

Diversity Measurement and Subset Selection for Instruction Tuning Datasets

Peiqi Wang¹ Yikang Shen² Zhen Guo¹ Matthew Stallone² Yoon Kim¹ Polina Golland¹ Rameswar Panda²

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 downstream 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 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-

¹MIT, Cambridge, MA, USA ²MIT-IBM Watson AI Lab, Cambridge, MA, USA. Correspondence to: Peiqi Wang <wpq@mit.edu>.

Preliminary work. Do not distribute.

world user interactions with these models (InT, 2023; 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 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 datasets that are less diverse.

In addition, we measure dataset diversity with *log determi-*

nant 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 dataset 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 and incorporating a variety of prompt templates to encourage task generalization and robustness to prompt wordings.

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 (InT, 2023). These datasets cover a wide range of topics and user intents, capturing real-world use scenarios.

Here 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 manual 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). Rather than propose new scoring functions, we focus on evaluating the efficacy of existing quality scores for instruction tuning subset selection. 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.

2.3. Selecting Natural Language Data

Our work is adjacent to research that selects datasets for pretraining LLMs. Unfiltered pretraining text corpora like

Common Crawl (Rana, 2010) are not ideal because they contain a large number of unintelligible documents and some documents are repeated many times. To discard low-quality documents, past work has relied on quality scores that scale well with the size of the dataset, e.g., ratings from Reddit users (Brown et al., 2020) or the perplexity of the document computed by a pretrained language model (Wenzek et al., 2020; Marion et al., 2023). Approximate string matching algorithms, e.g., MinHash (Broder, 1998), are employed to detect pairs of documents with high n-gram overlap (Lee et al., 2022). In contrast, properly curated instruction tuning datasets usually contain well-written texts with few repeated examples. Therefore, basic quality filters are less important. The challenge is to find informative axes of variations important for instruction following performance, e.g., topics, tasks, and user intents, which is the focus of our work.

One might want to select instruction tuning datasets for computing efficiency. 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). In contrast, 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 diversity of instruction tuning datasets.

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 (Kulesza & Taskar, 2012).

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 $|Y| > \text{rank}(L)$ 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 $\text{rank}(L)$. Instead, we use kernel functions that induce full rank Gram matrices, e.g., the radial basis function (RBF) kernel $K_{ij} = \exp\{-\gamma \|x_i - x_j\|^2\}$, 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 $K_{ij} = \exp\{2\gamma x_i^T x_j\}$.

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:

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

that corresponds to a DPP parameterized by the kernel matrix $L = \text{diag}(e^{\beta q}) K \text{diag}(e^{\beta q})$ 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:

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

Although this problem is NP-hard (Ko et al., 1995), the log probability in Equation (1) is submodular (Gillenwater et al., 2012) and therefore can be solved efficiently

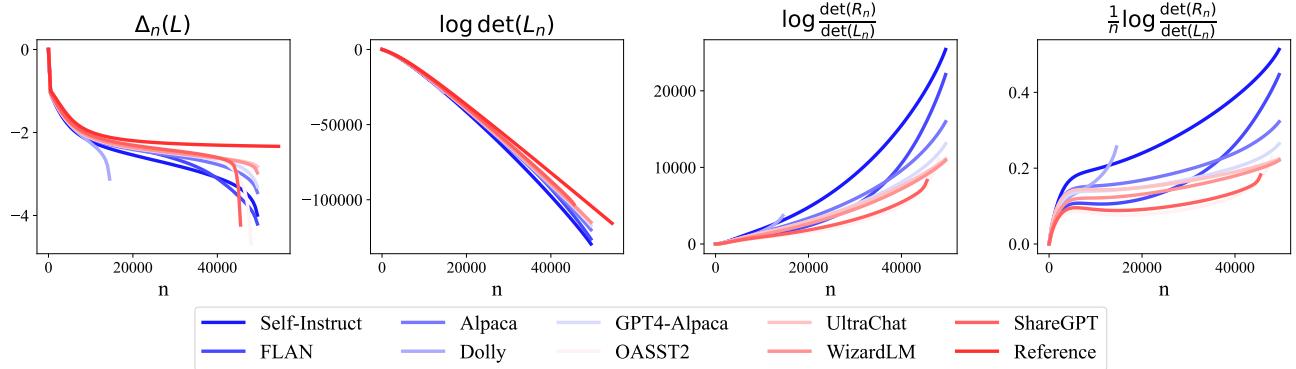


Figure 1. Step-by-step demonstration to compute the log determinant distance on a set of instruction tuning datasets of varying diversity. The marginal gain curve $\Delta_n(L)$ is derived from the greedy MAP algorithm for DPPs (1st figure). $\log \det(L_n)$ is the cumulative marginal gain curve (2nd figure). Note, scaling the kernel matrix L by a constant $c > 0$ shifts $\log \det(L_n)$ linearly by $n \log(c)$, complicating result interpretation. Moreover, $\log \det(L_n)$ is heavily influenced by dataset size, e.g., Dolly contains 15k examples and has a much larger $\log \det(L)$ compared to that of the other datasets subsampled to include roughly 50k examples despite it having a comparatively smaller $\log \det(L_{15k})$. To address these challenges, we compute the difference between the log determinant for a maximally diverse “reference” dataset and each dataset of interest (3rd figure) and then divide by dataset size (4th figure). The log determinant distance for a dataset is the value of the corresponding curve $(1/n) \log(\det(R_n)/\det(L_n))$ at the last iteration.

with a greedy algorithm (Nemhauser et al., 1978). We use Chen et al. (2018)’s implementation with $\mathcal{O}(\text{NMD})$ time and $\mathcal{O}(N(M+D))$ memory complexity, respectively. The costly evaluation of kernel matrix entries at each iteration can be parallelized on the GPU at the memory cost of $\mathcal{O}(ND)$. The algorithm is a feasible solution for selecting instruction tuning datasets at the current scale of several hundred thousand examples.

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

$$i^*(S) = \arg \max_{i \in [N] \setminus S} [\log \det(L_{S \cup \{i\}}) - \log \det(L_S)]$$

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

$$\begin{aligned} \Delta_n(L) &= \log \det(L_n) - \log \det(L_{n-1}) \\ &= \log \frac{\det(L_n)}{\det(L_{n-1})}, \quad n = 2, 3, \dots \end{aligned}$$

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, i.e.,

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

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. Figure 1 illustrates these problems.

To address the aforementioned challenges, we introduce a reference dataset that is maximally diverse for comparison. For example, we can generate the reference dataset by sampling at random on the hypersphere to ensure maximum diversity. We use R to denote the reference dataset’s kernel matrix computed using the same kernel function $k(\cdot, \cdot)$. We define *Log Determinant Distance* as

$$\text{LDD} \triangleq \frac{1}{N} \log \frac{\det(R)}{\det(L)}. \quad (3)$$

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:

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

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. Figure 1 illustrates a step-by-step computation of the log determinant distance from marginal gains.

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: $\text{LDD} \geq 0$. 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 two Johnson-Lindenstrauss (JL) transforms (Johnson & Lindenstrauss, 1984) consecutively 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

$$\nabla_B \ell = \nabla_h \ell \cdot z^T A^T = \nabla_W \ell \cdot A^T. \quad (4)$$

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.1 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.

Lemma 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 instruction 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 distinct aspects of the model: factual knowledge across various subjects like engineering and law 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 ALPACAEVAL % WIN to denote the proportion of times a model’s generation is preferred by GPT-4 over davinci-003’s response and ALPACAEVAL LEN as the average number of tokens in a model’s responses. We follow the evaluation procedure in (Wang et al., 2023a) closely to enable fair comparisons.

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 set $\gamma = 1$ for both LLAMA $\nabla_{\theta}\ell$ and MPNET EMB, $\gamma = 10$ for LLAMA EMB, and $\gamma = 0.01$ for LLAMA $\nabla_{\theta}\ell$ NOT NORM.. The choice of γ is not critical, as long as the greedy MAP inference algorithm does not terminate prematurely and that the kernel values do not underflow.

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 (3).

4.2. Results: Diversity Assessment

To assess the log determinant distance as a diversity measure, we compute the log determinant distance using weight

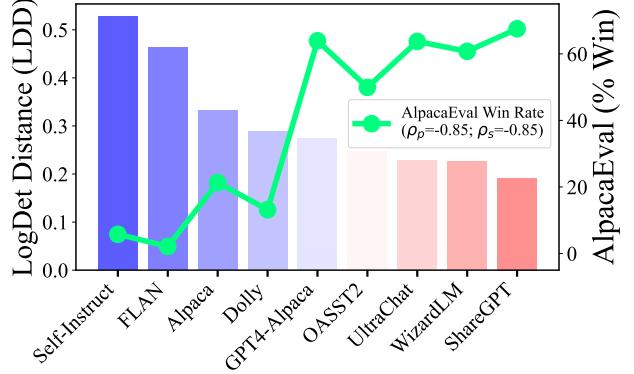


Figure 2. The log determinant distance in Equation (3) of instruction tuning datasets is correlated with instruction following performance when the model is finetuned on these datasets, with a Pearson correlation of $\rho_p = -0.85$ and a Spearman’s rank correlation of $\rho_s = -0.85$.

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.

The results Figure 2 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 3 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: (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 2).

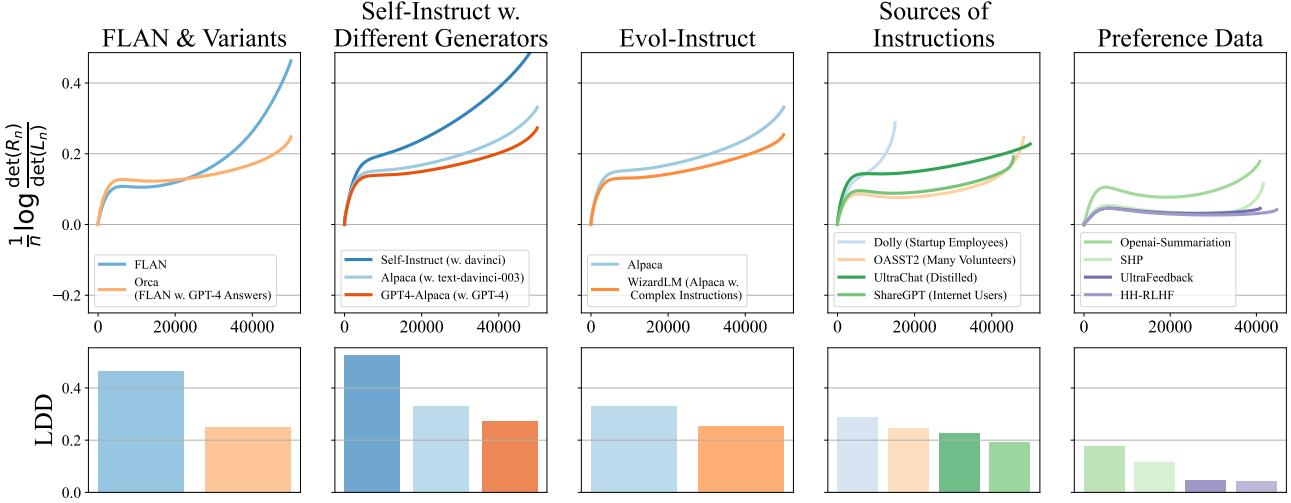


Figure 3. Studies of the log determinant distance as a measure of diversity of instruction tuning and preference learning datasets. More diverse datasets yield a log determinant distance curve that is closer to a horizontal line and closer to zero. Distilling responses from capable large language models 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).

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. Table 2 provides further details.

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, e.g., LLAMA EMB and MPNET EMB, on the Alpaca dataset. Figure 6 illustrates that this is true for different data budgets. On the more diverse UltraChat dataset, random selection is a strong baseline, and different data representations exhibit similar performance.

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 instruc-

tion following performance. Retaining examples with small $\|\nabla_{\theta}\ell\|_2$, instead of large $\|\nabla_{\theta}\ell\|_2$ typically used in active learning (Park et al., 2022), leads to significant improvement. Selecting examples with large #OUTPUT TOKENS yields the most substantial improvement in instruction tuning performance, doubling the win rate compared to random selection on the Alpaca dataset and match the performance of finetuning on 100% of the data on the UltraChat dataset.

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. We assign a higher weight ($\lambda = 0.9$) to #OUTPUT TOKENS. This approach leads to slightly improved instruction following performance compared to solely relying on #OUTPUT TOKENS on the Alpaca dataset and no improvement on the UltraChat dataset. Figure 5 illustrates the diversity-quality trade-off of DPP-based selection method on a length adjusted win rate metric: AlpacaEval’s win rate divided by the average output token lengths of the model’s output, multiplied by that of the reference model’s generations.

5. Discussion

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. Our approach of measuring data similarity in the normalized gradi-

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. (\uparrow) indicates that data points with higher quality scores are selected.

DATASETS METHODS	ALPACA			ULTRACHAT		
	BENCHMARK AVG	% WIN	LEN	BENCHMARK AVG	% WIN	LEN
100% DATA	23.0	21	91	22.9	64	215
RANDOM	21.6	18	90	22.6	57	213
DPP (LLAMA $\nabla_{\theta}\ell$)	21.7	26	104	23.3	58	217
DPP (LLAMA $\nabla_{\theta}\ell$ NOT NORM.)	22.7	8	34	22.7	54	197
DPP (LLAMA EMB)	22.5	21	92	23.3	53	204
DPP (LLAMA EMB NOT NORM.)	22.1	22	93	23.2	56	215
DPP (MPNET EMB)	22.9	20	88	23.3	58	211
DEDUP(MPNET EMB)	22.8	19	85	23.0	58	219
$\ \nabla_{\theta}\ell\ _2$ (\downarrow)	22.7	37	136	23.3	63	249
ALPAGASUS RATING (\uparrow)	21.6	21	95	-	-	-
EL2N (\downarrow)	22.8	21	99	22.7	56	220
IFD (\uparrow)	21.5	29	113	23.0	62	271
PERPLEXITY (\downarrow)	22.7	26	106	22.8	60	231
#INPUT TOKENS (\uparrow)	25.1	19	88	22.9	51	217
#OUTPUT TOKENS (\uparrow)	23.4	39	152	22.7	64	262
#TOTAL TOKENS (\uparrow)	20.4	38	149	22.7	62	251
DPP (LLAMA $\nabla_{\theta}\ell$ + #OUTPUT TOKS)	23.5	41	150	22.6	64	273

ent space improves instruction following performance over alternative data similarity measures on redundant dataset like Alpaca. More importantly, we proposed log determinant distance to quantify dataset diversity that is correlated with instruction following performance. We can use the proposed diversity metric to understand how much data should be kept when selecting data subsets. Figure 6 illustrates that the less diverse a dataset is (e.g., Dolly and Alpaca), the more data could be pruned without sacrificing performance. This implies that we should adjust the data budget proportional to its diversity. Furthermore, it provides us with a way to compare the diversity of existing instruction tuning & preference learning datasets to better understand the impact of the different curation strategies’ on dataset diversity.

When assessing the performance of finetuned models using AlpacaEval (Dubois et al., 2023), we corroborated existing observations that GPT-4 judge favors more verbose responses. Notably, we found that selecting data subsets with long responses yielded the most substantial improvement in win rate compared to alternative data selection methods. One might argue that longer responses contain more information that users prefer. However, we’d want to understand as well as to optimize for the model’s instruction following abilities after controlling for the length bias. Some work has started to address this issue of biased evaluation (Shen et al., 2023) that will be crucial for our problem of selecting optimal data subsets for instruction tuning.

Enforcing dataset diversity proves beneficial on less diverse datasets. This benefit diminishes when applied to more diverse datasets. In such cases, selecting datasets randomly

after basic text deduplication may be adequate. Our research on dataset diversity is useful to determine the placement of a new dataset on the diversity spectrum, helping us understand 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.

6. Conclusion

We present a DPP-based approach to select instruction tuning data subsets that prioritizes both diversity and quality. We propose to use log determinant distance to measure dataset diversity that is useful for analyzing datasets and selecting data subsets.

7. Impact

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

- (InThe)WildChat: 570K ChatGPT Interaction Logs In The Wild. In *The Twelfth International Conference on Learning Representations*, October 2023.
- Abbas, A., Tirumala, K., Simig, D., Ganguli, S., and Mornos, A. S. SemDeDup: Data-efficient learning at web-scale through semantic deduplication, March 2023.
- Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *International Conference on Learning Representations*, September 2019.
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., Das-Sarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback, April 2022.
- Broder, A. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp. 21–29, Salerno, Italy, 1998. IEEE Comput. Soc. ISBN 978-0-8186-8132-5. doi: 10.1109/SEQUEN.1997.666900.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.
- Bukharin, A. and Zhao, T. Data Diversity Matters for Robust Instruction Tuning, November 2023.
- Cao, Y., Kang, Y., and Sun, L. Instruction Mining: High-Quality Instruction Data Selection for Large Language Models, July 2023.
- Chen, L., Zhang, G., and Zhou, E. 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.
- Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., Tang, Z., Srinivasan, V., Zhou, T., Huang, H., and Jin, H. AlpagoS: Training A Better Alpaca with Fewer Data. In *International Conference on Learning Representations*, May 2024.
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating Large Language Models Trained on Code, July 2021.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling Instruction-Finetuned Language Models, December 2022.
- Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. 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.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training Verifiers to Solve Math Word Problems, November 2021.
- Conover, M., Hayes, M., Mathur, A., Xie, J., Wan, J., Shah, S., Ghodsi, A., Wendell, P., Zaharia, M., and Xin, R. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023.
- Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., and Sun, M. UltraFeedback: Boosting Language Models with High-quality Feedback. In *International Conference on Learning Representations*, May 2024.
- Ding, N., Chen, Y., Xu, B., Qin, Y., Hu, S., Liu, Z., Sun, M., and Zhou, B. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In Bouamor, H., Pino, J., and Bali, K. (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.
- Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., and Hashimoto, T. Alpaca-Farm: A Simulation Framework for Methods that Learn from Human Feedback. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.
- Ethayarajh, K., Choi, Y., and Swayamdipta, S. Understanding Dataset Difficulty with \$\mathcal{V}\$-Usable Information. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 5988–6008. PMLR, June 2022.
- Friedman, D. and Dieng, A. B. The Vendi Score: A Diversity Evaluation Metric for Machine Learning. *Transactions on Machine Learning Research*, June 2023.
- Gillenwater, J., Kulesza, A., and Taskar, B. Near-Optimal MAP Inference for Determinantal Point Processes. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, October 2020.
- Honovich, O., Scialom, T., Levy, O., and Schick, T. Unnatural Instructions: Tuning Language Models with (Almost) No Human Labor. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (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.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, January 2022.
- Huang, J., Child, R., Rao, V., Liu, H., Satheesh, S., and Coates, A. Active Learning for Speech Recognition: The Power of Gradients. *NIPS Workshop*, 2016.
- Johnson, W. B. and Lindenstrauss, J. Extensions of Lipschitz mappings into a Hilbert space. In Beals, R., Beck, A., Bellow, A., and Hajian, A. (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.
- Ko, C.-W., Lee, J., and Queyranne, M. An Exact Algorithm for Maximum Entropy Sampling. *Operations Research*, 43(4):684–691, 1995. ISSN 0030-364X.
- Köpf, A. and Kilcher, Y. OpenAssistant Conversations - Democratizing Large Language Model Alignment. In *Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.
- Kulesza, A. and Taskar, B. Structured Determinantal Point Processes. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.
- Kulesza, A. and Taskar, B. 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.
- Le, Q. and Mikolov, T. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1188–1196. PMLR, June 2014.
- Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating Training Data Makes Language Models Better. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577.
- Li, M., Zhang, Y., Li, Z., Chen, J., Chen, L., Cheng, N., Wang, J., Zhou, T., and Xiao, J. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. September 2023.
- Liu, W., Zeng, W., He, K., Jiang, Y., and He, J. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *International Conference on Learning Representations*, May 2024.
- Lu, K., Yuan, H., Yuan, Z., Lin, R., Lin, J., Tan, C., Zhou, C., and Zhou, J. #InsTag: Instruction Tagging for Analyzing Supervised Fine-tuning of Large Language Models. In *International Conference on Learning Representations*. arXiv, May 2024.
- Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., and Hooker, S. When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale, September 2023.
- Mukherjee, S., Mitra, A., Jawahar, G., Agarwal, S., Palangi, H., and Awadallah, A. Orca: Progressive Learning from Complex Explanation Traces of GPT-4, June 2023.
- Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. 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.

- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2022.
- Park, D., Papailiopoulos, D., and Lee, K. Active Learning is a Strong Baseline for Data Subset Selection. *NeurIPS HITY Workshop*, 2022.
- Paul, M., Ganguli, S., and Dziugaite, G. K. Deep Learning on a Data Diet: Finding Important Examples Early in Training. In *Advances in Neural Information Processing Systems*, November 2021.
- Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction Tuning with GPT-4, April 2023.
- Rana, A. Common crawl – building an open web-scale crawl using hadoop, 2010.
- Sanh, V., Webson, A., Raffel, C., Bach, S., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S. S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M. T.-J., Wang, H., Manica, M., Shen, S., Yong, Z. X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J. A., Teehan, R., Scao, T. L., Biderman, S., Gao, L., Wolf, T., and Rush, A. M. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*, January 2022.
- Sener, O. and Savarese, S. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*, February 2018.
- Settles, B. Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences, 2009.
- Shen, W., Zheng, R., Zhan, W., Zhao, J., Dou, S., Gui, T., Zhang, Q., and Huang, X. Loose lips sink ships: Mitigating Length Bias in Reinforcement Learning from Human Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2859–2873, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.188.
- Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. MP-Net: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867. Curran Associates, Inc., 2020.
- Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: Beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, December 2022.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020.
- Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them, October 2022.
- Swayamdipta, S., Schwartz, R., Lourie, N., Wang, Y., Hajishirzi, H., Smith, N. A., and Choi, Y. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In Webber, B., Cohn, T., He, Y., and Liu, Y. (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.
- Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model, 2023.
- Toneva, M. and Sordoni, A. An Empirical Study of Example Forgetting during Deep Neural Network Learning. *International Conference on Learning Representations*, 2019.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. LLaMA: Open and Efficient Foundation Language Models, February 2023.
- Tsang, I. W., Kwok, J. T., and Cheung, P.-M. Core Vector Machines: Fast SVM Training on Very Large Data Sets. *Journal of Machine Learning Research*, 6(13):363–392, 2005. ISSN 1533-7928.
- Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K. K., Patel, M., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R. S., Karia, R., Doshi, S., Sampat, S. K., Mishra, S., Reddy A. S., Patro, S., Dixit, T., and Shen, X. 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.
- Wang, Y., Ivison, H., Dasigi, P., Hessel, J., Khot, T., Chandu, K., Wadden, D., MacMillan, K., Smith, N. A., Beltagy, I., and Hajishirzi, H. 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.
- Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (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.
- Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*, January 2022.
- Welling, M. 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.
- Wenzek, G., Lachaux, M.-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4.
- Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., and Jiang, D. WizardLM: Empowering Large Language Models to Follow Complex Instructions. In *International Conference on Learning Representations*, May 2024.
- Zheng, L., Chiang, W.-L., Sheng, Y., Li, T., Zhuang, S., Wu, Z., Zhuang, Y., Li, Z., Lin, Z., Xing, E., Gonzalez, J., Stoica, I., and Zhang, H. RealChat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *International Conference on Learning Representations*, May 2024.
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., Zhang, S., Ghosh, G., Lewis, M., Zettlemoyer, L., and Levy, O. LIMA: Less Is More for Alignment. In *Thirty-Seventh Conference on Neural Information Processing Systems*, November 2023.

A. Appendix

A.1. Proof of Lemma 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 $p_1, \dots, p_m \in \mathbb{R}^n$ be rows of $\text{vec}(\nabla_W \ell)$ and $q_1, \dots, q_m \in \mathbb{R}^r$ be rows of $\text{vec}(\nabla_B \ell)$. Due to Equation (4), we have $q_k = Ap_k$ for $k = 1, \dots, m$. 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):

$$\mathbb{P} \left[\left| \|q_k\|_2^2 - \|p_k\|_2^2 \right| < \frac{\epsilon}{m} \right] \geq 1 - \frac{\delta}{m}. \quad (5)$$

By union bound,

$$\mathbb{P} \left[\bigcup_{k=1}^m \left\{ \left| \|q_k\|_2^2 - \|p_k\|_2^2 \right| > \frac{\epsilon}{m} \right\} \right] \leq \sum_{k=1}^m \mathbb{P} \left[\left| \|q_k\|_2^2 - \|p_k\|_2^2 \right| > \frac{\epsilon}{m} \right] \leq \sum_{k=1}^m \frac{\delta}{m} \leq \delta. \quad (6)$$

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

$$\left| \|q\|_2^2 - \|p\|_2^2 \right| = \left| \sum_{k=1}^m \|q_k\|_2^2 - \sum_{k=1}^m \|p_k\|_2^2 \right| \leq \sum_{k=1}^m \left| \|q_k\|_2^2 - \|p_k\|_2^2 \right| \leq \sum_{k=1}^m \frac{\epsilon}{m} = \epsilon. \quad (7)$$

Therefore,

$$\mathbb{P} \left[\left| \|q\|_2^2 - \|p\|_2^2 \right| \leq \epsilon \right] \geq \mathbb{P} \left[\bigcap_{k=1}^m \left\{ \left| \|p_k\|_2^2 - \|q_k\|_2^2 \right| < \frac{\epsilon}{m} \right\} \right] \geq 1 - \delta \quad (8)$$

where the last inequality is by Equation (6).

B. Appendix

This section of the appendix contains additional tables and figures from Section 4.

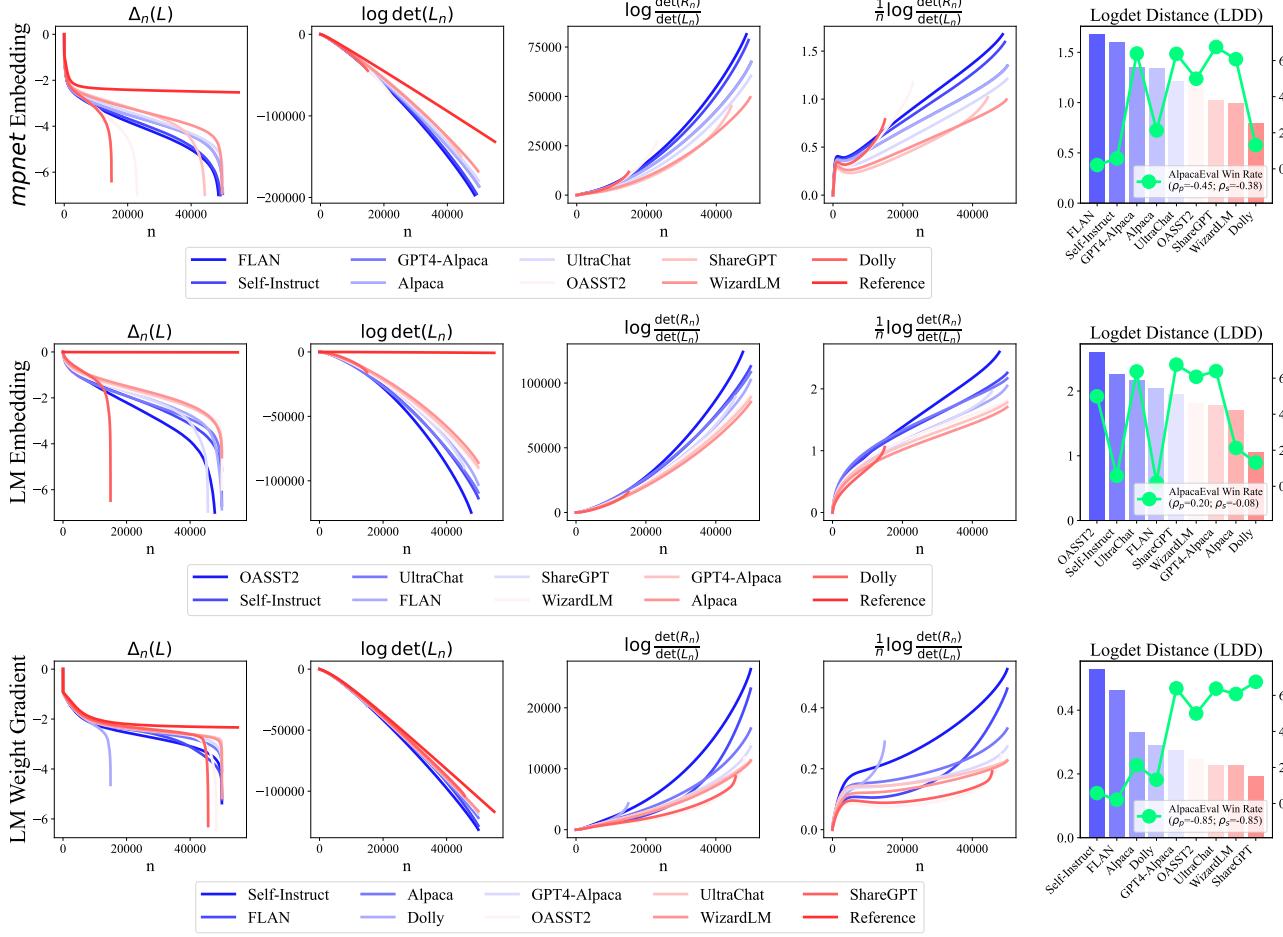


Figure 4. Comparison of the log determinant distance computed using different data representations: MPNET EMB (top row), Llama EMB (middle row), and Llama $\nabla_\theta \ell$ with respect to instruction tuning loss (bottom row). The log determinant distance based on weight gradient provides the strongest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

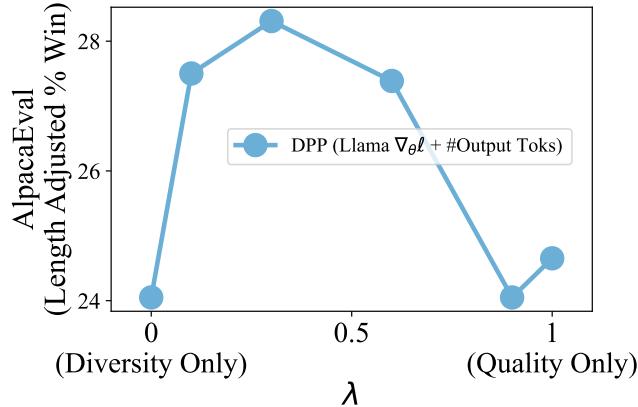


Figure 5. Vary λ interpolates between enforcing diversity and selecting for quality. Here we use the length adjusted win rate metric computed from AlpacaEval's win rate divide by the average length of the model's generations, then multiply by that of the reference model's generations.

Table 2. 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. (\uparrow) indicates that data points with higher quality score are selected.

METHODS	ACADEMIC BENCHMARKS						ALPACA EVAL	
	MMLU	GSM	BBH	TYDIQA	CODEX EVAL	Avg	% WIN	LENGTH
ALPACA								
RANDOM	32.3	4.8	33.4	22.4	8.5	21.6	18	90
100% DATA	41.7	4.5	32.2	20.2	9.8	23.0	21	91
DPP (LLAMA $\nabla_{\theta}\ell$ NOT NORM.)	41.3	5.3	26.3	25.0	8.5	22.7	8	34
DEDUP(MPNET EMB)	38.0	5.8	32.1	21.4	11.0	22.8	19	85
DPP (MPNET EMB)	39.1	5.8	33.7	20.0	8.5	22.9	20	88
DPP (LLAMA EMB)	37.9	6.9	29.9	21.0	11.2	22.5	21	92
DPP (LLAMA EMB NOT NORM.)	36.9	5.1	31.6	21.5	8.7	22.1	22	93
DPP (LLAMA $\nabla_{\theta}\ell$)	28.8	7.4	34.1	22.6	9.6	21.7	26	104
#INPUT TOKENS (\uparrow)	43.8	5.3	34.0	24.8	10.4	25.1	19	88
EL2N (\downarrow)	38.3	6.4	32.0	19.9	12.2	22.8	21	99
ALPAGASUS RATING (\uparrow)	36.2	5.2	31.3	19.8	9.1	21.6	21	95
PERPLEXITY (\downarrow)	37.7	4.7	32.9	21.0	11.6	22.7	26	106
IFD (\uparrow)	34.6	4.4	30.9	23.5	6.7	21.5	29	113
$\ \nabla_{\theta}\ell\ _2$ (\downarrow)	36.9	6.9	33.1	20.9	8.5	22.7	37	136
#TOTAL TOKENS (\uparrow)	25.1	7.2	33.5	21.7	8.5	20.4	38	149
#OUTPUT TOKENS (\uparrow)	36.3	7.1	35.4	21.8	9.1	23.4	39	152
DPP (LLAMA $\nabla_{\theta}\ell$ + #OUTPUT TOKS)	36.6	6.6	35.2	22.3	9.8	23.5	41	150
ULTRACHAT								
RANDOM	36.2	7.5	32.8	19.8	11.0	22.6	57	213
100% DATA	37.8	7.8	32.1	20.2	10.4	22.9	64	215
DPP (LLAMA EMB)	38.1	8.9	32.8	18.9	12.2	23.3	53	204
DPP (LLAMA $\nabla_{\theta}\ell$ NOT NORM.)	37.7	8.4	32.6	17.8	11.0	22.7	54	197
DPP (LLAMA EMB NOT NORM.)	38.0	8.6	34.2	19.6	8.5	23.2	56	215
DPP (MPNET EMB)	37.3	8.6	33.9	19.1	11.6	23.3	58	211
DEDUP(MPNET EMB)	37.4	8.7	31.5	21.2	9.1	23.0	58	219
DPP (LLAMA $\nabla_{\theta}\ell$)	36.2	7.6	34.5	20.1	12.8	23.3	58	217
#INPUT TOKENS (\uparrow)	37.7	8.5	33.6	18.5	9.1	22.9	51	217
EL2N (\downarrow)	38.6	7.5	32.4	18.2	11.0	22.7	56	220
PERPLEXITY (\downarrow)	36.9	7.6	33.0	19.4	11.6	22.8	60	231
#TOTAL TOKENS (\uparrow)	34.6	8.3	33.3	20.9	9.8	22.7	62	251
IFD (\uparrow)	34.5	10.2	33.4	21.4	7.9	23.0	62	271
$\ \nabla_{\theta}\ell\ _2$ (\downarrow)	34.7	10.3	32.2	21.4	12.2	23.3	63	249
#OUTPUT TOKENS (\uparrow)	34.5	10.3	31.9	20.8	9.1	22.7	64	262
DPP (LLAMA $\nabla_{\theta}\ell$ + #OUTPUT TOKS)	34.5	8.9	32.3	21.6	9.1	22.6	64	273

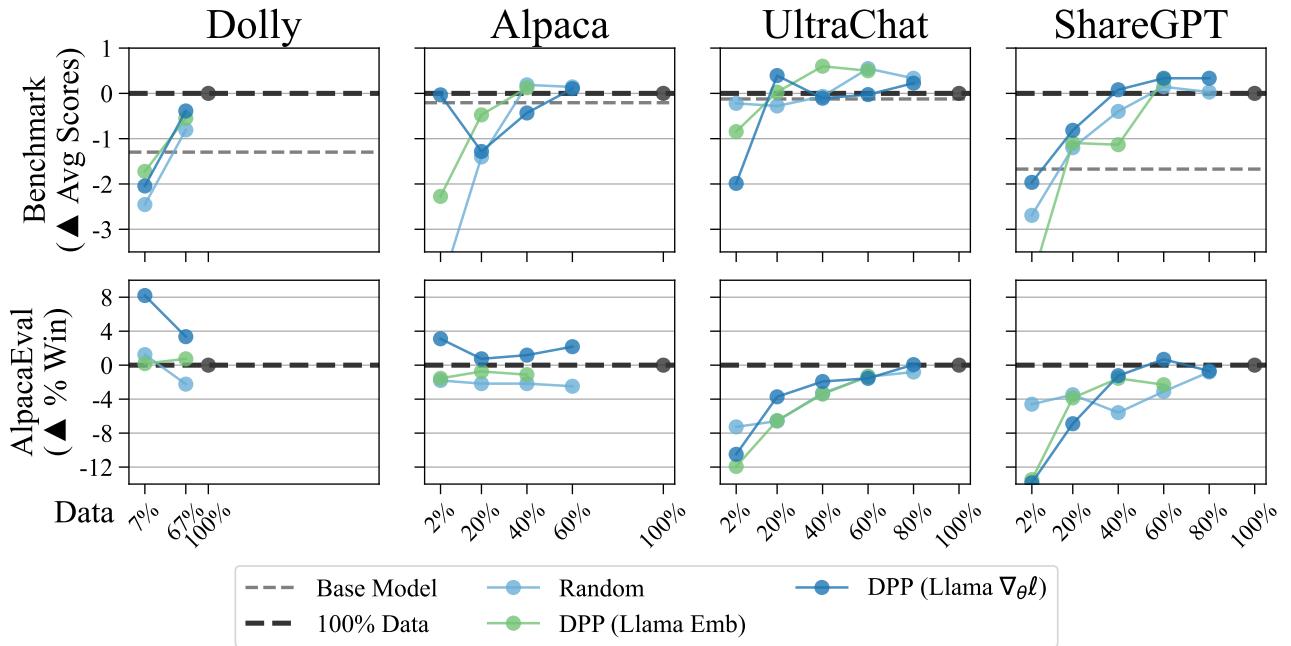


Figure 6. Performance of DPP data selection methods relative to the full finetuning of Llama-7b on 4 datasets with different data budget. Using the LLAMA $\nabla_{\theta} \ell$ as data representation is superior to LLAMA EMB on Dolly and Alpaca, while exhibiting comparable performance on UltraChat and ShareGPT. For (money) budget reasons, we use GPT-4-Turbo as the judge for AlpacaEval.