Swift Cross-Dataset Pruning: Enhancing Fine-Tuning Efficiency in Natural Language Understanding

Binh-Nguyen Nguyen[1,3] **and Yang He**[1,2*]
[1]CFAR, Agency for Science, Technology and Research, Singapore
[2]IHPC, Agency for Science, Technology and Research, Singapore
[3]VNU University of Engineering and Technology, Hanoi, Vietnam
21020526@vnu.edu.vn, he_yang@cfar.a-star.edu.sg

Abstract

Dataset pruning aims to select a subset of a dataset for efficient model training. While data efficiency in natural language processing has primarily focused on within-corpus scenarios during model pre-training, efficient dataset pruning for task-specific fine-tuning across diverse datasets remains challenging due to variability in dataset sizes, data distributions, class imbalance and label spaces. Current cross-dataset pruning techniques for fine-tuning often rely on computationally expensive sample ranking processes, typically requiring full dataset training or reference models. We address this gap by proposing **Swift Cross-Dataset Pruning (SCDP)**. Specifically, our approach uses TF-IDF embeddings with geometric median to rapidly evaluate sample importance. We then apply dataset size-adaptive pruning to ensure diversity: for smaller datasets, we retain samples far from the geometric median, while for larger ones, we employ distance-based stratified pruning. Experimental results on six diverse datasets demonstrate the effectiveness of our method, spanning various tasks and scales while significantly reducing computational resources. Source code is available at: https://github.com/he-y/NLP-Dataset-Pruning.

1 Introduction

Deep learning progress has been fueled by massive datasets (Tan et al., 2024; Gadre et al., 2024), but managing and training on such data poses computational and storage challenges (Yang et al., 2023). Dataset pruning, or coreset selection, aims to identify a subset that achieves comparable model performance to the full dataset (Mirzasoleiman et al., 2020; Killamsetty et al., 2021), reducing training and storage costs while maintaining model effectiveness (Huang et al., 2021; Xia et al., 2022).

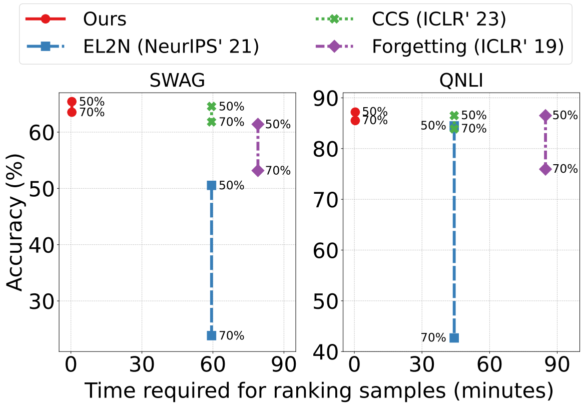

Figure 1: Accuracy and time required for ranking samples for SWAG, QNLI datasets with our proposed method, EL2N, CCS, Forgetting at 50% pruning rate and 70% pruning rate. Our method is significantly more time-efficient and yields higher accuracy.

This challenge is particularly evident in language model (LM) training, which involves two distinct scenarios: pre-training and fine-tuning, each requiring different data handling approaches. Data for pre-training, like those used for BERT, comprise large-scale, unlabeled, and diverse corpora like BookCorpus and English Wikipedia, collectively containing over 3,300 million words (Devlin et al., 2019). These corpora aim to facilitate the learning of broad language representations. In contrast, datasets for fine-tuning for downstream tasks, such as SWAG (113,000 examples) (Zellers et al., 2018), are smaller, labeled, and task-specific, designed to evaluate targeted abilities such as commonsense reasoning. Although dataset efficiency techniques such as language filtering, quality assessment, and deduplication (Albalak et al., 2024; Longpre et al., 2024) are proposed for large-scale pre-training corpora, they are not suitable for fine-tuning.

While single-dataset fine-tuning benefits from a narrow target distribution (Albalak et al., 2024), establishing general dataset pruning rules for cross-

[*]Corresponding Author
This work was completed during Binh-Nguyen Nguyen's internship at CFAR, A*STAR.

Proceedings of the 31st International Conference on Computational Linguistics, pages 726–739
January 19–24, 2025. ©2025 Association for Computational Linguistics

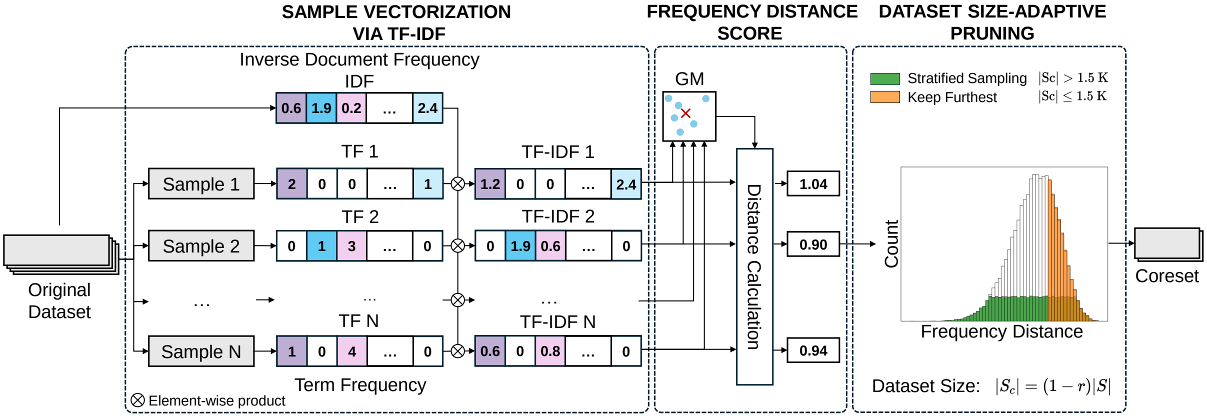

Figure 2: Overview of the proposed method. We introduce the Frequency Distance (FD) score, in which we leverage TF-IDF embeddings combined with geometric median calculations to swiftly assess sample importance. We propose dataset size-adaptive pruning to enhance adaptability in cross-dataset setting.

dataset fine-tuning remains challenging due to the diversity of natural language processing (NLP) tasks. Common benchmarks reveal significant variations in task types, dataset sizes, and domains. For instance, for fine-tuning datasets, training set sizes range from just 2.49k examples (RTE) to 105k examples (QNLI), while task types span from single-sentence classification (e.g., SST-2) to complex inference tasks (e.g., SWAG).

This heterogeneity is further complicated by the different data types, ranging from movie reviews to news reports and Wikipedia articles. Such diversity presents a unique challenge in developing pruning strategies that can effectively generalize across the spectrum of NLP tasks, underscoring the need for data pruning approaches in cross-dataset scenarios.

Existing cross-dataset pruning methods require computationally expensive sample ranking processes. These methods require training to be run on original data to collect pruning statistics, access to reference models and label information (Fayyaz et al., 2022; Zayed et al., 2023). As shown in Figure 1, these methods typically take around 60 minutes to process standard datasets like SWAG, with more complex approaches or larger datasets demanding even more time. In contrast, our method achieves comparable or superior performance in mere seconds, regardless of dataset size or task complexity. The superiority in time efficiency and performance of our method is shown in Figure 1.

To tackle these problems, we introduce Swift Cross-Dataset Pruning (SCDP). Specifically, our approach introduces **Frequency Distance (FD)** score, in which we leverage TF-IDF embeddings combined with geometric median calculations to swiftly assess sample importance. This technique offers two significant advantages. 1) **Cross-Dataset Generalizability**: By using TF-IDF embeddings, our method captures the semantic importance of words across various NLU tasks and domains. The geometric median calculation then provides a task-agnostic measure of centrality in the embedding space. This combination ensures that our approach is universally adaptable across multiple datasets and task types, from language inference to reasoning and beyond. 2) **Computational Efficiency**: Unlike existing cross-dataset pruning methods that often require computationally expensive processes such as model training, access to reference models, or label information (Fayyaz et al., 2022; Zayed et al., 2023), our approach allows for rapid evaluation of sample importance. The TF-IDF and geometric median calculations can be performed efficiently on raw text data, significantly reducing the computational overhead typically associated with sample ranking processes.

Furthermore, we apply **dataset size-adaptive pruning** to ensure diversity for two distinct scenarios. For smaller datasets, we retain samples far from the geometric median, preserving outliers and edge cases to maintain diversity by keeping "unusual" examples. For larger datasets, we select samples from each stratum to maintain a balanced representation of the data distribution while significantly reducing the dataset size.

Our main contributions are:

- We propose Frequency Distance, a score that uses TF-IDF embeddings and geometric median to swiftly rank samples.
- We propose dataset size-adaptive pruning to

enhance adaptability in cross-dataset setting.

- We conduct extensive experiments on six diverse datasets, encompassing various tasks such as paraphrase identification, natural language inference, and reasoning. Our experiments span a range of dataset sizes, demonstrating the superior efficiency and performance of our method in cross-dataset settings.

2 Related Work

Dataset Pruning for Vision Dataset. Early works in dataset pruning focus its application in computer vision tasks. Toneva et al. (2018) define forgetting as the number of transitions from correct prediction to incorrect prediction of a sample in training, and use this to rank samples. Paul et al. (2021) propose to use EL2N and GraNd scores obtained from training period to rank samples. AUM (Pleiss et al., 2020) is a metric that calculates the difference between the logit of the ground truth label and the highest other logit. Coleman et al. (2020) propose to use a small proxy model to obtain presentation for dataset pruning. These works require training on original data and a pre-trained model to obtain sample ranking metrics and are computationally expensive for large-scale models. On sampling from ranking metrics, Zheng et al. (2023) propose coverage-centric coreset selection, an algorithm based on stratified sampling for dataset pruning and achieve better performance at high pruning rates for image classification tasks. Xia et al. (2022) selects data points with scores that are close to the score median to build a moderate coreset.

Dataset Pruning for Language Model Pre-training. Most previous works on data efficiency for language tasks focus on the pre-training phase of LMs. Common approaches for this task are language filtering (Wenzek et al., 2020; Raffel et al., 2020; Xue et al., 2021; Laurençon et al., 2022), heuristic filtering (Rae et al., 2021; Xue et al., 2021), data quality assessment (Du et al., 2022; Marion et al., 2023), data deduplication (Lee et al., 2022; Abbas et al., 2023; Tirumala et al., 2024), toxic or explicit content filtering (Jansen et al., 2022; Subramani et al., 2023; Maharana et al., 2024). These methods address issues like irrelevant content and data redundancy before the model learns from the data, which is crucial in LM pre-training. However, it is difficult to apply these methods to cross-dataset scenarios, due to differences in task target, use-case, dataset pruning criterias. Dataset pruning in cross-dataset settings is more difficult due to the wide range of tasks, dataset sizes involved.

Dataset Pruning for Language Model Fine-tuning. For fine-tuning phase of pre-trained LMs, Attendu and Corbeil (2023) and Fayyaz et al. (2022) study the application of EL2N and GraNd scores to fine-tuning transformer-based language models. Zayed et al. (2023) propose a new metric based on EL2N which also use model logits to prune samples for fairness. Yang et al. (2024) use training trajectories from small models to select samples for dataset pruning. Chen et al. (2024) try to use strong LLMs such as ChatGPT to rate the quality of samples. Maharana et al. (2024) create a graph for the whole dataset, in which each node represents a sample and is initiliazed with difficulty score from model training. These methods require extensive training time on original data and depend on models to assess data importance, making them computationally expensive.

3 Methodology

3.1 Task Description and Formulation

Given a training set of a fine-tuning task $S = (x_i, y_i)_{i=1}^N$, where $N \in \mathbb{N}$, x_i represents the i-th input and y_i is its true label. Pruning rate $r \in \mathbb{R}$, where $0 < r < 1$ represents the portion of dataset that will be removed. The objective of dataset pruning is to identify a coreset $S_c \subset S$ where $|S_c| = (1 - r)|S|$, and when the model is trained on S_c, it may still retain the highest possible performance on the test set.

3.2 Sample Vectorization via TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) (Sparck Jones, 1972) is a numerical measure that indicates the importance of a word within a specific document in relation to its occurrence across a collection of documents, or corpora. TF-IDF is ideal in this scenario because it captures the significance of terms relative to the entire dataset while being fast and scalable. Unlike transformer-based embeddings like BERT (Devlin et al., 2019), which focus on contextual similarity and are computationally intensive, TF-IDF emphasizes term frequency and rarity, making it efficient for identifying unique and informative samples. This allows for effective pruning strategies that preserve diverse and representative samples.

We use TF-IDF for unigrams to assess the significance of a term in a sample by considering how frequently it appears in that sample and how infrequently it appears across other samples in the dataset. From the dataset S, we obtain the vocabulary $\mathcal{V} = \{w_0, w_1, ..., w_n\}$ which contains all terms from all samples in dataset S, w_j denotes the j-th term in the vocabulary.

The term frequency for term w_j in sample x_i, denoted as $\mathbf{tf}_{i,j} \in \mathbb{R}$, is calculated as:

$$\mathbf{tf}_{i,j} = \frac{f(w_j, x_i)}{\sum_{w_k \in x_i} f(w_k, x_i)}, \tag{1}$$

where $f(w_k, x_i)$ is the number of occurrences of term w_k in sample x_i.

By concatenating all the term frequencies of each term in sample x_i, we get the term frequency vector $\mathbf{tf}_i \in \mathbb{R}^n$ of sample x_i as follows:

$$\mathbf{tf}_i = \mathbf{tf}_{i,0} \oplus \mathbf{tf}_{i,1} \oplus ... \oplus \mathbf{tf}_{i,n}. \tag{2}$$

The inverse document frequency of a term in the dataset $\mathbf{idf}_j \in \mathbb{R}$ is computed as:

$$\mathbf{idf}_j = \log \frac{N}{1 + \text{df}(w_j)}. \tag{3}$$

where $\text{df}(w_j)$ is the number of samples from the dataset that contains the term w_j.

By concatenating all the inverse document frequencies of each term, we get the term frequency vector $\mathbf{idf} \in \mathbb{R}^n$ as follows:

$$\mathbf{idf} = \mathbf{idf}_0 \oplus \mathbf{idf}_1 \oplus ... \oplus \mathbf{idf}_n, \tag{4}$$

Each sample x_i is represented as a vector of TF-IDF scores. The TF-IDF vector $\mathbf{t}_i \in \mathbb{R}^n$ for sample x_i is given by:

$$\mathbf{t}_i = \mathbf{tf}_i \odot \mathbf{idf}, \tag{5}$$

where $\odot$ denotes the element-wise product operation of two vectors.

3.3 Frequency Distance Score

Due to the sparse nature of TF-IDF representations in datasets, clustering methods, as previously used for dataset pruning (Das and Khetan, 2023; Yang et al., 2024), cannot be applied. Therefore, we propose Frequency Distance (FD), a new distance-based scoring metric with geometric median, which calculates distance of each sample in the embedding space to the geometric median. This score represents the relative position of each sample with regard to the geometric median of the whole dataset in the embedding space.

Algorithm 1 Dataset-size adaptive pruning

```
Inputs: $S = \{(x_i, y_i)\}_{i=1}^n$: original dataset; $D = \{\text{FD}(x_i)\}_{i=1}^n$: set of calculated Frequency Distance scores; $r$: dataset pruning rate; $k$: the number of strata.
Outputs: $S_c$: the selected coreset
  function SizeAdaptivePruning($S$, $D$, $r$, $k$)
    if $(1-r)|S| > 1500$ then
      $R_1, R_2, ..., R_k \leftarrow$ splits scores from $D$ into $k$ ranges with even width
      $\mathcal{B} \leftarrow \{\mathbb{B}_i$: $\mathbb{B}_i$ consists of samples whose scores are in $R_i\}$
      $m \leftarrow n \times (1-r)$
      $S_c \leftarrow \varnothing$
      while $\mathcal{B} \neq \varnothing$ do
        $\mathcal{B}_{min} \leftarrow \arg\min_{\mathbb{B} \in \mathcal{B}} |\mathbb{B}|$
        $m_B \leftarrow \min\{|\mathbb{B}_{min}|, \lfloor \frac{m}{|\mathcal{B}|} \rfloor\}$
        $S_B \leftarrow$ randomly sample $m_B$ samples from $\mathbb{B}_{min}$
        $S_c \leftarrow S_c \cup S_B$
        $\mathcal{B} \leftarrow \mathcal{B} \setminus \{\mathbb{B}_{min}\}$
        $m \leftarrow m - m_B$
      end while
    else if $(1-r)|S| \leq 1500$ then
      $D' \leftarrow \texttt{argsort}(D)$
      $S_c \leftarrow D'[(1-r)|S| :]$
    end if
  end function
```

From the set of N points $\{\mathbf{t}_0, \mathbf{t}_1, ..., \mathbf{t}_N\}$ which represent the embeddings of each document, we find the geometric median point $\mathbf{g}^* \in \mathbb{R}^n$ that minimizes the sum of L2 distances to every point:

$$\mathbf{g}^* = \arg\min_{\mathbf{g} \in \mathbb{R}^n} f(\mathbf{g}), \\ \text{where } f(\mathbf{g}) = \sum_{i \in [1,N]} ||\mathbf{g} - \mathbf{t}_i||_2. \tag{6}$$

Computing the geometric median is challenging, and no linear time algorithm currently exists (Bajaj, 1988). Consequently, we employ an approximation technique proposed by Vardi and Zhang (2000) to estimate the geometric median. This approach yields an ϵ-accurate geometric median, satisfying the condition $f(\mathbf{g}_\epsilon) \leq (1 + \epsilon) f(\mathbf{g}^*)$.

For each sample x_i in the dataset, we obtain FD score by calculating the L2 distance of its embed-

Dataset	**Metric**	**Task**	**Size**
RTE	Accuracy	NLI	2.49k
MRPC	Accuracy	Paraphrase	3.67k
CoLA	Matthews corr.	Grammatical Acceptability	8.55k
SST-2	Accuracy	Sentiment Analysis	67.3k
SWAG	Accuracy	Reasoning	73.5k
QNLI	Accuracy	QA/NLI	105k

Table 1: Evaluation metric, task and original size of train set of the datasets used in experimental evaluation. QA: Question Answering, NLI: Natural Language Inference.

PR	EL2N	AUM	Forgetting	CCS	Ours
			RTE		
50%	45.84	45.72	47.53	50.78	**55.83**
70%	43.96	45.00	45.12	48.49	**57.40**
			MRPC		
50%	68.54	68.38	74.58	77.77	**83.00**
70%	68.38	68.38	70.09	71.64	**75.73**
			CoLA		
50%	12.90	0.00	43.94	**46.38**	45.30
70%	0.01	0.00	4.05	36.86	**43.39**
			SST-2		
50%	90.51	**90.97**	90.75	90.02	90.74
70%	88.79	88.99	89.90	89.52	**90.21**
			SWAG		
50%	55.76	50.54	61.40	64.57	**65.43**
70%	28.14	23.83	53.17	61.82	**63.54**
			QNLI		
50%	83.88	84.47	86.50	86.50	**87.19**
70%	66.53	42.68	75.94	83.88	**85.52**

Table 2: Overall result of baselines and our method. Pruning rate (PR) is the percentage of data that is removed from full training data during dataset pruning. The best results are highlighted in **bold**.

ding to the geometric median $d_i \in \mathbb{R}$ as follows:

$$\mathrm{FD}(x_i) = ||\mathbf{t}_i - \mathbf{g}_\epsilon||_2. \tag{7}$$

For each sample, we use Eq. 7 to calculate its score, to obtain the set of score of every sample $D = \{\mathrm{FD}(x_i)\}_{i=1}^{N}$. This metric is used to evaluate data points in the training set to perform dataset pruning.

3.4 Dataset size-adaptive pruning

In our cross-dataset setting, to ensure performance across diverse scales of datasets, we present dataset size-adaptive pruning, a novel sampling method to choose samples from set of FD scores, as presented in Algorithm 1. By applying dataset size-adaptive pruning, we ensure diversity for two distinct scenarios.

For smaller datasets, every sample could potentially carry unique information crucial for model performance. We retain samples far from the geometric median, preserving outliers and edge cases. This maintains diversity by keeping "unusual" examples that, while rare, are crucial for comprehensive model training and incentivize the understanding of rare or complex patterns.

For larger datasets, the challenge lies in maintaining a balanced representation of the data distribution while significantly reducing the dataset size. By selecting samples from each stratum following Zheng et al. (2023), we ensure a diverse range of examples in the pruned dataset, from typical central cases to unique peripheral ones. Dataset size-adaptive pruning is applied as follows:

- Case 1: For **small post-pruning coreset size**, where $(1-r)|S| \leq 1500$, keeping furthest samples is the preferred strategy. From the set of calculated FD scores D, this strategy keeps the samples furthest to geometric median as coreset S_c.
- Case 2: For **middle to large post-pruning coreset size**, where $(1-r)|S| > 1500$, since we have enough representative samples in each stratum, we use stratified sampling to ensure that the pruned dataset retains a representative mix of samples across different strata.

4 Experiments

4.1 Experiment Settings

Evaluation Datasets. We experimented on six natural language understanding datasets, including RTE, QNLI, CoLA, MRPC, SST-2 from the GLUE benchmark (Wang et al., 2019), SWAG (Zellers et al., 2018). The task and evaluation metric of each dataset is listed in Table 1. The selected datasets have a large variance in size of samples and has a wide variety of tasks, which demonstrates the universal effectiveness of our method.

Model settings. We fine-tune pre-trained DistilBERT (Sanh et al., 2019) in all experiments. A task-specific head is added to the final layer of DistilBERT. First, we use the proposed method to prune and obtain the remaining coreset and fine-

Pruning rate	10%	30%	50%	70%
		RTE		
Sentence-BERT	57.03	**57.88**	**55.83**	49.81
TF-IDF	**61.85**	57.39	**55.83**	**57.40**
		MRPC		
Sentence-BERT	84.47	83.33	81.78	72.87
TF-IDF	**84.55**	**83.41**	**83.00**	**75.73**
		CoLA		
Sentence-BERT	**49.53**	47.39	**45.65**	41.58
TF-IDF	49.43	**48.09**	45.30	**43.39**
		SST-2		
Sentence-BERT	90.63	90.71	90.21	89.44
TF-IDF	**90.97**	**91.13**	**90.74**	**90.21**
		SWAG		
Sentence-BERT	67.09	66.63	64.57	63.47
TF-IDF	**67.45**	**66.68**	**65.43**	**63.54**
		QNLI		
Sentence-BERT	89.01	87.80	**87.21**	85.30
TF-IDF	**89.10**	**88.27**	87.19	**85.52**

Table 3: Ablation study for embedding methods. The best results are highlighted in **bold**.

Pruning rate	10%	30%	50%	70%
		RTE		
Random	58.06	54.75	53.90	54.75
Our method	**61.85**	**57.39**	**55.83**	**57.40**
		SWAG		
Random	65.20	64.75	63.96	62.88
Our method	**67.45**	**66.68**	**65.43**	**63.54**
		QNLI		
Random	87.44	87.32	86.06	85.27
Our method	**89.10**	**88.27**	**87.19**	**85.52**

Table 4: Comparison with random selection. The best results are highlighted in **bold**.

Model	BERT	ALBERT	XLNet	RoBERTa
Random	57.03	52.70	**61.01**	59.92
EL2N	39.71	45.84	47.65	44.40
AUM	43.32	48.01	50.54	52.70
Forgetting	44.76	55.59	50.54	52.70
CCS	56.67	52.70	53.79	50.54
Ours	**61.37**	**57.40**	59.20	**56.67**

Table 5: Experiments on other models on RTE dataset at 70% pruning rate. The best and the second best results are highlighted in **bold** and underlined, respectively.

tune DistilBERT with the coreset. We fine-tune DistilBERT on 3 epochs with an initial learning rate of $5e-5$ with cosine annealing scheduler (Loshchilov and Hutter, 2016), batch size 32, using AdamW optimizer (Loshchilov and Hutter, 2019).

Pruning settings. For geometric median approximation, we use $\epsilon = 1e-5$. For stratified sampling, we set the number of strata to $k = 100$.

Baselines. We compare our method against five baselines, four of which are SOTA methods for dataset pruning. We use **(1) Random**: we randomly select samples to form the coreset, **(2) AUM** (Pleiss et al., 2020), **(3) EL2N** (Paul et al., 2021), **(4) Forgetting** (Toneva et al., 2018), **(5) CCS** (Zheng et al., 2023): coverage-centric coreset selection with AUM as the importance score. All experiments are run three times and average score is reported in this paper.

4.2 Experimental Results

Main Experiments. The evaluation results of the baseline methods are compared to our proposed method. We conduct experiments on multiple pruning rates to investigate how our method perform at different data compression rates. In Table 2, we present the performance on all datasets at 50% and 70% pruning rates. Full results at 10%, 30%, 50% and 70% over all datasets are presented in Appendix A.4. In overall, our proposed method performs the best compared to all other baselines methods and outperforms baseline methods.

When compared to SOTA baselines, our method has best performance overall. Our method consistently outperforms state-of-the-art baselines AUM, EL2N, Forgetting and CCS. For high post-pruning coreset size, distance-based stratified sampling is particularly effective because it ensures that the pruned dataset remains both diverse and informative, eliminating redundant data while preserving the core structure of the dataset. At small post-pruning coreset size, furthest samples prove to be effective, since most distant from the central tendency can help maintain the diversity and richness of the data.

Comparison between our method and random baseline is shown in Table 4. At 10% pruning rate, the overall improvement to random is 2.57%, and at 70% pruning rate, the overall improvement to random is 1.19%.

We tested our methods on other language models, specifically BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), XLNet (Yang et al., 2019) and RoBERTa (Liu, 2019) on the RTE dataset to evaluate the robustness of our methods, as shown in Table 5. Our method consistently perform well

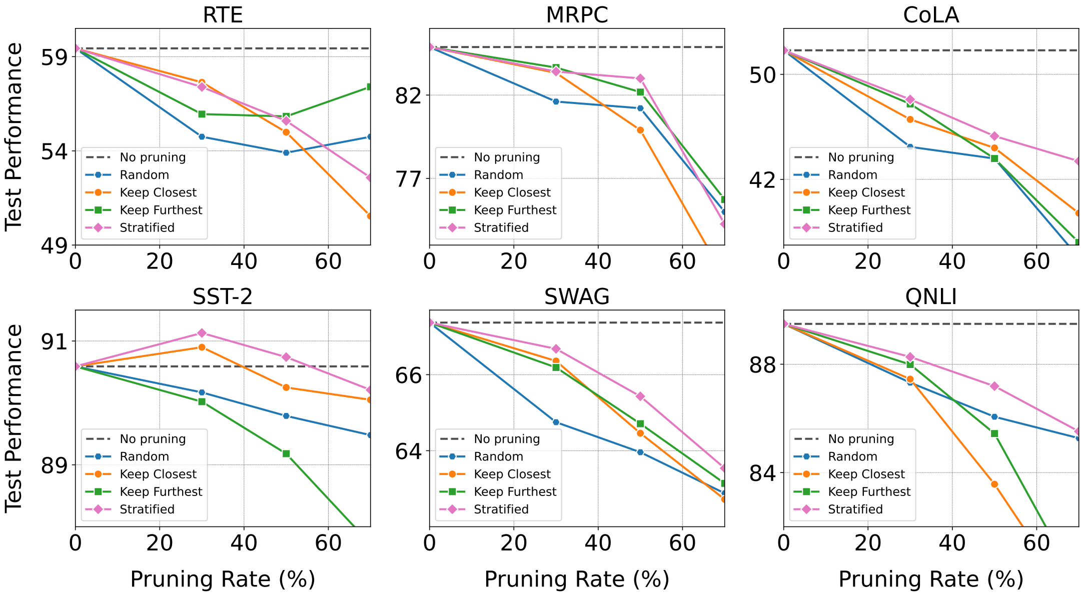

Figure 3: Results of pruning strategies from set of distance-based scores.

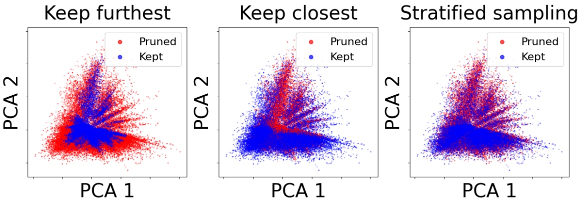

Figure 4: PCA Plot of selected data points with regard to full training set for SWAG dataset at 70% pruning rates for different pruning strategies.

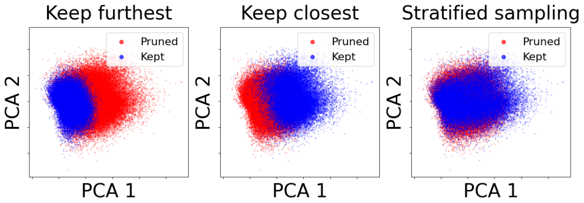

Figure 5: PCA Plot of selected data points with regard to full training set for QNLI dataset at 70% pruning rates for different pruning strategies.

across all models, which prove that our method is not a phenomenon with respect to a particular language model.

Dataset size-adaptive pruning. In Figure 3, we compare three different pruning strategies from distance-based scores: only keeping the closest samples to geometric median, only keeping the furthest samples to geometric median, and using stratified sampling to select the coreset.

With regard to the proposed dataset size-adaptive pruning, there are three cases where keeping furthest samples are applied: for MRPC dataset at 70% pruning rate, and for RTE dataset at 50% and 70% pruning rates. For these scenarios, furthest samples prove to be more effective coresets, producing higher performance than stratified sampling. For coresets with bigger size, stratified sampling prove to be the best strategy compared to the other methods, since it maintains the distribution characteristics of the original dataset.

Keeping closest samples method still retain good performance at 30% pruning rate. However, at higher pruning rates, its performance drop quickly, for most datasets it may perform worse than random pruning.

4.3 Further Analysis

TF-IDF embedding performs superior to Sentence-BERT embedding. To validate the effectiveness of TF-IDF embedding, we performed experiments with Sentence-BERT (Reimers and Gurevych, 2019) embedding of samples. The performance of TF-IDF is superior compared to Sentence-BERT across all datasets, as shown in Table 3. TF-IDF is superior to Sentence-BERT embeddings in this context because it directly captures the importance of individual words within each sample, emphasizing terms that are significant to the dataset. This makes TF-IDF particularly effective for identifying central and representative samples in the dataset, as it reflects the specific vocabulary and term frequency patterns of the data. In

ID	Example	FD	QuRating
2015	What does 基督徒(pinyin: jīdū tú) mean? (pinyin: jīdū tú), literally "Christ follower."	1.009 (99.99%)	-1.22 (6.72%)
22851	Which vowel remains distinct? ; /i/ remains distinct.	1.008 (99.99%)	-0.83 (11.90%)
10539	Name one of the individuals considered the founding fathers of modern Cypriot art. Arguably the two founding fathers of modern Cypriot art were Adamantios Diamantis (1900–1994) who studied at London's Royal College of Art and Christopheros Savva (1924–1968) who also studied in London, at Saint Martin's School of Art.	0.990 (53.77%)	3.92 (98.58%)
40675	Where was the University of Paris located? The Left Bank was the site of the University of Paris, a corporation of students and teachers formed in the mid-12th century to train scholars first in theology, and later in canon law, medicine and the arts.	0.966 (0.31%)	3.10 (95.28%)

Table 6: Comparison to QuRating. The percentage shows a sample's percentile rank in the dataset.

ID	Example	FD	PPL
2015	What does 基督徒(pinyin: jīdū tú) mean? (pinyin: jīdū tú), literally "Christ follower."	1.009 (99.99%)	34.19 (30.37%)
22851	Which vowel remains distinct? ; /i/ remains distinct.	1.008 (99.99%)	399.56 (99.81%)
10539	Name one of the individuals considered the founding fathers of modern Cypriot art. Arguably the two founding fathers of modern Cypriot art were Adamantios Diamantis (1900–1994) who studied at London's Royal College of Art and Christopheros Savva (1924–1968) who also studied in London, at Saint Martin's School of Art.	0.990 (53.77%)	31.06 (24.77%)
40675	Where was the University of Paris located? The Left Bank was the site of the University of Paris, a corporation of students and teachers formed in the mid-12th century to train scholars first in theology, and later in canon law, medicine and the arts.	0.966 (0.31%)	26.04 (16.06%)

Table 7: Comparison to perplexity (PPL) of GPT-2.

contrast, Sentence-BERT embeddings focus on capturing the overall semantic meaning of sentences, which obscure the importance of individual words and lead to less precise centrality measures.

Stratified sampling maintains good coverage in embedding space. Figure 4 and Figure 5 demonstrate the selected data points and the pruned data points for 70% pruning rate in 2 dimension space using principle component analysis (Wold et al., 1987). Stratified sampling method works very well for medium to large size datasets since it is able to keep a balanced representation of the data distribution. Compared to keep furthest strategy or keep closest strategy, stratified sampling coreset covers much wider regions of the plot. Interestingly, in the keep furthest samples strategy, selected samples are placed relatively close to the center of the 2 dimensional plot in the SWAG dataset. This is explained by the sparsity of TF-IDF vectors of the furthest samples, where rare words are included and their lengths are often shorter.

Comparison of our Frequency Distance score with previous quality-based score. To compare our method with quality-based methods in pre-training dataset pruning, we use QuRating (Wettig et al., 2024), a method that utilizes a fine-tuned Sheared-Llama-1.3B model (Xia et al., 2024) to judge the quality of text, and perplexity (PPL) of GPT-2 (Radford et al., 2019). Quality assessment is often used in pre-training dataset pruning to select data that resembles a high-quality corpus (Albalak et al., 2024; Wettig et al., 2024; Chowdhery et al., 2023; Xie et al., 2023). For QuRating, high scoring samples are preferred, while for perplexity, low scoring samples are preferred. However, this approach is not applicable to cross-dataset pruning, since it hurts diversity of the coreset and harms model performance in this scenario.

In Table 6 and Table 7, we show examples that are chosen by our stratified sampling strategy at 70% pruning rate in the QNLI dataset and compare FD score to QuRating and perplexity scores. Our method selects samples ranging from low quality to high quality, ensuring a more diverse set of samples. In contrast, quality-based methods like QuRating or perplexity-based quality filtering prioritizes high quality samples, which can lead to a loss of important linguistic and contextual variety.

5 Conclusion

In this paper, we introduce SCDP to enhance fine-tuning efficiency for NLP tasks in cross-dataset scenarios. We propose Frequency Distance, a sample ranking score that is computationally efficient and bypasses the need for expensive reference models or training on the full original data. Furthermore, we propose dataset size-adaptive pruning to improve adaptability across a diverse range of tasks and dataset sizes. Our extensive experiments across six datasets validate the effectiveness of this approach, demonstrating that our method achieves competitive performance while reducing computational costs. This work represents a significant step towards data-efficient training in NLP, particularly for fine-tuning in cross-dataset settings.

6 Limitations

While our method shows promising results, it has several limitations. 1) Task diversity: Although tested on six NLU tasks, the method's performance on more complex tasks such as text generation is unknown. 2) Theoretical grounding: While empirically effective, our work lacks a rigorous theoretical analysis explaining the superiority of TF-IDF with the geometric median for cross-dataset pruning.

References

Amro Kamal Mohamed Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Transactions on Machine Learning Research*. Survey Certification.

Jean-michel Attendu and Jean-Philippe Corbeil. 2023. Nlu on data diets: Dynamic data subset selection for nlp classification tasks. In *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 129–146.

Chanderjit Bajaj. 1988. The algebraic degree of geometric optimization problems. *Discrete & Computational Geometry*, 3:177–191.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2024. Alpagasus: Training a better alpaca with fewer data. In *International Conference on Learning Representations*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

C Coleman, C Yeh, S Mussmann, B Mirzasoleiman, P Bailis, P Liang, J Leskovec, and M Zaharia. 2020. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*.

Devleena Das and Vivek Khetan. 2023. Deft: Data efficient fine-tuning for large language models via unsupervised core-set selection. *arXiv preprint arXiv:2310.16776*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Mohammad Taher Pilehvar, Yadollah Yaghoobzadeh, and Samira Ebrahimi Kahou. 2022. Bert on a data diet: Finding important examples by gradient-based pruning. *arXiv preprint arXiv:2211.05610*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Jiawei Huang, Ruomin Huang, Wenjie Liu, Nikolaos Freris, and Hu Ding. 2021. A novel sequential coreset method for gradient descent algorithms. In *International Conference on Machine Learning*, pages 4412–4422. PMLR.

Tim Jansen, Yangling Tong, Victoria Zevallos, and Pedro Ortiz Suarez. 2022. Perplexed by quality: A perplexity-based method for adult and harmful content detection in multilingual heterogeneous web data. *arXiv preprint arXiv:2212.10440*.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. 2021. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8110–8118.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. 2024. D2 pruning: Message passing for balancing diversity & difficulty in data pruning. In *The Twelfth International Conference on Learning Representations*.

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 preprint arXiv:2309.04564*.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. 2020. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Q Weinberger. 2020. Identifying mislabeled data using the area under the margin ranking. *Advances in Neural Information Processing Systems*, 33:17044–17056.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220.

Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. 2024. Data pruning via moving-one-sample-out. *Advances in Neural Information Processing Systems*, 36.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2024. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. 2018. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*.

Yehuda Vardi and Cun-Hui Zhang. 2000. The multivariate l 1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Édouard Grave. 2020. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012.

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

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024. Sheared llama: Accelerating language model pre-training via structured pruning. In *The Twelfth International Conference on Learning Representations*.

Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. 2022. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36:34201–34227.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. 2023. Dataset pruning: Reducing training data by examining generalization influence. In *The Eleventh International Conference on Learning Representations*.

Yu Yang, Siddhartha Mishra, Jeffrey N Chiang, and Baharan Mirzasoleiman. 2024. Smalltolarge (s2l): Scalable data selection for fine-tuning large language models by summarizing training trajectories of small models. *arXiv preprint arXiv:2403.07384*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

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 AAAI Conference on Artificial Intelligence*, volume 37, pages 14593–14601.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104.

Haizhong Zheng, Rui Liu, Fan Lai, and Atul Prakash. 2023. Coverage-centric coreset selection for high pruning rates. In *International Conference on Learning Representations*.

A Experiments

A.1 Dataset Descriptions

We provide a detailed description of datasets used in our experiments below:

- **RTE**. In the input, two text fragments are given. The task is to recognize whether the meaning of one text is entailed from the other text.
- **MRPC**. This dataset consists of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent.
- **CoLA**. This dataset consists of English acceptability judgments drawn from books and journal articles on linguistic theory. Each example is a sequence of words annotated with whether it is a grammatical English sentence.
- **SST-2**. This dataset consists of sentences from movie reviews and human annotations of their sentiment. The task is to predict the sentiment of a given sentence.
- **SWAG**. Given a partial text description, the model has to reason about the situation and anticipate what comes next by choosing one of multiple-choice text options.
- **QNLI**. This dataset is a question-answering dataset consisting of question-paragraph pairs, where one of the sentences in the paragraph contains the answer to the corresponding question. The task is to determine whether the context sentence contains the answer to the question.

A.2 Data Preprocessing

For each dataset, we preprocess the input before using TF-IDF embedding as follows:

- **SWAG.** Concatenate sentence 1 and sentence 2 as input to get TF-IDF embedding, and not include 4 answer options.
- **SST-2.** The input sentence is the input to get TF-IDF embedding.
- **QNLI.** Concatenate question and answer as input to get TF-IDF embedding.
- **RTE.** Concatenate sentence 1 and sentence 2 as input to get TF-IDF embedding.
- **CoLA.** The input sentence is the input to get TF-IDF embedding.
- **MRPC.** The input sentence is the input to get TF-IDF embedding.

A.3 Baseline Experiment Settings

For **AUM**, **EL2N**, **CCS** baseline, we get training statistics in every epoch. That is, for each epoch, predictions on the training set will be used to calculate sample importance. Since we fine-tune DistilBERT on 3 epochs for every dataset, we obtain scores 3 times and use the mean as the final sample importance score. For **Forgetting** baseline, since 3 times are not enough to evaluate forgetting scores, we evaluate the model on the training set after iterations instead of epochs. That is, we evaluate the model on the training set after every fixed number of iterations to get the forgetting information of samples.

A.4 Detailed Results

Table 8 describes the detailed results over 10%, 30%, 50%, 70% pruning rates on all datasets. The result of fine-tuning on the full dataset is shown in the rows with 0% pruning rate

B Examples

B.1 Quality-based Comparison

In Table 9-13, we list examples chosen by our method at 70% pruning rate and compute their quality-based score with QuRating and Perplexity of GPT-2. Examples show that our method is able to keep the diversity of samples based on quality and perplexity, which benefits the model to generalize to linguistic and contextual features.

B.2 Examples of Furthest Samples and Closest Samples to Geometric Median

Text input of the samples furthest to the geometric median and samples closest to the geometric median of the QNLI dataset are displayed in Table 14. By applying FD score, we can obtain the most distinctive samples by choosing samples with highest FD score. On the other hand, samples nearest to geometric median tend to have similar terms and topics. Therefore, it is suitable to use our sampling method to keep the most diverse samples to benefit the model performance on the coreset.

Pruning rate	Random	EL2N	AUM	Forgetting	CCS	Ours
RTE						
0%	$59.44_{\pm 4.37}$					
10%	$58.06_{\pm 1.50}$	$53.67_{\pm 0.91}$	$54.63_{\pm 3.34}$	$55.95_{\pm 2.73}$	$58.12_{\pm 1.91}$	$\mathbf{61.85}_{\pm \mathbf{1.82}}$
30%	$54.75_{\pm 0.91}$	$48.49_{\pm 3.24}$	$46.08_{\pm 1.37}$	$48.13_{\pm 4.93}$	$55.71_{\pm 2.92}$	$\mathbf{57.39}_{\pm \mathbf{0.62}}$
50%	$53.90_{\pm 2.46}$	$45.84_{\pm 0.96}$	$45.72_{\pm 0.76}$	$47.53_{\pm 0.21}$	$50.78_{\pm 2.53}$	$\mathbf{55.83}_{\pm \mathbf{0.62}}$
70%	$54.75_{\pm 0.55}$	$43.96_{\pm 0.50}$	$45.00_{\pm 1.37}$	$45.12_{\pm 1.65}$	$48.49_{\pm 0.55}$	$\mathbf{57.40}_{\pm \mathbf{2.35}}$
MRPC						
0%	$84.88_{\pm 0.14}$					
10%	$82.43_{\pm 0.71}$	$84.15_{\pm 1.02}$	$\mathbf{85.53}_{\pm \mathbf{0.65}}$	$84.80_{\pm 0.74}$	$83.98_{\pm 0.99}$	$84.55_{\pm 1.91}$
30%	$81.61_{\pm 1.77}$	$80.06_{\pm 4.41}$	$78.84_{\pm 4.50}$	$82.84_{\pm 2.01}$	$\mathbf{84.31}_{\pm \mathbf{1.37}}$	$83.41_{\pm 0.14}$
50%	$81.21_{\pm 1.58}$	$68.54_{\pm 0.28}$	$68.38_{\pm 0.00}$	$74.58_{\pm 3.20}$	$77.77_{\pm 3.33}$	$\mathbf{83.00}_{\pm \mathbf{0.57}}$
70%	$74.99_{\pm 4.43}$	$68.38_{\pm 0.00}$	$68.38_{\pm 0.00}$	$70.09_{\pm 0.88}$	$71.64_{\pm 1.63}$	$\mathbf{75.73}_{\pm \mathbf{3.78}}$
CoLA						
0%	$51.83_{\pm 1.73}$					
10%	$47.31_{\pm 0.51}$	$\mathbf{51.46}_{\pm \mathbf{2.38}}$	$50.50_{\pm 0.93}$	$49.72_{\pm 2.82}$	$51.43_{\pm 2.54}$	$49.43_{\pm 1.09}$
30%	$44.48_{\pm 1.08}$	$45.54_{\pm 4.07}$	$44.51_{\pm 3.34}$	$48.57_{\pm 2.02}$	$47.95_{\pm 0.89}$	$\mathbf{48.09}_{\pm \mathbf{0.98}}$
50%	$43.59_{\pm 2.14}$	$12.90_{\pm 12.94}$	$0.00_{\pm 0.00}$	$43.94_{\pm 0.43}$	$\mathbf{46.38}_{\pm \mathbf{0.43}}$	$45.30_{\pm 0.43}$
70%	$36.05_{\pm 1.33}$	$0.01_{\pm 0.01}$	$0.00_{\pm 0.00}$	$4.05_{\pm 0.43}$	$36.86_{\pm 0.43}$	$\mathbf{43.39}_{\pm \mathbf{0.43}}$
SST-2						
0%	$90.59_{\pm 0.50}$					
10%	$90.32_{\pm 0.78}$	$90.71_{\pm 0.20}$	$90.73_{\pm 0.32}$	$90.44_{\pm 0.27}$	$\mathbf{91.05}_{\pm \mathbf{0.11}}$	$90.97_{\pm 0.48}$
30%	$90.17_{\pm 0.54}$	$\mathbf{91.43}_{\pm \mathbf{0.24}}$	$91.20_{\pm 0.33}$	$90.74_{\pm 0.48}$	$90.97_{\pm 0.78}$	$91.13_{\pm 0.58}$
50%	$89.79_{\pm 0.80}$	$90.51_{\pm 0.84}$	$\mathbf{90.97}_{\pm \mathbf{0.35}}$	$90.75_{\pm 0.18}$	$90.02_{\pm 0.53}$	$90.74_{\pm 0.27}$
70%	$89.48_{\pm 0.17}$	$88.79_{\pm 1.09}$	$88.99_{\pm 0.64}$	$89.90_{\pm 0.64}$	$89.52_{\pm 0.47}$	$\mathbf{90.21}_{\pm \mathbf{0.29}}$
SWAG						
0%	$67.37_{\pm 0.23}$					
10%	$65.20_{\pm 0.16}$	$67.20_{\pm 0.18}$	$66.85_{\pm 1.06}$	$67.22_{\pm 0.32}$	$67.14_{\pm 0.09}$	$\mathbf{67.45}_{\pm \mathbf{0.32}}$
30%	$64.75_{\pm 0.22}$	$65.16_{\pm 0.09}$	$65.80_{\pm 0.70}$	$66.18_{\pm 0.10}$	$66.32_{\pm 0.51}$	$\mathbf{66.68}_{\pm \mathbf{0.41}}$
50%	$63.96_{\pm 0.16}$	$55.76_{\pm 1.68}$	$50.54_{\pm 2.58}$	$61.40_{\pm 0.41}$	$64.57_{\pm 0.16}$	$\mathbf{65.43}_{\pm \mathbf{0.39}}$
70%	$62.88_{\pm 0.35}$	$28.14_{\pm 2.20}$	$23.83_{\pm 0.80}$	$53.17_{\pm 1.13}$	$61.82_{\pm 0.21}$	$\mathbf{63.54}_{\pm \mathbf{0.38}}$
QNLI						
0%	$89.49_{\pm 0.38}$					
10%	$87.44_{\pm 0.38}$	$89.10_{\pm 0.35}$	$89.41_{\pm 0.31}$	$\mathbf{89.19}_{\pm \mathbf{0.18}}$	$88.99_{\pm 0.73}$	$89.10_{\pm 0.06}$
30%	$87.32_{\pm 0.59}$	$87.38_{\pm 0.88}$	$88.61_{\pm 0.07}$	$\mathbf{88.80}_{\pm \mathbf{0.53}}$	$87.95_{\pm 0.33}$	$88.27_{\pm 0.34}$
50%	$86.06_{\pm 0.42}$	$83.88_{\pm 0.69}$	$84.47_{\pm 0.86}$	$86.50_{\pm 1.31}$	$86.50_{\pm 0.70}$	$\mathbf{87.19}_{\pm \mathbf{0.07}}$
70%	$85.27_{\pm 0.43}$	$66.53_{\pm 0.43}$	$42.68_{\pm 0.21}$	$75.94_{\pm 1.96}$	$83.88_{\pm 0.68}$	$\mathbf{85.52}_{\pm \mathbf{1.02}}$

Table 8: Overall result with standard deviation.

ID	Example	Frequency Distance	QuRating	Perplexity
496	Euro-Scandinavian media cheer Denmark v Sweden draw. Denmark and Sweden tie.	1.009 (99.91%)	-2.65 (18.21%)	162.14 (99.39%)
904	Rumsfeld said the Pentagon's annual assessment of China's military capabilities shows China is spending more than its leaders acknowledge, expanding its missile capabilities and developing advanced military technology. China was increasing its military spending and buying large amounts of sophisticated weapons.	0.994 (77.10%)	-2.56 (20.44%)	25.21 (36.30%)
787	Guggenheim Museum, officially Solomon R. Guggenheim Museum, was founded in 1939 as the Museum of Non-Objective Art. The Solomon R. Guggenheim Museum was opened in 1939.	0.998 (88.83%)	-3.18 (6.97%)	14.76 (8.39%)

Table 9: Comparison to QuRating and perplexity of GPT-2 in RTE.

ID	Example	Frequency Distance	QuRating	Perplexity
1019	Tonight a spokesman for Russia 's foreign ministry said the ministry may issue a statement on Thursday clarifying Russia 's position on cooperation with Iran 's nuclear-energy efforts . Tonight a spokesman for the Russian Foreign Ministry said it might issue a statement on Thursday clarifying Russia 's position on aiding Iran 's nuclear-energy efforts .	0.993 (71.01%)	-2.54 (21.15%)	13.58 (9.73%)
2857	Mr. Soros branded Mr. Snow ś policy shift a " mistake . " Soros criticised Snow ś policy shift as a " mistake " .	1.006 (99.97%)	-1.10 (69.98%)	94.34 (98.11%)
2208	AAA spokesman Jerry Cheske said prices may have affected some plans , but cheap hotel deals mitigated the effect . AAA spokesman Jerry Cheske said prices might have affected some plans , but cheap hotel deals made up for it .	0.998 (92.44%)	-2.81 (13.79%)	57.03 (91.27%)

Table 10: Comparison to QuRating and perplexity of GPT-2 in MRPC.

ID	Example	Frequency Distance	QuRating	Perplexity
145	It is important for the more you to eat, the more careful to be.	0.958 (0.01%)	-0.57 (4.63%)	124.01 (41.37%)
3576	Students intended to surprise the teacher.	0.989 (36.01%)	1.06 (52.08%)	228.78 (61.19%)
2940	Donna fixed a sandwich.	1.007 (99.71%)	1.23 (60.09%)	918.57 (89.70%)

Table 11: Comparison to QuRating and perplexity of GPT-2 in CoLA.

ID	Example	Frequency Distance	QuRating	Perplexity
22585	The woman hands the girl to someone. Someone and someone	0.924 (0.01%)	0.94 (39.75%)	187.70 (83.06%)
54915	A yellow cab speeds towards him, then skids to a halt. Someone	1.008 (99.99%)	1.60 (55.71%)	89.56 (53.67%)
53012	Someone slowly gets up, locks eyes with someone. Someone looks guilty, weakly shaking his head, it	0.980 (38.96%)	2.79 (81.53%)	80.09 (47.81%)

Table 12: Comparison to QuRating and perplexity of GPT-2 in SWAG.

ID	Example	Frequency Distance	QuRating	Perplexity
32532	d)	0.096 (0.01%)	-1.75 (2.87%)	1648.69 (72.21%)
10397	in jerking off in all its byzantine incarnations to bother pleasuring its audience	0.991 (30.17%)	0.40 (28.98%)	208.79 (25.30%)
2940	Donna fixed a sandwich.	1.007 (99.71%)	1.62 (56.28%)	20360.19 (95.74%)

Table 13: Comparison to QuRating and perplexity of GPT-2 in SST-2.

Sample	FD
Furthest samples to geometric median	
What does 基督徒(pinyin: jīdū tú) mean? (pinyin: jīdū tú), literally "Christ follower.	1.009
Who wrote Carmen? Georges Bizet's Carmen premiered 3 March 1875.	1.009
Which Tigranes successor composed Greek tragedies? Tigranes' successor Artavasdes II even composed Greek tragedies himself.	1.009
What mostly affects polarization? Reflections generally affect polarization.	1.009
What does Orthodoxy strongly condemn? Similarly, Orthodoxy strongly condemns intermarriage.	1.009
Nearest samples to geometric median	
On which side of the war were the Chinese? The major Allied participants were the United States, the Republic of China, the United Kingdom (including the armed forces of British India, the Fiji Islands, Samoa, etc.), Australia, the Commonwealth of the Philippines, the Netherlands (as the possessor of the Dutch East Indies and the western part of New Guinea), New Zealand, and Canada, all of whom were members of the Pacific War Council.	0.954
What was the name of the Philippines nation? The major Allied participants were the United States, the Republic of China, the United Kingdom (including the armed forces of British India, the Fiji Islands, Samoa, etc.), Australia, the Commonwealth of the Philippines, the Netherlands (as the possessor of the Dutch East Indies and the western part of New Guinea), New Zealand, and Canada, all of whom were members of the Pacific War Council.	0.956
The Battle of the Chernaya took place in what year? The deployment of Italian troops to the Crimea, and the gallantry shown by them in the Battle of the Chernaya (16 August 1855) and in the siege of Sevastopol, allowed the Kingdom of Sardinia to be among the participants at the peace conference at the end of the war, where it could address the issue of the Risorgimento to other European powers.	0.956
When were the economic laws passed in Mexico City? The politics pursued by the administrations of heads of government in Mexico City since the second half of the 20th century have usually been more liberal than those of the rest of the country, whether with the support of the federal government—as was the case with the approval of several comprehensive environmental laws in the 1980s—or through laws recently approved by the Legislative Assembly.	0.957
What political leaning does Mexico City take? The politics pursued by the administrations of heads of government in Mexico City since the second half of the 20th century have usually been more liberal than those of the rest of the country, whether with the support of the federal government—as was the case with the approval of several comprehensive environmental laws in the 1980s—or through laws recently approved by the Legislative Assembly.	0.958

Table 14: Nearest samples to geometric median and furthest samples to geometric median on the QNLI dataset.