DIVERSITY MEASUREMENT AND SUBSET SELECTION FOR INSTRUCTION TUNING DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

We aim to select data subsets for the fine-tuning of large language models to more effectively follow instructions. Prior work has emphasized the importance of diversity in dataset curation but relied on heuristics such as the number of tasks. In this paper, we use determinantal point processes to capture the diversity and quality of instruction tuning datasets for subset selection. We propose to measure dataset diversity with log determinant distance that is the distance between the dataset of interest and a maximally diverse reference dataset. Our experiments demonstrate that the proposed diversity measure in the normalized weight gradient space is correlated with instruction-following performance. Consequently, it can be used to inform when data selection is the most helpful and to analyze dataset curation strategies. We demonstrate the utility of our data selection approach on various instruction tuning datasets.

1 INTRODUCTION

Large language models (LLMs) are powerful but unwieldy for practical use. They often require demonstrations in context to elicit proper responses and even then may generate responses not intended by users. The base language model is typically "instruction tuned", i.e., finetuned to predict target responses given instructions. Instruction tuning enables the base language model to perform zero-shot tasks and follow users' intent more effectively, thus improving usability. Moreover, it is an indispensable step before additional preference learning to align the language model's output to human preference (Ouyang et al., 2022).

The number of instruction tuning datasets is rapidly growing, some with millions of data points (Ding et al., 2023; Zheng et al., 2024). This growth is facilitated by the ease of generating synthetic datasets by prompting LLMs (Wang et al., 2023b) and a growing effort to retain records of real-world user interactions with these models (Zhao et al., 2024; Zheng et al., 2024). Finetuning on ever-increasing data demands additional computational resources. As training on low quality data (e.g., incorrect responses) can lead to suboptimal models. Some data selection or pruning is required.

Practitioners in the field face an important challenge of selecting the optimal data subset for finetuning to maximize instruction following performance subject to a fixed computational budget. While various solutions have been proposed for finding representative subsets in active learning (Sener & Savarese, 2018), their applicability to natural language datasets remains underexplored. For instance, active learning methods that search for subsets with diverse weight gradients (Ash et al., 2019) were ineffective in our initial studies as they prioritized data points with short responses or those with large weight gradient norms. Most related methods aim to provide sufficient coverage of instruction tuning examples in the space of decoder-based language models' output token embeddings (Bukharin & Zhao, 2023; Liu et al., 2024) that lacks semantic structure (Le & Mikolov, 2014). Moreover, ensuring diversity in the embedding space of encoder-based masked language models is limited by encoders' short context length.

Practitioners also grapple with a closely related question of estimating how much data allocated for model finetuning would achieve comparable performance with that of the entire dataset. One approach involves assigning a score to each dataset that indicates the extent to which a dataset can be reduced without compromising performance after model finetuning. While various scoring methods exist, here we focus on dataset diversity. Common measures of dataset diversity often rely

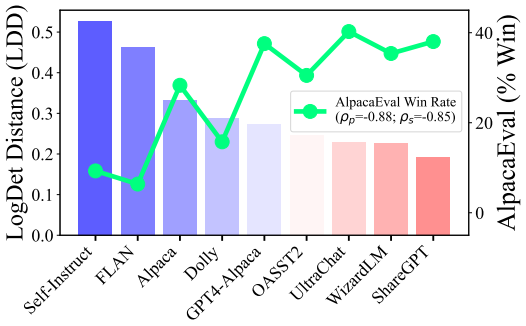

Figure 1: The log determinant distance in Equation (4) of instruction tuning datasets is correlated with instruction following performance when the model is finetuned on these datasets, with a Pearson correlation of -0.88 and a Spearman's rank correlation of -0.85.

on intuitive heuristics, e.g., the number of tasks (Wei et al., 2022; Sanh et al., 2022), topics and user intents (Lu et al., 2024), or do not scale well with the dataset size (Friedman & Dieng, 2023).

We turn to determinantal point processes (DPPs) (Kulesza & Taskar, 2012) to identify diverse subsets of high quality instruction tuning data. We investigate several choices of data representations that capture data points' similarity and find that the radial basis kernel applied to the *normalized* weight gradients of the model is particularly effective when selecting from less diverse datasets.

In addition, we measure dataset diversity with *log determinant distance* that is the difference between the log determinant of kernel matrix of a maximally diverse dataset and that of the dataset under consideration, normalized by the dataset size. Log determinant distance is readily computable from the MAP inference algorithm that identifies the optimal subset. We demonstrate that log determinant distance is correlated with instruction following performance when using weight gradients as the data representation. As a result, the diversity measure can be used to evaluate the utility of instruction tuning datasets for finetuning and to predict, before any finetuning takes place, the extent to which we can prune data without sacrificing model performance. In addition, we investigate the implications of curation strategies on diversity.

2 Related Work

2.1 Instruction Tuning Datasets

Diversity and quality are recurring themes in the curation of instruction tuning datasets. Early instruction tuning datasets, e.g., Super-NaturalInstructions (Wang et al., 2022) and FLAN (Wei et al., 2022; Chung et al., 2022), are adapted from existing natural language processing benchmarks, with a particular focus on scaling the number of tasks to encourage task generalization.

Some instruction tuning datasets are curated using Self-Instruct and its variants (Honovich et al., 2023; Wang et al., 2023b) that prompt a LLM to generate a wide array of instructions and high-quality responses. These datasets, e.g., Alpaca (Taori et al., 2023), are typically distilled from performant language models that underwent finetuning to generate user-preferred responses, e.g., variants of InstructGPT (Ouyang et al., 2022), and are well-suited for the purposes of creating a chat assistant. Moreover, they are distilled from increasingly powerful LLMs, e.g., GPT4-Alpaca (Peng et al., 2023), and contain more complex instructions, e.g., WizardLM (Xu et al., 2024), step-by-step explanations in the responses, e.g., Orca (Mukherjee et al., 2023), or multi-turn conversations, e.g., UltraChat (Ding et al., 2023).

Another family of instruction tuning datasets aims to better reflect LLMs' real-world use cases that include significant human authorship. Some are manually curated from sources with helpful responses such as Reddit, e.g., LIMA (Zhou et al., 2023), or from company employees, e.g., Dolly (Conover et al., 2023). Alternatively, real-world user interactions are curated with state-of-the-art LLMs from the internet, e.g., ShareGPT, RealChat-1M (Zheng et al., 2024), WildChat (Zhao et al., 2024). These datasets cover a wide range user intents, capturing real-world use scenarios.

We systematically study the relative diversity of aforementioned datasets and its impact on instruction following performance. Our experiments yield insights into the efficacy of the different curation approaches, e.g., distillation and human annotation.

2.2 Data Selection

Analogous to data selection for finetuning, active learning selects informative examples to label from a pool of unlabeled examples subject to a fixed labeling budget. Our approach is closely related to research that formulates active learning as core-set selection, i.e., finding the representative data subset (Tsang et al., 2005; Welling, 2009). Examples include searching for a covering of the full dataset with the smallest cover radius by solving the k-center problem (Sener & Savarese, 2018) and identifying subsets that are sufficiently spread out using `k-means++` initialization (Ash et al., 2019). Related, data pruning methods remove redundant data points that are too close to each other (Abbas et al., 2023) or to their respective cluster centroids (Sorscher et al., 2022). Similarly, our work uses DPPs to model data subsets and relies on a greedy MAP algorithm (Chen et al., 2018) to identify diverse subsets. The choice of distance metric and data representations is crucial. Prior works have employed the ℓ-2 distance between neural network activations (Sener & Savarese, 2018; Sorscher et al., 2022; Abbas et al., 2023) or between weight gradients of the log likelihood (Huang et al., 2016; Ash et al., 2019). Here we investigate several data similarity measures on instruction tuning datasets.

While choosing diverse subsets is driven by the notion that similar data points are redundant, an alternative approach is motivated by the assumption that certain data points provide more value than others. Specifically, many data selection algorithms define a quality score for each data point and select the portion of the dataset with the highest scores. Various quality scoring functions have been proposed for classification tasks, including the norm of the weight gradient (Settles, 2009; Huang et al., 2016; Paul et al., 2021), the number of times an example transitions from correctly classified to misclassified (i.e., "forgotten") during training (Toneva & Sordoni, 2019), the variability of the ground-truth label likelihood over the course of training (Swayamdipta et al., 2020), and the average ℓ-2 norm of the classification error vector (Paul et al., 2021). Our approach of modeling data subsets with DPPs accommodates arbitrary scores and aims to strike a balance between choosing data points with high quality scores and ensuring diversity within the selected subset.

Our work falls under a growing number of studies that select subsets of instruction tuning datasets to finetune LLMs. Many use quality scores to rank and select data points including simple natural language indicators like coherence (Cao et al., 2023) or perplexity (Li et al., 2023), and the LLM's rating of data points based on metrics such as helpfulness (Chen et al., 2024; Liu et al., 2024). Others select data subsets with sufficient coverage of topics and user intents (Lu et al., 2024). Our approach is closely related to methods that balance quality and diversity, e.g., by solving a variant of the facility location problem (Bukharin & Zhao, 2023) or prioritize high quality data points while avoiding duplicates (Liu et al., 2024). Different from prior approaches that solely focus on data selection, we also aim to identify ways to characterize the diversity of instruction tuning datasets, beyond the number of tasks (Wei et al., 2022), topics, and/or user intents (Lu et al., 2024). Specifically, we model data subsets with DPPs that naturally emit a diversity metric over datasets that correlates well with the downstream instruction following performance. This metric is useful for predicting improvements in the instruction following performance and for comparing the dataset diversity.

3 Method

3.1 Subset Selection with DPPs

A point process on a set of N items is a probability distribution over all subsets of $[N]$. A DPP P is a point process where the probability measure is parameterized by a positive semi-definite matrix $L \in \mathbb{R}^{N \times N}$, i.e., $P(Y) \propto \det(L_Y)$ for any subset $Y \subset [N]$. $L_Y \equiv [L_{ij}]_{i,j \in Y}$ is a sub-matrix of L indexed by Y in rows and columns. Intuitively, the diagonal elements of L are related to the marginal probability of including the particular items, i.e., $P(\{i\}) \propto L_{ii}$. The off-diagonal elements of L represents the similarity between items. Similar items are less likely to co-occur, i.e., $P(\{i,j\}) \propto L_{ii}L_{jj} - L_{ij}L_{ji}$.

Any positive semi-definite matrix L can be expressed as a Gram matrix VV^T for some matrix $V \in \mathbb{R}^{N \times D}$. Each row of V can be viewed as a feature vector for i-th item. The absolute value of the determinant of L_Y is the volume of the parallelepiped spanned by rows of V. Therefore, a high probability subset under V is a subset whose feature vectors span a large volume.

Given a dataset with N items $\{x_n\}_{n=1}^N$, we parameterize a determinantal point process P with a kernel matrix $K \in \mathbb{R}^{N \times N}$ that measures the similarity between data points and possibly a vector $q \in \mathbb{R}^N$ that indicates the quality of each data point. For instance, we can treat the cosine similarity between language models' output token embeddings as the similarity measure and the perplexity of the response conditioned on the instruction as data quality.

To select a moderately large subset of size M, the inner product kernel on features of dimension $D \ll M$ is unsuitable due to rank deficiency. Specifically, any subset Y with $|Y| > \mathrm{rank}(L)$ has zero probability mass $P(Y) \propto \det(L_Y) = 0$ and therefore the size of the most likely subset under P is upper bounded by $\mathrm{rank}(L)$. Instead, we use kernel functions that induce full rank Gram matrices, e.g., the radial basis function (RBF) kernel $K_{ij} = \exp\{-\gamma \left\|x_i - x_j\right\|^2\}$, where a larger value of γ implies that the repulsive force between data points is more local. For data representations that are normalized to unit length, the radial basis function kernel reduces to $K_{ij} = \exp\left\{2\gamma x_i^T x_j\right\}$.

Following Kulesza & Taskar (2010), we define $L_{ij} = K_{ij} q_i q_j$. This is equivalent to scaling the kernel feature map by a scalar quality score. As long as K is positive semi-definite, so is L. This structure enables us to model similarity and quality independently while considering both components during inference. Moreover, the probability of any subset $Y \subset [N]$ factors, i.e.,

$$\log P(Y) \propto \sum_{i \in Y} \log q_i^2 + \log\det(K_Y).$$

The log likelihood is maximized for subsets with high quality (1st term) and diversity (2nd term). Similar to Chen et al. (2018), we introduce a hyperparameter $\lambda \in [0, 1]$ to control the relative importance of diversity and quality:

$$\log P(Y) \propto \lambda \sum_{i \in Y} q_i + (1 - \lambda) \log\det(K_Y), \tag{1}$$

that corresponds to a DPP parameterized by the kernel matrix $L = \mathrm{diag}\left(e^{\beta q}\right) K \mathrm{diag}\left(e^{\beta q}\right)$ with $\beta = \lambda/(2(1-\lambda))$.

Given a data budget M, we pose subset selection as maximum a posteriori (MAP) inference under distribution P with a cardinality constraint:

$$Y^* = \underset{Y \subset [N]:\ |Y| = M}{\arg\max} \det(L_Y). \tag{2}$$

Although this problem is NP-hard (Ko et al., 1995), the log probability in Equation (1) is submodular (Gillenwater et al., 2012) and therefore Equation (2) can be solved efficiently with a greedy algorithm (Nemhauser et al., 1978) with at least (1-1/e)-approximation guarantee, e.g., near optimal under certain assumptions (Sharma et al., 2015).

We use the greedy MAP inference algorithm (Chen et al., 2018) that grows the set of indices $S_1 \subset \cdots, S_N \subset [N]$ by adding

$$i^*(S) = \underset{i \in [N] \setminus S}{\arg\max} \left[\log\det(L_{S \cup \{i\}}) - \log\det(L_S)\right] \tag{3}$$

to the set at each iteration. We define $L_n \triangleq L_{S_n}$ as shorthand for the kernel matrix L indexed by the greedy solution S_n at the n-th iteration ($L_N \equiv L$). The marginal gains $\Delta_1(L) = \log\det(L_1)$ and

$$\Delta_n(L) = \log\det(L_n) - \log\det(L_{n-1})$$

for $n = 2, 3, \cdots$ approximate the rate of change in diversity of selected subsets $\{S_n\}$ over the iterations. Larger marginal gains means the selected item contributes more to the diversity of the already selected subset. The unnormalized probability for the whole dataset is the sum of the marginal gains:

$$\log\det(L) = \sum_{n=1}^{N} \Delta_n(L).$$

3.2 Log Determinant Distance as a Measure of Diversity

We propose a novel way to measure dataset diversity that is a byproduct of solving the MAP inference problem in Equation (2). The measure of diversity depends entirely on the kernel that defines the DPP. For a fixed kernel function, we can compare dataset diversity quantitatively.

While $\log\det(L)$ may seem like a natural choice, it is unsuitable for measuring dataset diversity for two reasons. First, $\log\det(L)$ is not invariant to scaling of the kernel matrix, leading to widely different values that complicate the interpretation of results. For instance, if a kernel matrix is scaled by a constant $c > 0$, $\log\det(cL) = N\log(c) + \log\det(L)$ changes by $N\log(c)$ for the same dataset. Second, $\log\det(L)$ depends heavily on the dataset size, particularly when there are significant marginal gains for each item selected.

To address the aforementioned challenges, we devise a metric that measures how far $\det(L)$ is from the largest possible value than its absolute value. We operationally define a maximally diverse dataset as a set of vectors on the hypersphere in $\mathbb{R}^D$ that achieves the highest $\det(R)$. We use R to denote the reference dataset's kernel matrix computed using the same kernel function $k(\cdot,\cdot)$. To our best knowledge, there is not easy solution to obtain this set of vectors. In practice, we approximate such a set of vectors by sampling vectors randomly on the hypersphere. This yielded $\det(R)$ that is in absolute value larger than $\det(L)$ derived from every real datasets we tried, and therefore served our purpose of having a common reference to compare the diversity of different datasets to. We define *Log Determinant Distance* as

$$\text{LDD} \triangleq \frac{1}{N}\log\frac{\det(R)}{\det(L)}. \tag{4}$$

The log determinant distance measures the average deviation of the volume of the parallelepiped spanned by the rows of the Gram matrix decomposition of L, i.e., $|\det(L)|$, from the largest possible volume, e.g., $|\det(R)|$. A smaller log determinant distance implies that the dataset is closer to the maximally diverse reference dataset, and therefore is more diverse. Alternatively, we can interpret the log determinant distance as the deficit in the average contribution of a data point to dataset diversity from optimum:

$$\text{LDD} \equiv \frac{1}{N}\sum_{n=1}^{N}\left(\Delta_n(R) - \Delta_n(L)\right).$$

The log determinant distance can be readily computed from the determinants of the kernel matrices $\det(L)$ and $\det(R)$ obtained by running the greedy MAP algorithm (Chen et al., 2018) twice.

Given our assumption that the reference dataset is maximally diverse, i.e., $|\det(R)| \geq |\det(L)|$ for any kernel matrix L, the non-negativity property holds: LDD ≥ 0. Moreover, it is straightforward to show that the log determinant distance is invariant to scaling of kernels. The log determinant distance is also invariant to permutation of datasets since it is based on matrix determinants. In summary, the log determinant distance possesses favorable properties for measuring the dataset diversity.

3.3 Weight Gradient as Data Representation

We use the language model's weight gradient $\nabla_\theta \ell(x;\theta)$ of scalar-valued loss function ℓ as the data representation for data point x. As an example, ℓ can be the average log likelihood of the response conditioned on the instructions. For LLMs, the full weight gradient consists of billions of elements, rendering kernel computation infeasible. In this work, we apply Johnson-Lindenstrauss (JL) transforms (Johnson & Lindenstrauss, 1984) twice on weight gradients to reduce their dimensionality.

We first apply JL transform implicitly via Low-Rank Adaptation (LoRA) (Hu et al., 2022) to reduce memory as well as computation since most derivatives are neither stored nor computed. For weight matrix $W \in \mathbb{R}^{m\times n}$ in a fully connected layer, LoRA enforces a rank $r \ll min(m,n)$ update to the weight matrix that is a composition of two matrices: $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$. For input activation $z \in \mathbb{R}^n$, the output activation $h \in \mathbb{R}^m$ after the update is

$$h = (W + \Delta W)z = Wz + BAz.$$

We initialize A to $\mathcal{N}(0, r^{-1})$ to construct a distance preserving random projection matrix and B to zero to preserve the forward pass activations. To obtain a lower-dimensional representation of the

full weight gradient $\nabla_W \ell$, we use LoRA at initialization to compute

$$\nabla_B \ell = \nabla_h \ell \cdot z^T A^T = \nabla_W \ell \cdot A^T. \tag{5}$$

Here, we use $\ell \equiv \ell(x; \theta)$ for brevity. We also explored approaches that use LoRA to project $\nabla_W \ell$ onto a vector of size r, instead of m vectors of size r. For example, we can sum over the rows of $\nabla_B \ell$ or rows of $\nabla_B \ell$ after shifting the i-th row by i positions. We found these approaches induce a larger pairwise distance error.

Note that A is not applied to the entire weight gradient $\nabla_W \ell$. Instead, each row in $\nabla_W \ell$ of dimension n is projected to the corresponding row in $\nabla_B \ell$ of dimension r. The typical Johnson-Lindenstrauss Lemma also holds in this case.

Lemma 3.1. *Let $\epsilon, \delta > 0$. If $r = \mathcal{O}(\log(1/\delta)/\epsilon^2)$, then*

$$\left| \|\text{vec}(\nabla_B \ell)\|_2^2 - \|\text{vec}(\nabla_W \ell)\|_2^2 \right| \leq \epsilon$$

with probability at least $1 - \delta$.

The proof in Appendix A.2 involves simple application of the union bound.

We then apply the sparse JL transform to the concatenation of $\text{vec}(\nabla_B \ell)$ for every fully connected layer in the neural network to further reduce storage and compute cost. Using a sparse projection matrix is necessary since concatenated $\text{vec}(\nabla_B \ell)$ is still too costly to work with as it contains mlr entries where l is the number of fully connected layer in the network.

Theorem 3.1 can be extended trivially to include the second JL transform. It immediately follows that the two JL transforms together preserve the pairwise distance between weight gradients, in the same way that a single JL transform does on the entire weight gradient.

4 Experiments

4.1 Implementation Details

Dataset We employ a collection of instruction tuning datasets to understand the effect of data on model's intruction following performance and to evaluate their relative diversity: FLAN (Wei et al., 2022), Self-Instruct (Wang et al., 2023b), Dolly (Conover et al., 2023), Alpaca (Taori et al., 2023), GPT-4Alpaca (Peng et al., 2023), OASST2 (Köpf & Kilcher, 2023), Orca (Mukherjee et al., 2023), UltraChat (Ding et al., 2023), WizardLM (Xu et al., 2024), and ShareGPT. We also evaluate the diversity of preference datasets: OpenAI-Summarization (Stiennon et al., 2020), SHP (Ethayarajh et al., 2022), UltraFeedback (Cui et al., 2024), and HH-RLHF (Bai et al., 2022). For each dataset, we remove examples with sequence lengths greater than 2,048 to ensure the language model learns to generate the end-of-sequence token properly. Except for Dolly and OASST2 that contain fewer examples, the aforementioned datasets are subsampled to 50,000 examples to control for the effect of dataset size.

Model & Training In all experiments, we finetune Llama-7b (Touvron et al., 2023) for 3 epochs with learning rate of 2e-5 and a batch size of 128. We use AdamW optimizer with no weight decay and linearly decay the learning rate after warmup for 3% of the total number of training steps.

Evaluations We evaluate the performance of instruction-following models using a few benchmarks that measure model capabilities: factual knowledge across various subjects with Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020), reasoning on math problems using Grade School Math (GSM) (Cobbe et al., 2021) and on general reasoning problems with the Big-Bench Hard benchmark (BBH) (Suzgun et al., 2022), multilinguality with TydiQA (Clark et al., 2020), and coding skills with Codex-Eval (Chen et al., 2021). We use BENCHMARKS AVG to denote the average performance across all aforementioned benchmarks. We use Alpaca-Eval (Dubois et al., 2023) to evaluate instruction following. Specifically, We use length-controlled win rates, i.e., ALPACAEVAL LC-% WIN, to denote the proportion of times a model's generation is preferred by GPT-4 over davinci-003's response, adjusting for for bias towards longer outputs (Dubois et al., 2024). We follow the evaluation procedure in (Wang et al., 2023a) closely.

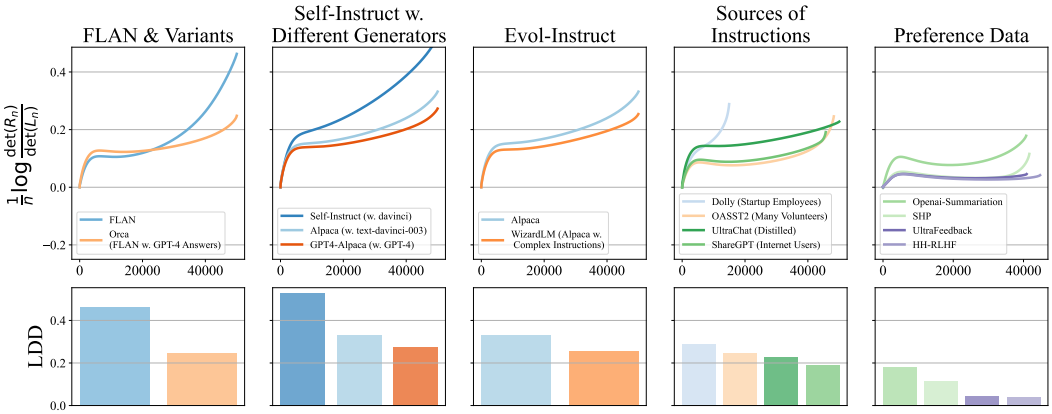

Figure 2: Studies of the log determinant distance as a measure of diversity of instruction tuning and preference learning datasets. More diverse datasets yield a log determinant distance curve that is closer the LDD $= 0$ line. Distilling responses from capable large language models improves diversity (1st panel). The diversity of synthetic datasets generated using Self-Instruct (Wang et al., 2023b) increase with better teacher model (2nd panel). Using LLMs to re-write instructions to be more complex (Xu et al., 2024) also improves diversity (3rd panel). Curating instructions from diverse sources like ShareGPT & OASST2 yields consistently higher average marginal gains compared to those curated with less human involvement (4th panel). Preference datasets overall are a lot more diverse than instruction tuning datasets, some with no apparent drop off in average marginal gains (5th panel).

Kernel Function To compute the kernel matrix L, we fix the kernel function to radial basis kernel and vary the data representations. We employ Llama-7b representing decoder-only language model to compute the average output token embeddings (LLAMA EMB) & the weight gradients vectors (LLAMA $\nabla_\theta \ell$), and MPNet (Song et al., 2020) representing encoder-only masked language model to compute the average output token embeddings of instructions (MPNET EMB). We normalize these data representations to unit length and use the abbreviation NOT NORM. to imply unnormalized vectors. We will refer readers to Appendix B.1 for details on how γ is selected.

Log Determinant Distance To compute the log determinant distance of a dataset, we generate a reference dataset by sampling vectors randomly on the surface of a D dimensional hypersphere; $D = 4096$ for LLAMA EMB and LLAMA $\nabla_\theta \ell$, and $D = 768$ for MPNET EMB. We then use the greedy MAP algorithm (Chen et al., 2018) to obtain the determinants of the kernel matrices $\det(L)$ and $\det(R)$, from which we compute the log determinant distance in Equation (4).

4.2 RESULTS: DIVERSITY ASSESSMENT

To assess the log determinant distance as a diversity measure, we compute the log determinant distance using weight gradient vectors $\nabla_\theta \ell$ on 9 instruction tuning datasets and 4 preference learning datasets detailed in Section 4.1. For instruction tuning datasets, ℓ is the average log likelihood of tokens in the response conditioned on the instruction. For preference learning datasets, ℓ is the log odds of the preferred response over an alternative worse response. Appendix C.1 demonstrates that we can compute LDD in reasonable time. We also verified that the stochasticity from sampling random vectors on the hypersphere to construct the reference dataset does not affect the estimate significantly, e.g., average LDD of the Alpaca dataset over 5 runs has a standard deviation of 6e-7.

Figure 1 demonstrate that the log determinant distance of instruction tuning datasets is correlated with instruction following performance of models finetuned on these datasets. Figure 4 compares the log determinant distance of datasets computed across different data representations: MPNET EMB, LLAMA EMB, and LLAMA $\nabla_\theta \ell$, and illustrates that LLAMA $\nabla_\theta \ell$ is the only data representation that provides a useful predictor of instruction following performance.

Figure 2 compares the log determinant distance of instruction tuning datasets and preference learning datasets. Using log determinant distance as a proxy for dataset diversity, there are a few takeaways:

Table 1: The performance of Llama-7b finetuned on 10k (20%) data subset obtained using our DPP-based and alternative baseline data selection methods. We evaluate finetuned language models' generic abilities BENCHMARK AVG and instruction following abilities ALPACAEVAL on Alpaca and UltraChat. (↑) indicates that data points with higher quality scores are selected.

DATASETS	ALPACA		ULTRACHAT	
METHODS	BENCHMARK AVG	ALPACAEVAL LC-% WIN	BENCHMARK AVG	ALPACAEVAL LC-% WIN
100% DATA	23.2	28	22.9	40
RANDOM	21.6	27	22.6	36
DPP (LLAMA $\nabla_\theta \ell$)	21.7	29	23.3	37
DPP (LLAMA $\nabla_\theta \ell$) NOT NORM.	22.7	15	22.7	36
DPP (LLAMA EMB NOT NORM.)	22.1	27	23.2	38
DPP (LLAMA EMB)	22.5	27	23.3	35
DPP (MPNET EMB)	22.9	25	23.3	35
DEDUP(MPNET EMB)	22.8	24	23.0	39
$\|\nabla_\theta \ell\|_2$ (↓)	22.7	31	23.3	38
ALPAGASUS RATING (↑)	21.6	27	-	-
EL2N (↓)	22.8	24	22.7	37
IFD (↑)	21.5	27	23.0	38
PERPLEXITY (↓)	22.7	29	22.8	37
#INPUT TOKENS (↑)	**25.1**	23	22.9	35
#OUTPUT TOKENS (↑)	23.4	31	22.7	38
#TOTAL TOKENS (↑)	20.4	30	22.7	38
DPP (LLAMA $\nabla_\theta \ell$ + #OUTPUT TOKS)	22.8	**32**	22.6	**41**

(1) the diversity of datasets improves from distilling responses or both instructions and responses from a performant LLM, (2) distilling from better teacher models improves dataset diversity even more, (3) rephrase instructions to be more complex also improves dataset diversity, (3) curating instructions from diverse sources, e.g., from real users on the internet or large-scale crowdsourcing, promotes dataset diversity, and (5) preference learning datasets are overall more diverse than instruction tuning datasets.

4.3 RESULTS: DATA SELECTION WITH DPPS

In this section, we benchmark our DPP data selection approach on two instruction tuning datasets of varying diversity: Alpaca (Taori et al., 2023) and UltraChat (Ding et al., 2023). The latter is more diverse than the former (Figure 1). The data budget is 20% (10,000) of the total dataset size.

Baselines include random selection (RANDOM), set-cover based deduplication (DEDUP) (Abbas et al., 2023). We also include several rank-and-select approaches based on the norm of the weight gradient ($\|\nabla_\theta \ell\|_2$), ChatGPT ratings of examples (ALPAGASUS RATING) (Chen et al., 2024), (EL2N) (Paul et al., 2021), instruction following difficulty (IFD) (Li et al., 2023), the perplexity of the response conditioned on the instruction (PERPLEXITY), and token counts (#INPUT TOKENS, #OUTPUT TOKENS, #TOTAL TOKENS).

Table 1 reports the performance of models finetuned on data subsets selected using our method and baselines on BENCHMARK AVG for generic abilities and ALPACAEVAL for instruction following. We provide additional results demonstrating that our DPP-based data selection method works well with different base language models (Table 5), scales to large dataset size (Table 6), handles varying data budgets (Table 7).

We investigate the effect of data representation choice for our DPP-based approach. Using LLAMA $\nabla_\theta \ell$ as the data representation yields the largest improvement in instruction following performance compared to alternative data representations on the Alpaca dataset. While simple data deduplication DEDUP(MPNET EMB) yields the largest improvement on UltraChat. Appendix C.7 provides a qualitative analysis of examples selected by the DPP-based approach across different data representations. One interesting take-away is that examples that are close in the weight gradient space tend to have similar answer structures (e.g., contain lists or long-form writing), even if they do not share the same topics or keywords.

We also assess data selection methods based on quality scores. In general, data selection with EL2N and #INPUT TOKENS results in subsets that perform worse than random subsets while using all other quality scores improve instruction following performance. Retaining examples with small $\|\nabla_\theta \ell\|_2$,

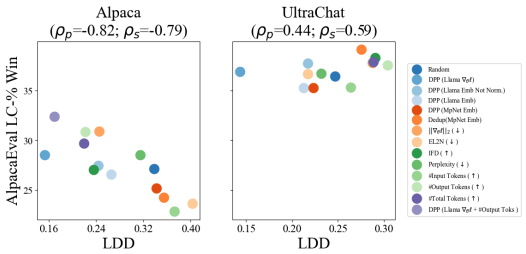

Figure 3: Performance (AlpacaEval LC-% Win) vs. diversity (LDD using $\nabla_\theta \ell$) for 20% subsets of the Alpaca dataset obtained from different data selection methods.

instead of large $\|\nabla_\theta \ell\|_2$ typically used in active learning (Park et al., 2022), leads to significant improvement. Similar gains on length-controlled win rates are observed when selecting examples with large #Output Tokens, suggesting these gains aren't simply due to longer outputs.

To investigate how DPP-based approach balances diversity and quality, we balance most effective quality score #Output Tokens with diversity in the normalized weight gradient space. Figure 5 illustrates the trade-off of our DPP-based selection approach and demonstrates that careful balancing of diversity and quality is necessary to achieve optimal performance. DPP (Llama $\nabla_\theta \ell$ + #Output Toks) leads to slightly improved instruction following performance compared to alternatives on both Alpaca and UltraChat, exceeding the performance of finetuning on the full dataset with 20% of the data.

Figure 3 shows the performance (AlpacaEval LC-%Win) vs. diversity (LDD using $\nabla_\theta \ell$) for 20% subsets of the Alpaca dataset obtained from different data selection methods. AlpacaEval length-controlled win rate and LDD is negatively correlated (e.g., $\rho_{\text{pearson}} = -0.82$) on Alpaca; While such correlation diminishes on the more diverse UltraChat dataset (e.g., $\rho_{\text{pearson}} = 0.44$). This finding has an intuitive explanation: it's worthwhile to enforce diversity only if there is room for improvement.

5 Conclusion

We introduced a DPP-based approach to select instruction tuning data subsets that provide a flexible framework to integrate different notions of data similarity and quality. We explored several choices of data representations and quality scores and determined their utility. We demonstrated that our approach outperforms baselines in various ablation studies. More importantly, we proposed log determinant distance (LDD) to quantify dataset diversity and demonstrated that LDD is correlated with instruction following performance. We can use LDD to (1) gauge how much data should be kept when selecting data subsets (2) whether we should care to implement algorithms to enforce diversity, and (3) the impact of different data curation strategy on dataset diversity.

6 Limitations

Enforcing dataset diversity using DPP-based approach proves beneficial on less diverse datasets (e.g., Alpaca). This benefit diminishes when applied to more diverse datasets (e.g., UltraChat). Instruction tuning datasets that are based on crowdsourcing user interactions with strong LLMs (e.g., WildChat) are pretty diverse to begin with. Enforcing diversity may provide limited gains and random selection after basic text deduplication may be adequate. However, the log determinant distance can still be used to understand the diversity of these datasets and to determine whether it is worthwhile to implement more sophisticated ways to encourage diversity.

Our work suggests how to improve dataset diversity. We emphasize the importance of curating datasets with realistic instructions from diverse sources, e.g., internet user interactions with LLMs. If extensive human involvement is cost-prohibitive, an alternative approach is to distill the dataset entirely or re-write partially using the most capable LLMs. Surprisingly, preference learning datasets exhibit greater diversity compared to instruction tuning datasets, even if derived from the same source (e.g., UltraFeedback is curated from FLAN, UltraChat etc.). More work is required to better understand this phenomenon and its implications.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication, March 2023.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *International Conference on Learning Representations*, September 2019.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022.

Gantavya Bhatt, Yifang Chen, Arnav M. Das, Jifan Zhang, Sang T. Truong, Stephen Mussmann, Yinglun Zhu, Jeffrey Bilmes, Simon S. Du, Kevin Jamieson, Jordan T. Ash, and Robert D. Nowak. An Experimental Design Framework for Label-Efficient Supervised Finetuning of Large Language Models, May 2024.

Alexander Bukharin and Tuo Zhao. Data Diversity Matters for Robust Instruction Tuning, November 2023.

Yihan Cao, Yanbin Kang, and Lichao Sun. Instruction Mining: High-Quality Instruction Data Selection for Large Language Models, July 2023.

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*, volume 31. Curran Associates, Inc., 2018.

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 *International Conference on Learning Representations*, May 2024.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code, July 2021.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models, December 2022.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020. doi: 10.1162/tacl_a_00317.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems, November 2021.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting Language Models with High-quality Feedback. In *International Conference on Learning Representations*, May 2024.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators, April 2024.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 5988–6008. PMLR, June 2022.

Uriel Feige, Vahab S. Mirrokni, and Jan Vondrák. Maximizing Non-monotone Submodular Functions. *SIAM Journal on Computing*, 40(4):1133–1153, January 2011. ISSN 0097-5397, 1095-7111. doi: 10.1137/090779346.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*, June 2023.

Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. Near-Optimal MAP Inference for Determinantal Point Processes. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, October 2020.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14409–14428, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.806.

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 *International Conference on Learning Representations*, January 2022.

Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. Active Learning for Speech Recognition: The Power of Gradients. *NIPS Workshop*, 2016.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2, November 2023.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In Richard Beals, Anatole Beck, Alexandra Bellow, and Arshag Hajian (eds.), *Contemporary Mathematics*, volume 26, pp. 189–206. American Mathematical Society, Providence, Rhode Island, 1984. ISBN 978-0-8218-5030-5 978-0-8218-7611-4. doi: 10.1090/conm/026/737400.

Chun-Wa Ko, Jon Lee, and Maurice Queyranne. An Exact Algorithm for Maximum Entropy Sampling. *Operations Research*, 43(4):684–691, 1995. ISSN 0030-364X.

Andreas Köpf and Yannic Kilcher. OpenAssistant Conversations - Democratizing Large Language Model Alignment. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.

Alex Kulesza and Ben Taskar. Structured Determinantal Point Processes. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2-3):123–286, 2012. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000044.

Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1188–1196. PMLR, June 2014.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. September 2023.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *International Conference on Learning Representations*, May 2024.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. #InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models. In *International Conference on Learning Representations*. arXiv, May 2024.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive Learning from Complex Explanation Traces of GPT-4, June 2023.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—I. *Mathematical Programming*, 14(1):265–294, December 1978. ISSN 0025-5610, 1436-4646. doi: 10.1007/BF01588971.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022.

Dongmin Park, Dimitris Papailiopoulos, and Kangwook Lee. Active Learning is a Strong Baseline for Data Subset Selection. *NeurIPS HITY Workshop*, 2022.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In *Advances in Neural Information Processing Systems*, November 2021.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction Tuning with GPT-4, April 2023.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen,

Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*, January 2022.

Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*, February 2018.

Burr Settles. Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Dravyansh Sharma, Ashish Kapoor, and Amit Deshpande. On Greedy Maximization of Entropy. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1330–1338. PMLR, June 2015.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867. Curran Associates, Inc., 2020.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: Beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, December 2022.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them, October 2022.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9275–9293, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.746.

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, 2023.

Mariya Toneva and Alessandro Sordoni. An Empirical Study of Example Forgetting during Deep Neural Network Learning. *International Conference on Learning Representations*, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023.

Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core Vector Machines: Fast SVM Training on Very Large Data Sets. *Journal of Machine Learning Research*, 6(13):363–392, 2005. ISSN 1533-7928.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby

Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5085–5109, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023a.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13484–13508, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754.

Jason Wei, Maarten Bosma, Vincent 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 *International Conference on Learning Representations*, January 2022.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1121–1128, New York, NY, USA, June 2009. Association for Computing Machinery. ISBN 978-1-60558-516-1. doi: 10.1145/1553374.1553517.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. WizardLM: Empowering Large Language Models to Follow Complex Instructions. In *International Conference on Learning Representations*, May 2024.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. (InThe)WildChat: 570K ChatGPT Interaction Logs In The Wild. In *International Conference on Learning Representations*, May 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph Gonzalez, Ion Stoica, and Hao Zhang. RealChat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *International Conference on Learning Representations*, May 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

A THEORY

A.1 APPROXIMATION GUARANTEES OF DPP-BASED DATA SELECTION THAT INCLUDES DIVERSITY AND QUALITY

For completeness sake, we will provide a (1-1/e)-approximation guarantee proof for the greedy maximization of the log determinant function. Specifically, we can make any log determinant function $f(X) = \log\det(K_X)$ monotone by scaling the matrix up such that the minimum eigenvalue is at least 1, as proved & discussed in Proposition 2 in (Sharma et al., 2015), thereby obtaining the (1-1/e) approximation guarantee for the greedy maximization of non-negative, monotone, and submodular functions. In practice, the greedy objective in Equation (3) is invariant to the scaling of the kernel matrix. Therefore, we would have obtained a solution with (1-1/e)-approximation guarantee without actually scale K.

Now, we argue that this approximation guarantee holds when we takes into account of both diversity and quality of data points. Specifically, as long as L is a valid kernel matrix, then the LogDet function $f(X) = \log\det(L_X)$ can be made monotone shown previously. We assumed that k is a valid kernel function, therefore $L_{ij} = K_{ij}e^{\beta q_i}e^{\beta q_j}$ is also a valid kernel function since (1) product of a real-valued function $g(x_i) = e^{\beta q_i}$ is a valid kernel and (2) kernels are closed under product (Refer to Proposition 3.22 in (Shawe-Taylor & Cristianini, 2004)).

A.2 PROOF OF THEOREM 3.1

Let ϵ, δ be given. For notation convenience, let $p \equiv \text{vec}(\nabla_W \ell)$ and $q \equiv \text{vec}(\nabla_B \ell)$. Let $p_1, \cdots, p_m \in \mathbb{R}^n$ be rows of $\text{vec}(\nabla_W \ell)$ and $q_1, \cdots, q_m \in \mathbb{R}^r$ be rows of $\text{vec}(\nabla_B \ell)$. Due to Equation (5), we have $q_k = Ap_k$ for $k = 1, \cdots, m$. Provided $A \sim \mathcal{N}(0, \frac{1}{r})$ and $r = \mathcal{O}(\log(1/\delta)/\epsilon^2)$, the following holds due to Johnson-Lindenstrauss Lemma (Johnson & Lindenstrauss, 1984):

$$\mathbb{P}\left[\left|\|q_k\|_2^2 - \|p_k\|_2^2\right| < \frac{\epsilon}{m}\right] \geq 1 - \frac{\delta}{m}. \tag{6}$$

By union bound,

$$\mathbb{P}\left[\bigcup_{k=1}^{m}\left\{\left|\|q_k\|_2^2 - \|p_k\|_2^2\right| > \frac{\epsilon}{m}\right\}\right] \leq \sum_{k=1}^{m}\mathbb{P}\left[\left|\|q_k\|_2^2 - \|p_k\|_2^2\right| > \frac{\epsilon}{m}\right] \leq \sum_{k=1}^{m}\frac{\delta}{m} \leq \delta. \tag{7}$$

If $\left|\|q_k\|_2^2 - \|p_k\|_2^2\right| < \frac{\epsilon}{m}$ for all $k = 1, \cdots, m$, then

$$\left|\|q\|_2^2 - \|p\|_2^2\right| = \left|\sum_{k=1}^{m}\|q_k\|_2^2 - \sum_{k=1}^{m}\|p_k\|_2^2\right| \leq \sum_{k=1}^{m}\left|\|q_k\|_2^2 - \|p_k\|_2^2\right| \leq \sum_{k=1}^{m}\frac{\epsilon}{m} = \epsilon. \tag{8}$$

Therefore,

$$\mathbb{P}\left[\left|\|q\|_2^2 - \|p\|_2^2\right| \leq \epsilon\right] \geq \mathbb{P}\left[\bigcap_{k=1}^{m}\left\{\left|\|p_k\|_2^2 - \|q_k\|_2^2\right| < \frac{\epsilon}{m}\right\}\right] \geq 1 - \delta \tag{9}$$

where the last inequality is by Equation (7).

B IMPLEMENTATION DETAILS

B.1 DETAILS ON HOW THE KERNEL HYERPARAMETER γ IS SELECTED

The goal is to find γ large enough such that the returned subset by f is close to full rank but not too large such that the kernel matrix becomes the identity matrix.

- If γ is very small, the gain is also very small negative number. Depending on how the stopping criterionr's tolerance parameter ϵ is set, the algorithm terminates with a subset S of size $|S| \ll N$. This would be problematic if we want to select a larger subset.

Table 2: The instruction following performance of Llama-7b finetuned on 20% data subset of Alpaca with different values for the kernel hyperparamter γ. The subset selection algorithm pretty is robust to the choice of γ.

γ	AlpacaEval LC-%Win
0.1	29.4
1.0	28.5
10.0	28.7

- If γ is very large, the gain is close to 0, and therefore there is a risk of choosing examples close to arbitrary order.

We use Newton's method to get a rough idea on a γ large enough to cover a significant portion of the dataset. Specifically, we treat the greedy algorithm as a black box function $f : \gamma \mapsto S$ that takes in the kernel hyperparameter γ and returns a subset S. The greedy algorithm will return a subset with size at most equal to the numerical rank of the kernel matrix, and therefore the size of the subset obtained depends on γ. We use Newton's method to find the root of $g : \gamma \mapsto |f(\gamma)| - N$ where N is the dataset size. We pick the final γ to be larger than the γ obtained from the Newton's method, usually a magnitude larger to ensure we can use the same γ for different datasets/data budget, and is some exponent of 10, e.g., 1, 10, 1e-3. For example, when using normalized weight gradient as the data representation, Newton's method returns $\gamma \approx 0.05$ and in the paper we just picked $\gamma = 1$.

We emphasize that the range of possible γ is pretty wide and as long as γ does not lie in the two extreme cases, the algorithm is pretty robust to the choice of γ, as illustrated in Table 2.

C Additional Experiments and Results

Table 3: Time for a single run of greedy DPP MAP algorithm for various N and D on a Nvidia V100 40GB GPU.

N	**D**		
	256	1024	4096
10k	11 sec	10 sec	11 sec
50k	5.7 min	5.8 min	6.4 min
100k	17.1 min	16.3 min	17.7 min

C.1 Runtime cost of computing the log determinant distance

We use Chen et al. (2018)'s implementation with $\mathcal{O}(\text{NMD})$ time and $\mathcal{O}(\text{N(M+D)})$ memory complexity, respectively, where N is size of the original dataset, M is the subset size desired, and D is dimension of data representation. The costly evaluation of kernel matrix entries at each iteration can be parallelized on the GPU at the memory cost of $\mathcal{O}(\text{ND})$. The algorithm is a feasible solution for selecting datasets at the scale of several hundred thousand examples.

To compute LDD, the time complexity is $\mathcal{O}(\text{N}^2\text{D})$. Table 3 reports the time of a single run of the greedy DPP MAP algorithm for a few choices of N and D on a Nvidia V100 40GB GPU. Note $\det(R)$ is computed only once and acts as a drop-in placement when computing LDD of each dataset. Therefore, the compute cost of obtaining $\det(R)$ can be amortized.

C.2 Comparisons of data representations used to compute the log determinant distance

Figure 4 compares the log determinant distance (LDD) of datasets computed across different data representations: MPNET EMB, LLAMA EMB, and LLAMA $\nabla_\theta \ell$. The log determinant distance based on weight gradient provides the strongest correlation with instruction following performance. This

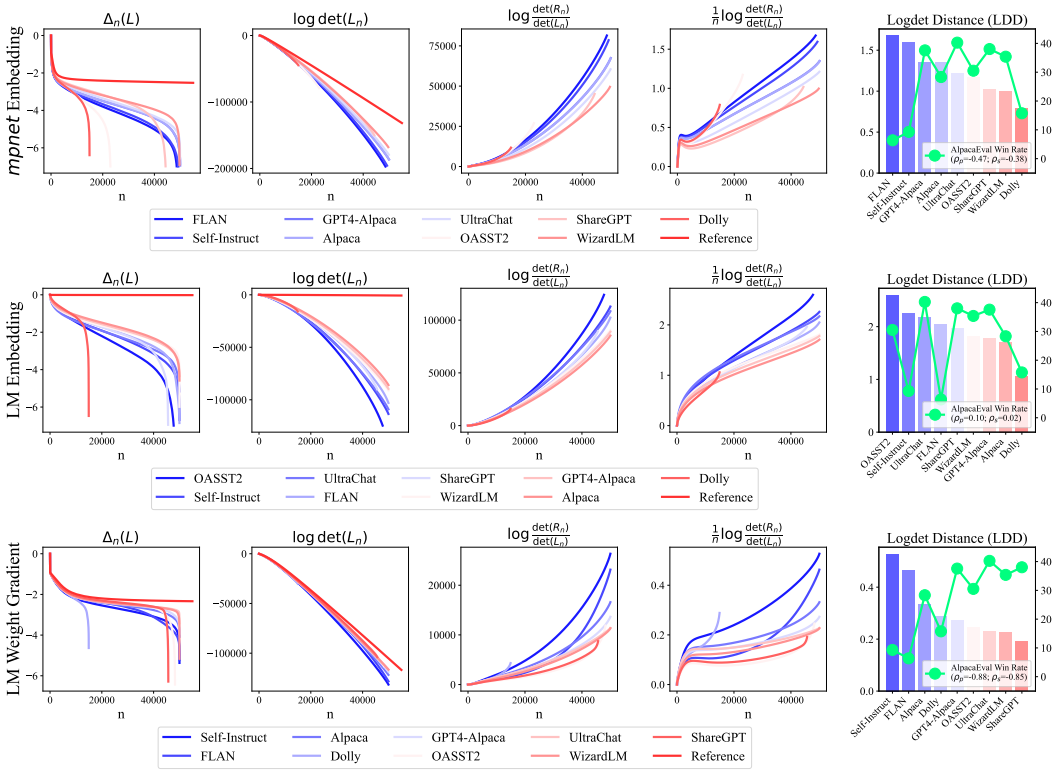

Figure 4: Comparison of the log determinant distance computed using different data representations: MPNET EMB (top row), LLAMA EMB (middle row), and LLAMA $\nabla_\theta \ell$ with respect to instruction tuning loss (bottom row). The log determinant distance based on weight gradient provides the strongest correlation with instruction following performance.

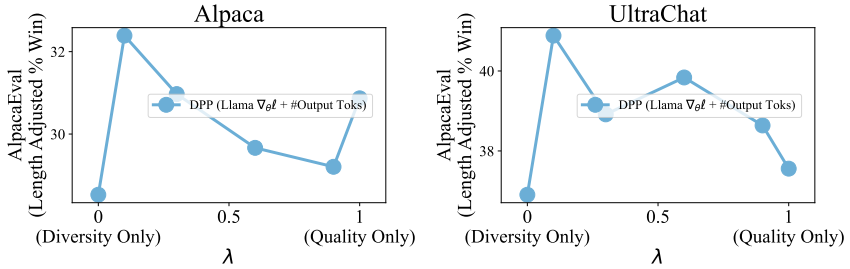

Figure 5: Vary λ interpolates between enforcing diversity and selecting for quality. The results indicate that the optimal performance for both datasets is achieved when $\lambda = 0.1$, highlighting the importance of accounting for both diversity and quality.

suggests that the weight gradient representation is the most informative data representation for measuring the diversity of instruction tuning datasets.

C.3 PERFORMANCE TRADE-OFF BETWEEN DIVERSITY AND QUALITY

Figure 5 shows the performance of the proposed selection approach DPP(LLAMA$\nabla_\theta \ell$ + #OUTPUT TOKENS) with varying λ. The λ parameter balances the contribution of examples to diversity in the normalized weight gradient space with the selection of examples with longer responses. The results indicate that the optimal performance for both datasets is achieved when $\lambda = 0.1$, highlighting the importance of accounting for both diversity and quality.

Table 4: Academic benchmark and instruction following evaluation of Llama-7b finetuned on 10k (20%) data subset of Alpaca (top block) and UltraChat (bottom block). This table compares random selection and full finetuning baseline as well as data selection methods that ensure diversity, quality, or both. (↑) indicates that data points with higher quality score are selected.

METHODS	ACADEMIC BENCHMARKS						ALPACAEVAL
	MMLU	GSM	BBH	TYDIQA	CODEXEVAL	AVG	LC-% WIN
ALPACA							
100% DATA	41.3	5.0	32.5	19.9	11.0	23.2	28
RANDOM	32.3	4.8	33.4	22.4	8.5	21.6	27
DPP (LLAMA $\nabla_\theta \ell$)	28.8	**7.4**	34.1	22.6	9.6	21.7	29
DPP (LLAMA $\nabla_\theta \ell$) NOT NORM.	41.3	5.3	26.3	**25.0**	8.5	22.7	15
DPP (LLAMA EMB NOT NORM.)	36.9	5.1	31.6	21.5	8.7	22.1	27
DPP (LLAMA EMB)	37.9	6.9	29.9	21.0	11.2	22.5	27
DPP (MPNET EMB)	39.1	5.8	33.7	20.0	8.5	22.9	25
DEDUP(MPNET EMB)	38.0	5.8	32.1	21.4	11.0	22.8	24
$\|\nabla_\theta \ell\|_2$ (↓)	36.9	6.9	33.1	20.9	8.5	22.7	31
ALPAGASUS RATING (↑)	36.2	5.2	31.3	19.8	9.1	21.6	27
EL2N (↓)	38.3	6.4	32.0	19.9	**12.2**	22.8	24
IFD (↑)	34.6	4.4	30.9	23.5	6.7	21.5	27
PERPLEXITY (↓)	37.7	4.7	32.9	21.0	11.6	22.7	29
#INPUT TOKENS (↑)	**43.8**	5.3	34.0	24.8	10.4	**25.1**	23
#OUTPUT TOKENS (↑)	36.3	7.1	35.4	21.8	9.1	23.4	31
#TOTAL TOKENS (↑)	25.1	7.2	33.5	21.7	8.5	20.4	30
DPP (LLAMA $\nabla_\theta \ell$ + #OUTPUT TOKS)	34.1	6.5	34.2	22.2	11.6	22.8	**32**
ULTRACHAT							
100% DATA	37.8	7.6	31.9	20.4	10.4	22.9	40
RANDOM	36.2	7.5	32.8	19.8	11.0	22.6	36
DPP (LLAMA $\nabla_\theta \ell$)	36.2	7.6	**34.5**	20.1	**12.8**	23.3	37
DPP (LLAMA $\nabla_\theta \ell$) NOT NORM.	37.7	8.4	32.6	17.8	11.0	22.7	36
DPP (LLAMA EMB NOT NORM.)	38.0	8.6	34.2	19.6	8.5	23.2	38
DPP (LLAMA EMB)	38.1	8.9	32.8	18.9	12.2	23.3	35
DPP (MPNET EMB)	37.3	8.6	33.9	19.1	11.6	23.3	35
DEDUP(MPNET EMB)	37.4	8.7	31.5	21.2	9.1	23.0	39
$\|\nabla_\theta \ell\|_2$ (↓)	34.7	**10.3**	32.2	21.4	12.2	23.3	38
EL2N (↓)	**38.6**	7.5	32.4	18.2	11.0	22.7	37
IFD (↑)	34.5	10.2	33.4	21.4	7.9	23.0	38
PERPLEXITY (↓)	36.9	7.6	33.0	19.4	11.6	22.8	37
#INPUT TOKENS (↑)	37.7	8.5	33.6	18.5	9.1	22.9	35
#OUTPUT TOKENS (↑)	34.5	**10.3**	31.9	20.8	9.1	22.7	38
#TOTAL TOKENS (↑)	34.6	8.3	33.3	20.9	9.8	22.7	38
DPP (LLAMA $\nabla_\theta \ell$ + #OUTPUT TOKS)	34.8	7.7	32.9	21.5	9.8	22.6	**41**

C.4 THE PERFORMANCE OF DATA SELECTION APPROACHES ON DIFFERENT DATASETS AND BASE LANGUAGE MODELS

We provide additional experimental details on the performance of our proposed data selection method and baselines. Table 4 reports the performance of Llama-7b fintuned on 20% data subset of the Alpaca and UltraChat datasets. Table 5 reports the performance of Mistral-7b finetuned on 20% data subset of the Alpaca dataset. Together, these tables provide a comprehensive view of the performance characteristics of our proposed data selection method and baselines, and demonstrate that our claims hold across different datasets and base language models used.

C.5 THE PERFORMANCE OF DATA SELECTION APPROACHES ON LARGE-SCALE DATASET

We provide additional experiments on a data mix that consists of a balanced mixture of a subset of the datasets listed in Figure 1 of the paper, i.e., all but Self-Instruct and GPT4-Alpaca. The resulting mix has approximately 313k data points, with at most 50k data points from each sub-dataset. This is similar to the scale of the Tulu-v2 mixture (with 326k data points) that is a well-known data mixture for finetuning LLMs (Ivison et al., 2023). Table 6 reports the performance of Llama-7b fintuned on different 10k subsets (3% data budget) of the full dataset, demonstrating that our approach scales to a relatively large dataset.

Table 5: Academic benchmark and instruction following evaluation of Mistral-7b finetuned on 10k (20%) data subset of Alpaca. This table compares random selection and full finetuning baseline as well as data selection methods that ensure diversity, quality, or both. (↑) indicates that data points with higher quality score are selected.

METHODS	ACADEMIC BENCHMARKS						ALPACAEVAL
	MMLU	GSM	BBH	TYDIQA	CODEXEVAL	AVG	LC-% WIN
100% DATA	46.8	18.1	**50.1**	31.5	29.3	35.8	39
RANDOM	**61.1**	21.4	49.4	31.2	30.5	39.6	39
DPP (LLAMA $\nabla_\theta \ell$)	60.3	21.2	49.1	33.5	**34.8**	40.3	44
DPP (LLAMA $\nabla_\theta \ell$) NOT NORM.	60.1	19.6	49.9	30.0	31.7	39.0	40
DPP (LLAMA EMB NOT NORM.)	58.8	21.1	49.5	29.6	31.1	38.8	39
DPP (LLAMA EMB)	57.6	19.8	49.2	28.3	30.5	37.8	45
DPP (MPNET EMB)	58.3	20.8	48.1	29.1	29.3	38.0	41
DEDUP(MPNET EMB)	60.3	20.4	49.2	31.8	27.4	39.0	40
$\|\nabla_\theta \ell\|_2$ (↓)	59.8	26.3	49.4	21.9	30.5	38.4	45
ALPAGASUS RATING (↑)	59.8	21.7	49.3	**35.9**	29.3	40.3	37
EL2N (↓)	58.6	21.6	48.0	29.4	31.7	38.5	42
IFD (↑)	56.3	20.0	48.5	33.8	29.3	38.5	41
PERPLEXITY (↓)	59.3	21.4	48.9	27.5	**34.8**	38.8	41
#INPUT TOKENS (↑)	60.5	21.1	48.7	34.2	31.1	40.0	41
#OUTPUT TOKENS (↑)	60.7	**26.7**	**50.1**	22.4	32.3	39.1	47
DPP (LLAMA $\nabla_\theta \ell$ + #OUTPUT TOKS)	60.6	25.0	49.5	30.4	**34.8**	**40.6**	**48**

Table 6: Comparison of different data selection methods on a relatively large dataset (313k data points). Our proposed approach scales with the size of the dataset.

Method	BENCHMARK AVG	ALPACAEVAL LC-%WIN
Random	23.7	29.7
DPP(LLAMA $\nabla_\theta \ell$)	**24.6**	28.4
DPP (LLAMA EMB)	24.1	27.2
Perplexity (↓)	24.7	35.8
$\|\nabla_\theta \ell\|_2$ (↓)	23.4	38.2
#OUTPUT TOKENS (↑)	22.6	39.7
DPP(LLAMA $\nabla_\theta \ell$ + #OUTPUT TOKS)	24.0	**42.5**

C.6 THE PERFORMANCE OF DATA SELECTION APPROACHES WITH VARYING DATA BUDGETS

Table 7: Performance of Llama-7b finetuned on subsets of Alpaca with varying data budgets. Our proposed approach outperforms the baseline at low data budgets.

Method	**Data Budget**				
	10%	20%	40%	60%	100%
Random	27.6	25.0	25.9	27.1	28.3
#OUTPUT TOKENS (↑)	29.3	30.9	30.4	**30.4**	28.3
DPP(LLAMA $\nabla_\theta \ell$ + #OUTPUT TOKS)	**31.9**	**32.4**	**32.6**	29.6	28.3

Table 7 reports the performance of Llama-7b finetuned on subsets of Alpaca with varying data budgets. We select a strong quality-based baseline #OUTPUT TOKENS for comparison. We demonstrate that our proposed approach provides strong performance at different data budgets and outperforms the baselines at low data budgets.

C.7 QUALITATIVE ANALYSIS OF SELECTED EXAMPLES

We visualize nearest neighbors of examples using different data representations to understand what data points would be removed (e.g., those that are similar) if we enforce diversity during selection.

The following data pair would be considered close by all data representations we investigated (MPNET EMB, LLAMA EMB, and LLAMA $\nabla_\theta \ell$) because they share keywords and belong to similar tasks:

> **Instruction**: As a personal trainer, create a nutrition plan for an athlete ...
>
> **Instruction**: As a nutritionist, design a meal plan for a client with a rare genetic condition ...

Unlike other data representations, LLAMA $\nabla_\theta \ell$ tends to group examples with similar answer structures (e.g., lists or long-form writing) together, even if they do not share the same topics or keywords. The following pair is an example:

Instruction: Rename the following folder: Documents_1940. **Response**: 1940_Documents

Instruction: Generate a username for Jamie that must contain the letter J and between 8-12 characters. **Response**: JaimeJr89

D ANSWERS TO POTENTIAL QUESTIONS

D.1 WHY NOT USE DIVERSITY METRICS SUCH AS VENDI SCORE (FRIEDMAN & DIENG, 2023) ?

Vendi Score $\exp(-\sum_{i=1}^{N} \lambda_i \log \lambda_i)$ depends on eigenvalues $\lambda_1, \cdots, \lambda_N$ of the kernel matrix $\frac{1}{N}K$, where N is the number of examples. Computing the Vendi Score for arbitrary kernel function (e.g., radial basis kernel that we used in our work) takes $\mathcal{O}(N^3)$ time for eigendecomposition whereas computing LDD requires $\mathcal{O}(N^2)$ time. This implies that computing Vendi Score on moderately large datasets would be more prohibitive than LDD. Additionally, Vendi Score introduces additional constraints on the kernel, i.e., $k(x, x) = 1$, to ensure invariance to the scaling of the kernel matrix. In contrast, LDD maintains scale invariance without imposing this constraint on the kernel function.

In addition to proposing a diversity measure, we investigate different data representations (or the kernel functions) to determine their utility in the context of instruction tuning. Specifically, Figure 4 demonstrated that LDD based on normalized weight gradient provides the best predictor of instruction following performance. This is an orthogonal contribution that previous works have not tried to tackle. Therefore, even if we could feasibly compute the Vendi Score, we would still rely on the findings in our paper to pick the kernel function.

D.2 WHY NOT USE FACILITY LOCATION FUNCTION FOR SUBSET SELECTION ?

Table 8: Comparison of DPP and Facility Location based data selection approaches. We report ALPACAEVAL LC-%WIN of Llama-7b finetuned on different 20% data budget subset of the Alpaca dataset. The two algorithms are pretty comparable.

Data Representation	**Algorithm**	
	DPP	Facility Location
LLAMA $\nabla_\theta \ell$	28.5	28.2
LLAMA EMB	26.6	26.3
MPNET EMB	25.2	24.5

Facility location and DPPs share many commonalities for subset selection. For example, they both (1) can be solved with a (1-1/e)-greedy algorithm, (2) can incorporate quality scores, and (3) have the same time complexity $\mathcal{O}(N^2)$ if using the greedy algorithm on the entire datasets. We consider our work concurrent to (Bukharin & Zhao, 2023; Bhatt et al., 2024) that uses facility location to select instruction tuning datasets.

We emphasize that we are not attached to the exact data selection framework, instead we care more about using some framework, be it DPP or facility location, to better understand how some properties of the dataset, e.g., diversity, quality or both, influence model performance. In addition to improving upon baselines, our work aims to answer questions that have practical implications, informing practitioners: (1) what diversity metric and quality scores are worth trying for selecting instruction tuning datasets, (2) alternative data representations (e.g., normalized weight gradient)

that can be used to better quantify similarity between long-form text, (3) how diverse a new dataset is compared to existing ones, and if we should care to implement algorithms to enforce diversity, (4) the effect of different data curation strategies on dataset diversity.

For completeness sake, we will provide a comparison of DPP and facility location function based subset selection approaches. We use `apricot`'s implementation of facility location with exact lazy greedy algorithm (Feige et al., 2011). We use the same kernel function for facility location function as used in DPP. Table 8 reports ALPACAEVAL LC-%WIN of Llama-7b finetuned on different 20% data budget subset of the Alpaca dataset. The two algorithms are pretty comparable.

D.3 IS ENFORCING DIVERSITY ALL WE NEED ?

While Figure 1 shows there exists a strong correlation between dataset diversity (LDD) and performance (ALPACAEVAL LC-% WIN) across datasets. We do not claim that diversity is correlated with performance all the time. For instance, diversity is strongly correlated with performance on 20% subsets of a less diverse dataset (e.g., Alpaca) while such correlation diminishes on that of a more diverse dataset (e.g., UltraChat), as shown in Figure 3.

Diversity represents one axis of variation of a dataset, alongside with other factors such as quality. Knowing when such correlation exists for a specific dataset can help us determine if diversity matters and if we need to spend effort to enforce diversity. We can get an idea of where a dataset lie in this diversity spectrum with our proposed diversity measure: LDD.