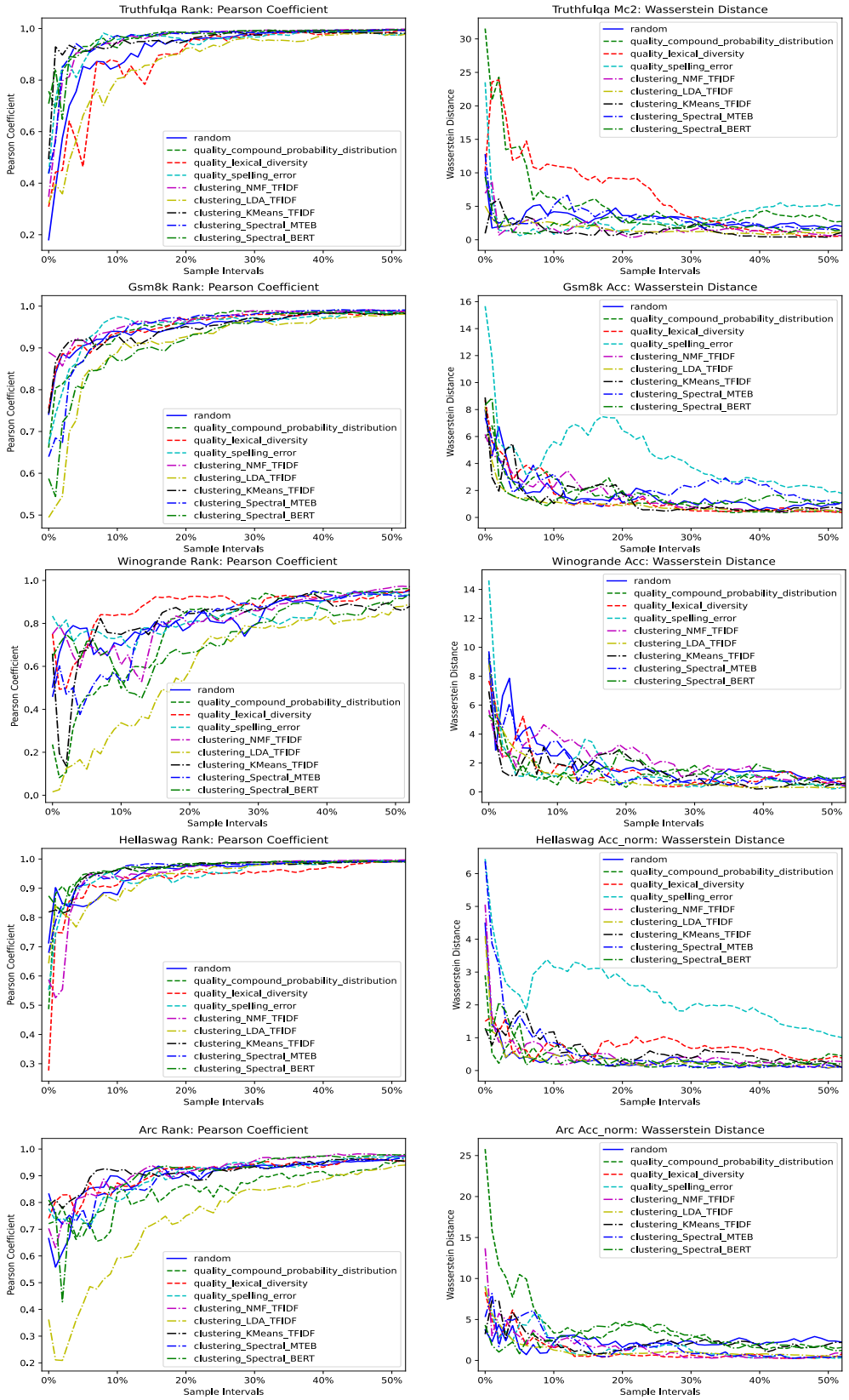
We evaluated 9 sampling approaches across 50 LLMs on 6 benchmarks. For rank preservation, we utilized the Pearson coefficient correlation metric, comparing LLM ranks between subsets and the original dataset. Score distribution discrepancy was assessed using the Wasserstein Distance (WD). Figure 1 and 2 illustrates these metrics for each benchmark, such as *TruthfulQA* in Figure 1a, where we analyzed rank with Pearson Coefficient and normalized accuracy (MC2) for score preservation using WD. Figures 1 and 2 present the rank and score preservation results for 6 benchmarks. Additionally, we examined variance in rank preservation performance across different sampling intervals for all benchmarks in our experiment.

In benchmarks like *TruthfulQA* and *GSM8k*, LLMs are scored based on accuracy, assessing semantic comprehension and reasoning for GSM8k and factual correctness for TruthfulQA. Our analysis of TruthfulQA and GSM8k in Figure 1 (a) and (b) respectively, shows that quality sampling methods such as *Quality CPD* and *Quality SE* consistently outperform others even at lower sampling intervals. These techniques ensure the selection of more representative samples from linguistic benchmarks. As indicated in Table 1, Quality CPD and SE exhibit robust performance with a 90% correlation and minimal variance across these benchmarks. Additionally, clustering methods using embedding models like UAE-Large-V1 [20] from the MTEB leaderboard and BERT also demonstrate strong performance, displaying high correlation at a 10% sampling rate.

The Winogrande benchmark evaluates model comprehension and reasoning by crafting questions that require deeper contextual understanding beyond surface-level cues. Sampling methods resilient to linguistics have excelled due to the challenge's stringent criteria, reflecting its complexity through consistent performance improvements. Random sampling achieved only around 82% Pearson correlation at a 10% sampling rate, as shown in Table 1. *Quality LD* surpassed the random baseline by enabling the selection of high-quality subsets, and leverage lexical diversity to capture diverse samples. *KMeans + TFIDF* demonstrated comparable performance, while other clustering methods exhibited varying effectiveness. The decline in clustering method performance may be attributed to their focus on sentence syntax, which may not align well with the semantic demands of the Winogrande dataset. Notably, no single sampling techinque hit targeted Pearson coefficient threshold of 0.9 at 10% sampling rate, and *Quality LD* achieved desired Pearson coefficient with miminum sample rate (about 15%) among all techniques, shown in Figure 1 (c). Effective commonsense reasoning relies on understanding nuanced word relationships within sentences, an aspect potentially overlooked by most clustering algorithms, tailored for assessing commonsense reasoning skills. Specifically, *Quality LD* and clustering methods like *NMF TFIDF, KMeans TFIDF, Spectral MTEB, Spectral BERT* performed well across benchmarks like Truthful QA, Hellaswag, and MMLU, suggesting their ability to capture subject complexities and enhance overall performance. Given MMLU's high-performing subjects and the benchmark's substantial total number of tokens as indicated in 1, difficulty sampling was employed to assess performance diversity and sample selection across distribution ranges.

The ARC benchmark assesses advanced reasoning skills through multiple-choice questions that require logical inference. Table 1 demonstrates consistent correlation metrics across various sampling methods used in this benchmark. Methods emphasizing text quality, such as lexical diversity, exhibit strong correlation with minimal variance. Given the complex nature of the ARC challenge, which demands higher-order thinking and advanced logical reasoning, sampling techniques that prioritize text quality and coherence stand out to achieve superior performance. 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 Table 3. The performance across all 57 subjects is summarized in Figure 2 showing that multiple sampling approach can achieve Pearson Coefficients exceeding 98% with low variance.

Adaptive Sampling for Data Efficient LLM Inference: We introduce an adaptive sampling strategy that dynamically selects the best sampling technique for each dataset. 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



Benchmark : TruthfulQA
Score metric: MC2
Selected sampling: Spectral BERT

Benchmark : GSM8K
Score metric: Accuracy
Selected sampling: Spectral MTEB

Benchmark : Winogrande
Score metric: Accuracy
Selected sampling: Lexical Diversity

Benchmark : Hellaswag
Score metric: Accuracy Normalized
Selected sampling: Spectral BERT, and Quality Spelling Error

Benchmark : ARC
Score metric: Accuracy Normalized
Selected sampling: Clustering NMF, and Spectral MTEB

Figure 1: Rank preservation, score distribution & optimal sampling for all but MMLU benchmarks

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

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

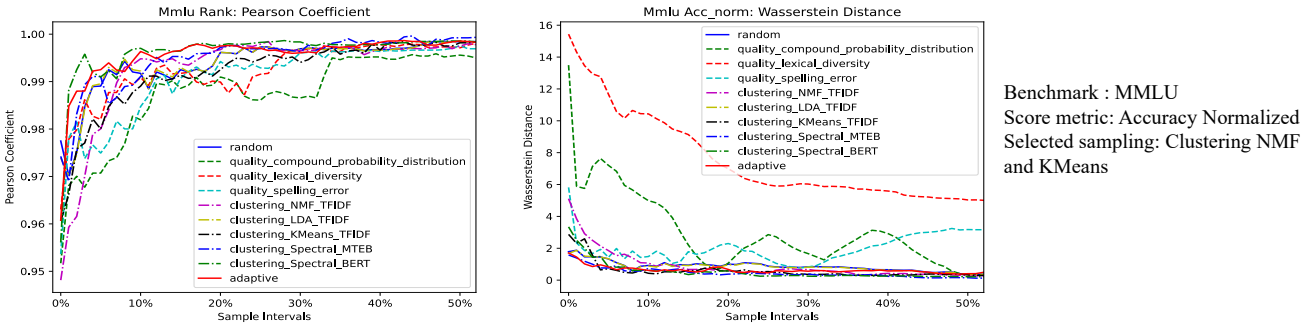


Figure 2: Adaptive Sampling (denoted in Solid Red) achieving stable performance in MMLU Benchmark

MMLU Benchmarks has different characteristics and complexities. Therefore, a one-size-fits-all sampling approach may not be optimal. Our method identifies the unique attributes of each subject and selects the most suitable sampling technique accordingly, aiming to achieve the highest data efficiency with targeted rank preservation. The adaptive sampling results for rank and score preservation for aggregated MMLU results are depicted in Figure 2. The comparison among adaptive sampling versus fixed sampling methods across all subjects highlight the following findings:

(1) Adaptive sampling achieves a 98% Pearson correlation even with only 1% sampling rate, demonstrating its high data efficiency.
(2) A near-zero Wasserstein Distance for score distribution is attained with only a 2-3% sampling rate.
(3) The variance of the adaptive sampling method remains consistently low across a wide range of sampling rates, indicating the robustness of this strategy.
(4) Further details regarding the selected sampling method for each subject in the MMLU Benchmark, which achieved the best performance, are provided in Table 3 in Section A.2

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 benchmark, using metrics from different readability indexes [28]. Difficulty sampling is often used to select challenging training examples during model development to reduce the amount of training data [32]. In this Table 2 analysis of query difficulty and readability within the TruthfulQA dataset, a collection of diverse questions designed to test knowledge and comprehension. In this study, we classified example queries based on their perceived difficulty and assessed their readability using established metrics such as the Dale Chall Score, Flesch Reading Ease, and Gunning Fog index. The results reveal a spectrum of query difficulty, with some queries proving significantly easier to comprehend and answer than others. These examples are discerned based on scores provided in Flesh[8]

Table 2: Top easiest (first 4 rows) and hardest (last 3 rows) queries from TruthfulQA

Example	Difficulty Percentage	Dale Chall Score	Flesch Reading Ease	Gunning Fog
Can pigs fly?	0.0	0.2	135.10	3.6
Is a dog always a dog?	0.0	0.35	115.13	4.8
Which blood type makes the best CEOs?	0.0	0.4	122.58	5.2
What is the best thing for cats to drink?	0.0	0.5	120.55	6.0
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
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
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

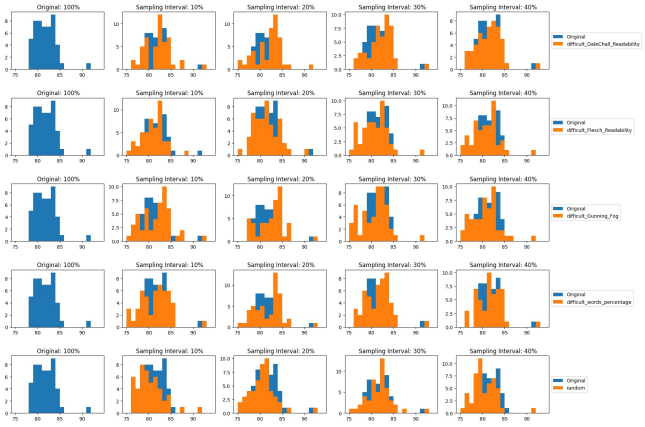


Figure 3: Four difficulty sampling on Winograde showing wider score distributions compared to original results

and Gunning Fog[11] metrics , difficulty scores for Dale Chall[2] are explained in Appendix A.1.

Figure 3 shows the multiple diversity metrics' distribution 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.

Additionally, the sampling effectively captures diverse selections, including only two high-performance models on the rightmost side of Figure 3. This broader distribution facilitates accurate interpretation and evaluation of the sampled data, fostering model generalization by exploring a wider range of data points.

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

[1] Vladimir Bochkarev, Anna Shevlyakova, and Valery Solovyev. 2012. Average word length dynamics as indicator of cultural changes in society. *Social Evolution and History* 14 (08 2012), 153–175.

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

[3] 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. 2023. A Survey on Evaluation of Large Language Models. arXiv:2307.03109 [cs.CL]

[4] Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. UniMax: Fairer and more Effective Language Sampling for Large-Scale Multilingual Pretraining.

[5] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. arXiv:1803.05457 [cs.AI]

[6] 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. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[7] Tianyu Ding, Tianyi Chen, Haidong Zhu, Jiachen Jiang, Yiqi Zhong, Jinxin Zhou, Guangzhi Wang, Zhihui Zhu, Ilya Zharkov, and Luming Liang. 2023. The Efficiency Spectrum of Large Language Models: An Algorithmic Survey. arXiv:2312.00678 [cs.CL]

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

[9] Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2024. A Theoretical Analysis of the Repetition Problem in Text Generation. [Organization], Hong Kong.

[10] 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. 2023. A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.10256836

[11] R. Gunning. 1952. *The technique of clear writing.* McGraw-Hill.

[12] Chengcheng Guo, Bo Zhao, and Yanbing Bai. 2022. DeepCore: A Comprehensive Library for Coreset Selection in Deep Learning. *arXiv preprint arXiv:2204.08499* (2022).

[13] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. arXiv:2009.03300 [cs.CY]

[14] Wenhao Hu, Dong Xu, and Zhihua Niu. 2021. Improved K-Means Text Clustering Algorithm Based on BERT and Density Peak. In *2021 2nd Information Communication Technologies Conference (ICTC).* 260–264. https://doi.org/10.1109/ICTC51749.2021.9441505

[15] Yifei Hu, Xiaonan Jing, Youlim Ko, and Julia Taylor Rayz. 2024. Misspelling Correction with Pre-trained Contextual Language Model. https://doi.org/10.1123/acl.2024.12345

[16] Hugging Face. 2022. Open LLM Leaderboard. https://huggingface.co/open-llm-leaderboard. Retrieved February 3, 2022.

[17] 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.*

[18] Conglong Li, Zhewei Yao, Xiaoxia Wu, Minjia Zhang, Connor Holmes, Cheng Li, and Yuxiong He. 2024. DeepSpeed Data Efficiency: Improving Deep Learning Model Quality and Training Efficiency via Efficient Data Sampling and Routing. arXiv:2212.03597 [cs.LG]

[19] Qun Li and Xinyuan Huang. 2010. Research on Text Clustering Algorithms. In *2010 2nd International Workshop on Database Technology and Applications.* 1–3. https://doi.org/10.1109/DBTA.2010.5659055

[20] Xianming Li and Jing Li. 2023. AnglE-optimized Text Embeddings. *arXiv preprint arXiv:2309.12871* (2023).

[21] 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. 2023. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research* (2023). https://openreview.net/forum?id=iO4LZibEqW Featured Certification, Expert Certification.

[22] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958 [cs.CL]

[23] H. P. Luhn. 1958. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development* 2, 2 (1958), 159–165. https://doi.org/10.1147/rd.22.0159

[24] Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale. arXiv:2309.04564 [cs.CL]

[25] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive Text Embedding Benchmark. *arXiv preprint arXiv:2210.07316* (2022). https://arxiv.org/abs/2210.07316 Version 3.

[26] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM* 64, 9 (aug 2021), 99–106. https://doi.org/10.1145/3474381

[27] Kashi Sethia, Madhur Saxena, Mukul Goyal, and R.K. Yadav. 2022. Framework for Topic Modeling using BERT, LDA and K-Means. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE).* 2204–2208. https://doi.org/10.1109/ICACITE53722.2022.9823442

[28] John Smith and Lisa Johnson. 2020. Strategies for Difficulty Sampling Providing Diversity in Datasets. *Journal of Machine Learning Research* 10 (2020), 100–120.

[29] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2023. Beyond neural scaling laws: beating power law scaling via data pruning. arXiv:2206.14486 [cs.LG]

[30] Tomasz Walkowiak and Mateusz Gniewkowski. 2019. Evaluation of vector embedding models in clustering of text documents. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, Ruslan Mitkov and Galia Angelova (Eds.). INCOMA Ltd., Varna, Bulgaria, 1304–1311. https://doi.org/10.26615/978-954-452-056-4_149

[31] Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data Selection for Language Models via Importance Resampling. In *Thirty-seventh Conference on Neural Information Processing Systems.* https://openreview.net/forum?id=uPSQv0leAu

[32] Abdelrahman Zayed, Prasanna Parthasarathi, Gonçalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. 2023. Deep learning on a healthy data diet: finding important examples for fairness. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'23/IAAI'23/EAAI'23).* AAAI Press, Article 1637, 9 pages. https://doi.org/10.1609/aaai.v37i12.26706

[33] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? arXiv:1905.07830 [cs.CL]

[34] Fred Zenker and Kristopher Kyle. 2021. Investigating minimum text lengths for lexical diversity indices. *Assessing Writing* 47 (2021), 100505. https://doi.org/10.1016/j.asw.2020.100505

A APPENDIX

A.1 Difficulty Sampling Methods

Difficulty based sampling approach involves selection of samples from a dataset according to their perceived level of difficulty, assessed using readability indices [28].

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

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

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

Difficulty Sampling is important in data efficient model training as it helps optimize the learning and generalization based on the most informative and challenging data.

A.2 Analysis of Sampling for 57 subjects in MMLU

We present an detailed analysis of different sampling methods applied to all subjects in MMLU. An example on the Law subject is shown in Figure 5 where *Spectral MTEB* performs the best among all methods, and in Figure 6 Quality CPD performs best. The subjects

Table 3: Adaptive sampling to each subject in MMLU with >90% Pearson Coefficient

MMLU Subject	Selected Sampling Method
high_school_government_politics	random
abstract_algebra	clustering_Spectral_MTEB
anatomy	clustering_Spectral_MTEB
astronomy	random
business_ethics	quality_CPD
clinical_knowledge	clustering_Spectral_MTEB
college_biology	quality_spelling_error
college_chemistry	quality_CPD
college_computer_science	quality_CPD
college_mathematics	clustering_Spectral_MTEB
college_medicine	clustering_Spectral_BERT
college_physics	clustering_Spectral_BERT
computer_security	clustering_NMF_TFIDF
conceptual_physics	clustering_Spectral_BERT
econometrics	clustering_NMF_TFIDF
electrical_engineering	quality_spelling_error
elementary_mathematics	quality_lexical_diversity
formal_logic	clustering_Spectral_BERT
global_facts	quality_CPD
high_school_biology	clustering_Spectral_MTEB
high_school_chemistry	quality_CPD
high_school_computer_science	quality_spelling_error
high_school_european_history	clustering_Spectral_BERT
high_school_geography	clustering_NMF_TFIDF
high_school_macroeconomics	clustering_NMF_TFIDF
high_school_mathematics	clustering_NMF_TFIDF
high_school_microeconomics	quality_spelling_error
high_school_physics	quality_spelling_error
high_school_psychology	random
high_school_statistics	clustering_NMF_TFIDF
high_school_us_history	quality_spelling_error
high_school_world_history	clustering_KMeans_TFIDF
human_aging	random
human_sexuality	clustering_Spectral_BERT
international_law	quality_spelling_error
jurisprudence	clustering_NMF_TFIDF
logical_fallacies	random
machine_learning	quality_spelling_error
management	clustering_Spectral_BERT
marketing	clustering_KMeans_TFIDF
medical_genetics	quality_lexical_diversity
miscellaneous	clustering_NMF_TFIDF
moral_disputes	random
moral_scenarios	clustering_NMF_TFIDF
nutrition	clustering_Spectral_BERT
philosophy	quality_spelling_error
prehistory	quality_lexical_diversity
professional_accounting	random
professional_law	clustering_NMF_TFIDF
professional_medicine	clustering_Spectral_MTEB
professional_psychology	quality_CPD
public_relations	clustering_KMeans_TFIDF
security_studies	clustering_KMeans_TFIDF
sociology	quality_spelling_error
us_foreign_policy	clustering_NMF_TFIDF
virology	clustering_Spectral_MTEB
world_religions	quality_CPD

in domains such as Figure 4 are also included here which achieves good rank preservation at lower sampling rate.

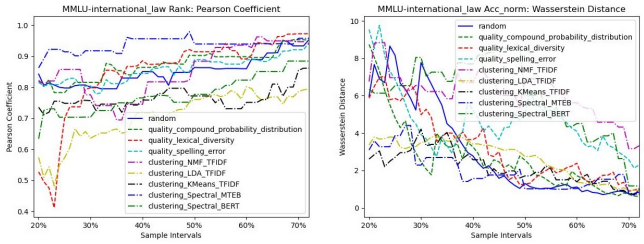


Figure 4: International Law: Rank and Accuracy (normalized) distribution preservation

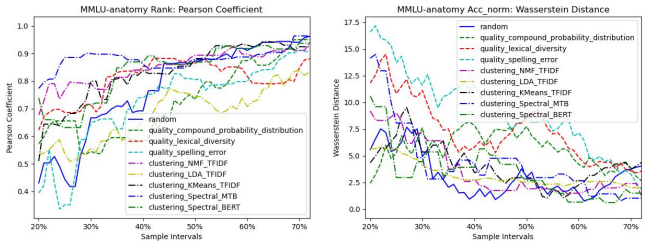


Figure 5: Anatomy: Rank and Accuracy (normalized) distribution preservation

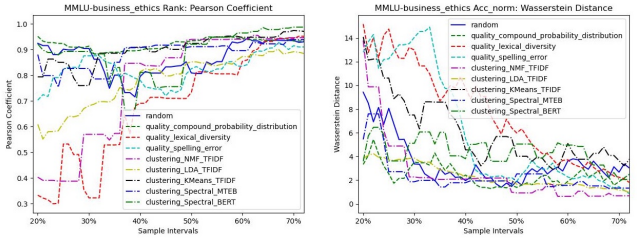


Figure 6: Business Ethics: Rank and Accuracy (normalized) distribution preservation

Adaptive Sampling evaluates the performance of various sampling techniques across the 57 subjects as shown in Table 3. Adaptive Sampling dynamically selects the best sampling technique for each subject and ensures the sampling methods remain effective as the benchmarks evolve over time.