Helpful or Harmful Data?
Fine-tuning-free Shapley Attribution for Explaining Language Model Predictions

Jingtan Wang [* 1 2]   **Xiaoqiang Lin** [* 1]   **Rui Qiao** [* 1]   **Chuan-Sheng Foo** [2]   **Bryan Kian Hsiang Low** [1]

Abstract

The increasing complexity of foundational models underscores the necessity for explainability, particularly for fine-tuning, the most widely used training method for adapting models to downstream tasks. Instance attribution, one type of explanation, attributes the model prediction to each training example by an instance score. However, the robustness of instance scores, specifically towards dataset resampling, has been overlooked. To bridge this gap, we propose a notion of robustness on the sign of the instance score. We theoretically and empirically demonstrate that the popular leave-one-out-based methods lack robustness, while the Shapley value behaves significantly better, but at a higher computational cost. Accordingly, we introduce an efficient fine-tuning-free approximation of the Shapley value (FreeShap) for instance attribution based on the neural tangent kernel. We empirically demonstrate that FreeShap outperforms other methods for instance attribution and other data-centric applications such as data removal, data selection, and wrong label detection, and further generalize our scale to large language models (LLMs). Our code is available at `https://github.com/JTWang2000/FreeShap`.

1. Introduction

Modern deep learning is primarily driven by pretraining foundation models (Bommasani et al., 2021) such as BERT (Devlin et al., 2019), GPT3 (Brown et al., 2020), and CLIP (Radford et al., 2021) on massive datasets with increasingly more model parameters, then fine-tuning them with labeled training data for downstream tasks. However, as larger foundation models have more complexity and opaqueness, their predictions are hard to explain and justify. The lack of explainability reduces the trustworthiness of foundation models, especially for systems in areas that involve high-stakes decision-making such as healthcare, finance, and legal justice (Rudin, 2019). As a result, *instance attribution*, also known as example-based explanation, has been proposed in response to the increasing need for explainable AI. Different from other types of explanation (e.g., feature attribution (Li et al., 2016; Madsen et al., 2022; Ribeiro et al., 2016; Sundararajan et al., 2017)), instance attribution traces the effect of the labeled training examples on the resultant model parameters (Barshan et al., 2020; Guo et al., 2021; Han et al., 2020; Pezeshkpour et al., 2021). On a high level, this approach offers a better understanding of the fine-tuned model by sifting the influential training examples with larger effects on the model performance. On a detailed level, it explains the model prediction on each data point by highlighting the most relevant training examples. Besides, instance attribution can also be practically used for data selection (Xia et al., 2024), the identification of dataset artifacts (Han & Tsvetkov, 2021), model debugging (Lertvittayakumjorn & Toni, 2021), wrong label detection, adversarial example crafting (Koh & Liang, 2017), and data pricing in collaborative machine learning (Sim et al., 2022).
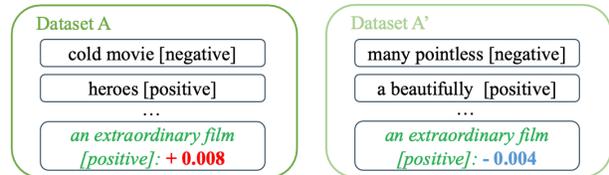

Figure 1. An example of non-robust instance attribution. The *same* training example receives *different* signs of the instance score when it is placed in different datasets sampled from the same task.

Given a dataset for a downstream task, instance attribution calculates an instance score for each training example. A positive instance score deems the example "helpful" for the task as it improves the model performance, while a

---
[*]Equal contribution   [1]Department of Computer Science, National University of Singapore   [2]Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR). Correspondence to: Bryan Kian Hsiang Low <lowkh@comp.nus.edu.sg>.

negative score reflects the opposite. Thus, the sign of the scores is a widely used criterion for the usefulness of the data (Han et al., 2020; Yeh et al., 2018). It is important to realize that approaches for instance attribution are dataset-dependent, meaning that the instance scores can be different when the dataset is resampled. If the instance score for one example flips signs frequently when the rest of the training examples change as shown in Fig. 1, this score results in a confusing and less trustworthy analysis. Ideally, the instance score for a "helpful" example should be consistently positive even if the rest of the training data are resampled from the same task, the same applies to a "harmful" example. Thus, we define the *robustness of instance attribution* as the ability of an approach to maintain the sign consistency of the instance scores across different resampled datasets. Our robustness definition is motivated from the perspective of data distribution, hence complementary to existing perturbation-based robustness measures (Alvarez-Melis & Jaakkola, 2018; Bhatt et al., 2020; Ghorbani et al., 2019; Tang et al., 2022).
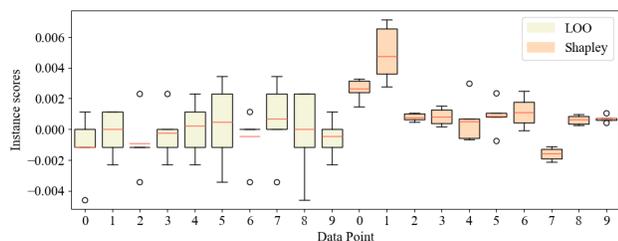

Figure 2. Mean and variance for instance scores of 10 examples when computed using LOO or the Shapley value.

Most of the popular instance attribution approaches (Barshan et al., 2020; Han et al., 2020; Guo et al., 2021; Garima et al., 2020) are based on the leave-one-out (LOO) scheme, which evaluates the marginal contribution (i.e., the gain in model performance by adding the target data point) to the rest of the training set. However, LOO tends to produce instance scores with small magnitude and high variance (Kwon & Zou, 2022). Therefore, LOO-based scores are more likely to have inconsistent signs when the dataset is resampled (see Tab. 2), leading to poor robustness. On the contrary, the Shapley value (Ghorbani & Zou, 2019; Jia et al., 2019; Wang & Jia, 2023) considers a weighted average of marginal contributions to all training subsets. As a result, it tends to enjoy a relatively larger magnitude and smaller variance (see Fig. 2). Moreover, we theoretically and empirically show that, under mild assumptions, the Shapley value possesses noticeably better robustness compared to LOO in Sec. 3.2 and Sec. 4.2.

However, the time complexity to compute the exact Shapley value grows exponentially w.r.t. the number of training data,

which is substantially amplified by the computational cost of fine-tuning. Although existing approaches have attempted to reduce the exponential time w.r.t. dataset size (Ghorbani & Zou, 2019; Jia et al., 2019; Kolpaczki et al., 2023), the cost of fine-tuning has been overlooked. As the size of foundation models grows larger, there is an increasing need to reduce the running time for fine-tuning in order to efficiently approximate the Shapley value. To this end, we propose an efficient fine-tuning-<u>free</u> approximation of the <u>Shap</u>ley value (FreeShap). Specifically, it has been shown that kernel regression using empirical neural tangent kernel (eNTK) (Jacot et al., 2018) resembles the fine-tuning process (Wei et al., 2022; Malladi et al., 2023). An observation is that the eNTK matrices for different data subsets are submatrices of the eNTK matrices of the whole dataset. We exploit this observation to accelerate FreeShap. Specifically, FreeShap amortizes the cost of calculating the Shapley values by pre-computing the eNTK matrix to replace all subsequent training with kernel regression (i.e., fine-tuning-free) which is significantly cheaper in computational cost than fine-tuning. Our approach generally applies to *any* foundation model, but we focus our demonstrations in the natural language processing (NLP) domain in light of the breakthroughs in language models (LMs), especially LLMs. Our contributions are:

- Introducing β-robustness, a notion of robustness for *instance attribution* (Sec. 3.1), and theoretically analyzing and empirically demonstrating that the Shapley value is more robust than LOO (Sec. 3.2 and Sec. 4.2).
- Proposing an efficient fine-tuning-free approximation of Shapley value, FreeShap, by *replacing fine-tuning with kernel regression*, increasing the scalability of the Shapley value significantly for NLP tasks (Sec. 3.3 and Sec. 4.1).
- Demonstrating the effectiveness of the FreeShap through extensive experiments using pre-trained LMs and LLMs for instance attribution, data removal, data selection, and wrong label detection on real-world datasets (Sec. 5).

2. Background and Preliminaries

Let $D_N := \{z_i = (x_i, y_i)\}_{i=1}^n$ denote the training set, where $N := \{1, \ldots, n\}$ is the set of indices, and z_i consisting of the input $x_i \in \mathcal{X}$ and the label $y_i \in [C]$ is a data point sampled from a data distribution $\mathcal{P}$. Similar notations apply to the test set $D_T := \{z_t = (x_t, y_t)\}_{t=n+1}^{n+m}$, the test indices $T := \{n+1, \ldots, n+m\}$, and the test example z_t. An attribution function $g(z_i, D_T, D_N)$ quantifies the contribution of a training example z_i to the model predictions on the test set D_T (or the test example $\{z_t\}$). We define the utility for a subset of the training data $D_S := \bigcup_{j \in S} \{z_j \in D_N\}, S \subseteq N$ as the test accuracy on D_T of its resultant model f_S parameterized by θ:

$$U(S, D_T) := \frac{1}{|D_T|} \sum_{(x_t, y_t) \in D_T} \mathbb{1}[f_S(x_t; \theta) = y_t] . \quad (1)$$

2.1. Prompt-based Fine-tuning

The LMs are pretrained on tasks such as masked token prediction (Devlin et al., 2019) or next-token prediction (Brown et al., 2020). Prompt-based fine-tuning (Gao et al., 2021; Schick et al., 2021) is an advanced technique that appends natural language cues like "This is [MASK]." (i.e., prompt) to the end of x and then employs the pretrained model to predict the word for [MASK] from words that are relevant to the task. For example, binary sentiment analysis can be formulated as predicting only the task-specific terms including "great" and "terrible" at the "[MASK]" location. Prompt-based fine-tuning has been shown to have better performance than fine-tuning with sentence representations (e.g., the last hidden-layer representation for the [CLS] token) without prompts. Consequently, we focus on prompt-based fine-tuning in this work.

2.2. LOO and Shapley Value for Instance Attribution

Most of the instance attribution approaches can be viewed as some forms of marginal contributions (Kwon & Zou, 2022). Define the marginal contribution of a data point z_i to the subset with size k of all other training data as:

$$\Delta_{z_i}^{D_N}(k, D_T) := \frac{1}{\binom{n-1}{k}} \sum_{\substack{S \subseteq N \setminus \{i\} \\ |S|=k}} U(S \cup \{i\}, D_T) - U(S, D_T) \tag{2}$$

Since LOO (Koh & Liang, 2017) considers the change in model accuracy by removing the target data point from the training set, it can be viewed as the marginal contribution of z_i to $D_N \setminus \{z_i\}$. Specifically, LOO can be written as the following instance scoring function:

$$g^{\text{LOO}}(z_i, D_T, D_N) = \Delta_{z_i}^{D_N}(n-1, D_T). \tag{3}$$

Many existing instance attribution schemes including the influence function (Han et al., 2020) and its subsequent works (Barshan et al., 2020; Guo et al., 2021) approximate the LOO scores in different ways. However, LOO only considers the marginal contribution of z_i to one subset of D_N with size $n-1$. The instance score of an example drops significantly when another similar example appears in the training set (Ghorbani & Zou, 2019), leading to the weak reliability and robustness of the score.

Instead of considering only one subset, the Shapley value (Shapley, 1953) generalizes it by considering the impact of z_i in all subsets (Ghorbani & Zou, 2019) and can be written as the following instance scoring function:

$$g^{\text{Shap}}(z_i, D_T, D_N) = n^{-1} \sum_{k=0}^{n-1} \Delta_{z_i}^{D_N}(k, D_T). \tag{4}$$

This can be viewed as the average of the marginal contributions of data point z_i to subsets with all possible sizes. However, calculating the Shapley value on the training set with size n requires an exponential number (i.e., $n!$) of LM fine-tuning, which necessitates acceleration for improved scalability on real-world datasets and models.

2.3. Neural Tangent Kernel

The neural tangent kernel (NTK) theory is proposed to study the training dynamics of infinite-width neural networks (Jacot et al., 2018), and has been extended to other applications, such as active learning (Hemachandra et al., 2023; Lau et al., 2024). Existing works (Arora et al., 2019; Liu et al., 2020) show that training a fully connected and sufficiently wide neural network is equivalent to solving kernel regression with the NTK at random initialization. However, there are two challenges when applying the NTK theory to fine-tuning LMs: (1) the prevalence of using pretrained weights instead of random initialization; (2) the usage of prompts. Malladi et al. (2023) extends the analysis to show that solving the kernel regression with the empirical NTK (eNTK) calculated from the pretrained weights can resemble prompt-based fine-tuning. Moreover, kernel regression using eNTK has been shown to have similar performance as fine-tuning in both computer vision (Wei et al., 2022) and NLP (Malladi et al., 2023), which indicates its applicability as a surrogate to fine-tuning.

Specifically, the eNTK is calculated using the Jacobian of the model output when taking a data point x_i: $\psi(x_i) := \frac{\partial f(x_i; \theta_0)}{\partial \theta_0} \in \mathbb{R}^{C \times P}$, where $\theta_0 \in \mathbb{R}^P$ consists of the pretrained weights. For a training dataset D_S with indices $S := \{1, \ldots, k\}$, denote the input matrix by $X_S := [x_1, \ldots, x_k]^\top$ and the corresponding labels in one-hot by $Y_S := [y_1^1, \ldots, y_1^C, \ldots, y_k^1, \ldots, y_k^C]^\top \in \{0, 1\}^{kC}$, where y_k^j is 1 if $y_i = j$. A test input x_t can be predicted using the eNTK regression model as follows:

$$f_S^{\text{entk}}(x_t) = K(x_t, X_S)^\top K(X_S, X_S)^{-1} Y_S \tag{5}$$

where $K(x_t, X_S) = \left[\psi(x_t)\psi(x_i)^\top\right]_{i=1}^k \in \mathbb{R}^{kC \times C}$ and $K(X_S, X_S) = \left[\psi(x_i)\psi(x_j)^\top\right]_{i,j=1}^k \in \mathbb{R}^{kC \times kC}$.

3. Methodology

3.1. Definition of Robustness

Denote $D_N \sim \mathcal{P}^{n-1}|z_i$ as sampling a new dataset D_N with z_i fixed and other $n-1$ data points i.i.d. sampled from $\mathcal{P}$.

Definition 3.1 (Expected marginal contribution). Let $\tau_k := \mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i} \left[\Delta_{z_i}^{D_N}(k, D_T)\right]$ be the expected marginal contribution of z_i to the subsets of D_N with size k.

Definition 3.2 (Consistently helpful/harmful data point). A training data point z_i is *consistently helpful* to test set D_T if $\tau_k \geq 0, \forall k \geq \{0, \ldots, n-1\}$. Similarly, it is *consistently harmful* if $\tau_k < 0$.

If z_i is a *consistently harmful* data point, including z_i in any subset of the dataset is expected to hurt the model performance on the test set. Intuitively, a *consistently harmful* data point z_i should receive a negative instance score. However, due to the randomness of the other data points when

sampling the dataset, the attribution score might flip to non-negative. This is undesirable, especially for high-stakes applications, since a harmful data point can be interpreted as helpful due to its positive score at one specific dataset and hence not being removed. Therefore, it is preferable to have a robust instance attribution scheme that could assign the correct signs for the training examples more consistently regardless of the companion dataset D_N. We emphasize the *sign* as it is the natural threshold that determines the usefulness of the data (Han et al., 2020; Yeh et al., 2018). Therefore, we define the following:

Definition 3.3 (Robustness of instance attribution). A sign function is defined as $\text{sgn}(x) = -1$ if $x < 0$, and $\text{sgn}(x) = 1$ if $x \geq 0$. For a *consistently harmful* point z_i, let the $\text{sgn*}(z_i) = -1$, while for a *consistently helpful* point z_i, let the $\text{sgn*}(z_i) = 1$. An instance attribution approach is β-robust in giving instance score for a data point if:

$$\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\bigg(\text{sgn}\Big(g(z_i, D_T, D_N)\Big) \neq \text{sgn*}(z_i)\Big)\bigg) = \beta.$$

In other words, given a training example z_i, an instance attribution scheme is β-robust if it has a probability of β in giving an instance score with the opposite sign when resampling the dataset. Overall, a robust instance attribution scheme should maintain a small β and ensure the *sign-consistency* of the instance scores for a data point across different datasets drawn from the same data distribution.

3.2. Robustness of Instance Attribution

Theorem 3.4 (Robustness for Shapley value & LOO). *Let* $\delta_k := Var_{D_N \sim \mathcal{P}^{n-1}|z_i}(\Delta_{z_i}^{D_N}(k, D_T)), \forall k \in \{0, \ldots, n-1\}$. *Shapley value is* β^{Shap}-*robust and LOO is* β^{LOO}-*robust where*

$$\beta^{Shap} \leq \frac{n^{-1}\sum_{k=0}^{n-1}\delta_k}{(n^{-1}\sum_{k=0}^{n-1}\tau_k)^2} \quad and \quad \beta^{LOO} \leq \frac{\delta_{n-1}}{\tau_{n-1}^2}.$$

The proof is in App. A. On the variance δ_k, Theorem 1 (Kwon & Zou, 2022) shows that the asymptotic of $\delta_k \sim O(Ck^2/n)$ (i.e., the upper bound of the δ_k scales quadratically with the subset size, k), it is reasonable to assume that $n^{-1}\sum_{k=0}^{n-1}\delta_k \leq \delta_{n-1}$. Additionally, as the dataset size grows, the contribution of each training example on the model typically diminishes, because the abundance of other data points is more likely to compensate for the performance drop. Hence, we also assume diminishing influence, i.e., $|\tau_0| \geq \cdots \geq |\tau_{n-1}|$, which coincides with the assumptions in Killamsetty et al. (2021); Wang et al. (2021) and is empirically verified in App. G.1. As a result:

Corollary 3.5 (Robustness Analysis between the Shapley value and LOO). *Assume that* $n^{-1}\sum_{k=0}^{n-1}\delta_k \leq \delta_{n-1}$. *Additionally, for any consistently harmful (or helpful) contribut-*

ing data point z_i, *assume* $|\tau_0| \geq \cdots \geq |\tau_{n-1}|$. *Then,*

$$\frac{n^{-1}\sum_{k=0}^{n-1}\delta_k}{(n^{-1}\sum_{k=0}^{n-1}\tau_k)^2} \leq \frac{\delta_{n-1}}{\tau_{n-1}^2}. \tag{6}$$

This implies that the upper bound of the β-robustness of the Shapley value is no more than that of LOO. The upper bounds are valid proxies to the β-robustness for comparison because they are *derived with the same non-trivial techniques*. Having a better upper bound suggests that the Shapley value can be less likely to give sign-inconsistent scores, and hence more robust. Note that the robustness analysis applies to *every* training example z_i. We provide empirical evidence to demonstrate that the Shapley value is indeed more robust than LOO in Sec. 4.2, w, which further validates our theoretical insights.

The numerator in the upper bounds of β is related to the variance of the instance score, while the denominator is related to the expectation of the instance score. Thus, the robustness of the Shapley value can also be interpreted from the expectation and variance. Without loss of generality, we analyze for a *consistently helpful* data point.

Remark 3.6 (Relative relationship of expectation and variance between Shapley and LOO). For a *consistently helpful* point z_i, with the assumption that $|\tau_0| \geq \cdots \geq |\tau_{n-1}|$:

$$\mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big) = n^{-1}\sum_{k=0}^{n-1}\tau_k \\ \geq \tau_{n-1} = \mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{LOO}}(z_i, D_T, D_N)\big). \tag{7}$$

With the assumption that $n^{-1}\sum_{k=0}^{n-1}\delta_k \leq \delta_{n-1}$:

$$\text{Var}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big) \leq n^{-1}\sum_{k=0}^{n-1}\delta_k \\ \leq \delta_{n-1} = \text{Var}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{LOO}}(z_i, D_T, D_N)\big). \tag{8}$$

Overall, the Shapley value can have a larger magnitude of expectation and smaller variance than LOO. For a *consistently helpful* data point, the Shapley value has a lower probability of fluctuating around the borderline 0 due to the larger expectation and smaller variance. The remark is further empirically verified in the App. G.2.

3.3. Fine-tuning-free Shapley Value (FreeShap)

Shapley value computation requires marginal contribution computations. For each subset S, we need to retrain a model f_S. As there are 2^n subsets for D_N, exponential times of fine-tuning are needed. Prior works have adopted Monte-Carlo (MC) sampling (Castro et al., 2009; Maleki et al., 2013; Ghorbani & Zou, 2019; Lin et al., 2023) to avoid the exponential factor. Specifically, the marginal contribution of each data point is evaluated in each MC iteration where a random permutation is generated. The average result of each iteration is the final estimation of the Shapley value.

Algorithm 1 Fine-tuning-free Shapley Value (FreeShap)

1: **Require:** Training set D_N, test set D_T, utility function U, maximum MC iteration M, eNTK function K
2: **Output:** Instance score of training examples computed by the Shapley value: $\phi_1, \ldots, \phi_n$
3: Initialize $\phi_i \leftarrow 0$ for $i = 1, \ldots, n$, $t \leftarrow 0$
4: Precompute the eNTK matrix $\tilde{K} := K(X_{T \cup N}, X_N)$
5: **while** $t \leq M$ **do**
6:     $t \leftarrow t + 1$
7:     π^t: A permutation of the training indices N
8:     $u_0^t \leftarrow U(\emptyset, D_T)$
9:     **for** $p \in \{1, \ldots, n\}$ **do**
10:        $S_p^t \leftarrow \{\pi^t[1], \ldots, \pi^t[p]\}$
11:        $f_{S_p^t}^{\text{entk}}(D_T) = \tilde{K}[T, S_p^t]\tilde{K}[S_p^t, S_p^t]^{-1}Y_{S_p^t}$
               ▷See Eqn. (9)
12:        $u_p^t \leftarrow U(S_p^t, D_T | f_{S_p^t} = f_{S_p^t}^{\text{entk}})$
13:        $\phi_{\pi^t[p]} \leftarrow \frac{t-1}{t}\phi_{\pi^{t-1}[p]} + \frac{1}{t}(u_p^t - u_{p-1}^t)$
14:     **end for**
15: **end while**

MC sampling gives an unbiased estimate of the Shapley value with a tolerable error bound (Maleki et al., 2013).

Although MC sampling has reduced the exponential time to polynomial time, it is still a large factor that multiplies with the fine-tuning cost. Fine-tuning a large pre-trained LM itself is computationally expensive. To further speed up the Shapley value for large NLP datasets, we aim to improve the complexity of computing the utility function which relies on the fine-tuned model f_S. Instead of fine-tuning, we adopt kernel regression on the *empirical* NTK (eNTK) (Wei et al., 2022), using f_S^{entk} to resemble f_S. Note that D_N denotes the training dataset of size n, while D_T denotes the testing dataset of size m, and D_S denotes a subset of the training dataset. We just need to precompute the eNTK matrix $\tilde{K}$ once, as shown in line 4 of Alg. 1, and then reuse it for all marginal contributions by taking the submatrices of $\tilde{K}$ for corresponding kernel regression. Specifically,

$$\tilde{K}[T, S_p^t] = K(X_T, X_{S_p^t}), \quad \tilde{K}[S_p^t, S_p^t] = K(X_{S_p^t}, X_{S_p^t}). \quad (9)$$

We refer to this method as the fine-tuning-free Shapley value (FreeShap), which ensures the scalability of computing the Shapley value with pretrained LMs on sizable datasets. We further enhance the computational efficiency by truncated Monte-Carlo (TMC) sampling (Ghorbani et al., 2019) and employing efficient kernel approximation and blockwise inversion (BI), which are discussed in App. B.

4. Experiments and Results

We first empirically demonstrate that the proposed FreeShap effectively approximates the de facto implementation of the Shapley value. Hence, it justifies the usage of FreeShap for

all subsequent experiments at scale. We then empirically compare the β-robustness between the FreeShap and LOO and show that the Shapley value is more robust, which is aligned with our theoretical insights in Sec. 3.2.

Datasets: We conduct experiments on both single-sentence tasks and sentence pair tasks for comprehensiveness. In particular, we select Stanford Sentiment Treebank v2 (SST-2) and Rotten Tomatoes Movie Review (MR) for the single-sentence task, and Microsoft Research Paraphrase Corpus (MRPC) and Recognizing Textual Entailment (RTE) for the sentence pair task. Detailed information about the datasets is in App. H. We use BERT for this experiment. More hyperparameter settings are in the App. D.1.

4.1. FreeShap Approximates the Shapley Value Well

Setup. Due to the exponential time complexity for the exact Shapley value, we use the Monte-Carlo (MC) Shapley value as the reference. We perform prompt-based fine-tuning with the Adam optimizer (Kingma & Ba, 2015) when calculating its marginal contribution. We benchmark against gradient Shapley (G-Shapley) (Ghorbani & Zou, 2019), which also minimizes fine-tuning costs by limiting training to a single epoch. For evaluation metrics, we utilize Pearson correlation coefficients to capture linear relationships and Spearman correlation coefficients to measure the degree of similarity between rankings. For each dataset, we randomly sample a subset of 500 training examples, then measure the similarity of the Shapley values calculated by different methods on 50 random data points from the subset. The scale is limited here because the reference algorithm MC-Shapley is extremely slow. All experiments are repeated for 3 trials.

Table 1. Correlations with MC-Shapley when the dataset size is 500. The higher, the better.

Dataset	FreeShap		G-shapley	
	Pearson	Spearman	Pearson	Spearman
Single Sentence Task				
SST-2	**0.70±0.08**	**0.60±0.03**	0.44±0.08	0.48±0.14
MR	**0.57±0.15**	**0.53±0.04**	0.32±0.22	0.32±0.23
Sentence Pair Task				
MRPC	**0.84±0.04**	**0.73±0.09**	0.68±0.08	0.58±0.11
RTE	**0.78±0.04**	**0.72±0.02**	0.28±0.04	0.36±0.05

Results. Tab. 1 demonstrates that our approach is consistently more effective in approximating the MC-Shapley compared to G-Shapley on all four NLP datasets. In particular, FreeShap has relatively high positive linear correlations and ranking correlations, which are also positively correlated with each other. The competitive correlations of FreeShap persist when the subset size grows to 1000 (App. E.1).

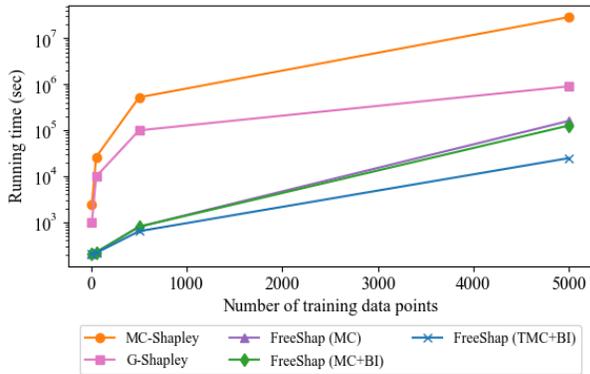FreeShap is also more computationally efficient than other

Figure 3. Running time comparison. The time for 5k points for G-Shapley and 500/5k points for MC-Shapley are projected.

methods, as shown in Fig. 3 which estimates the running time on the SST-2 dataset. Due to the better efficiency and more reliable performance, we use FreeShap for computing the Shapley value in the following sections.

4.2. The Shapley Value is More Robust than LOO

Setup. To empirically compare the β-robustness of the Shapley value and the LOO, we evaluate the consistency of the signs for a data point's instance score when the point lies in different datasets drawn from the same distribution. Due to the necessity of sampling numerous datasets when determining if a data point is *consistently harmful* or *helpful*, it is impossible to calculate the exact β. To empirically estimate the β-robustness, we design the experiments as follows: For a data point, we place it into 5 sampled training sets. This process yields 5 corresponding instance scores for the data point. We use the majority sign of these scores to determine whether the data point is *consistently harmful* or *helpful*. If any score's sign contradicts the majority, this is noted as a non-robust occurrence. We evaluate 50 training examples with the same 5 complementary dataset samples. We use the percentage of non-robust occurrence out of the 50 training examples as the empirical metric for β-robustness. An illustrative figure of the setup is in Fig. 7 in the Appendix. We approximate the Shapley value with FreeShap and calculate the exact LOO score. The experiment is performed comprehensively on datasets of various sizes.

Results. Tab. 2 shows that for each size of the training set, the Shapley value generally has fewer non-robust occurrences, hence its instance scores are generally more robust than the one from LOO. Results on more dataset sizes are in App. E.2. We subsequently analyze a few exemplary examples where the Shapley value succeeds but LOO fails in computing consistent instance scores in Tab. 3. The training example "it's not going to be everyone's bag of popcorn, but it gives you something to chew on" reflects

a positive sentiment, but it lacks clear polarity. Including such challenging training points can adversely impact test performance. It might be better to categorize the example as low-quality for further curation. Nonetheless, LOO-based instance attribution might sometimes identify this sample as helpful. Consequently, users might incorrectly conclude that this sample is beneficial for the task and retain it in the dataset, which impairs the performance. Furthermore, the explanations could influence future data collection, resulting in the accumulation of training instances with ambiguous polarity. Overall, the Shapley value yields instance scores that are much more sign-consistent and better aligned with human intuition: negative scores indicate ambiguity, while non-negative scores are associated with better clarity and polarity. These results are further demonstrated in App E.2, where we show more case studies of different dataset sizes.

Table 2. The empirical comparison of the β-robustness between the Shapley value and LOO by the percentage of non-robust occurrences. The lower, the better.

| dataset | $|D_N| = 500$ | | $|D_N| = 1000$ | | $|D_N| = 2000$ | |
	LOO	Shapley	LOO	Shapley	LOO	Shapley
SST-2	86%	**30%**	38%	50%	90%	**46%**
MR	98%	**18%**	100%	**28%**	100%	**44%**
MRPC	96%	**28%**	92%	**36%**	90%	**48%**
RTE	86%	**56%**	98%	**54%**	78%	**66%**

5. Applications of Instance Attribution

In this section, we compare the quality of instance scores calculated using various baselines on the 4 aforementioned NLP datasets in Sec. 4. We adopt qualitative and quantitative evaluation criteria including *test prediction explanation*, *data removal*, *data selection*, and *wrong label detection*. The other baselines include the influence function (Koh & Liang, 2017), TracIn (Garima et al., 2020), and representer point (Yeh et al., 2018). For the influence function, we use a more accurate approximation with conjugate gradients (Martens et al., 2010), which is perceived as the exact influence function with enough iterations and often performs better than the most used approximation, LiSSA, as shown in App. G.4. Due to the improved scalability, we adopt the modern practical prompt-based fine-tuning regime with sizable pre-trained models including BERT and RoBERTa (Liu et al., 2019). More experimental set details are in App. D.2.

Test Prediction Explanation. To explain a test prediction, we analyze the helpful and harmful training examples categorized by the sign of the instance scores. Tab. 4 shows the *most* helpful/harmful examples (with the highest/lowest scores) of the instance attribution approaches. FreeShap's explanations align more closely with the test example syntactically, discussing related topics about films, while the explanations provided by other approaches are less related.

Table 3. Case study: We demonstrate the examples where Shapley value succeeds but LOO fails in computing sign-consistent instance scores. [pos] means positive sentiment class, while [neg] means negative sentiment class. We report the mean and std of the scores scaled by a factor of $1e-3$ for readability. The dataset size n=2000.

Text example	Label	Shap	Shapley value	LOO	LOO value
confident filmmaking and a pair of fascinating performances	[pos]	5+/0-	0.45 ±0.23	4+/1-	0.23±1.52
seriously, rent the Disney version	[neg]	5+/0-	1.85±0.72	3+/2-	-0.69±2.36
it's not going to be everyone's bag of popcorn, but it definitely gives you something to chew on'	[pos]	0+/5-	-1.91±0.57	3+/2-	0.23±1.34
would fit chan like a $ 99 bargain-basement special	[neg]	0+/5-	-0.36±0.20	2+/3-	0.00±1.45

Table 4. Correctly predicted test samples with positive sentiment in SST-2 interpreted by different instance attribution approaches. [pos] means positive sentiment class, while [neg] means negative sentiment class. Words annotated in blue indicate negative sentiment, while those in red suggest positive sentiment. These highlights are author-chosen for improved readability.

Test input: a quiet, pure, elliptical *film*				[pos]
Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	a funny little *film*	[pos]	a particularly slanted , *** */* *fantasy*	[neg]
Influence	goes down easy	[neg]	a guiltless *film*	[pos]
TracIn	the only young people who possibly will enjoy it	[neg]	earnhart 's *film* is more about the optimism of a group of people who are struggling to give themselves a better lot in life than the ones	[pos]
Representer	a guiltless *film*	[pos]	seagal	[neg]

Moreover, the helpful examples provided by FreeShap are semantically akin to the test example, whereas the identified harmful examples exhibit opposing semantics. Notably, by looking at more top-ranked examples, FreeShap provides consistently more meaningful explanations, as shown in Tab. 11 in App. F.1. The advantages of FreeShap are likely due to its consideration of the subset's marginal contribution. When trained with smaller subsets, the semantically and syntactically relevant examples are likely to have more impact and score higher. Additional case studies such as error explanations and attribution analyses are also in App. F.1.

Data Removal. To quantitatively assess the correlation between the instance scores and the contributions to the test performance increase, we perform data removal experiments. This involves sequentially removing the data points 10% at a time in *order* from the training set, followed by evaluating the performance of the model retrained on the remaining data. The order is either from the highest to the lowest instance scores or vice versa, and we demonstrate for both, which is more comprehensive than the setting of Pezeshkpour et al. (2021) (i.e., only removing the highest score instances). As depicted in Fig. 4, the left four columns show that when using the instance scores from FreeShap, the performance degrades comparably faster than other baselines when discarding the examples starting from the highest scores. Conversely, when discarding from the lowest scores, using FreeShap results in slower degradation and even improves the test performance in 3 out of the 4 datasets except for SST-2. As other baselines are essentially based on the LOO framework, the consideration of more comprehensive marginal contributions in FreeShap leads to

a more accurate estimation of the instance scores.

Data Selection. With the effectiveness of FreeShap, we further conduct an ablation study to examine its potential in data selection settings. Specifically, we sequentially add the data points with the highest score (computed on the test set) and evaluate the generalized performance on a held-out set. We empirically evaluate the instance attribution approaches on the MR dataset with BERT model. From Tab. 5, using FreeShap leads to faster performance improvements, validating its potential for data selection. More results on other models and removal settings are in App. F.2.

Table 5. **Data Selection:** Difference between the model performance of no training data (i.e., 0-shot) and the model performance when trained on a subset that contains data points with high instance scores. The subset size is given as a percentage of the full set. The higher the difference, the better the approach.

Subset Size	2%	4%	6%	8%	10%
FreeShap	0.1951	**0.2111**	**0.2148**	**0.2167**	**0.2223**
Influence	0.0272	0.0647	0.0393	0.0647	0.0750
TracIn	0.1370	0.1904	0.2008	0.1548	0.1970
Representer	0.1182	0.1351	0.1388	0.0619	0.1238
Random	**0.1970**	0.1951	0.2073	0.2017	0.1529

Wrong Label Detection. Datasets can be mislabeled due to crowd-sourcing errors (Frénay & Verleysen, 2013) or data tampering (Steinhardt et al., 2017), which may have detrimental effects on model training such as increased reliance on spurious correlations (Qiao & Low, 2024). We compare the effectiveness of identifying mislabeled data using different instance attribution approaches. We poison each
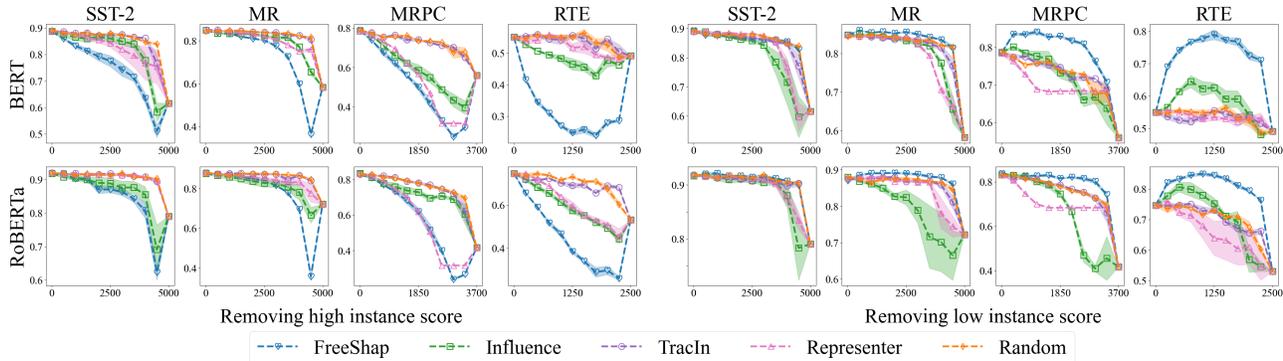
Figure 4. **Data Removal**: The test accuracy on models retrained on subsets obtained by iteratively removing 10% of the data, either from the highest or the lowest instance score. Faster degradation is preferable for high-score removals, while improvement or slower degradation is ideal for low-score removals. Overall, the scores from FreeShap are better correlated with test performance.

training set by flipping 10% of the labels, then automatically examine the poisoned data by reviewing the data points in order of their instance scores, starting from the lowest to the highest. As shown in Fig. 5, FreeShap is significantly more effective in identifying mislabeled data in 7 out of 8 experiments, while being comparable otherwise. The superiority of FreeShap in detecting mislabeled data is likely attributed to its comprehensive evaluation of the marginal contributions for the subsets as smaller data subsets are more vulnerable to the harmful effects of mislabeled data.
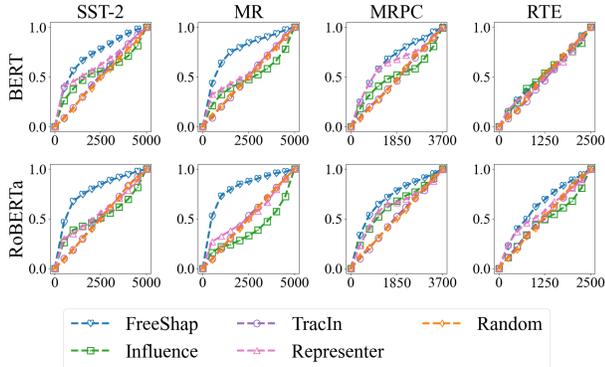

Figure 5. **Wrong Label Detection**: It shows the detected percentage of poisoned data when inspecting data from lowest to highest instance score. In most cases, FreeShap leads to the earliest identification of incorrectly labeled instances.

Extension to LLMs. In light of the recent breakthroughs in LLM (Achiam et al., 2023), we extend our empirical study to assess the efficacy of FreeShap on Llama2 (Touvron et al., 2023), which is made possible by the scalability of Free-Shap. We perform data removal and wrong label detection experiments on the prompt-based fine-tuning for Llama2 with LoRA (Hu et al., 2022). Details on experimental settings are in App. D.3. Since the influence function with conjugate gradients is not computationally feasible at the

scale of the Llama2, we use the LiSSA approximation.

We present the results for the test prediction explanation in Tab. 6. FreeShap consistently provides more insightful explanations in terms of both syntactic similarity to the test example and semantic alignment. We also present the results for data removal (of high instance score) and wrong label detection in Fig. 6. The high-score examples identified by FreeShap have more impact on model performance, especially for RTE, and the low-score examples are more accurate in pinpointing mislabeled data. We also demonstrate more case studies when FreeShap offers explanations, data removal of low instance score, time efficiency, and an extension to Llama2-13B in App. F.5.

6. Related work

We discuss related works on instance attribution and different existing notions of robustness. More discussions on Shapley value approximation and its usage in NLP explainability are in App. C.

Instance attribution in NLP. Various methods have been developed to assess the influence of training instances on model predictions (Pezeshkpour et al., 2021). The influence function (Koh & Liang, 2017) identifies the impact of a training instance by approximating the LOO score. Its explainability and time complexity are subsequently improved by Barshan et al. (2020); Grosse et al. (2023); Guo et al. (2021); Kwon et al. (2023). Han et al. (2020) applies the influence function to the BERT model and shows its faithfulness and the ability to reveal the reasoning process. TracIn (Garima et al., 2020) quantifies the effect of each training example on a test prediction based on the overall loss reduction using the first-order Taylor approximation. Its NLP variant, TracIn-WE (Yeh et al., 2022), exploits the word embedding layer for better instance scores. Representer point (Tsai et al., 2023; Yeh et al., 2018) enables the

Table 6. A correctly predicted example with positive sentiment in SST-2 interpreted by different attribution approaches on prompt-based fine-tuned **Llama2** model. [pos] means positive sentiment class and predicted correctly, while [pos(neg)] means a positively labeled training example being wrongly predicted as negative.

| Test input: ... a magnificent *drama* well worth tracking down .				[pos]
Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	a consummate *actor* incapable of being boring	[pos]	others will find their humor-seeking dollars best spent elsewhere.	[neg]
Influence	most of it given to children	[pos(neg)]	4ever has the same sledgehammer appeal as pokemon *videos* , but	[pos(neg)]
TracIn	superb	[pos]	the book 's irreverent energy , and scotches most	[neg]
Representer	you 've endured a long workout without your pulse ever racing	[pos(neg)]	self-deprecating, biting and witty feature	[neg]
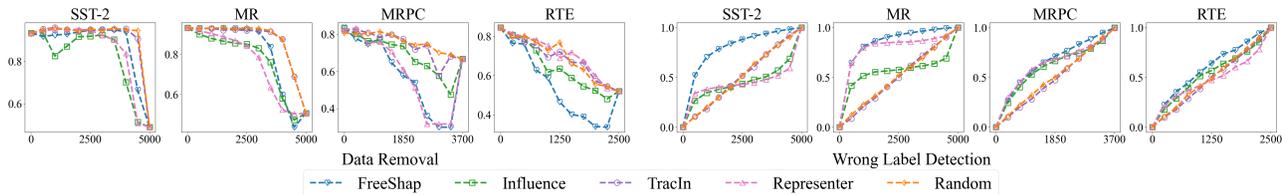
Figure 6. **Data Removal and Wrong Label Detection for Llama2**: When removing data from the highest instance scores, the data points identified by FreeShap cause relatively faster degradation (more significant on the RTE dataset) and hence are more essential for the model. For wrong label detection, FreeShap identifies the incorrectly labeled instances earlier.

model diagnosis for misclassified instances. These works all fall under the LOO scheme. In contrast, our work focuses on the Shapley value and we demonstrate its robustness and quality for data over the LOO methods. Source attribution (Wang et al., 2023), another attribution scheme that finds the sources of the contents present in a sentence generated from an LLM, is different from the instance attribution and hence is omitted from our discussion here.

Robust Explanations. Robustness determines the stability of explanations when inputs are slightly altered (Subhash et al., 2022). The explanations are expected to remain similar when inputs are insignificantly modified. Robust explanations have been shown to increase user trust (Yeh et al., 2019). Robustness to three types of perturbations are discussed: general perturbations which modify the input randomly and locally (Alvarez-Melis & Jaakkola, 2018; Bhatt et al., 2020; Tang et al., 2022); adversarial perturbations that mislead explanations (Ghorbani et al., 2019; Ivankay et al., 2022); and group-based perturbations that change the membership of sensitive features in inputs, examining fairness (Dai et al., 2021). The prior work of robustness focuses more on feature attributions and perturbation-based robustness, while we raise a complimentary notion of robustness, targeting instance attribution w.r.t. training data's distribution. The work of Lin et al. (2024) also studies the notion of distributional robustness in data valuation. However, the robustness result is w.r.t. validation distribution,

hence different from our work.

7. Conclusion

We propose a new notion of robustness for instance attribution and demonstrate the superiority of Shapley value in achieving robustness compared to other methods. To mitigate the high computational cost of the Shapley value, we introduce FreeShap and empirically demonstrate its efficiency and effectiveness in providing better instance attribution in prompt-based fine-tuning for LMs and LLMs. Our approach facilitates improving the trustworthiness of the model and can inspire future research in model interpretation, data curation, and data selection. However, our study is limited to the NLP domain with the assumption of having white-box access to the models. Additionally, we concentrate on explaining predictions oriented toward classification. Extending FreeShap to text generation tasks is non-trivial due to their autoregressive nature, which is left for future work.

Acknowledgement

Impact Statement

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

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alvarez-Melis, D. and Jaakkola, T. S. Towards robust interpretability with self-explaining neural networks. In *NeurIPS*, pp. 7786–7795, 2018.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In *NeurIPS*, volume 32, 2019.

Barshan, E., Brunet, M.-E., and Dziugaite, G. K. Relatif: Identifying explanatory training samples via relative influence. In *AISTATS*, volume 108, pp. 1899–1909, 2020.

Bentivogli, L., Clark, P., Dagan, I., and Giampiccolo, D. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1, 2009.

Bhatt, U., Weller, A., and Moura, J. M. F. Evaluating and aggregating feature-based model explanations. In *IJCAI-20*, pp. 3016–3022, 2020.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the shapley value based on sampling. *Comput. Oper. Res.*, 36(5):1726–1730, 2009.

Chen, H., Zheng, G., and Ji, Y. Generating hierarchical explanations on text classification via feature interaction detection. In *ACL*, pp. 5578–5593, 2020.

Dai, J., Upadhyay, S., Bach, S. H., and Lakkaraju, H. What will it take to generate fairness-preserving explanations? *arXiv e-prints*, pp. arXiv–2106, 2021.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.

Dolan, W. B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *IJCNLP*, 2005.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., and Tang, J. GLM: General language model pretraining with autoregressive blank infilling. In *ACL*, pp. 320–335, 2022.

Frénay, B. and Verleysen, M. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners. In *ACL-IJCNLP*, pp. 3816–3830, 2021.

Garima, Liu, F., Kale, S., and Sundararajan, M. Estimating training data influence by tracing gradient descent. In *NeurIPS*, NIPS'20, Red Hook, NY, USA, 2020. ISBN 9781713829546.

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning. In *ICML*, volume 97, pp. 2242–2251, 2019.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. *AAAI*, 33(01):3681–3688, 2019.

Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., Steiner, B., Li, D., Durmus, E., Perez, E., et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

Guo, H., Rajani, N., Hase, P., Bansal, M., and Xiong, C. FastIF: Scalable influence functions for efficient model interpretation and debugging. In *EMNLP*, pp. 10333–10350, 2021.

Han, X. and Tsvetkov, Y. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4398–4409, 2021.

Han, X., Wallace, B. C., and Tsvetkov, Y. Explaining black box predictions and unveiling data artifacts through influence functions. In *ACL*, pp. 5553–5563, 2020.

Hemachandra, A., Dai, Z., Singh, J., Ng, S.-K., and Low, B. K. H. Training-free neural active learning with initialization-robustness guarantees. In *ICML*, pp. 12931–12971. PMLR, 2023.

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 *ICLR*, 2022.

Ivankay, A., Girardi, I., Marchiori, C., and Frossard, P. Fooling explanations in text classifiers. In *ICLR*, 2022.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, volume 31. Curran Associates, Inc., 2018.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. J. Towards efficient data valuation based on the shapley value. In *AISTATS*, volume 89, pp. 1167–1176, 2019.

Jia, R., Wu, F., Sun, X., Xu, J., Dao, D., Kailkhura, B., Zhang, C., Li, B., and Song, D. Scalability vs. utility: Do we have to sacrifice one for the other in data importance quantification? In *CVPR*, pp. 8235–8243, 2021.

Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., and Iyer, R. Glister: Generalization based data subset selection for efficient and robust learning. In *AAAI*, volume 35, pp. 8110–8118, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *ICML*, ICML'17, pp. 1885–1894, 2017.

Kolpaczki, P., Bengs, V., and Hüllermeier, E. Approximating the shapley value without marginal contributions. *arXiv preprint arXiv:2302.00736*, 2023.

Kwon, Y. and Zou, J. Beta shapley: a unified and noise-reduced data valuation framework for machine learning. In *AISTATS*, volume 151, pp. 8780–8802, 2022.

Kwon, Y., Wu, E., Wu, K., and Zou, J. Datainf: Efficiently estimating data influence in lora-tuned llms and diffusion models. *arXiv preprint arXiv:2310.00902*, 2023.

Lau, G. K. R., Hemachandra, A., Ng, S.-K., and Low, B. K. H. PINNACLE: PINN adaptive collocation and experimental points selection. In *ICLR*, 2024.

Lertvittayakumjorn, P. and Toni, F. Explanation-based human debugging of NLP models: A survey. *TACL*, 9: 1508–1528, 2021.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. Visualizing and understanding neural models in NLP. In *NAACL*, pp. 681–691, 2016.

Lin, X., Xu, X., Ng, S.-K., Foo, C.-S., and Low, B. K. H. Fair yet asymptotically equal collaborative learning. In *Proc. ICML*, 2023.

Lin, X., Xu, X., Wu, Z., Ng, S.-K., and Low, K. H. B. Distributionally robust data valuation. In *Proc. ICML*, 2024.

Liu, C., Zhu, L., and Belkin, M. Toward a theory of optimization for over-parameterized systems of non-linear equations: the lessons of deep learning. *arXiv preprint arXiv:2003.00307*, 7, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *NeurIPS*, NIPS'17, pp. 4768–4777. Curran Associates Inc., 2017.

Madsen, A., Reddy, S., and Chandar, S. Post-hoc interpretability for neural nlp: A survey. *ACM Comput. Surv.*, 55(8), 2022.

Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., and Rogers, A. Bounding the estimation error of sampling-based shapley value approximation. *arXiv preprint arXiv:1306.4265*, 2013.

Malladi, S., Wettig, A., Yu, D., Chen, D., and Arora, S. A kernel-based view of language model fine-tuning. In *ICML*, volume 202, pp. 23610–23641, 2023.

Martens, J. et al. Deep learning via hessian-free optimization. In *ICML*, volume 27, pp. 735–742, 2010.

Mosca, E., Szigeti, F., Tragianni, S., Gallagher, D., and Groh, G. SHAP-based explanation methods: A review for NLP interpretability. In *COLING*, pp. 4593–4603, 2022.

Pang, B. and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *ACL*, pp. 115–124, 2005.

Pezeshkpour, P., Jain, S., Wallace, B., and Singh, S. An empirical comparison of instance attribution methods for NLP. In *NAACL*, pp. 967–975, 2021.

Qiao, R. and Low, B. K. H. Understanding domain generalization: A noise robustness perspective. In *ICLR*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ribeiro, M., Singh, S., and Guestrin, C. "why should I trust you?": Explaining the predictions of any classifier. In *NAACL*, pp. 97–101, 2016.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019. ISSN 2522-5839.

Schick, T., Udupa, S., and Schütze, H. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *TACL*, 9:1408–1424, 2021.

Schoch, S., Mishra, R., and Ji, Y. Data selection for fine-tuning large language models using transferred shapley values. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 266–275, 2023.

Shapley, L. *A value for n-person games*, volume 2, pp. 307–317. Princeton University Press, 1953.

Sim, R. H. L., Xu, X., and Low, B. K. H. Data valuation in machine learning: "ingredients", strategies, and open challenges. In *IJCAI*, pp. 5607–5614, 2022.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, pp. 1631–1642, 2013.

Steinhardt, J., Koh, P. W. W., and Liang, P. S. Certified defenses for data poisoning attacks. *NeurIPS*, 30, 2017.

Subhash, V., Chen, Z., Havasi, M., Pan, W., and Doshi-Velez, F. What makes a good explanation?: A harmonized view of properties of explanations. In *Progress and Challenges in Building Trustworthy Embodied AI*, 2022.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *ICML*, ICML'17, pp. 3319–3328, 2017.

Tang, R., Chen, H., and Ji, Y. Identifying the source of vulnerability in explanation discrepancy: A case study in neural text classification. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 356–370, 2022.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tsai, C.-P., Yeh, C.-K., and Ravikumar, P. K. Sample based explanations via generalized representers. In *NeurIPS*, 2023.

Wang, J., Lu, X., Zhao, Z., Dai, Z., Foo, C.-S., Ng, S.-K., and Low, B. K. H. Wasa: Watermark-based source attribution for large language model-generated data. *arXiv preprint arXiv:2310.00646*, 2023.

Wang, J. T. and Jia, R. Data banzhaf: A robust data valuation framework for machine learning. In *AISTATS*, volume 206, pp. 6388–6421, 2023.

Wang, T., Zeng, Y., Jin, M., and Jia, R. A unified framework for task-driven data quality management. *arXiv preprint arXiv:2106.05484*, 2021.

Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. In *ICML*, volume 162, pp. 23549–23588, 2022.

Williams, A., Nangia, N., and Bowman, S. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, June 2018.

Wu, Z., Shu, Y., and Low, B. K. H. DAVINZ: Data valuation using deep neural networks at initialization. In *ICML*, 2022.

Wu, Z., Xu, X., Sim, R. H. L., Shu, Y., Lin, X., Agussurja, L., Dai, Z., Ng, S.-K., Foo, C.-S., Jaillet, P., Hoang, T. N., and Low, B. K. H. Data valuation in federated learning. In *Federated Learning*, pp. 281–296. Elsevier, 2024.

Xia, M., Malladi, S., Gururangan, S., Arora, S., and Chen, D. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*, 2024.

Yeh, C.-K., Kim, J., Yen, I. E.-H., and Ravikumar, P. K. Representer point selection for explaining deep neural networks. In *NeurIPS*, volume 31, 2018.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in)fidelity and sensitivity of explanations. In *NeurIPS*, volume 32, 2019.

Yeh, C.-K., Taly, A., Sundararajan, M., Liu, F., and Ravikumar, P. First is better than last for language data influence. 2022.

A. Proof for the Main Paper

A.1. Proof of Theorem 3.4

Proof. When z_i is a *consistently helpful* point: $\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\left(\mathrm{sgn}\left(g^{\mathrm{Shap}}(z_i, D_T, D_N)\right) \neq \mathrm{sgn}^*(z_i) \right)$ is equivalent to $\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\left(g^{\mathrm{Shap}}(z_i, D_T, D_N) < 0 \right)$.

$$
\begin{aligned}
&\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\left(g^{\mathrm{Shap}}(z_i, D_T, D_N) \leq \mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\left(g^{\mathrm{Shap}}(z_i, D_T, D_N)\right) - t \right) \\
&\leq \mathbb{P}\left(\left| g^{\mathrm{Shap}}(z_i, D_T, D_N) - \mathbb{E}\left(g^{\mathrm{Shap}}(z_i, D_T, D_N)\right) \right| \geq t \right) \\
&\leq \frac{\mathrm{Var}\left(g^{\mathrm{Shap}}(z_i, D_T, D_N)\right)}{t^2}
\end{aligned}
\tag{10}
$$

The second inequality of Eqn. (10) holds according to Chebyshev's inequality.

$$
\begin{aligned}
\mathrm{Var}\left(g^{\mathrm{Shap}}(z_i, D_T, D_N)\right) &= \frac{1}{n^2}\Bigg[\sum_{k=0}^{n-1} \mathrm{Var}\left(\Delta_{z_i}^{D_N}(k, D_T)\right) + \\
&\qquad \sum_{p,q \in \{0,\ldots,n-1\}, p \neq q} \mathrm{Cov}\left(\Delta_{z_i}^{D_N}(p, D_T), \Delta_{z_i}^{D_N}(q, D_T)\right) \Bigg] \\
&\leq \frac{1}{n^2}\Bigg[\sum_{k=0}^{n-1} \mathrm{Var}\left(\Delta_{z_i}^{D_N}(k, D_T)\right) + \\
&\qquad \sum_{p,q \in \{0,\ldots,n-1\}, p \neq q} \frac{\mathrm{Var}\left(\Delta_{z_i}^{D_N}(p, D_T)\right) + \mathrm{Var}\left(\Delta_{z_i}^{D_N}(q, D_T)\right)}{2} \Bigg] \\
&= \frac{1}{n^2}\Bigg[\sum_{k=0}^{n-1} \frac{\mathrm{Var}(\Delta_{z_i}^{D_N}(k, D_T)) + \mathrm{Var}\left(\Delta_{z_i}^{D_N}(k, D_T)\right)}{2} + \\
&\qquad \sum_{p,q \in \{0,\ldots,n-1\}, p \neq q} \frac{\mathrm{Var}(\Delta_{z_i}^{D_N}(p, D_T)) + \mathrm{Var}\left(\Delta_{z_i}^{D_N}(q, D_T)\right)}{2} \Bigg] \\
&= \frac{1}{n^2}\Bigg[\sum_{p,q \in \{0,\ldots,n-1\}} \frac{\mathrm{Var}(\Delta_{z_i}^{D_N}(p, D_T)) + \mathrm{Var}\left(\Delta_{z_i}^{D_N}(q, D_T)\right)}{2} \Bigg] \\
&= \frac{1}{n^2}\Bigg[n \sum_{p=0}^{n-1} \mathrm{Var}\left(\Delta_{z_i}^{D_N}(p, D_T)\right) \Bigg] \\
&= \frac{1}{n} \sum_{k=0}^{n-1} \delta_k
\end{aligned}
\tag{11}
$$

The first inequality of Eqn. (11) holds because:

$$
\begin{aligned}
\mathrm{Cov}(\Delta_{z_i}^{D_N}(p, D_T), \Delta_{z_i}^{D_N}(q, D_T)) &= \mathbb{E}\left[\left(\Delta_{z_i}^{D_N}(p, D_T) - \tau_p\right)\left(\Delta_{z_i}^{D_N}(q, D_T) - \tau_q\right) \right] \\
&\leq \mathbb{E}\left[\frac{\left(\Delta_{z_i}^{D_N}(p, D_T) - \tau_p\right)^2 + \left(\Delta_{z_i}^{D_N}(q, D_T) - \tau_q\right)^2}{2} \right] \\
&= \frac{\mathrm{Var}\left(\Delta_{z_i}^{D_N}(p, D_T)\right) + \mathrm{Var}\left(\Delta_{z_i}^{D_N}(q, D_T)\right)}{2}
\end{aligned}
\tag{12}
$$

Substituting t in Eqn. (10) to $\mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big)$, we have:

$$\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\Big(g^{\text{Shap}}(z_i, D_T, D_N) < 0\Big) \leq \mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\Big(g^{\text{Shap}}(z_i, D_T, D_N) \leq 0\Big)$$

$$\leq \frac{n^{-1}\sum_{k=0}^{n-1}\delta_k}{\Big(\mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big)\Big)^2}$$

$$\leq \frac{n^{-1}\sum_{k=0}^{n-1}\delta_k}{(n^{-1}\sum_{k=0}^{n-1}\tau_k)^2} \ .$$

Similarly, we have:

$$\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\Big(g^{\text{LOO}}(z_i, D_T, D_N) < 0\Big) \leq \frac{\delta_{n-1}}{\tau_{n-1}^2} \ .$$

Consider two edge cases when $\mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{LOO}}(z_i, D_T, D_N)\big) = 0$ while the expectation of Shapley is not 0 and when $\mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big) = 0$ while the expectation of LOO is 0: For the first case, the upper bound for Shapley value is still well defined, and the upper bound for LOO is infinite. For the latter case, both upper bounds for Shapley and LOO are infinite. Even if the bounds derived from Chebyshev's inequality are not well-defined, the bounds still apply to both Shapley and LOO. The subsequent bound comparison is valid as well.

When z_i is an *consistently harmful* data point: $\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\Big(\text{sgn}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big) \neq \text{sgn}^*(z_i)\Big)$ is equivalent to

$\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\Big(g^{\text{Shap}}(z_i, D_T, D_N) \geq 0\Big)$

$$\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\Big(g^{\text{Shap}}(z_i, D_T, D_N) \geq \mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big) + t\Big)$$
$$\leq \mathbb{P}\big(\big|g^{\text{Shap}}(z_i, D_T, D_N) - \mathbb{E}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big)\big| \geq t\big) \tag{13}$$
$$\leq \frac{\text{Var}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big)}{t^2}.$$

The second inequality of Eqn. 13 holds according to Chebyshev's inequality. Substituting t in Eqn. (10) to $-\mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big)$, we have:

$$\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\Big(g^{\text{Shap}}(z_i, D_T, D_N) \geq 0\Big) \leq \frac{n^{-1}\sum_{k=0}^{n-1}\delta_k}{\Big(-\mathbb{E}_{D_N \sim \mathcal{P}^{n-1}|z_i}\big(g^{\text{Shap}}(z_i, D_T, D_N)\big)\Big)^2}$$

$$\leq \frac{n^{-1}\sum_{k=0}^{n-1}\delta_k}{(n^{-1}\sum_{k=0}^{n-1}\tau_k)^2} \ . \tag{14}$$

Similarly, we have:

$$\mathbb{P}_{D_N \sim \mathcal{P}^{n-1}|z_i}\Big(g^{\text{LOO}}(z_i, D_T, D_N) \geq 0\Big) \leq \frac{\delta_{n-1}}{\tau_{n-1}^2} \ . \tag{15}$$

$\square$

B. Improving the Efficiency of FreeShap

We adopt two tricks when using the eNTK to emulate fine-tuning: one for efficient kernel approximation and the other for efficient kernel regression.

B.1. Efficient Kernel Approximation

For eNTK matrix $K(X_N, X_N)$ where $N := \{1, \ldots, n\}$, the matrix belongs to $\mathbb{R}^{nC \times nC}$. The complete eNTK can be approximated by $I_C \otimes K_c$, where $K_c \in \mathbb{R}^{n \times n}$ is a kernel matrix concerning the output logit of class c, and I_C is the $C \times C$

identity matrix. This approximation allows kernel regression for $K(X_N, X_N)$ to be split into C separate kernel regression tasks, each concerning K_c. If the model has randomly initialized output layers, Wei et al. (2022) further approximates the kernel matrix by $I_C \otimes K_0$ where K_0 is the eNTK concerning a random output logit. As we adopt prompt-based fine-tuning, our output layer is based on the embedding layer which is not randomly initialized. Thus, we experiment with computing only one K_0 (which is more time efficient) as well as computing different K_c. Following the setting in Sec 4.1 using SST-2 dataset, we compare the correlation for one seed with MC-Shapley when using each class's K_c for kernel regression and using K_0 (the eNTK concerning a random output logit). It shows a similar correlation, the former is 0.69 while the latter is 0.71. Considering the similar approximation performance and computing only one K_0 shows an even better approximation, we use the same K_0 which is the eNTK concerning one of the class's outputs to conduct C separate kernel regression tasks in all our experiments for computational efficiency.

B.2. Efficient Kernel Regression

When generalizing to the dataset with a larger size, the computation of the kernel regression can be more computationally expensive as it requires the inversion of the matrix $K(X_S, X_S)$, as shown in Eqn. (5). Instead of inverting the full matrix from scratch, we use a blockwise inversion (BI) approach: we make use of the last step's inverse matrix and reduce the average time complexity of kernel regression from $O(n^3)$ to $O(n^2)$. Specifically, when computing $K(X_S, X_S)^{-1}$, we partition this inverse into four blocks:

$$K(X_S, X_S)^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix} \quad (16)$$

Technically, BI is a dynamic programming approach that efficiently computes the exact inverse. Nonetheless, to check the possible numerical instability in practice, we still empirically compare the Shapley value of the SST-2 dataset computed using the BI and without the BI (the exact inverse). We compute the Shapley value for 5k SST-2 points. For each test point, there is a (5000,1) vector, denoting the instance score for each training point. For each test point, there is a correlation value between computing using the BI and without the BI. We average all test points' correlation values and get a 0.9984 Pearson correlation and 0.9963 Spearman correlation. The decent correlation demonstrates that kernel regression using BI offers an efficient yet accurate computation of the inverse. This justifies our later usage of BI when the dataset is relatively larger in Sec. 5.

C. More Related Works

Approximating the Shapley Value. Numerous works have focused on improving the scalability of the Shapley value by reducing the number of computations of the subset's utility, such as using Monte-Carlo sampling (Ghorbani & Zou, 2019), adopting group testing for fewer subsets utility computation (Jia et al., 2019), or updating multiple estimates simultaneously with one time of utility computation (Kolpaczki et al., 2023). On the other hand, some works aim to reduce the computation of the utility function as ours. Gradient Shapley (Ghorbani & Zou, 2019) approximates the full training with just one epoch, while All-S Influence Shapley and Largest-S Influence Shapley (Jia et al., 2019) leverage the influence function to approximate the marginal contribution. The influence-function-based approximations require repeated Hessian calculations or consider only a single subset, losing some of Shapley value's essential properties. Consequently, they are not selected as baselines in our analysis. We show that our work provides a better approximation to the original Shapley values with comparable efficiency. The work of Wu et al. (2022) proposes a domain-aware approximation of generalization error of neural network model based on NTK and distributional distance. However, it is unclear how it can be used to approximate validation accuracy in the setting of prompt-based fine-tuning. Other works extract fixed representations from the last hidden state and then conduct KNN (Jia et al., 2021) or last-layer fine-tuning (Schoch et al., 2023). We did not make further comparisons with the previous two works as they focus solely on using representations and fine-tuning the last hidden layer, which Yeh et al. (2022) suggest can lead to suboptimal results for language models. In contrast, we aim to resemble a more comprehensive fine-tuning phase. FreeShap uses validation accuracy as the utility function to define Shapley value and more discussion on other utility functions can be found in Wu et al. (2024).

Shapley Value in NLP Explainability. Shapley value has been employed for feature-attribution explanations (Lundberg & Lee, 2017; Mosca et al., 2022), treating each word as a player in a game, with the model's performance as the outcome. In terms of NLP, this approach has been adapted for text classification, identifying keywords or phrases influencing classification decisions. Various strategies for efficient Shapley value approximation have been developed, such as KernelSHAP, LinearSHAP, and DeepSHAP (Lundberg & Lee, 2017), targeting different model types. However,

these methods oversimplify text data by not accounting for word interactions and context. HEDGE (Chen et al., 2020) offers hierarchical explanations, considering phrases and multiple relevance levels. In contrast to these feature-attribution applications, recent works have started using Shapley value for instance scores in NLP (Schoch et al., 2023), focusing on selecting training points with high Shapley values. Our work differs from these in both purpose and technique, aiming at robust instance attribution and utilizing eNTK to resemble fine-tuning.

D. Detailed Experiment Setup

D.1. Detailed Experiment Setup for Sec. 4

In terms of the model, we adopt the BERT (Devlin et al., 2019), comprised of 12 layers of encoders. To address time complexity, we freeze the first eight layers of encoders and fine-tune the remaining layers. We adopt Adam optimizer during the training process. We disable the dropout in the BERT model to reduce the other uncertainty. We also disable the data shuffling when feeding data to the model. In terms of implementation, we adopt the huggingface transformers package to implement BERT's fine-tuning. In terms of hyperparameters, we use a learning rate of $1e-5$ (G-Shapley has a $2e-5$ learning rate), batch size 16, and 10 epochs. The max sequence length set for SST-2 dataset and MR dataset is 64, for MRPC is 128, and for RTE is 256. The prompts for each dataset are listed in Tab. 7, which is the same as Malladi et al. (2023)'s. We adopt 200 Monte-Carlo iterations for experiments in this section. In terms of eNTK, we adopt the standard neural tangent kernel described in the main paper (Sec. 2.3), different from the suggested signGD kernel (Malladi et al., 2023). The analysis for the signGD kernel applies only to "early-stage" training with Adam, and previous empirical work is limited to few-shot classification. We experiment with both kernels: We compare the correlation between MC-Shapley and FreeShap when using the standard neural tangent kernel or signGD kernel and for seed 2023 on four datasets, and find that the standard neural tangent kernel overall has a relatively higher correlation compared with the signGD kernel. Therefore, we proceed with the standard neural tangent kernel. In terms of hardware, fine-tuning and eNTK precomputation are conducted on L40 GPUs. Notably, kernel regression can be efficiently computed with just CPU.
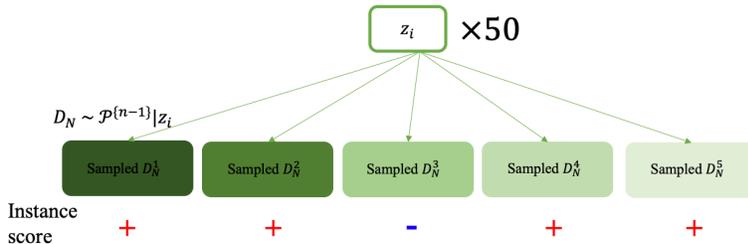

Figure 7. Experiment setup for robustness comparison.

Table 7. Prompts used in experiments. [x] in the single sentence task means the training input, while [x0] and [x1] are two sentences in the sentence pair task. The prompts for SST-2 and MR are the same for encoder-only or decoder-only models.

Dataset	Type	Prompts	Task-specific term
encoder-only model			
SST-2	sentiment	[x] It was mask.	terrible, great
MR	sentiment	[x] It was mask.	terrible, great
MRPC	paraphrase	[x0] mask, [x1]	No, Yes
RTE	NLI	[x0]? mask, [x1]	No, Yes
decoder-only model			
MRPC	paraphrase	[x0] Question: [x1] Yes or No? mask	No, Yes
RTE	NLI	[x0]? [x1] Entailment or not? mask	No, Yes

D.2. Detailed Experiment Setup for Sec. 5

In terms of the model, we adopt BERT and RoBERTa (Liu et al., 2019). The practical implementation and hyperparameters for both models are the same as the setup described in App. D.1. One difference is that we adopt TMC sampling (Ghorbani & Zou, 2019) here instead of MC sampling for FreeShap, which has a performance tolerance of 0.05 and 200 Monte-Carlo iterations. Another difference is that we adopt BI as described in App. B for the experiments in this section as the size of the dataset is larger.

In terms of other instance attributions: for the influence function, the damping factor is 0.003, while the maximum iterations are 1000; for TracIn, the checkpoint step is 3; for representer point, the λ is $1e-4$. Influence function and TracIn are computed on unfrozen parameters.

D.3. Detailed Experiment Setup for Llama2

In terms of the model, we adopt Llama2-7B, comprised of 32 layers of decoders. We do not freeze the decoders as we adopt LoRA for efficient fine-tuning. In the LoRA experiment setup, the rank is 16 while the alpha value is also 16. Targeted modules for LoRA adjustments include query, key, value, output, gate, down, and up projections in the transformer's attention mechanism. We adopt no bias and dropout. Additionally, the final predictive layer is fine-tuned. In terms of hyperparameters, we use a learning rate of $1e-5$ or $1e-6$ and 5 epochs. The batch size for the SST-2 dataset and MR dataset is 4, and for MRPC and RTE is 2. For Llama2, a decoder-only model, slightly different prompts are adopted. The prompts employed for Llama2 are similar to those used in transformer structure models in Du et al. (2022). The prompts for each dataset for Llama2 are also listed in Tab. 7. The SST-2 dataset and MR dataset share similar prompts as the previous setting so it is not repeated showing. The setup for FreeShap is the same as the setup described in App. D.2. In terms of other instance attributions: for the influence function, the damping factor is 0.003, repeat time is 1, depth is 2500 and the scaling factor is $1e4$; for TracIn, the checkpoint step is 1; for representer point, the λ is $1e-4$. Influence function and TracIn are computed on unfrozen parameters.

E. More Experimental Results

E.1. FreeShap Approximates the Shapley Value Well

Tab. 8 compliments the results of Tab. 1, further demonstrating that our approach is more effective in approximating MC-Shapley under different sizes of the dataset. Additionally, comparing Tab. 1 and Tab. 8, we also observe that when the dataset size is 500, both FreeShap and G-Shapley have a generally better approximation. Shapley values in size of 500 points and 1000 points share similar variance as they use the same number of Monte-Carlo iterations; while Shapley values in size of 500 points have a larger magnitude, and hence they may fluctuate less and better at maintaining the correlation.

Table 8. Correlation between MC-Shapley when the dataset size is 1k. Each correlation is computed using three trials. The higher the better and FreeShap consistently has a higher correlation.

Dataset	FreeShap		G-Shapley	
	pearson	spearman	pearson	spearman
Single Sentence Task				
SST-2	**0.65±0.06**	**0.50±0.01**	0.35±0.08	0.39±0.16
MR	**0.55±0.08**	**0.41±0.06**	0.17±0.07	0.17±0.04
Sentence Pair Task				
MRPC	**0.79±0.05**	**0.75±0.03**	0.51±0.06	0.42±0.02
RTE	**0.61±0.10**	**0.59±0.13**	0.16±0.02	0.24±0.07

E.2. The Shapley Value Is More Robust than LOO

Table 9 is a more comprehensive version of Tab. 2, demonstrating the Shapley value is more robust in instance scores compared to LOO. Across various training set sizes, the Shapley value generally exhibits fewer non-robust occurrences. Note that for the RTE dataset, the maximum size of the training data is 2,490. Therefore, the column $|D_N| = 3000$ represents the scenario where $|D_N|$ is at its maximum.

Table 10 presents more exemplary examples across various dataset sizes, where the Shapley value consistently has sign-consistent scores but LOO fails. This mirrors the main paper's findings: the Shapley value produces sign-consistent scores for training instances that resonate with human intuition, where negative scores signal confusing examples, and non-negative scores indicate more straightforward and stronger polarized examples. In the main paper, we have discussed an example with a negative instance score, while here we use "looks great, has solid acting and a neat premise" when dataset size is 3000 as a case study. This example, with its clear positive sentiment, demonstrates that including clear-polarity training points can enhance test performance. The Shapley value consistently assigns a positive score to this example. However, LOO may incorrectly label this example as harmful, leading to potential misunderstandings about its influence or underrepresentation in future data collection and curation.

Table 9. A complete version of the empirical comparison of the β-robustness between the Shapley value and LOO by the percentage of non-robust occurrences when the complementary dataset is resampled. The lower, the better.

size dataset	$|D_N| = 300$ LOO	Shapley	$|D_N| = 500$ LOO	Shapley	$|D_N| = 1000$ LOO	Shapley	$|D_N| = 2000$ LOO	Shapley	$|D_N| = 3000$ LOO	Shapley
SST-2	**44%**	56%	86%	**30%**	**38%**	50%	90%	**46%**	100%	**52%**
MR	44%	**24%**	98%	**18%**	100%	**28%**	100%	**44%**	98%	**36%**
MRPC	100%	**26%**	96%	**28%**	92%	**36%**	90%	**48%**	98%	**4%**
RTE	98%	**58%**	86%	**56%**	98%	**54%**	78%	**66%**	100%	**48%**

Table 10. Case study: we demonstrate the examples with consistent instance scores computed by Shapley value and sign-inconsistent instance scores computed by LOO. [pos] means positive sentiment class, while [neg] means negative sentiment class. We report the mean and std of the scores scaled by a factor of $1e-3$ for readability. The dataset size n=1000, 2000, 3000.

Text example	Label	Shap	Shapley value	LOO	LOO value
1000 points					
prefer to keep on watching	[pos]	0+/5-	-1.01±0.26	4+/1-	0.00±1.26
it seems impossible that an epic four-hour indian musical about a cricket game could be this good	[pos]	0+/5-	-1.09±0.29	3+/2-	0.69±2.96
2000 points					
confident filmmaking and a pair of fascinating performances	[pos]	5+/0-	0.45 ±0.24	4+/1-	0.23±1.52
seriously, rent the Disney version	[neg]	5+/0-	1.85±0.72	3+/2-	-0.69±2.36
it's not going to be everyone's bag of popcorn, but it definitely gives you something to chew on'	[pos]	0+/5-	-1.91±0.57	3+/2-	0.23±1.34
would fit chan like a $ 99 bargain-basement special	[neg]	0+/5-	-0.36±0.20	2+/3-	0.00±1.45
3000 points					
looks great, has solid acting and a neat premise.	[pos]	5+/0-	0.58 ±0.40	3+/2-	0.46±1.72
seriously, rent the Disney version	[neg]	5+/0-	1.17±0.60	2+/3-	-0.69±3.04
it's not going to be everyone's bag of popcorn, but it definitely gives you something to chew on '	[pos]	0+/5-	-1.27±0.68	2+/3-	0.00±2.71
they shovel into their mental gullets to simulate sustenance	[neg]	0+/5-	-0.28±0.14	3+/2-	0.23±2.22

F. More Results for Applications of Instance Attributions

F.1. More Test Prediction Explanations

For the tables in this section: [pos] means positive sentiment class, while [neg] means negative sentiment class. Each training example shown here is predicted correctly so we do not include its prediction result. Words annotated in blue indicate negative sentiment, while those in red suggest positive sentiment. These highlights are author-chosen for improved readability. The dataset used in this section is SST-2.

We provide a more complete example for the **most** helpful/harmful examples (with the highest/lowest scores) of the correct prediction in Tab. 11. The Shapley value yields consistently more insightful explanations for the top three most significant training examples, both in terms of syntactic similarity to the test example and semantic alignment. Specifically, helpful examples share identical semantics with the test example, whereas harmful examples demonstrate contrasting semantics. Other instance attribution methods might offer a film-related example in their explanations but fail to provide consistently meaningful explanations.

We further list the explanations for an incorrect prediction in Tab. 12. The helpful examples provided by the Shapley value are syntactically similar to the test example in that they contain a mix of words with positive and negative connotations. This mixture could potentially confuse a deep learning model. The representer point also provides a meaningful explanation, "peril", a negative sentiment yet labeled under the positive class. This aligns with the findings of Yeh et al. (2018) that representer points can offer insights into the neural network's tendency to predict the positive sentiment for this test example. This tendency stems from its reliance on examples that, despite conveying a negative sentiment, are labeled within the positive class.

Analyzing training data with high instance scores relative to the test set, as shown in Table 13, guides new data collection by targeting similar profiles. Specifically, the Shapley value indicates that adding examples with longer sequences and clear polarity could significantly enhance model performance. Conversely, the model struggles with sarcasm, as evident from the most harmful examples. The influence function effectively identifies examples with incorrect labels. Interestingly, the most supportive example per the influence function aligns with the representer point but lacks clear positive polarity, with the term "odd" suggesting uniqueness rather than positivity. This example is assigned a negative score by Shapley, suggesting it may hinder performance in smaller subsets. Additionally, TracIn may be biased towards the same training examples when they have large gradients.

Table 12. A wrongly predicted test example with negative sentiment in SST-2 interpreted by different instance attribution approaches.

Test input: unflinchingly bleak and desperate				[neg]
Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	self-defeatingly decorous	[neg]	excessively quirky	[pos]
	oddly moving	[pos]	it tight and nasty	[pos]
	vicious and absurd	[neg]	damning and damned compelling	[pos]
Influence	is completely at sea	[neg]	of businesses	[pos]
	more slowly	[neg]	retread , with the emphasis on self-empowering schmaltz and big-wave surfing that gives pic its title an afterthought	[neg]
	a guiltless film	[pos]	are more interesting ways of dealing with the subject	[neg]
TracIn	by surrounding us with hyper-artificiality	[neg]	effective film	[pos]
	standard thriller and drag audience enthusiasm to crush depth	[neg]	are crisp and purposeful without overdoing it	[pos]
	i did n't care.	[neg]	lend some dignity to a dumb story	[neg]
Representer	a guiltless film	[pos]	seagal	[neg]
	their own mortality	[pos]	more slowly	[neg]
	peril	[pos]	puzzling	[neg]

Table 11. A correctly predicted example with positive sentiment in SST-2 interpreted by different instance attribution approaches.

	Test input: a quiet, pure, elliptical *film*			[pos]
Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	a funny little *film*	[pos]	a particularly slanted , *** */* *fantasy*	[neg]
	pretty good little *movie*	[pos]	overly-familiar and poorly-constructed *comedy*	[neg]
	a well-put-together *piece*	[pos]	a shaky, uncertain *film*	[neg]
Influence	goes down easy	[neg]	a guiltless *film*	[pos]
	is like nothing we westerners have seen before	[pos]	money-oriented	[neg]
	retread, with the emphasis on self-empowering schmaltz and big-wave surfing that gives *pic* its title an afterthought	[neg]	is not as well-conceived as either of those *films*	[neg]
TracIn	the only young people who possibly will enjoy it	[neg]	earnhart 's *film* is more about the optimism of a group of people who are struggling to give themselves a better lot in life than the ones	[pos]
	're down	[neg]	ditched the saccharine sentimentality of bicentennial man	[pos]
	authentic to the core of his being	[pos]	highlander	[pos]
Representer	a guiltless *film*	[pos]	seagal	[neg]
	it is different from others in its genre in that it is does not rely on dumb gags, anatomical humor, or *character* cliches; it primarily relies on *character* to tell its story	[pos]	macaroni and cheese	[neg]
	between flaccid satire and what	[pos]	grasping actors' workshop that it is	[neg]

Table 13. The most helpful/harmful examples for SST-2 test datasets prediction interpreted by different attribution methods.

Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	a dramatic comedy as pleasantly dishonest and pat as any hollywood fluff.	[neg]	the paranoid claustrophobia of a submarine movie with the unsettling spookiness of the supernatural – why did n't hollywood think of this sooner ?	[pos]
	it 's anchored by splendid performances from an honored screen veteran and a sparkling newcomer who instantly transform themselves into a believable mother/daughter pair .	[pos]	kids who are into this thornberry stuff will probably be in wedgie heaven	[pos]
Influence	is about something very interesting and odd that	[pos]	it tight and nasty	[pos]
	exercise in formula crash-and-bash action	[neg]	its characterization of hitler and the contrived nature of its provocative conclusion	[pos]
TracIn	the only young people who possibly will enjoy it	[neg]	earnhart 's film is more about the optimism of a group of people who are struggling to give themselves a better lot in life than the ones	[pos]
	're down	[neg]	ditched the saccharine sentimentality of bicentennial man	[pos]
Representer	is about something very interesting and odd that	[pos]	4ever has the same sledgehammer appeal as pokemon videos , but '	[pos]
	is the one	[pos]	a guiltless film	[pos]

We further demonstrate the explanations for natural language inference tasks in Tab. 14. [yes] means entailment, while [no] means not entailment. Each training example shown here is predicted correctly. The dataset used in this section is RTE, while the model used here is RoBERTa. Similar to the previous examples, the Shapley value yields consistently more insightful explanations since among the top three most significant training examples, there exists one that has almost the same semantics as the test example.

Table 14. A correctly predicted example with an entailment pair in RTE interpreted by different instance attribution approaches.

	Test input:	
	sentence1: As spacecraft commander for Apollo XI, the first manned lunar landing mission, Armstrong was the first man to walk on the Moon. "That's one small step for a man, one giant leap for mankind." With these historic words, man's dream of the ages was fulfilled. sentence2: Neil Armstrong was the first man who landed on the Moon.	[yes]

Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	sentence1: In 1867, Nobel obtained a patent on a special type of nitroglycerine, which he called "dynamite". The invention quickly proved its usefulness in building and construction in many countries. sentence2: Alfred Nobel is the inventor of dynamite.	[yes]	sentence1: Gabriel Garcia Marquez was a liberal thinker whose left-wing politics angered many conservative politicians and heads of state. His job as a reporter for the Cuban news agency Prensa Latina, in 1960, and and friendship with Fidel Castro resulted in his being ultimately denied entry to the United States for political reasons sentence2: Gabriel Garcia Marquez was a conservative politician.	[no]
	sentence1: Neil Armstrong was an aviator in the Navy and was chosen with the second group of astronauts in 1962. Made seven flights in the X-15 program 1960 photo), reaching an altitude of 207,500 feet. Was backup command pilot for Gemini 5, command pilot for Gemini 8, backup command pilotfor Gemini 11, backup commander for Apollo 8, and commander for Apollo 11: successfully completing the first moonwalk. sentence2: Neil Armstrong was the first man who landed on the Moon.	[yes]	sentence1: Evan Wolfson, the founder of the modern gay marriage movement, tells the waiter he would like an iced decaf and "the usual." Wolfson, one of Time Magazineś Most Influential People in the World, is a man who unflinchingly knows what he wants and stays his course, whether it be in his choice of restaurant or in his choice of battle. And others always know when they see Evan coming what it is that he wants. sentence2: Time Magazine was created by Evan Wolfson.	[no]
	sentence1: Located just three miles from Tullamore and only 45 minutes from the K Club, venue of the 2006 Ryder Cup, is Esker Hills, a genuine hidden gem and one of Irish golf's best kept secrets. sentence2: The K Club will host the 2006 Ryder Cup.	[yes]	sentence1: Hugh Jackman has done it all - almost. He's hosted the Oscars, danced on Broadway, flashed his winning grin and flexed his acting muscles, but Friday is the first time the Australian actor will claw his way into movie theaters as the main attraction. In "X-Men Origins: Wolverine," there's no Halle Berry or Nicole Kidman by his side to buoy the box office. This time itś just Jackman, his claws and a heap of Hollywood expectations to the tune of $150 million. sentence2: Hugh Jackman will play the magician Harry Houdini.	[no]
Influence	sentence1: Rather than deterring crime, capital punishment actually increases the level of brutality in society. sentence2: Capital punishment is a deterrent to crime.	[no]	sentence1: A male rabbit is called a buck and a female rabbit is called a doe, just like deer. sentence2: A female rabbit is called a doe.	[yes]
	sentence1: On Feb. 1, 1945, the Polish government made Warsaw its capital, and an office for urban reconstruction was set up. sentence2: Warsaw remained Poland's capital after the war.	[no]	sentence1: Bountiful arrived after war's end, sailing into San Francisco Bay 21 August 1945. Bountiful was then assigned as hospital ship at Yokosuka, Japan, departing San Francisco 1 November 1945. sentence2: Bountiful reached San Francisco on 1 November 1945.	[no]
	sentence1: Vera Beers, of Olivet, recently was named employee of the month at Standard Printing in Marshall. sentence2: Vera Beers works for Standard Printing.	[yes]	sentence1: 'In other words, with its 2 million inhabitants, Slovenia has only 5.5 thousand professional soldiers. sentence2: Slovenia has 5.5 million inhabitants.	[no]
TracIn	sentence1: Martin was taken to New Orleans Ochsner Foundation Hospital, where nurse Jinny Resor said he was treated for dehydration. sentence2: Martin works for the Ochsner Foundation Hospital.	[no]	sentence1: Prime Minister Mahmoud Abbas has offered 'the hand of peace' to Israel after his landslide victory in Sunday's presidential election. sentence2: Mahmoud Abbas has claimed victory in the presidential elections.	[yes]
	sentence1: In other words, with its 2 million inhabitants, Slovenia has only 5.5 thousand professional soldiers. sentence2: Slovenia has 5.5 million inhabitants.	[no]	sentence1: Canada is officially a bilingual country but, with nearly 60% of the population speaking English as their mother-tongue, and only 24% speaking French as their first language, some people are questioning whether Canada is truly a bilingual nation or rather, a bilingual nation on paper only. French is not the only linguistic minority in Canada, and some of the languages spoken, in order of popularity, are Chinese, Italian, German, Polish, Spanish, Portuguese, Punjabi, Ukrainian, Arabic, Dutch, Tagalog, Greek, Vietnamese, Cree and Inuktitut. sentence2: French is the most widely spoken language in Québec.	[no]
	sentence1: With half of the vote counted, the ANC has 66% of the vote. Its nearest rival, the Democratic Alliance, has 16%, while ANC splinter-group the Congress of the People trails with 8%. The results will see ANC leader Jacob Zuma elected as President of South Africa when the National Assembly reconvenes in May.Provincial elections are also being held, and the ANC looks likely to lose power in the province of Western Cape to the Democratic Alliance. This will be the first time an opposition party has won control of a provincial parliament since the end of apartheid. The election campaign has focused on crime, poverty, and the suitability of Zuma to be President. Zuma was acquitted of rape in 2006, and corruption charges against him were withdrawn shortly before the election after prosecutors found the charges had been politically motivated. sentence2: Jacob Zuma is the leader of the ANC party.	[yes]	sentence1: Christopher Reeve, an actor and director who became an inspiration worldwide after being paralyzed in a horse riding accident, died Sunday of heart failure. sentence2: Christopher Reeve had an accident.	[yes]
Representer	sentence1: A male rabbit is called a buck and a female rabbit is called a doe, just like deer. sentence2: A female rabbit is called a doe.	[yes]	sentence1: In other words, with its 2 million inhabitants, Slovenia has only 5.5 thousand professional soldiers. sentence2: Slovenia has 5.5 million inhabitants.	[no]
	sentence1: A small bronze bust of Spencer Tracy sold for £174,000. sentence2: A small bronze bust of Spencer Tracy made £174,000.	[yes]	sentence1: Bountiful arrived after war's end, sailing into San Francisco Bay 21 August 1945. Bountiful was then assigned as hospital ship at Yokosuka, Japan, departing San Francisco 1 November 1945. sentence2: Bountiful reached San Francisco on 1 November 1945.	[no]
	sentence1: In other words, with its 2 million inhabitants, Slovenia has only 5.5 thousand professional soldiers. sentence2: Slovenia has 2 million inhabitants.	[yes]	grasping actors' workshop that it is	[no]

22

F.2. Extension to Data Selection

As elaborated in the main paper, we examine the potential of FreeShap in data selection. In this setting, we have a training set, a separate test set (validation set in the conventional sense, we refer to it as the test set to be consistent with the main paper) used for computing instance scores, and a held-out set (test set in the conventional sense) used for reporting the performance of the model trained on the selected data subset. Note that the setting of data selection is different from data removal as described in Sec. 5 in that data selection uses a separate held-out set to evaluate the performance of the selected data subset and hence is more applicable to selecting data in real-life applications. The setting of data removal is meant to validate the effectiveness of the instance attribution and hence is designed for analytical purposes. We conduct experiments on the MR dataset as its labeled held-out set is easily accessible. To evaluate FreeShap for data selection, we conduct two types of data selection tasks: data selection with addition and data selection with removal. The addition involves sequentially adding the data points with the highest score (included in the main paper), followed by evaluating the performance of the model on the held-out dataset retrained on the selected data; while the removal involves sequentially removing the data points with the lowest scores in order from the training set. Intuitively, in data selection with addition, a good data selection scheme should result in faster improvements in performance; while in data selection with removal, a good data selection scheme should result in improved performance or slower degradation of performance. We supplement the results for data addition in LLama 2 in Tab. 15. FreeShap achieves faster improvements, validating its effectiveness as a data selection scheme in both models. However, for the removal setting in Tab. 16, FreeShap is less competitive. We hypothesize that FreeShap is computed primarily based on the marginal contributions of small subsets due to TMC sampling and hence is less effective in removing data points from large datasets w.r.t. the held-out set. Enhancing FreeShap for such scenarios remains a future research.

Table 15. **Data Selection with Addition**: Difference between the model performance of *no data* (i.e., 0-shot) and the model performance when training on a *subset that contains data points with high instance scores*. The subset size is given as a percentage of the full set. The higher the difference, the better the approach.

BERT					
Subset size	2%	4%	6%	8%	10%
FreeShap	0.1951	**0.2111**	**0.2148**	**0.2167**	**0.2223**
Influence	0.0272	0.0647	0.0393	0.0647	0.0750
TracIn	0.1370	0.1904	0.2008	0.1548	0.1970
Representer	0.1182	0.1351	0.1388	0.0619	0.1238
Random	**0.1970**	0.1951	0.2073	0.2017	0.1529

Llama2					
Subset size	2%	4%	6%	8%	10%
FreeShap	**0.0816**	**0.1398**	**0.1754**	**0.2186**	**0.2458**
Influence	-0.0038	-0.0038	-0.0038	-0.0038	-0.0038
TracIn	0.0638	0.1144	0.1304	0.1773	0.2092
Representer	-0.0019	0.0000	0.0019	0.0066	0.0056
Random	0.0638	0.1041	0.1529	0.1792	0.1951

F.3. More Experiments for Multi-class on MNLI Dataset

For multi-class classification, we perform experiments on MNLI datasets (with 3 classes). Regarding data removal, as shown in Fig. 8, performance drops more rapidly when removing examples with the highest scores from FreeShap, and removing data based on the lowest scores from FreeShap even improves the performance. In wrong label detection, as shown in Fig. 9, FreeShap is highly efficient at pinpointing mislabeled data.

F.4. Scalability to Larger Datasets

To verify the scalability of FreeShap on larger datasets, we have extended our experiments on data removal and wrong label detection (described in Sec. 4) to 25k SST2 training data, demonstrating the practical advantages of FreeShap over other baselines at a larger scale. TracIn is not included due to the computational inefficiency and its weak performance in main paper experiments (Fig. 4, 5). To further improve the efficiency of experiments, the hyperparameter of tolerance for

Table 16. **Data Selection with Deletion**: Difference between the model performance of *all data* and the model performance when training on a *subset that removes data points with low instance scores*. The subset size is given as a percentage of the full set. The smaller the difference, the better the approach. If the value is positive, which means the removal of low-value data improves the performance and is even better.

	BERT				
Subset size	98%	96%	94%	92%	90%
FreeShap	-0.0066	**-0.0009**	-0.0019	-0.0150	-0.0122
Influence	-0.0038	**-0.0009**	-0.0019	-0.0056	-0.0141
TracIn	**-0.0009**	-0.0056	**-0.0009**	-0.0093	-0.0113
Representer	-0.0028	-0.0066	-0.0046	-0.0066	**-0.0066**
Random	-0.0046	-0.0094	-0.0103	**-0.0047**	-0.0188

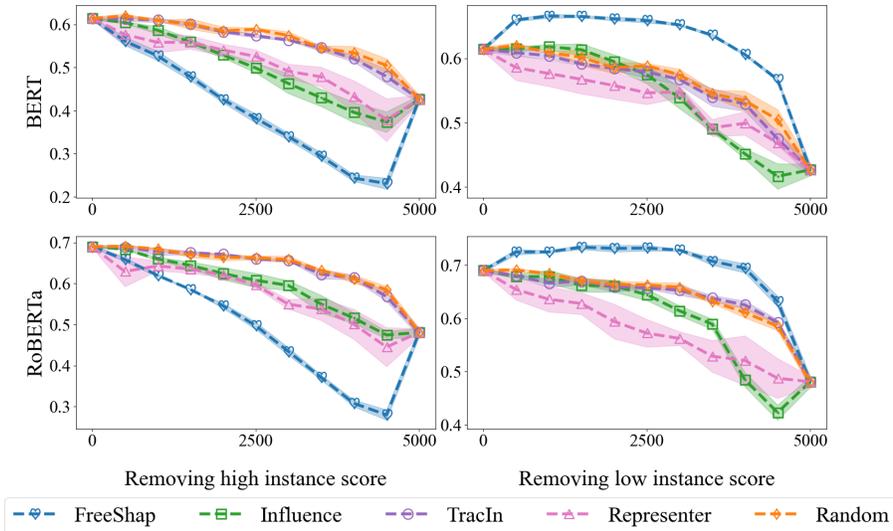|	Llama2				
Subset size	98%	96%	94%	92%	90%
FreeShap	-0.0018	-0.0009	-0.0009	-0.0018	0.0009
Influence	**0.0028**	**0.0019**	**0.0**	**0.0**	**0.0019**
TracIn	0.0	-0.0009	-0.0009	-0.0019	-0.0009
Representer	-0.0028	-0.0047	-0.0084	-0.0103	-0.0122
Random	-0.0056	-0.0103	**0.0**	-0.0009	-0.0009

Figure 8. **Data Removal**: The test accuracy on models retrained on subsets obtained by iteratively removing 10% of the data, either from the highest or the lowest instance score. Faster degradation is preferable for high-score removals, while improvement or slower degradation is ideal for low-score removals. Overall, the scores from FreeShap are better correlated with test performance.

FreeShap is further increased to 0.1 and the Monte-Carlo iterations are increased to 800. The influence function is computed with LiSSA and depth is 1000. From Fig. 10, when removing data with high instance scores, FreeShap outperforms other approaches. When removing data with low instance scores, the performance of FreeShap is comparable to other approaches. This is because the task is relatively simple as even random removal performs well. In addition, FreeShap performs well in identifying wrong labels.

F.5. More Experiments for Llama2

We further empirically show FreeShap's effectiveness as explanations, data removal, time efficiency, and a further upscaling to Llama2-13B here.
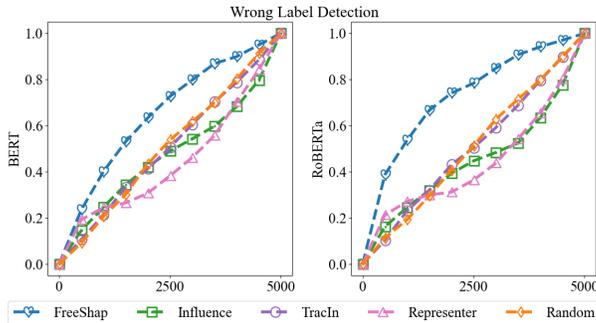
Figure 9. **Wrong Label Detection**: It shows the detected percentage of poisoned data from the subsets with the lowest instance scores. For the MNLI dataset, FreeShap leads to the earliest identification of mislabeled instances.
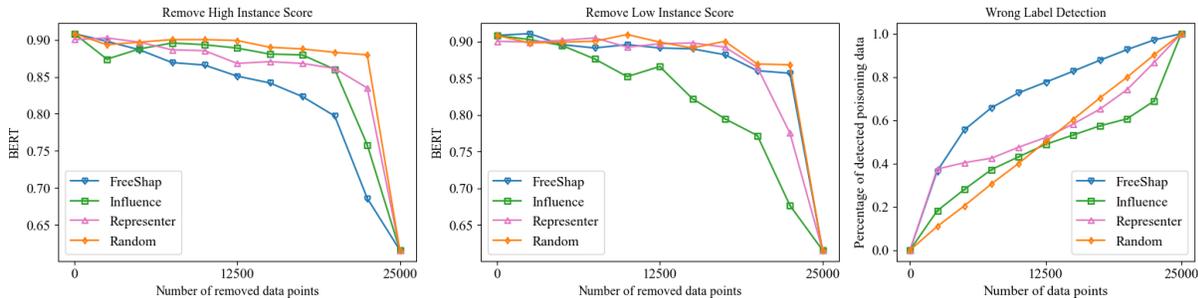

Figure 10. **Data Removal and Wrong Label Detection**: The two figures on the left show the test accuracy on models retrained on subsets obtained by iteratively removing 10% of the data. The rightmost figure shows the detected percentage of poisoned data from the subsets with the lowest instance scores. FreeShap is either superior or comparable in data removal while remaining effective at identifying wrong labels.

F.5.1. TEST PREDICTION EXPLANATIONS

For the tables in this section: [pos] means positive sentiment class and predicted correctly, while [pos(neg)] means a positively labeled training example being wrongly predicted as negative, and vice versa. We use SST-2 in this section.

We list more explanations using different instance attributions for the **most** helpful/harmful examples (with the highest/lowest scores) of a correct prediction in Tab. 17. Like previous case studies, FreeShap consistently offers insightful explanations for the top three examples. Other methods may present drama-related examples but lack consistently decent performance.

We provide explanations for an incorrect prediction in Table 18, noting that the most supportive examples identified by FreeShap are syntactically similar to the test case, particularly in film-related content, and feature a mix of positive and negative words.

Additionally, we highlight training data with the highest/lowest instance scores relative to the test set in Table 19. Interestingly, most harmful training examples are incorrectly predicted. FreeShap and representer point both identify potentially wrongly labeled examples among the harmful examples, while the Influence function identifies potentially wrongly labeled examples among the helpful examples. Additionally, there's an overlap between FreeShap's helpful examples in the current table and those in Table 13, as well as between the represented point's harmful examples across the two tables.

Last but not least, we also notice that the influence function and representer point tend to provide wrongly predicted training examples in explanations, aligning with Garima et al. (2020)'s experiments on the computer vision dataset.

Table 17. A correctly predicted example with positive sentiment in SST-2 interpreted by different attribution approaches on prompt-based fine-tuned **Llama2** model.

Test input: ... a magnificent *drama* well worth tracking down .		[pos]
Attribution	Most **helpful** training examples	Most **harmful** training examples

Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	a consummate *actor* incapable of being boring	[pos]	others will find their humor-seeking dollars best spent elsewhere.	[neg]
	moving and solidly entertaining	[pos]	and undemanding armchair tourists	[neg]
	increasingly diverse french *director*	[pos]	bogs down in a surfeit of characters and	[neg]
Influence	most of it given to children	[pos(neg)]	4ever has the same sledgehammer appeal as pokemon *videos* , but	[pos(neg)]
	's because the laughs come from fairly basic *comedic* constructs	[pos(neg)]	kids who are into this thornberry stuff will probably be in wedgie heaven	[pos(neg)]
	is consistently	[pos(neg)]	definitely funny stuff, but it's more of the 'laughing at' variety than the 'laughing with.'	[pos(neg)]
TracIn	superb	[pos]	the book 's irreverent energy , and scotches most	[neg]
	what–and has all the dramatic weight of a raindrop	[neg]	develop her own *film* language	[pos]
	scored and powered by a set of heartfelt *performances*	[pos]	something of a public service –	[pos]
Representer	you 've endured a long workout without your pulse ever racing	[pos(neg)]	self-deprecating, biting and witty feature	[neg]
	an hour 's worth of actual material	[pos(neg)]	on its taut *performances* and creepy atmosphere	[neg]
	beyond the cleverness, the weirdness and the pristine camerawork, one hour photo is a sobering meditation on why we take pictures.	[pos(neg)]	the ethos of the chelsea hotel may shape hawke 's artistic aspirations , but he has n't yet coordinated his own dv poetry with the beat he hears in his soul	[neg]

Table 18. A wrongly predicted example with positive sentiment in SST-2 interpreted by different attribution approaches on prompt-based fine-tuned **Llama2** model.

Test input: like leon, it's frustrating and still oddly likable.		[pos(neg)]
Attribution	Most **helpful** training examples	Most **harmful** training examples

Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	it's an effort to watch this movie, but it eventually pays off and is effective if you stick with it	[pos(neg)]	ended so damned soon	[neg]
	this is lightweight filmmaking, to be sure, but it's pleasant enough – and oozing with attractive men.	[pos]	if even the filmmakers didn't know what kind of movie they were making	[neg]
	, it still manages to string together enough charming moments to work.	[pos]	the story passes time until it's time for an absurd finale of twisted metal, fireballs and revenge.	[neg]
Influence	4ever has the same sledgehammer appeal as pokemon videos , but	[pos(neg)]	blood-curdling family intensity	[pos]
	kids who are into this thornberry stuff will probably be in wedgie heaven.	[pos(neg)]	to feel-good , follow-your-dream hollywood fantasies	[pos(neg)]
	have strip-mined the monty formula mercilessly since 1997.	[pos(neg)]	an hour 's worth of actual material	[pos(neg)]
TracIn	superb	[pos]	the book 's irreverent energy , and scotches most	[neg]
	what – and has all the dramatic weight of a raindrop	[neg]	develop her own film language	[pos]
	scored and powered by a set of heartfelt performances	[pos]	something of a public service –	[pos]
Representer	self-deprecating, biting and witty feature	[neg]	feel the screenwriter at every moment	[pos(neg)]
	on its taut performances and creepy atmosphere	[neg]	the sentimental cliches mar an otherwise excellent film .	[pos(neg)]
	the ethos of the chelsea hotel may shape hawke's artistic aspirations, but he hasn't yet coordinated his own dv poetry with the beat he hears in his soul	[neg]	kids who are into this thornberry stuff will probably be in wedgie heaven.	[pos(neg)]

26

Table 19. The most helpful/harmful examples for SST-2 test datasets prediction interpreted by different attribution methods on prompt-based fine-tuned **Llama2** model.

Attribution	Most **helpful** training examples		Most **harmful** training examples	
FreeShap	a culture clash comedy only half as clever as it thinks it is	[neg]	lush and beautifully photographed (somebody suggested the stills might make a nice coffee table book)	[pos(neg)]
	it's anchored by splendid performances from an honored screen veteran and a sparkling newcomer who instantly transform themselves into a believable mother/daughter pair.	[pos]	finding the characters in slackers or their antics amusing, let alone funny	[pos(neg)]
Influence	far too polite to scale the lunatic heights of joe dante 's similarly styled gremlins	[pos(neg)]	is as delightful as it is derivative.	[pos(neg)]
	4ever has the same sledgehammer appeal as pokemon videos , but	[pos(neg)]	his rental car	[pos(neg)]
TracIn	superb	[pos]	the book 's irreverent energy , and scotches most	[neg]
	what – and has all the dramatic weight of a raindrop	[neg]	develop her own film language	[pos]
Representer	a masterpeice	[pos(neg)]	have strip-mined the monty formula mercilessly since 1997	[pos(neg)]
	to the nonconformist in us all	[pos(neg)]	4ever has the same sledgehammer appeal as pokemon videos , but	[pos(neg)]

F.5.2. REMOVE LOW INSTANCE SCORE

We include the results for removing low-instance sore points for Llama2. Fig. 11 shows using FreeShap leads to a slower decrease in performance (most significant on the MRPC and RTE datasets) and even slight improvement for performance (in SST-2 and MR datasets).
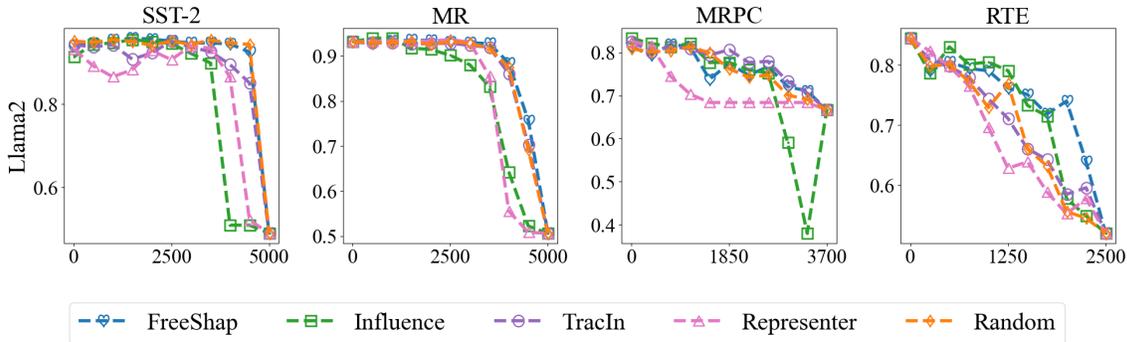

Figure 11. **Data Removal for Llama2**: When removing data from the lowest instance scores, the data points identified by FreeShap cause relatively slower degradation (slower is better).

F.5.3. TIME EFFIENCY

Fig. 12 shows the performance of BERT and Llama2-7B on the SST-2 dataset when using FreeShap with TMC and BI. The results indicate that FreeShap is still time efficient even with LLM. Initially, FreeShap takes a longer time to build the eNTK matrix for Llama2 due to the larger model size, but since the kernel regression avoids fine-tuning, the computational cost becomes independent of the model size. Interestingly, FreeShap on Llama2 can be even faster than BERT when the dataset becomes larger. This is because when TMC is applied, LoRA-finetuned Llama2 has better representation, which results in a more representative eNTK, and hence faster convergence (more pronounced diminishing marginal contribution) w.r.t. the number of training data and early termination for each Monte Carlo iteration when computing the Shapley value.
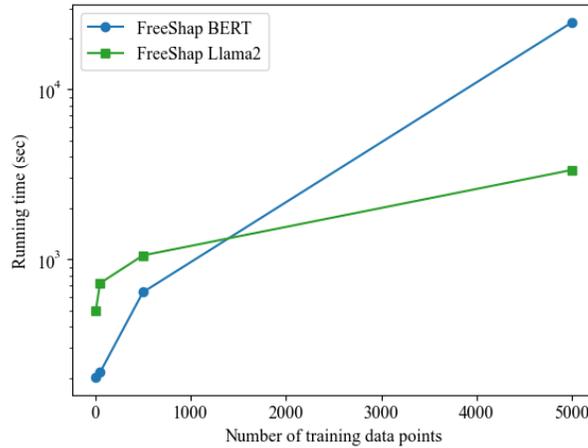
Figure 12. Running time comparison for dataset SST-2 when the model is BERT and Llama2.

F.5.4. SCALING UP FROM LLAMA2-7B TO LLAMA2-13B

To demonstrate the scalability of FreeShap w.r.t. the size of the model, we extend our experiments to the Llama2-13B model which is larger than the Llama2-7B model for wrong label detection on the SST2. We use a dataset with a size of 1k to demonstrate the scalability of the proposed method. As shown in Fig. 13, FreeShap remains more effective in identifying mislabeled data with a larger model. We also demonstrate the time complexity when using Llama2-7B and Llama2-13B in Fig. 14, showing the efficiency and scalability of FreeShap w.r.t. model size. Hence, our approach has the potential to be applied to larger models (70B or even more), and we encourage the community to explore this further with the right resources.
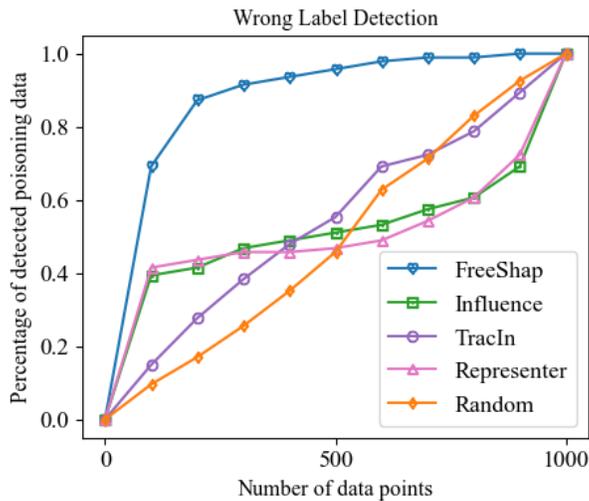

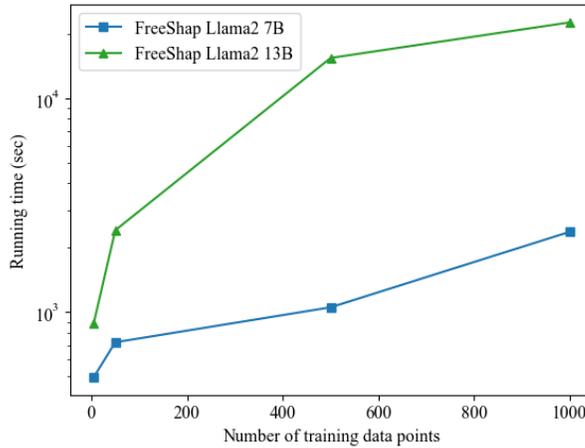*Figure 13.* Wrong Label Detection for dataset 1k SST-2 on Llama2-13b.

Figure 14. Running time comparison for dataset SST-2 when the model is Llama2-7B and Llama2-13B.

G. Ablation Studies

G.1. Diminishing Influence of Data Point

We empirically examine the assumption of diminishing influence of a data point. Specifically, for a *consistently harmful (or helpful) contributing* data point z_i, we examine if $|\tau_0| \geq \cdots \geq |\tau_{n-1}|$. We randomly select a data point and compute the marginal contribution of this data point within subsets of varying sizes, ranging from 0 to 300. For each subset of size k, we calculate the mean marginal contribution from three sampled subsets. Figure 15 plots the absolute mean marginal contribution of the data point across different subset sizes. The fitted curve illustrates a decreasing trend in the absolute marginal contribution, which is in accordance with our hypotheses and corroborated by previous studies (Killamsetty et al., 2021; Wang et al., 2021).
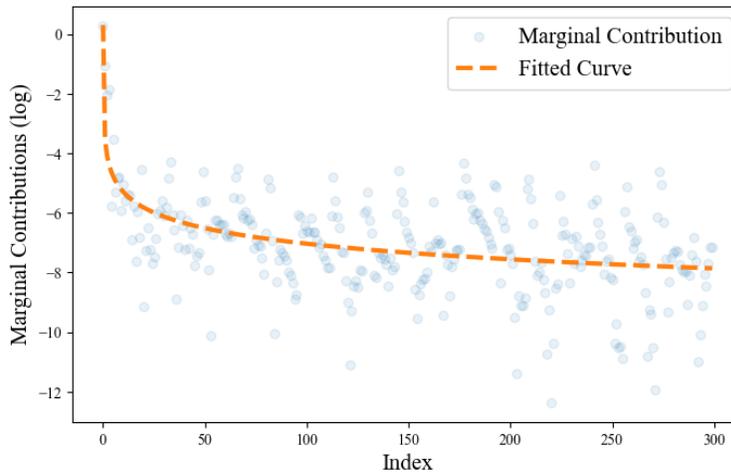

Figure 15. **Diminishing Marginal Contribution**.

G.2. Expectation and Variance for Shapley value and LOO

When comparing the robustness of Shapley and LOO, the upper bound is bounded by variables related to the expectation and variance. We can intuitively explain that the Shapley value is more robust as the Shapley value has a larger expectation and smaller variance, as illustrated in the remark 3.6. The relative relationship for expectation and variance between the Shapley value and LOO is further verified in the figure 16, where the figure on the left demonstrates that the Shapley value

generally has a larger expectation, and the figure on the right demonstrates that the Shapley value has a smaller variance.
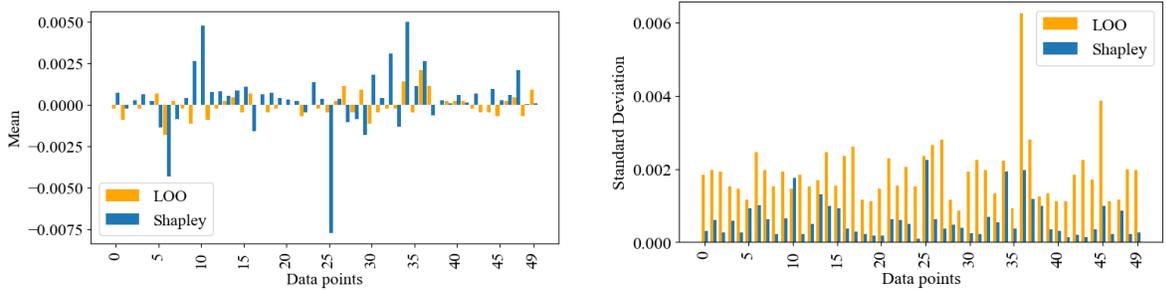
(a) Mean of instance scores when the dataset is of size 500. The average of absolute mean among 50 points is $0.50\mathrm{e}{-3}$ for LOO and $1.20\mathrm{e}{-3}$ for Shapley.

(b) Standard deviation of instance scores when the dataset is of size 500. The average of standard deviation among 50 points is $1.89\mathrm{e}{-3}$ for LOO and $0.61\mathrm{e}{-3}$ for Shapley.

Figure 16. This figure shows each point's mean and standard deviation of five instance attribution scores computed by Shapley and LOO. The figure on the left demonstrates each point's mean of instance scores. Each index denotes a point's average instance score among five instance scores computed when the point is placed in five datasets. The figure on the right demonstrates each point's standard deviation of instance scores. Each index denotes a point's standard deviation of five instance scores. Matching with our analysis, the Shapley value has a higher expectation and a lower standard deviation.

G.3. Kernel Behavior of Prompt-based Fine-tuning

Previous work investigates the kernel behavior of prompt-based fine-tuning using experiments but on a few shot settings (Malladi et al., 2023). To further empirically verify the kernel behavior of prompt-based fine-tuning for each training data size, we conduct the following ablation study. We perform both prompt-based fine-tuning and eNTK kernel regression on training points with sizes ranging from 1 to 1k. We record test performance for each size and plot them in Figure 17. At each training size, the eNTK (orange line) consistently provides a decent approximation to the fine-tuning method (blue line), evident from a small gap between the two curves. As the number of training points increases, both the eNTK and fine-tuning trends rise almost in tandem, indicating that eNTK offers a reliable approximation throughout the training process.
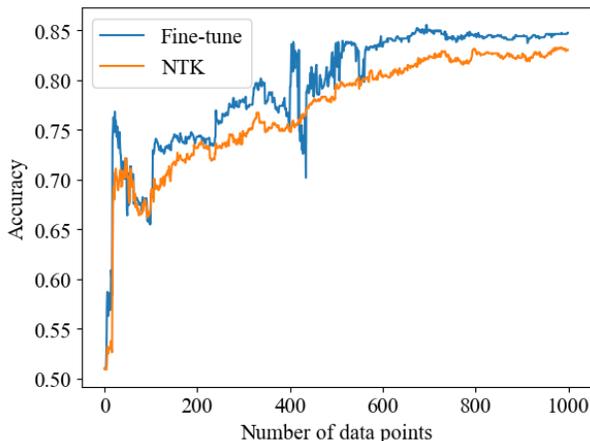

Figure 17. We plot the test performance as the number of training points increases when using eNTK kernel regression and fine-tuning. The two lines have a Pearson correlation of 0.97 and a Spearman correlation of 0.94.

G.4. Ablation Studies for FreeShap and Influence on Different Approximations

We compare the data removal performance (Fig.18) and wrong label detection (Fig.19) between FreeShap with TMC and FreeShap with MC, finding similar efficacy in identifying both high/low instance scores and mislabeled data.

Further, we evaluated the performance of the influence function with conjugate gradients (CG) (as used in the main paper) against the influence function with LiSSA in data removal (Fig.18) and wrong label detection (Fig.19). The CG-based influence function demonstrated superior or comparable effectiveness in detecting low instance scores (slower degradation of performance is better) and mislabeled data, but it was comparable or less effective in identifying high instance scores (faster degradation of performance is better). Due to CG's proficiency in two of the three tasks, we choose CG for a more equitable comparison.
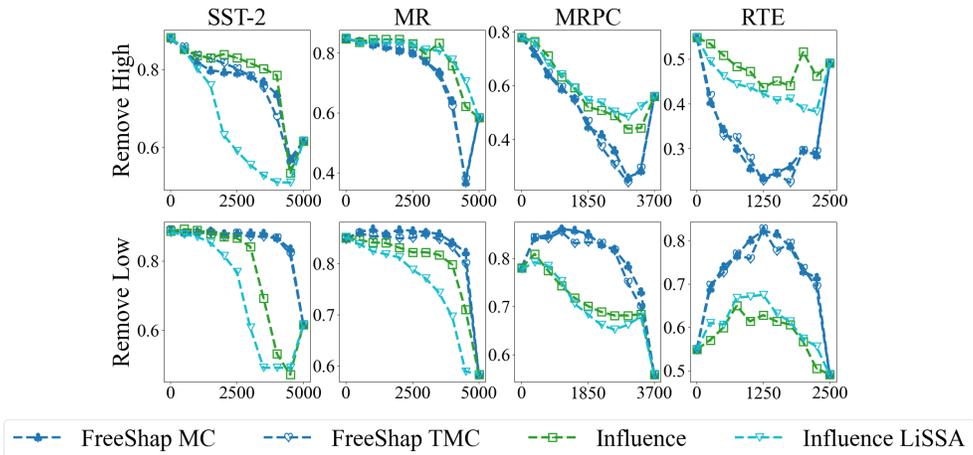

Figure 18. **Data Removal**: Data removal performance comparison: FreeShap with TMC vs. FreeShap with MC show similar results. In contrast, the influence function with CG (used in the main paper) versus the influence function with LiSSA reveals disparities. The CG-approximated influence function excels at identifying low instance scores in most datasets but performs comparably or poorly in identifying high instance scores.
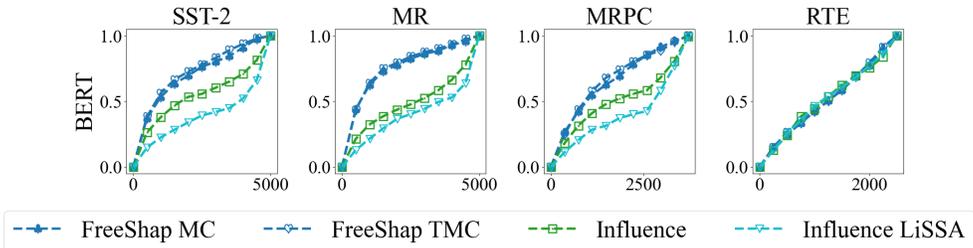

Figure 19. **Wrong label detection**: Wrong label detection comparison: FreeShap with TMC vs. FreeShap with MC demonstrates similar efficacy in identifying mislabeled data. Further comparison between the influence function with CG (as used in the main paper) and the influence function with LiSSA shows the CG-approximated influence function to be more effective in pinpointing mislabeled examples.

H. Dataset Information

Single sentence task involves classifying a single sentence independently to produce an output, like sentiment analysis (SST-2). The sentence pair task involves taking two sentences as input to determine a relationship between them, such as textual entailment or paraphrasing.

For the single-sentence task, we utilize two datasets:

1. The SST-2 (Socher et al., 2013) dataset is used for binary sentiment classification, aiming to categorize text snippets as either positive or negative. The dataset consists of movie reviews annotated with their corresponding sentiment labels. Dataset is retrieved from `https://huggingface.co/datasets/sst2`.

2. The Rotten Tomatoes movie review(MR) dataset is a collection of film reviews used for sentiment analysis tasks (Pang

& Lee, 2005). The reviews are labeled as either positive or negative, based on the sentiment expressed by critics and audience members. The dataset is retrieved from `https://huggingface.co/datasets/rotten_tomatoes`.

For the sentence-pair task, we utilize three datasets:

1. The MRPC (Microsoft Research Paraphrase Corpus) (Dolan & Brockett, 2005) dataset is a collection of sentence pairs labeled as either paraphrases of each other or not. It is commonly used for evaluating natural language processing tasks related to semantic textual similarity and paraphrase identification. Dataset is retrieved from `https://huggingface.co/datasets/glue/viewer/mrpc`.

2. The RTE (Recognizing Textual Entailment) (Bentivogli et al., 2009) dataset consists of sentence pairs labeled to indicate if the second sentence entails, contradicts, or is neutral concerning the first. It is widely used for natural language understanding tasks like textual entailment and inference. The dataset is retrieved from `https://huggingface.co/datasets/glue/viewer/rte`.

3. The MultiNLI (MNLI) (Williams et al., 2018) dataset is a large-scale, multi-genre natural language inference dataset designed for the training and evaluation of machine learning models on the NLI task. It contains pairs of sentences labeled as "entailment," "contradiction," or "neutral," drawn from a variety of sources, including fiction, government documents, and forums. The dataset is retrieved from `https://huggingface.co/datasets/glue/viewer/mnli`.

As our objective is to interpret the model predictions from the training data and obtain a better understanding of the dataset such as dataset debugging and mislabeled data detection, we directly evaluate all experiments on the test set. Since the test set label is not fully available (SST-2 and RTE), we use the validation set as the test set for simplicity. The necessary information for single sentence task datasets is in Tab. 20, while the information for the sentence pair task datasets is in Tab. 21, and the test column in tables below is the information of validation datasets.

Table 20. Detailed information about single sentence task dataset

Dataset	SST-2		MR	
	train	test	train	test
length	67.3k	872	8.53k	1.07k
min length of tokens	3	4	3	5
max length of tokens	66	55	78	67
average length of tokens	13.32	25.16	27.37	27.66

Table 21. Detailed information about sentence pair task dataset

Dataset	MRPC		RTE		MNLI	
	train	test	train	test	train	test
length	3668	1725	2490	3000	392k	9.8k
min length of premise	9	10	7	6	3	3
max length of premise	52	51	282	240	428	219
average length of premise	27.13	26.95	57.6	54.39	26.72	26.06
min length of hypothesis	10	10	5	4	3	3
max length of hypothesis	52	54	53	65	77	56
average length of hypothesis	27.12	27.03	13.59	13.24	14.19	14.12