Diversity-based Data Subset Selection with Deep Reinforcement Learning

Jianhong (Tovi) Tu, Anxu (Ben) Wang

12.9.2024

1 Introduction

High-quality datasets are fundamental to the success of modern machine-learning approaches. Prior studies in domains such as computer vision and natural language processing have highlighted the impact of carefully curated datasets, emphasizing diversity, reduced noise, and task relevance in improving large-scale learning algorithms [1, 2, 3]. Classical active learning methods, which assign quality scores to data points and selectively include the best data, have consistently demonstrated performance gains [4].

In natural language processing (NLP), particularly in the training of Large Language Models (LLMs), quantifying dataset diversity and data quality has proven critical for developing effective AI assistants. For example, [5] and [6] show that constructing user-assistant conversation datasets spanning diverse disciplines and task categories enhances LLMs' instruction-following abilities through Instruction Tuning (IT). With IT datasets either annotated by human labor or synthesized by models, fine-tuning on a small subset of high-quality data is sufficient to achieve comparable or even superior performance to full-dataset training [7]. However, excessive fine-tuning on synthetic data risks catastrophic forgetting or model collapse [8], underscoring the need for automated methods to select moderately sized, high-quality subsets for instruction tuning.

As IT datasets grow through large-scale crowdsourcing and machine annotation, efficient subset selection becomes increasingly challenging. Existing manual or algorithmic approaches often suffer from poor scalability or high computational complexity, limiting their practicality. Concurrently, a reinforcement learning-driven approach is proposed to learn dataset selection strategy, but its success is limited to the relatively simple problem of supervised classification [9].

Motivated by the restrictions above, this research proposes a novel RL-based dataset subsection selection method by formulating the problem as a simple Markov Decision Process and leveraging well-established policy gradient methods. Our approach allows subset selection of arbitrary size by learning a diversity score for each datum while achieving similar or better performance as compared to a theory-driven greedy method. Though motivated by instruction tuning, this method is readily applicable to other domains as long as a few basic assumptions are met. We demonstrate its effectiveness and efficiency on image and text datasets and validate its applicability in fine-tuning LLM.

2 Related Works

Data Valuation with RL [9] proposes a novel subset selection framework with the reinforcement learning paradigm by estimating a quality score independently for every datum in the training split, which assesses the datum's impact on a classifier's accuracy on the test split. The method, or DVRL, utilizes a usual deep classifier and instantiates a data value estimator as the agent, which is a multilayer perception that takes in the data point and outputs a scalar score. During training, mini-batches are sampled from the training set, and the classifier weights are updated via gradient steps. The reward signal is derived from the negative difference between the current classifier's loss and the moving-average loss on a reserved validation set. The data value estimator is optimized using a modified REINFORCE algorithm [10]. During inference, the value estimator assigns a quality score for every datum, and top-k sampling selects the best subset from the sorted dataset where k is the size constraint.

Although this method is successful in robust classification, its online learning strategy is not suitable for large-scale LLM training: 1) continual training of an LLM disrupts reward assignment, as its zero-shot ability primarily depends on the transfer of knowledge from the pre-training stage, which may be lost due to subsequent training. This effect is also known as catastrophic forgetting. 2) Gradient calculation and evaluation are expansive due to the massive size of LLMs. Therefore, our approach practically applies the inspiration from DVRL to LLM fine-tuning by decoupling agent training from LLM training and using a more tractable diversity metric as a proxy for instruction tuning performance. Additionally, we simplify the learning paradigm by designing an MDP compatible with common RL algorithms, thus allowing the direct application of more advanced algorithms with no modification, albeit with a minor redundancy introduced by this strategy.

Subset Selection with Greedy Methods Another line of work in the NLP communities adopts various greedy methods for subset selection. These approaches typically define selection criteria by combining diversity and quality measures additively or multiplicatively, applying them in iterative greedy algorithms. However, there is no consensus on how to evaluate the quality of individual data points. Common approaches include ChatGPT ranking [11] and LLM loss [12]. In its simplest form, [6] selects the best data point that is maximally dissimilar to all selected data points in the remaining dataset. In particular, [12] leverages a theorem-driven approach, known as the determinantal point process, and alternates between selecting the best data point and adjusting the likelihood of being selected for the rest of the dataset. The basic intuition is that a better data point is associated with a higher inclusion probability. Although both works have demonstrated empirical success, the overarching problem is the poor computational complexity: Given a IT dataset of size N, the former and the latter method ranks all datum in $O(N^4)$ and $O(N^2)$ time.

The primary computation costs for the greedy methods come from the need for repeated computation of the pair-wise distances and iteration through the dataset to find the best candidate. Our method accelerates selection by taking advantage of the online diversity algorithm, which does not apply to the mentioned methods, and selecting data points in one iteration. The most computation-heavy part of our method is training, but it is not explicitly tied to the dataset size and well benefits from GPU acceleration. In our experimentation, we demonstrate that our method achieves comparable or even better results while increasing the speed by 10× as compared to greedy DPP used in [12].

3 Background & Preliminaries

3.1 Dataset Assumptions

Before introducing our main approach, we first introduce the prerequisites on the datasets. Though majorly intended for instruction-tuning (text) datasets, this framework has basic assumptions and is compatible with common machine learning datasets. Generally, we suppose a dataset $D = \{x_i, y_i\}_{i=1}^{N} \sim P$, where x and y are commonly defined to be the inputs and targets respectively, but we primarily focus on the inputs only. Without imposing particular restrictions, inputs x can adopt arbitrary forms, including images and texts. Instead, we assume a feature transformation function $f : \chi \to \mathbb{R}^d$ that encodes $x \in \chi$ into high-dimensional vector representation $\tilde{x} \in \mathbb{R}^d$. For example, an appropriate transformation for any common image dataset is a pre-trained ResNet model [13] with the final layer removed. Finally, we consider a diversity measure $g : \mathbb{R}^{k\times d} \to \mathbb{R}$ that quantifies the spread of a finite set of feature vectors $\tilde{X} = [\tilde{x}_1, ...\tilde{x}_k]^T$. Common choices include the determinant or the trace of a covariance matrix and the average Euclidean or Cosine distances.

3.2 Problem Definition

Formally, the problem of diversity-promoting subset selection can be formulated as a maximization problem under a cardinality constraint M:

$$\begin{aligned} &\max_{D^-} g(\tilde{X}_{f,D^-}) \\ \text{s.t. } &|D^-| = M \leq |D| \\ &D^- \subseteq D \end{aligned}$$

where $\tilde{X}_{f,D^-}$ is the feature matrix of D^- with the transformation function f. Finding the solution D^{-*} to the optimization problem is non-trivial because of its non-differentiability and the combinatorial nature.

3.3 Online Diversity Calculation

A crucial ingredient of our method in achieving high efficiency is the effective usage of online algorithms for diversity measurement. The diversity metrics of interest, including calculating the covariance matrix and all-pair distances, involve computationally expensive operators. The online estimation for both statistics shares the same idea: For an additional data point, the algorithm 1) updates the mean estimate and 2) updates the statistics according to the temporal difference. For covariance matrix, the update rule can be characterized as:

$$\begin{aligned} \mu_t &= \mu_{t-1} + \frac{1}{t}\left(x_t - \mu_{t-1}\right), \\ \Sigma_t &= \Sigma_{t-1} + \frac{1}{t}\left((x_t - \mu_{t-1})(x_t - \mu_{t-1})^\top - \Sigma_{t-1}\right), \end{aligned}$$

where $\mu_1 = x_1$ and $\Sigma_1 = 0$ for $t \in \{1, ..., M\}$. Notably, when a datum arrives, the algorithm reduces the per-iteration time complexity from $O(d^2t)$ to $O(d^2)$ for re-evaluating the subset diversity and from $O(dt^2)$ to $O(dt)$ for mean distances.

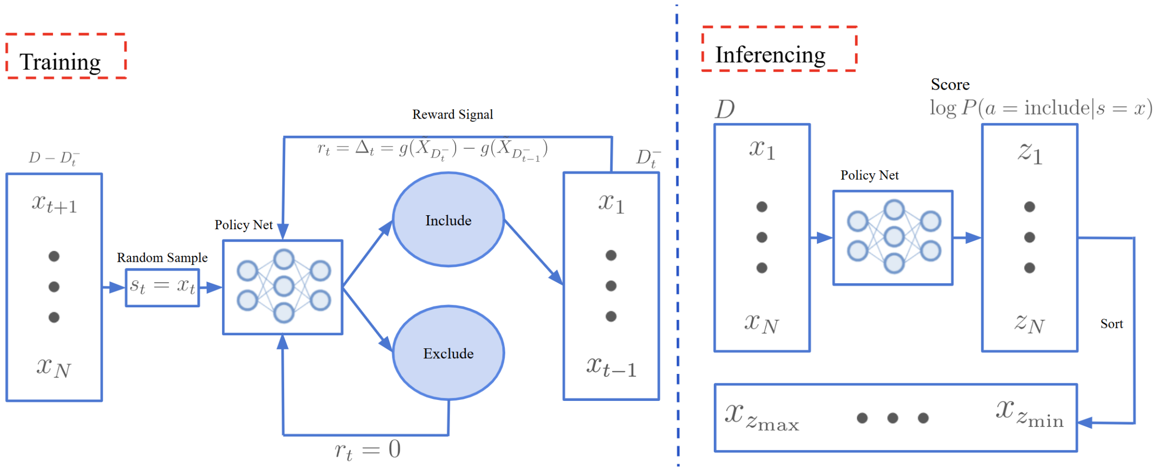

Figure 1: Overview for training and inferencing procedure. A single data point is fed into the agent, which decides whether or not to include it in the subset. However, the subset itself is not visible to the agent. The reinforcement learning signal guided by the marginal change in the diversity metric is used to update the agent. During the inference stage, the agent assigns a score for all data points. The top-k samples of the sorted data points are selected as the maximally diverse subset.

3.4 Proximal Policy Optimization

Proximal Policy Optimization (PPO) [14] is a performant policy gradient method based on the Actor-Critic framework. Both policy and value networks are updated by taking a gradient step to maximize a clipped surrogate objective:

$$L(\theta) = \hat{\mathbb{E}}_t[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]$$

where the advantage is weighted by a probability ratio $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$. In practice, the advantage $\hat{A}_t$ is obtained via Generalized Advantage Estimation:

$$\hat{A}_t = \sigma_t + (\gamma\lambda)\sigma_{t+1} + \ldots + (\gamma\lambda)^{T-t+1}\sigma_{T-1}$$
$$\sigma_t = r_t + \gamma V(s_{t+1}) - V(s_t)$$

for discount factor γ and the balance parameter λ in the range $[0, 1]$. The parameter λ balances between estimation bias and variance by weighting the history of temporal-difference error.

4 Method

Our method solves subset selection with a deep RL algorithm, focusing on PPO, by converting the optimization problem into a Markov decision process. The RL policy learns to maximize the final diversity during training and ranks all data points by assigning a diversity score. Figure 1 presents an overview of the method.

Training We define the state to be a **single** datum x_t drawn from the remaining dataset $D - D_{t-1}^-$, the action space to be either *include* or *exclude* the datum x_t, the transition function to be $P(s_{t+1}|s_t) \sim Uniform(D - D_{t-1}^-)$, and the reward to be the marginal change in the diversity measure $\Delta_t = g(\tilde{X}_{D_t^-}) - g(\tilde{X}_{D_{t-1}^-})$ for $D_t^- = D_{t-1}^- \cup x_t$ if $a =$*include*. The MDP terminates whenever there is no data point left in the remaining dataset $D = \emptyset$ or the selected subset reaches the cardinality constraint $t = M = |D^-|$. The designed MDP has a discrete and deterministic binary action space, and the state space contains finite vector representations, making it readily compatible with common RL algorithms. The reward signal is dense, as a non-zero value is returned whenever a new data point is included. Since the reward is the marginal change, the un-discounted sum rewards is exactly the diversity measure of the final subset D^-. In practice, we use the introduced online algorithm to estimate the reward value efficiently.

Inferencing For inference, inspired by DVRL [9], we adopt a distinct and more flexible strategy to assign a *diversity score* calculated as the log inclusion probability $P(a = \text{include}|s = x)$. This is possible because the inclusion probability only depends on the current sample x regardless of the selected subset D^-. The diversity score serves as a relative measure of the x's impact on the diversity metric across all potential subsets s.t. $x \notin D^-$. With predicted scores, the subset selection is possible for arbitrary size limits and is no longer restricted by M. This is achieved by first sorting the data points according to their scores in decreasing order and taking the top-k samples, where k is the desired subset size. Notably, minimizing the diversity is possible by reversing the order of the sorted data points and then taking the top-k samples.

Impact on the Algorithm Due to the design of the MDP, the state representation poorly reflects the current state of subset selection as the agent does not know D_t^- when making a decision on x_t. Therefore, we acknowledge that the usage of PPO algorithm introduces redundancy, especially concerning the state value estimation: 1) the state representation provides no information to predict the expected total reward $V(s_t)$ (discounted diversity measure) accurately. 2) the generalized advantage estimation barely benefits from the history of temporal differences σ_t. We empirically find that the learned values admit a normal distribution centered at 0, and the effectiveness of maximizing the diversity remains the same even when setting $\lambda = 0$, where $A_t = r_t - V$ for $V \sim N(0, \sigma)$.

Mechanism Analysis We argue that using $A_t = r_t$ provides sufficient signal for policy learning by considering a distribution of subsets $P_D(D^-)$ similar in the Determinantal Point Process theory [15]. Suppose $L \in \mathbb{R}^{n \times n}$ be a positive semi-definite covariance matrix defined by some kernel function, e.g. Radial Basis Function. DPP models subset selection as a Maximum Posterior inference such that $P(D^-) \propto \det(L_{D^-})$ for $D^- \in \mathcal{P}(\mathcal{D})$ and $D^- \sim P_D$, where the determinant provides a measure of the diversity. In the case of MDP, we similarly define state x_t and subset D_t^- as random variables that depend on the step variable t. Recall that the advantage is equivalent to the reward $A_t = r(x_t|D_t^-) = g(D_t^- \cup \{x_t\}) - g(D_t^-)$, if we ignore the value estimation. For simplicity, we also re-define the reward to only account for the direction of the impact:

$$r(x_t|D_t^-) = \begin{cases} +1, & \text{if } g(D_t^- \cup \{x_t\}) \geq g(D_t^-) \\ -1, & \text{otherwise} \end{cases}$$

During trajectory roll-out, the algorithm collects multiple samples of $r(x_t|D_t^-)$ for different D^-. For batch updates, the aggregate advantage of a particular x is a point estimate of the expected impact on the diversity measure $\mathbb{E}_{D_t^-}[r(x_t|D_t^-)]$. Observe that the expected impact is equivalent

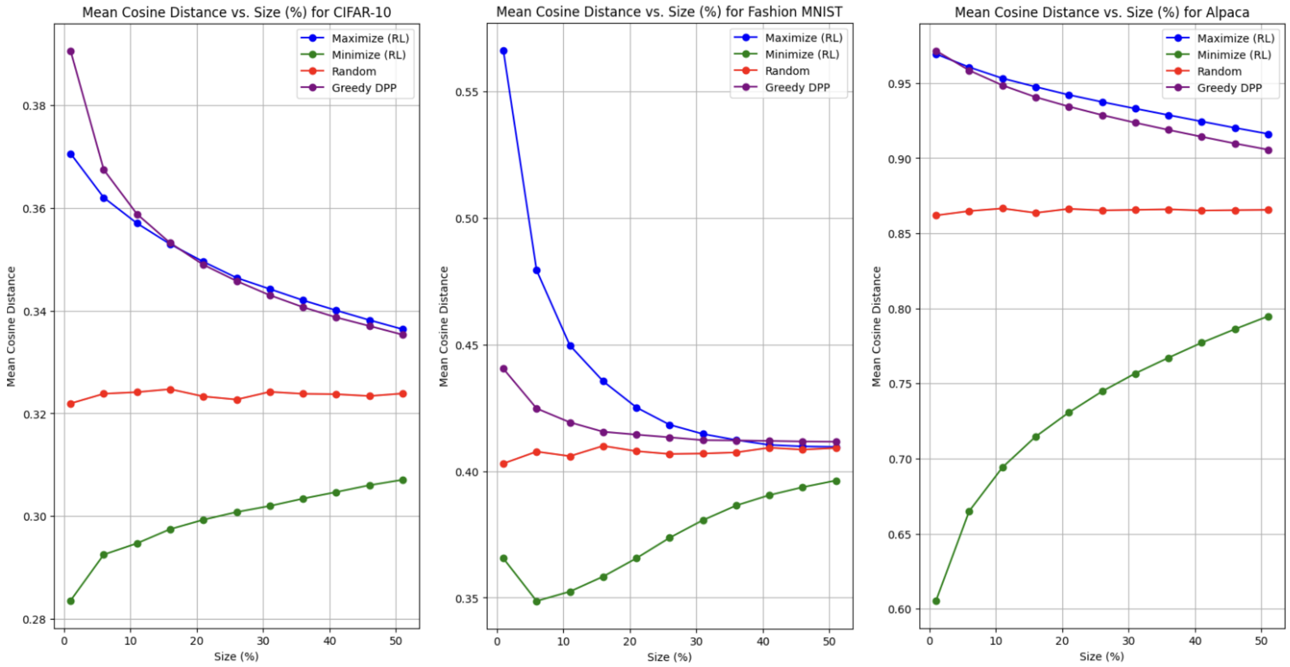

Figure 2: Diversity measure for subsets selected by four methods w.r.t. increasing sizes. Our method, denoted by RL, selects the most and the least diverse subset. Greedy DPP selects the most diverse subset. The random baseline (red) is added as a reference.

to the probability that including x will increase the diversity after linear transformation:

$$\begin{aligned}\mathbb{E}_{D^-}[r(x_t|D_t^-)] &= 2\mathbb{E}_{D^-}[\mathbb{1}[g(D_t^- \cup \{x_t\}) \geq g(D_t^-)]] - 1\\ &= 2P(g(D_t^- \cup \{x_t\}) \geq g(D_t^-)|x_t) - 1\end{aligned}$$

Therefore, optimizing the policy gradient is equivalent to encouraging the policy to include samples x_t, which will most likely promote subset diversity:

$$\arg\max_\theta \mathbb{E}_t[\pi_\theta(a_t|x_t)A_t] = \arg\max_\theta \mathbb{E}_t[\pi_\theta(a_t|x_t)P(g(D_t^- \cup \{x_t\}) \geq g(D_t^-)|x_t)]$$

5 Experiments

Though inspired by LLM fine-tuning, our method flexibly applies to general machine-learning datasets. Because of the recent development trend of multimodal LLMs, we majorly evaluate our method on two image datasets and one instruction-tuning dataset and compare its effectiveness against Greedy DPP[16, 12]. Finally, as a proof-of-concept, we validate the impact of diversity on instruction-tuned LLM's zero-shot abilities.

5.1 Diversity Optimization

Implementation We experiment our method on FashionMNIST [17], CIFAR-10 [18], and Alpaca [19]. Table 2 summarizes the definition of our MDP environments. We use the PPO [14] to optimize the RL agents with a learning rate of 0.0003 and a batch size of 64 for $1e5$ steps. The size limit M is set to 20% of the total data size. After training, we used the agent to sort the data points based on the predicted scores. We then take the top-k or bottom-k samples as

	FashionMNIST	CIFAR-10	Alpaca
Greedy DPP	113 mins	81 mins	56 mins
(Our) DSS-RL	7 mins	5.5 mins	4 mins

Table 1: Run time comparison of greedy DPP and our method on various datasets. Both algorithms are run on the same platform. Our method benefits from GPU acceleration.

Dataset	**Format**	**Feature Extractor** (f)	**Metric** (g)
FashionMNIST	60k 28^2 gray-scale images	Image Flattening	Trace(Cov($\cdot$))
CIFAR-10	60k 32^2 colored images	ResNet-18	Trace(Cov($\cdot$))
Alpaca	52k IT text paragraphs	all-mpnet-base-v2	Mean Cos Distance

Table 2: Summary of Feature Extractors and Metrics for Various Datasets

the subset that maximizes or minimizes the diversity. We denote them as "Maximize (RL)" and "Minimize (RL)."

Results For evaluation, we consider the mean cosine distance for all three datasets. Figure 2 illustrates the trend of the mean distances for subsets selected by our method, greedy DPP, and the random baseline. Compared with greedy DPP, We observe that our method has a competitive performance on CIFAR-10 and superior performance on both FashionMNIST and Alpaca, whereas the RL method finds a more diverse subset across various sizes. Meanwhile, our method also effectively finds the least diverse subsets, illustrated by the green curves. As expected, all diversity curves approach the random baseline as the size increases. However, the trend for "Minimize (RL)" on FashionMNIST is not monotonic as the diversity decreases from size 1% to 6%, while we anticipate it to increase. We list the total computation time to rank 50% of the data in Table 1. Using the same platform, our method benefits from an A6000 GPU and is around 10 times more efficient than greedy DPP. Since the major computation cost for our method occurs in training, it has an even greater advantage when ranking the full datasets.

Effect of Training Steps To better understand how training time steps affect the diversity of the selected subsets, we apply our method to the Alpaca dataset and train for 500,000 steps. We evaluate the agent for every 5,000 steps by plotting the total rewards in the training environment (Figure 3 Left) and diversity curves w.r.t. to various subset sizes (Figure 3 Right). We observe that the training reward steadily increases over all 500,000 steps. The rewards have 0 standard deviation as the agent takes an independent and deterministic action for every data point. Figure 3 right demonstrates the impact of training steps on inference, where we notice a strong diminishing return after 300,000 steps. Since our method beats greedy DPP after 100,000 training steps, this result shows that a stronger performance is possible with further training. In practice, one would choose an appropriate total training step to balance efficiency and performance, especially when a strong GPU is unavailable.

5.2 Effect on LLM Fine-tuning

To validate the dataset diversity's effect on LLM fine-tuning, we utilize the trained agent to select the 3,000 most or least diverse samples from the Alpaca dataset for LLM training. We follow the zero-shot evaluation protocol as in [6] and evaluate the fine-tuned LLM on four out-of-distribution benchmarks.

Implementation For training, we initialize a pre-trained LLM, Llama-2-7B [20] and train

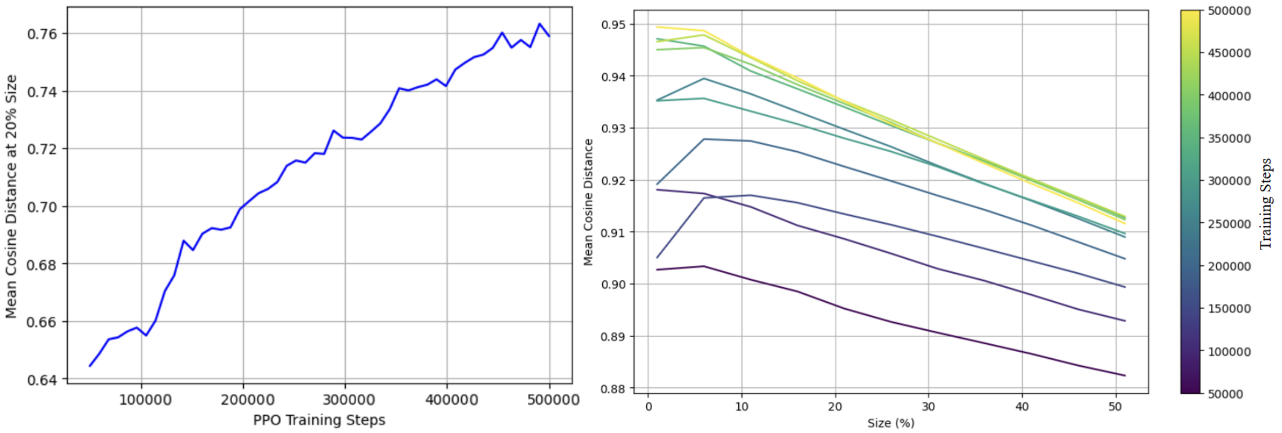

Figure 3: Impact of training steps on agent performance. Higher cosine distances are better. Left: Total training rewards vs. training steps. Right: The diversity curves vs. training steps. The diversity curves are generated by taking top-k samples of the full dataset with increasing k. The color denotes the number of training steps.

	MMLU				ARC		HellaSwag	BoolQ	AVG
	Humanity	Social Sci	STEM	Other	Easy	Challenge	-	-	-
Full	34.3	44.0	37.0	**50.2**	74.5	48.6	72.9	**82.1**	55.4
Random	34.2	44.5	37.2	49.7	72.4	46.2	71.9	79.2	54.4
Max	**34.4**	45.2	**37.2**	50.0	70.0	45.7	71.8	77.8	54.0
Min	34.2	**45.2**	36.4	50.1	**76.0**	**48.6**	**73.2**	81.9	**55.7**

Table 3: Comparison of Llama-2-7B fine-tuned on different subsets on four standard benchmarks in the zero-shot setting. The best scores are in **bold**.

it on four data splits: our Max subset, our Min subset, Random sample, and Full dataset. Specifically, we train the LLM using the AdamW optimizer at a learning rate of $2e-4$, a batch size of 128 for 3 epochs. We uses the Low Rank Adaption technique [21], with $r = 128$, to decrease the memory footprint. All training sessions are completed within 10 minutes on 4xA100 GPUs.

Results The fine-tuned LLMs are evaluated on four benchmarks: MMLU [22], ARC [23], HellaSwag [24], and BoolQ [25]. Respectively, they correspond to task types testing multidisciplinary knowledge, reasoning, commonsense, and reading comprehension. Table 3 lists the evaluation results. Surprisingly, despite previous evidence that zero-shot performance favors diverse IT datasets [6, 5, 12], we observe that using the Max subset yields the worst performance, while training on the Min subset achieves the best average score of 55.7, outperforming both random baseline and full dataset training. The LLM train on the Min subset sets the best scores on four of eight benchmark splits and performs competitively on the others except for the STEM split in MMLU. We reason that the reversed trend may be due to the noise in the Alpaca dataset. For example, the Clean-Alpaca project points out that the quality of Alpaca's training sequences varies, and some include factual errors, hallucination problems, and corrupted outputs. As we ignore the data point's quality, it is reasonable that the noisy samples are considered "novel" and thus preferred in the Max subset. Therefore, minimizing the diversity functions as anomaly detection, the Min subset contributes to better performance by including clean and consistent training data.

To qualitatively evaluate the subsets, we show examples from the Min subset in Table 4. We notice that most of the training paragraphs fall into the category of text summarization, thus verifying that our method indeed finds the most semantically homogeneous (least diverse) subset.

Instruction
Take the input and summarise it in 8 words.
Sum up the sentence using one or two words.
Select the sentence that best summarizes the following text.
Create a tagline that summarizes the idea of the product or service presented in the sentence.

Table 4: Qualitative samples of instructions from our least diverse subset.

6 Conclusions

In this project, we propose a novel method to efficiently find the maximally diverse subset under a size constraint by framing the optimization problem as a Markov Decision Process solvable with a standard reinforcement learning algorithm. Our comprehensive experimentations on three machine learning datasets across different domains demonstrate that our approach achieves comparable or superior performance while being 10x more computationally efficient (Figure 2, Table 1) compared with the greedy DPP.

Moreover, our method allows flexibility in selecting either the most or least diverse subsets during inference. Interestingly, contrary to our initial hypothesis, we found that the least diverse instruction-tuning dataset yielded the best performance for the pre-trained LLM, Llama-2-7B (Table 3). This finding highlights that diversity alone may not be a sufficient metric for developing a comprehensive data estimator.

7 Limitations & Future Works

Despite the advantageous speed, further improvement is possible by removing the redundant value estimation network in policy gradient algorithms. Both empirical and theoretical analyses show that the raw reward signal, marginal change in the diversity metric, is sufficient, and the value estimation performs no better than random noise. Furthermore, while dataset diversity has a significant impact, it may be mislead the noise and corrupted samples. This limitation underscores the need for incorporating additional metrics that better reflect data quality, such as relevance, representativeness, or alignment with task-specific objectives.

Moving forward, we aim to develop a more comprehensive framework for evaluating datasets. This framework would integrate multiple quality metrics, such as a reward model from RLHF [26], to provide a nuanced understanding of truthfulness and comprehensiveness. Additionally, we propose designing an expanded state space that accounts for the diversity and interactions of different data types, such as mathematical problems, summarization tasks, and question answering.

References

[1] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin, "Sharegpt4v: Improving large multi-modal models with better captions," in *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XVII*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, Eds. 2024, vol. 15075 of *Lecture Notes in Computer Science*, pp. 370–387, Springer.

[2] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, Eds., 2023.

[3] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev, "LAION-5B: an open large-scale dataset for training next generation image-text models," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, Eds., 2022.

[4] Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar, "Batch active learning at scale," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, Eds., 2021, pp. 11933–11944.

[5] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le, "Finetuned language models are zero-shot learners," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. 2022, OpenReview.net.

[6] Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang, "Data diversity matters for robust instruction tuning," in *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, Eds. 2024, pp. 3411–3425, Association for Computational Linguistics.

[7] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy, "LIMA: less is more for alignment," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, Eds., 2023.

[8] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross J. Anderson, and Yarin Gal, "AI models collapse when trained on recursively generated data," *Nat.*, vol. 631, no. 8022, pp. 755–759, 2024.

[9] Jinsung Yoon, Sercan Ömer Arik, and Tomas Pfister, "Data valuation using reinforcement learning," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. 2020, vol. 119 of *Proceedings of Machine Learning Research*, pp. 10842–10851, PMLR.

[10] Richard S. Sutton, David A. McAllester, Satinder Singh, and Yishay Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, Eds. 1999, pp. 1057–1063, The MIT Press.

[11] Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin, "Alpagasus: Training a better alpaca with fewer data," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. 2024, OpenReview.net.

[12] Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar Panda, "Diversity measurement and subset selection for instruction tuning datasets," *CoRR*, vol. abs/2402.02318, 2024.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2016, pp. 770–778, IEEE Computer Society.

[14] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[15] Alex Kulesza and Ben Taskar, "Determinantal point processes for machine learning," *Found. Trends Mach. Learn.*, vol. 5, no. 2-3, pp. 123–286, 2012.

[16] Laming Chen, Guoxin Zhang, and Eric Zhou, "Fast greedy MAP inference for determinantal point process to improve recommendation diversity," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, Eds., 2018, pp. 5627–5638.

[17] Han Xiao, Kashif Rasul, and Roland Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *CoRR*, vol. abs/1708.07747, 2017.

[18] Alex Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

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

[20] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom, "Llama 2: Open foundation and fine-tuned chat models," *CoRR*, vol. abs/2307.09288, 2023.

[21] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. 2022, OpenReview.net.

[22] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt, "Measuring massive multitask language understanding," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2021, OpenReview.net.

[23] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord, "Think you have solved question answering? try arc, the AI2 reasoning challenge," *CoRR*, vol. abs/1803.05457, 2018.

[24] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi, "Hellaswag: Can a machine really finish your sentence?," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez, Eds. 2019, pp. 4791–4800, Association for Computational Linguistics.

[25] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio, Eds. 2019, pp. 2924–2936, Association for Computational Linguistics.

[26] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, Eds., 2022.