On the Feasibility of In-Context Probing for Data Attribution

Cathy Jiao[1] **Weizhen Gao**[2] **Aditi Raghunathan**[2] **Chenyan Xiong**[1]
[1] Language Technologies Institute, Carnegie Mellon University
[2] Computer Science Department, Carnegie Mellon University
{cljiao, aditirag, cx}@cs.cmu.edu, wgao2@andrew.cmu.edu

Abstract

Data attribution methods are used to measure the contribution of training data towards model outputs, and have several important applications in areas such as dataset curation and model interpretability. However, many standard data attribution methods, such as influence functions, utilize model gradients and are computationally expensive. In our paper, we show in-context probing (ICP) – prompting a LLM – can serve as a fast proxy for gradient-based data attribution for data selection under conditions contingent on data similarity. We study this connection empirically on standard NLP tasks, and show that ICP and gradient-based data attribution are well-correlated in identifying influential training data for tasks that share similar *task type* and *content* as the training data. Additionally, fine-tuning models on influential data selected by both methods achieves comparable downstream performance, further emphasizing their similarities. We also examine the connection between ICP and gradient-based data attribution using synthetic data on linear regression tasks. Our synthetic data experiments show similar results with those from NLP tasks, suggesting that this connection can be isolated in simpler settings, which offers a pathway to bridging their differences.

1 Introduction

Data attribution methods aim to identify specific training data that contribute to the outputs of a model (Worledge et al., 2024). Methods for data attribution have numerous useful applications; for instance, dataset curation (Ilyas et al., 2022; Xia et al., 2024; Yu et al., 2024), model interpretability (Han and Tsvetkov, 2021; Akyurek et al., 2022; Li et al., 2024a), and data valuation (Ghorbani and Zou, 2019; Yang et al., 2024a; Zhang et al., 2025). While data attribution methods are useful, they can be computationally demanding. For instance, influence functions (Koh and Liang, 2017) are a classic tool for gradient-based data attribution (i.e., methods that utilize model gradients in their computation), but are challenging to scale for large deep learning models with billions of parameters (Grosse et al., 2023; Choe et al., 2024).

Recently, data selection using in-context probing (ICP) – prompting a LLM – to determine the quality of a training data sample has become an important avenue for curating high-quality training datasets (Rubin et al., 2022; Nguyen and Wong, 2023; Wettig et al., 2024). Yet, it is unclear why ICP is effective at training data selection since there are multiple factors to consider for determining the quality of training data, such as mixtures, utility, and the quantity of data (Lee et al., 2022; Xie et al., 2024; Goyal et al., 2024).

In this paper, we offer an explanation for this phenomenon by drawing a connection between ICP and gradient-based data attribution. To study the robustness of this connection, we empirically analyze the agreement between both methods for identifying influential training data for in-domain target tasks (i.e., tasks that share similar *task type* and *content* as the training data), and out-of-domain target tasks. On standard NLP tasks (including instruction-following and QA), our experiments reveal that ICP can approximate gradient-based data attribution for identifying influential training data in the in-domain setting. Further fine-tuning on the influential training data selected by either method — in particular, using data from the Alpaca Dataset (Taori et al., 2023) — results in similar model performance in instruction-following on Alpaca Eval (Li et al., 2023; Dubois et al., 2024b). This is advantageous since, unlike gradient-based attribution methods, ICP enables cost-effective data selection; it requires no access to model parameters, and can even be performed via API calls, making it ideal for black-box models.

In addition to standard NLP tasks, we study the connection between ICP and gradient-based data

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 5155–5170

attribution in a controlled setting using synthetic data, specifically linear regression tasks. In this setting, the *task type* (i.e., the specifically linear relation) and *content* (i.e., input distance) of the training data and target task are clearly defined, making them easy to adjust. Similar to standard NLP tasks, our findings on synthetic data show that ICP can approximate gradient-based data attribution in the in-domain setting. Furthermore, our synthetic data results show that this connection can be isolated, which paves way for future research bridging the gap between the two methods. Our contributions are summarized as follows:

1. We draw a connection between ICP and gradient-based data attribution and show they agree in identifying influential training data for in-domain target tasks.

2. To further highlight ICP as an effective proxy for gradient-based data attribution, we use both methods for dataset curation, and show that fine-tuning models on data highly-ranked by either method leads to similar performance.

3. We explore the relationship between ICP and gradient-based attribution using synthetic data in a controlled setting. Our results show that the connection between these two methods can be isolated in toy settings, making it a potential path to bridge their gaps.

2 Related Work

Obtaining high-quality training data is important for efficient model training (Lee et al., 2022; Sorscher et al., 2022; Ye et al., 2024; Albalak et al., 2024). One class of data attribution methods is gradient-based methods, such as influence functions (Koh and Liang, 2017), which utilize model gradients that estimate the influence of a training sample on model predictions. Despite being computationally expensive in LLM settings (Grosse et al., 2023), gradient-based methods are effective for curating subsets of high-quality training data (Pruthi et al., 2020; Park et al., 2023; Han et al., 2023; Xia et al., 2024; Engstrom, 2024).

Based on the phenomenon of transformers having in-context learning capabilities (Min et al., 2022; Han et al., 2023; Bhattamishra et al., 2023; Liu et al., 2024), recent works have used ICP for training data selection (Rubin et al., 2022; Nguyen and Wong, 2023; Iter et al., 2023; Wettig et al., 2024). These methods involve measuring the model output likelihoods of the task given an in-context train sample, or prompting an LLM with questions to identity high-quality training data. For example, Li et al. (2024b) demonstrated that training on subsets of high-quality data using ICP leads to better performance than training on the entire dataset.

Since both gradient-based data attribution methods and ICP can be used effectively for data selection, a key component to connecting these ideas lies in a recent body of work which suggests that in-context learning implicitly performs gradient descent by constructing meta-gradients (Irie et al., 2022; Dai et al., 2023; Von Oswald et al., 2023). Specifically, these studies highlight the duality between a forward pass through a transformer attention head and linear layers trained by gradient descent, but rely on major assumptions; including linear attention, and limited analysis on this phenomena on MLP layers. Despite these assumptions, transformer outputs for synthetic in-context tasks, such as linear regression, mirror the predictions of algorithms that implement gradient descent (Akyürek et al., 2023; Garg et al., 2022; Mahankali et al., 2024), making the relationship between ICP and gradient descent an open research area.

3 Preliminaries

In order to draw a connection between ICP and gradient-based data attribution, we first present three methods for data selection: influence functions (Koh and Liang, 2017), local datamodeling (Iter et al., 2023; Yu et al., 2024) and ICP scoring (Li et al., 2024b). We begin by defining some notation: let $\mathcal{D}_{train} = \{z_i\}_{i=1}^{N}$ be a set of training samples, where a sample $z_i = (x_i, y_i)$ contains an input and output. Similarly, let $\mathcal{D}_{test} = \{z'_i\}_{j=1}^{M}$ be a set of test samples.

Method 1: Influence Functions (Koh and Liang, 2017) approximate changes in model predictions when samples are added/removed from the model's training data. To measure the influence of train sample $z \in \mathcal{D}_{train}$, the change in model parameters θ^* is approximated when z is up-weighting by a small value ϵ. Thus, the empirical risk minimization is:

$$\theta^*(\epsilon) = \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(z_i; \theta) + \epsilon \mathcal{L}(z; \theta), \quad (1)$$

which is also called the response function. We wish to find the change in parameters $\Delta\theta = \theta^*(\epsilon) - \theta^*$,

which can be done via a first-order Taylor approximation to the response function at $\epsilon = 0$, which yields $\theta^*(\epsilon) - \theta^* \approx \epsilon \frac{\mathrm{d}\theta^*(\epsilon)}{\mathrm{d}\epsilon}\Big|_{\epsilon=0}$. Moreover, using the Implicit Function theorem, we get the influence of z on θ^*.

$$I_{\theta^*}(z) = \frac{\mathrm{d}\theta^*(\epsilon)}{\mathrm{d}\epsilon}\Big|_{\epsilon=0} = -H^{-1}\nabla_\theta \mathcal{L}(z;\theta^*), \quad (2)$$

where $H = \frac{1}{N}\sum_{i=1}^{N} \nabla_\theta^2 \mathcal{L}(z_i;\theta^*)$ and $z_i \in \mathcal{D}_{train}$. To quantify the influence of z specifically on z', we can measure influence with respect to $\mathcal{L}(z';\theta)$, the loss on z', which via the chain rule results in:

$$\text{Infl}(z', z) = \nabla\mathcal{L}(z';\theta)H^{-1}\nabla\mathcal{L}(z;\theta). \quad (3)$$

Computing H^{-1} expensive and unstable in non-convex loss function settings, such as for large deep learning models (Basu et al., 2021). A simpler and more cost effective alternative (Pruthi et al., 2020; Xia et al., 2024) is to drop the Hessian and only keep the *inner product*:

$$\text{Infl}_{\text{IP}}(z', z) = \nabla\mathcal{L}(z';\theta) \cdot \nabla\mathcal{L}(z;\theta). \quad (4)$$

In particular, (Yang et al., 2024b) showed that despite dropping the Hessian, Infl_{IP} exhibits good order-consistency with Infl.

Method 2: Local Data Influence. The influence of training sample $z \in \mathcal{D}_{train}$ towards a test sample $z' = (x', y')$ can also be measured using a one-step training score (Pruthi et al., 2020; Iter et al., 2023; Yu et al., 2024). Formally, this score is defined as:

$$\text{Infl}_{\text{Loc}}(z, z') = s_{zs}(z';\hat{\theta}) - s_{zs}(z';\theta). \quad (5)$$

where $\hat{\theta} = \theta - \eta\nabla\mathcal{L}(\theta, z)$ denotes the parameters of a model trained on z for single step with learning rate η. We denote $\text{s}_{\text{zs}}(z';\theta) = \log p(y'|x';\theta)$ as the zero-shot score (i.e., the model likelihood for the test sample output). Alternatively, the contribution of train sample z towards an entire test set $\mathcal{D}_{test}$ can be aggregated as:

$$\text{Infl}_{\text{Loc}}(z, \mathcal{D}_{test}) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}[s_{zs}(z'_j;\hat{\theta}) > s_{zs}(z'_j;\theta)]. \quad (6)$$

Infl_{Loc} performs *local* datamodeling since the influence of z is measured by a single training step on an existing pre-trained model, rather than fully re-training the model with z.

Method 3: In-Context Probing Score. Leveraging the in-context learning abilities of LLMs, the importance of training sample z can also be measured using a one-shot quality score introduced in Li et al. (2024b). Formally, the ICP score is:

$$\text{ICP}(z, \mathcal{D}_{test}) = \frac{1}{M}\sum_{j=1}^{M} \mathbb{1}[s_{os}(z'_j|z;\theta) > s_{zs}(z'_j;\theta)], \quad (7)$$

where for a test sample $z' = (x', y')$, the one-shot score is defined as $\text{s}_{\text{os}}(z', z;\theta) = \log p(y'|z, x';\theta)$, which is the model likelihood for the output of test sample z' with z as an in-context demonstration.

Connecting ICP, Infl_{Loc}, and Infl_{IP}. While all three methods can be used to measure the importance of training samples for a test task, they differ in computational efficiency. Notably, ICP is convenient since, unlike Infl_{Loc}, it requires no training, and, unlike Infl_{IP}, it does not access model gradients. Given the advantages of using ICP, we note the connection between ICP and Infl_{IP} through Infl_{Loc}, which shows how ICP can be an efficient proxy for gradient-based data attribution.

First, we draw a connection between ICP and Infl_{Loc} from recent works (Irie et al., 2022; Dai et al., 2023; Von Oswald et al., 2023) which show that a linear attention head performs an implicit gradient descent update on in-context demonstrations. We present this construction below (details in Appendix B):

$$Attn(K, V, q) \approx (W_{z'} + \Delta W_z)q, \quad (8)$$

where $W_{z'}$ represents the attention head weights for a test query z'. Notably, ΔW_z is the update for an in-context demonstration z, which is applied to attention head weights $W_{z'}$. Given the update of z onto $W_{z'}$, this construction shares similarities with performing an actual gradient descent update of z onto the model parameters. A resulting hypothesis is that for a model parameterized by θ, we have:

$$s_{os}(z'|z;\theta) \propto s_{zs}(z';\hat{\theta}), \quad (9)$$

where $\hat{\theta} = \theta - \eta\nabla\mathcal{L}(z;\theta)$. That is, taking a training step on z has similar effects on the model output likelihoods for z' as using it as an in-context demonstration. Thus, if equation 9 holds, then we have $\text{ICP}(z', z) \propto \text{Infl}_{\text{Loc}}(z', z)$. Moreover, connecting ICP and Infl_{IP} is straightforward since $\text{Infl}_{\text{IP}}(z', z)$ is an approximation of $\text{Infl}_{\text{Loc}}(z', z)$, as noted in Pruthi et al. (2020). As a result, we have:

$$\text{ICP}(z', z) \propto \text{Infl}_{\text{Loc}}(z', z) \approx \text{Infl}_{\text{IP}}(z', z). \quad (10)$$

The full derivation of this result is in Appendix A. An implication of equation 10 is that we expect positive correlation between ICP and Infl_{IP} scores with respect to how they rank training data for test samples. In the next sections, we empirically explore this correlation.

4 Experiments on NLP Datasets

Given the connection between ICP and Infl_{IP}, we describe our experimental setup to analyze how well these two methods correlate in their rankings of influential training data. As noted in the previous section, a key component in connecting $\text{ICP}(z', z)$ with $\text{Infl}_{\text{IP}}(z', z)$ is the hypothesis that ICP performs a process akin to a gradient descent step on a train sample z (i.e., $\text{ICP}(z, z') \propto \text{Infl}_{\text{Loc}}(z', z)$). A key question is whether this process occurs for any arbitrary z and z', since this would affect the correlation between $\text{ICP}(z', z)$ and $\text{Infl}_{\text{IP}}(z', z)$ rankings. To investigate this empirically, we vary z and z' to be "in-domain" and "out-of-domain", and examine the correlation step-by-step between ICP, Infl_{Loc}, and Infl_{IP}. Although "in-domain" is loosely defined in NLP, two features that commonly define whether z is in the same domain as z' involve *task* and *content* similarity (Ramponi and Plank, 2020).

Formally, we define a set of tasks $\{t_i\}_{i=1}^{T} \in \mathcal{T}$. Each task maps an input $x \in \mathcal{X}$ into a output $y \in \mathcal{Y}$ (i.e., $t : \mathcal{X} \rightarrow \mathcal{Y}$) to create a data sample $z = (x, y)$. We consider a set of train samples for task t, which we denote as $\mathcal{D}^{t}_{train}$, and a set of test samples for task t', which we denote as $\mathcal{D}^{t'}_{test}$. We are interested in how the correlation between ICP and Infl_{IP} changes as the following features vary between $\mathcal{D}^{t}_{train}$ and $\mathcal{D}^{t'}_{test}$:

1. **Task Similarity**: as the train task t and test task t' change. In our experiments, we heuristically define the train and test tasks to be standard NLP tasks. In particular, we fix the test task be to instruction-following, and vary the train tasks to be instruction-following, QA/-DocQA, and pretrain tasks, which we describe in detail in Section 4.1.

2. **Content Similarity**: as the semantic similarity between train sample z and test sample z' change. In our experiments, we fix the train task t and test task t' to be the same, and vary the content using BertScore (Zhang et al., 2020), a popular evaluation metric which measures similarity between two sequences using pretrained BERT embeddings.

Next, given $\mathcal{D}^{t}_{train}$ and $\mathcal{D}^{t'}_{test}$, we measure the Spearman correlation between ICP and Infl_{IP} rankings. First, we obtain ICP and Infl_{IP} scores of the all samples in $\mathcal{D}^{t}_{train}$ for test set $\mathcal{D}^{t'}_{test}$, which we denote as:

$$\mathcal{S}_{\text{ICP}}(\mathcal{D}^{t}_{train}, \mathcal{D}^{t'}_{test}) = \{\text{ICP}(z_i, \mathcal{D}^{t'}_{test}) | z_i \in \mathcal{D}^{t}_{train}\}_{i=1}^{N}, \quad (11)$$

$$\mathcal{S}_{\text{Infl}_{\text{IP}}}(\mathcal{D}^{t}_{train}, \mathcal{D}^{t'}_{test}) = \{\text{Infl}_{\text{IP}}(z_i, \mathcal{D}^{t'}_{test}) | z_i \in \mathcal{D}^{t}_{train}\}_{i=1}^{N}. \quad (12)$$

Finally, we calculate the Spearman correlation between the ICP and Infl_{IP} scores:

$$\text{Spearman}(\mathcal{S}_{\text{ICP}}(\mathcal{D}^{t}_{\text{train}}, \mathcal{D}^{t'}_{test}), \mathcal{S}_{\text{Infl}_{\text{IP}}}(\mathcal{D}^{t}_{\text{train}}, \mathcal{D}^{t'}_{\text{test}})). \quad (13)$$

4.1 Datasets and Models

In this section, we define a set of NLP tasks used in our experiments, which differ in objective and structure. We describe the datasets used for each task (see Table 3 in Appendix C for examples), and also describe our models.

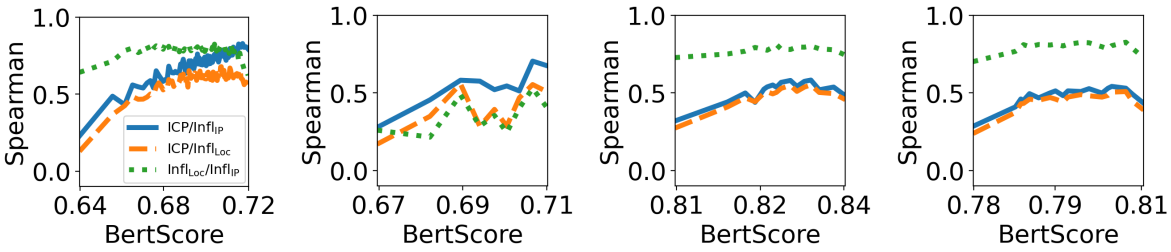
Instruction Tasks: Instruction-following requires a language model to generate an appropriate response by following an instruction (e.g., "Write a poem about the Autumn"), making it a key component in LLM research and real-world applications (Ouyang et al., 2022; Zhang et al., 2024). For instruction tasks, we use the Alpaca dataset (Taori et al., 2023), which contains 52K instruction demonstrations generated by GPT-4 following the Self-Instruct method (Wang et al., 2023a).

QA/DocQA Tasks: QA tasks are simple question-answering tasks without any context in the input (e.g., "What is the capital city of the U.S?"). They differ from instruction tasks since they may not explicitly provide an instruction in the task. DocQA tasks are question-answering tasks with additional context in the input (e.g., "Read following movie review and rate it: ..."). We sourced QA and DocQA tasks from PromptSource (Bach et al., 2022) dataset, which contains human-written prompts. We split the dataset into QA and DocQA datasets, using 9K and 8K examples, respectively.

Pretrain Tasks: Unlike the previous tasks, pretrain data is unstructured and does not contain any explicit questions or instructions. We sourced

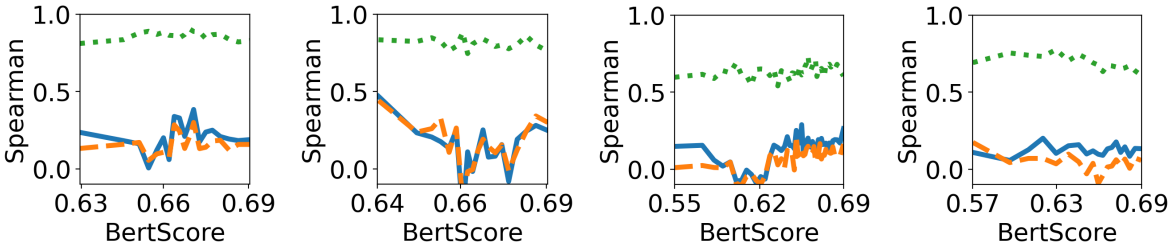
Table 1: Spearman correlation between ICP, Infl_{Loc}, and Infl_{IP} using test samples from the Alpaca dataset and train samples from Alpaca, UltraChat, QA, DocQA, and pretrain datasets. All p-values are $< .05$.

	Alpaca		**QA**		**DocQA**		**Pretrain**	
	Pythia-1b	Llama-3.2-3B	Pythia-1b	Llama-3.2-3B	Pythia-1b	Llama-3.2-3B	Pythia-1b	Llama-3.2-3B
$\text{ICP/Infl}_{\text{IP}}$	0.73	0.54	0.10	0.10	0.21	0.13	0.07	0.12
$\text{ICP/Infl}_{\text{Loc}}$	0.61	0.36	0.10	0.13	0.17	0.26	0.03	0.05
$\text{Infl}_{\text{Loc}}/\text{Infl}_{\text{IP}}$	0.78	0.57	0.78	0.87	0.86	0.88	0.65	0.63



(1a) Alpaca (Pythia) (1b) Alpaca (Llama) (1c) QA (Pythia) (1d) DocQA (Pythia)

Figure 1: Correlation analysis between ICP, Infl_{Loc}, and Infl_{IP} (aggregated across groups of 500 samples) with respect to content similarity (BertScore) using test and train samples from the same task.



(2a) DocQA (Pythia) (2b) DocQA (Llama) (2c) Pretrain (Pythia) (2d) Pretrain (Llama)

Figure 2: Correlation analysis between ICP, Infl_{Loc}, and Infl_{IP} (aggregated across groups of 500 samples) with respect to content similarity (BertScore) using test samples from Alpaca and training samples from DocQA/Pretrain datasets. Additional analysis in Appendix C.

pretrain data from Minipile (Kaddour, 2023), which is a subset of the Pile (Gao et al., 2020) dataset curated for data diversity. We split the Minipile dataset into sequences of 256 tokens, and take a subset of 25K pretrain sequences.

Train and Test Set Splits: For each task, we use each dataset as the train set and take 100 samples from each dataset to form their respective test sets.

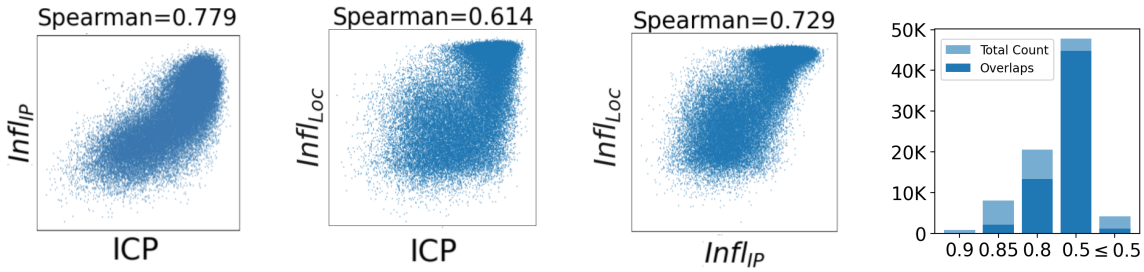
Models: Across all experiments, we calculate ICP, Infl_{Loc}, and Infl_{IP} scores using Pythia-1b-deduped (Biderman et al., 2023) and Llama-3.2-3B. For calculating Infl_{Loc}, we set the learning rate to 2e-5.

4.2 Task Similarity Results

Table 1 shows that $\text{ICP/Infl}_{\text{IP}}$ correlation is high when the task types of the train and test datasets are the same. This correlation decreases significantly when the task type of the train and test datasets differ. Note that $\text{Infl}_{\text{Loc}}/\text{Infl}_{\text{IP}}$ correlation remains high overall regardless of task difference between the train and test datasets. This suggests that Infl_{Loc} is a close approximation for Infl_{IP}, and that the breaking point lies between ICP and Infl_{Loc}. Thus, the hypothesis introduced in section 3 that ICP performs a gradient descent-like step does not hold when the train and test task types differ.

4.3 Content Similarity Results

Figure 1 shows that $\text{ICP/Infl}_{\text{IP}}$ correlation decreases as the content similarity (i.e., BertScore) decreases between the train and test samples. Moreoever, $\text{Infl}_{\text{Loc}}/\text{Infl}_{\text{IP}}$ correlation remains high overall, and does not change much as content similarity decreases, which implies that that Infl_{Loc} is a robust



(3a) ICP/Infl_{IP} scores. (3b) ICP/Infl_{Loc} scores. (3c) Infl_{IP}/Infl_{Loc} scores. (3d) ICP/Infl_{IP} overlaps.

Figure 3: Correlation analysis between rankings on the instructions from the Alpaca dataset assigned by ICP, Infl_{Loc}, and Infl_{IP}. All p-values are $< .05$.

Table 2: Results (winrates) evaluated on the Alpaca Eval dataset after being finetuned on data selected by ICP and Infl_{IP}. The highest winrate in each column is marked with $*$ for ICP and † for Infl_{IP}.

Score Bin	Method	Helpful Base	Koala	Self Instruct	Oasst	Vicunna	Overall
≤ 0.5	ICP	54.26	62.99	56.04	56.15	51.25	56.65
	Infl_{IP}	51.56	62.18	54.76	54.55	43.75	54.54
> 0.5	ICP	57.03	62.99	**61.35**	63.10	68.75	62.12
	Infl_{IP}	60.94	65.16	**65.08**†	60.42	52.50	62.09
> 0.8	ICP	61.42	59.35	58.73	**64.17***	67.50	61.42
	Infl_{IP}	62.79	**68.18**†	62.3	61.70	**68.75**†	**64.02**†
> 0.85	ICP	**62.79***	**62.58***	60.16	61.17	**70.00***	**62.26***
	Infl_{IP}	**65.12**†	67.95	56.35	**65.42**†	65.00	62.98
> 0.9	ICP	52.34	60.13	49.79	49.46	46.25	51.77
	Infl_{IP}	61.24	57.14	53.60	55.08	55.00	56.00

approximation for Infl_{IP} regardless of content similarity. Similar to task type, the breaking point again lies between ICP and Infl_{Loc} when the content similarity between the train and test samples differ.

4.4 Task vs. Content Similarity Results

Since both task type and content affects ICP/Infl_{IP} correlation, we vary both features simultaneously and observe its impact. In Figure 2, we examine ICP/Infl_{IP} correlation as the BertScore decreases between the Alpaca test samples and train samples from the other previously defined tasks. Figure 2 shows no increase in ICP/Infl_{IP} correlation as BertScore increases. Therefore, if the test and train tasks are different, then increasing content similarity does not result in better ICP/Infl_{IP} correlation.

5 ICP for Data Selection

In the previous section, we showed that ICP and Infl_{IP} correlate well in how they rank influential training data when the train set shares the same task type as the test set. This is is strongly reflected in the Alpaca instruction-following dataset (see Table 1 and Figure 3). In this case, it is possible that ICP can serve as a proxy for Infl_{IP}. This has promising implications: compute costs for ICP is significantly cheaper than Infl_{IP}. For instance, to score the entire Alpaca dataset with Pythia-1b, ICP incurred a total of 10 GPU hours while Infl_{IP} incurred 90 hours.

However, since ICP and Infl_{IP} rankings are not entirely aligned, we further compare them by using both methods to curate datasets for instruction-tuning. Following the same setup as Li et al. (2024b), we first we obtained ICP scores (and in our case, Infl_{IP} scores as well) for all training samples in the Alpaca dataset using the K-Means-100 dataset (a subset of 100 diverse instructions from the Alpaca dataset created by Li et al. (2024b)) as the test set. We use the ICP and Infl_{IP} scored training samples for instruction-tuning according the following procedure:

Finetuning Datasets: After obtaining ICP and Infl_{IP} scores (reminder: ICP $\in [0, 1]$) for the Alpaca dataset, we create ICP score bins of $\leq 0.5, > 0.5, > 0.8, > 0.85, > 0.9$. We used the number of samples in each score bin as

threshold cutoffs for Infl_{IP}. For example, if the > 0.9 ICP score bin had k training samples, then we also treated the top k samples from Infl_{IP} as the equivalent bin. We treat all bins as separate datasets, and randomly sample 700 demonstrations from each dataset for fine-tuning.

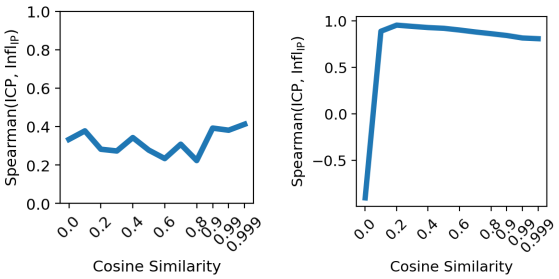
Training: We use the Adam optimizer with a batch size of 64 and lr=2e-7 to fine-tune Pythia-1b-deduped for 3 epochs. This is done separately for ICP and Infl_{IP} for each score bin.

Evaluation: We use the Alpaca Eval dataset (Li et al., 2023; Dubois et al., 2024b), which has 805 instruction demonstrations (details in Appendix C). The evaluation metric for the Alpaca Eval dataset is winrate (Li et al., 2023), which is the expected preference of a human (or LLM) annotator for a model's response compared to a baseline model's response. We follow the same setup as Li et al. (2024b), and use GPT-4 Turbo as the annotator. Winrates are calculated by comparing our fine-tuned models to Pythia-1b-deduped.

Results: First, we note that Figure 3d shows good overlap between instructions selected by both methods across different score bins, which suggests that ICP and Infl_{IP} have high agreement on instruction quality and valuation. Next, our results in Table 2 shows that fine-tuning on instruction data selected by ICP and Infl_{IP} result in similar model performance among different score ranking bins, and overall performance for ICP and Infl_{IP} both peaked around similar score bins (i.e., > 0.8 and > 0.85). Examples of top-ranked instructions selected by ICP and Infl_{IP} are shown in Table 5 in Appendix C. Overall, our findings highlight the consistency between ICP and Infl_{IP} in selecting high-quality instructions for when task type between the training data and target task are similar, which shows that bridging the gap between ICP and Infl_{IP} has promising implications.

6 Synthetic Study

While the results in Section 4 show that task and content similarity affects how well ICP approximates Infl_{IP} for NLP tasks, in this section, we further study the correlation between ICP and Infl_{IP} in a constrained and well-defined setting using linear regression tasks. Studying the correlation between ICP and Infl_{IP} in this setting offers a significant advantage: unlike standard NLP tasks, we can easily isolate and control both the task and content similarity of a linear regression task, allowing for a more granular examination of the correlation.



(4a) Task Similarity (4b) Content Similarity

Figure 4: Correlation analysis between ICP and Infl_{IP} as the task/content similarity of a single training demonstration vary with respect to the test query.

In this setup, we first sample a function parameter $w \in \mathbb{R}^d$ and an input $x \in \mathbb{R}^d$ separately from an isotropic Gaussian distribution $\mathcal{N}(0, I_d)$. The output $y = f(x) = w^T x$. Thus, the function parameter w is the *task* since it defines the relationship between the input and output. Given k in-context demonstrations $\{(x_1, y_1), ..., (x_k, y_k)\} \in \mathcal{D}^w_{train}$ and a test sample $(x', y') \in \mathcal{D}^{w'}_{test}$, we prompt the model to predict $y' = f(x') = w'^T x'$ using input prompt sequence $(x_1, y_1 ..., x_k, y_k, x')$. Next, we describe our experiments where we isolate and vary *task type* and *content* to examine how well ICP rankings correlate with Infl_{IP} rankings.

Task Similarity Experiment: Given a test sample (x', y') with function parameter w', we randomly draw sets of k in-context demonstrations $\{(x_1, y_1), ..., (x_k, y_k)\}$ where $y_i = w^T x_i$ and $cos\_sim(w, w') = c$ for $i = 1, ..., k$. We vary c from 0 to 9 with increments of 0.1, and also set $c = 0.99$ and $c = 0.999$ to examine cases where the training inputs are very close to the test input.

Content Similarity Experiment: Given a test sample (x', y') with function parameter w', we randomly draw sets of k in-context demonstrations $\{(x_1, y_1), ..., (x_k, y_k)\}$ where $y_i = w'^T x_i$ and $cos\_sim(x_i, x') = c$ for $i = 1, ..., k$. We test for the same values of c, and use the same dataset generation process as mentioned above.

Dataset Generation: For both task and content similarity, we generate datasets using the following

process: we first create 10 test inputs and 10 test function parameters. Using each test input and parameter pairing, we generate 1200 prompt sequences with k demonstrations with varying task or content similarity with respect to the test query. This is repeated for each $k \in \{1, 5, 10, 20, 40\}$.

Model: We use the model provided by Garg et al. (2022), a decoder-only Transformer architecture (9.5M parameters), which is pre-trained on linear function classes. The model was trained for 500k steps and batch size of 64, where prompts sequences were randomly sampled for each step.

Evaluation: Given a test sample $z' = (x', y')$ and train sample $z = (x, y)$, we evaluate the model output $\hat{y}'$ against y' using mean squared error (MSE) loss. For the synthetic data experiments, we set $\text{ICP}(z', z) = \text{MSE}(\hat{y}'; \theta) - \text{MSE}(\hat{y}'|y; \theta)$, where θ denotes the model parameters. Similarity, for Infl_{IP} we set the loss function to be MSE such that $\text{Infl}_{\text{IP}}(z', z) = \nabla_\theta MSE(\hat{y}'; \theta) \cdot \nabla_\theta MSE(\hat{y}; \theta)$.

Results: We observe the effects of task and content similarity and ICP/Infl_{IP} correlation in Figure 4. When the train and test function parameters (i.e., task) are the same, ICP/Infl_{IP} correlation is high, given that the content similarity is not low (Fig. 4b). However, when the train and test tasks are different, ICP/Infl_{IP} correlation is low (Fig. 4a). In the case where both task and content similarity are varied (Fig. 5), having greater content similarity can offset task disparity between the train and test samples. In addition, we note that as the number of in-context demonstrations in the prompt sequences increases (see Fig. 7 in Appendix C), the connection between ICP and Infl_{IP} breaks. This can be due to group effects, where ICP and Infl_{IP} provide different rankings for a group of training samples versus individual training samples.

Overall, our synthetic experiments show that ICP correlates well with Infl_{IP} when the train and test samples share the same task (i.e., function parameter), which is similar to our observation for NLP tasks in Section 4. Given that this trend appears in both standard NLP and synthetic data settings, this highlights that the relationship between ICP and Infl_{IP} can be studied from different angles. For instance, future work can explore additional cases for when connection between ICP and Infl_{IP} breaks using more complex function classes, from both theoretical and empirical perspectives.

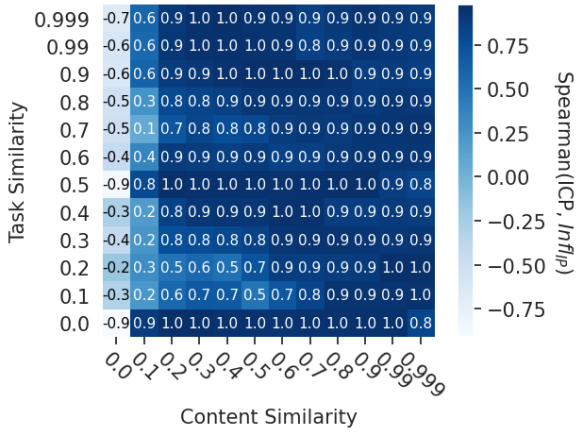


Figure 5: Correlation analysis between ICP and Infl_{IP} as *both* the task and content similarity of a single training demonstration vary with respect to the test query.

7 Conclusion

In this paper we have examined the connection between ICP and gradient-based data attribution. Empirically, we have shown that ICP can serve as a proxy for influence functions in an in-domain data setting, where the train and target data have similar task type and content. As a result, this offers a possible explanation for why in-context probing is effective for data selection. We show that fine-tuning on influential data selected by both methods lead to similar downstream performance on instruction-following, which highlights a use case of ICP as a proxy for gradient-based data attribution. We furthermore explore their connection in a synthetic data setting, and observe similar results as the standard NLP data setting, paving the way for future work to explore this connection from theoretical angles. There are several lines of work that can further explore this phenomenon. For instance, finding methods to check whether ICP approximates gradient-based data attribution methods for black-box models. In addition, an important problem is how these two methods compare for selecting groups of training samples.

8 Ethics and Limitations

First, we highlight limitations to our work. Our experiments are conducted using Pythia-1b deduped and LlaMa-3.2 3B. As model sizes change, the question of whether one data selection method triumphs over the other is an area for exploration. We also note our evaluation metric (winrate) for our instruction-tuning experiments rely on LLM annotation, and may be subject to LLM bias as

mentioned in Dubois et al. (2024a). Since our work involves understanding data valuation in language models, we note that language models themselves can be susceptible to biases. We hope that this work can lead to future work in understanding the mechanisms of LLMs. Further insight in that realm may be beneficial in understanding model predictions, especially when considering LLM safety, toxicity, and biases.

Acknowledgements

We would like to thank Juhan Bae for providing insight on adapting influence function computations for large language models. We would also like to thank Jacob Springer for his insights.

References

Ekin Akyurek, Tolga Bolukbasi, Frederick Liu, Binbin Xiong, Ian Tenney, Jacob Andreas, and Kelvin Guu. 2022. Towards tracing knowledge in language models back to the training data. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2429–2446, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *Preprint*, arXiv:2402.16827.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

S Basu, P Pope, and S Feizi. 2021. Influence functions in deep learning are fragile. In *International Conference on Learning Representations (ICLR)*.

Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. 2023. Understanding in-context learning in transformers and llms by learning to learn discrete functions. *Preprint*, arXiv:2310.03016.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. 2024. What is your data worth to gpt? llm-scale data valuation with influence functions. *Preprint*, arXiv:2405.13954.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024a. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *Preprint*, arXiv:2404.04475.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024b. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Logan Engstrom. 2024. Dsdm: Model-aware dataset selection with datamodels. In *Forty-first International Conference on Machine Learning*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR.

Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter. 2024. Scaling laws for data filtering–data curation cannot be compute agnostic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22702–22711.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R. Bowman. 2023. Studying large language model generalization with influence functions. *Preprint*, arXiv:2308.03296.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12660–12673, Toronto, Canada. Association for Computational Linguistics.

Xiaochuang Han and Yulia Tsvetkov. 2021. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4398–4409, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9525–9587. PMLR.

Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. 2023. Practical computational power of linear transformers and their recurrent and self-referential extensions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9455–9465, Singapore. Association for Computational Linguistics.

Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. 2022. The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention. *Preprint*, arXiv:2202.05798.

Dan Iter, Reid Pryzant, Ruochen Xu, Shuohang Wang, Yang Liu, Yichong Xu, and Chenguang Zhu. 2023. In-context demonstration selection with cross entropy difference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1150–1162, Singapore. Association for Computational Linguistics.

Jean Kaddour. 2023. The minipile challenge for data-efficient language models. *Preprint*, arXiv:2304.08442.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

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

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. 2024a. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14919–14935, Bangkok, Thailand. Association for Computational Linguistics.

Yunshui Li, Binyuan Hui, Xiaobo Xia, Jiaxi Yang, Min Yang, Lei Zhang, Shuzheng Si, Junhao Liu, Tongliang Liu, Fei Huang, and Yongbin Li. 2024b. One shot learning as instruction data prospector for large language models. *Preprint*, arXiv:2312.10302.

Fuxiao Liu, Paiheng Xu, Zongxia Li, Yue Feng, and Hyemi Song. 2024. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *Preprint*, arXiv:2307.05052.

Arvind V Mahankali, Tatsunori Hashimoto, and Tengyu Ma. 2024. One step of gradient descent is provably the optimal in-context learner with one layer of linear

self-attention. In *The Twelfth International Conference on Learning Representations*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *Preprint*, arXiv:2302.11042.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Mądry. 2023. Trak: attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023a. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

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

Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. 2024. Unifying corroborative and contributive attributions in large language models. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 665–683. IEEE.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2024. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36.

Jiaxi Yang, Wenglong Deng, Benlin Liu, Yangsibo Huang, James Zou, and Xiaoxiao Li. 2024a. Gmvaluator: Similarity-based data valuation for generative models. *Preprint*, arXiv:2304.10701.

Ziao Yang, Han Yue, Jian Chen, and Hongfu Liu. 2024b. Revisit, extend, and enhance hessian-free influence functions. *Preprint*, arXiv:2405.17490.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yunhua Zhou, Jun Zhan, and Xipeng Qiu. 2024. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *Preprint*, arXiv:2403.16952.

Zichun Yu, Spandan Das, and Chenyan Xiong. 2024. Mates: Model-aware data selection for efficient pretraining with data influence models.

Luyang Zhang, Cathy Jiao, Beibei Li, and Chenyan Xiong. 2025. Fairshare data pricing for large language models. *Preprint*, arXiv:2502.00198.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Appendix

The appendix covers supporting information for our paper. In Section A we provide details for the connection between ICP and Infl_{IP}. In Section C provide all additional tables or figures referred to throughout the paper.

A Connecting ICP, Infl_{Loc} and Infl_{IP}

In this section, we provide the full details for how ICP connects to Infl_{IP}. As described in section 3, for a train sample z and test sample z', we assume the the hypothesis $s_{os}(z'|z;\theta) \propto s_{zs}(z';\hat{\theta})$ holds, where $\hat{\theta} - \theta - \eta\nabla\mathcal{L}(z;\theta)$. As a result of this hypothesis, we have:

$$\text{ICP}(z', z) \propto \text{Infl}_{\text{Loc}}(z', z) \approx \text{Infl}_{\text{IP}}(z', z) \tag{14}$$

The first step is to show $\text{ICP}(z', z) \propto \text{Infl}_{\text{Loc}}(z', z)$. As noted in section 3, given train sample z and test sample z', we assume that $s_{os}(z'|z;\theta) \propto s_{zs}(z';\hat{\theta})$, where $\hat{\theta} - \theta - \eta\nabla\mathcal{L}(z;\theta)$. As a result, we have

$$\begin{aligned}
\text{ICP}(z', z) &= s_{os}(z'|z;\theta) - s_{zs}(z;\theta) \\
&\propto s_{zs}(z';\hat{\theta}) - s_{zs}(z';\theta) \quad \text{by } s_{os}(z'|z;\theta) \propto s_{zs}(z';\hat{\theta}) \\
&= \text{Infl}_{\text{Loc}}(z', z)
\end{aligned} \tag{15}$$

Next, to connect $\text{Infl}_{\text{Loc}}(z', z)$ with $\text{Infl}_{\text{IP}}(z', z)$, we begin by noting a derivation from Pruthi et al. (2020):

Lemma 1. *(Pruthi et al., 2020) Suppose we have a LLM with parameters θ. We perform a gradient descent step with training sample z with learning rate η such that $\hat{\theta} = \theta - \eta\nabla\mathcal{L}(z;\theta)$. Then,*

$$\mathcal{L}(z';\theta) - \mathcal{L}(z';\hat{\theta}) \approx \nabla\mathcal{L}(z';\theta) \cdot \nabla\mathcal{L}(z;\theta) \tag{16}$$

Proof: First, we consider the change in loss of z' using a first-order approximation:

$$\mathcal{L}(z';\hat{\theta}) = \mathcal{L}(z';\theta) + \nabla\mathcal{L}(z';\theta)\dot{(}\hat{\theta} - \theta) + \mathcal{O}(||\hat{\theta} - \theta||^2) \tag{17}$$

$$\mathcal{L}(z';\theta) - \mathcal{L}(z';\hat{\theta}) = -\nabla\mathcal{L}(z';\theta)\dot{(}\hat{\theta} - \theta) + \mathcal{O}(||\hat{\theta} - \theta||^2) \tag{18}$$

Next, suppose a gradient descent step is taken on training sample z, and the model parameters are updated as: $\hat{\theta} = \theta - \eta\nabla\mathcal{L}(z;\theta)$. Thus, we have $\hat{\theta} - \theta = -\eta\nabla\mathcal{L}(z;\theta)$, and the change in loss can be written as

$$\mathcal{L}(z';\theta) - \mathcal{L}(z';\hat{\theta}) \approx \eta\nabla\mathcal{L}(z';\theta) \cdot \nabla\mathcal{L}(z;\theta) \propto \nabla\mathcal{L}(z';\theta) \cdot \nabla\mathcal{L}(z;\theta) \tag{19}$$

Given that η is a constant.

Next, to connect $\text{Infl}_{\text{Loc}}(z', z)$ with $\text{Infl}_{\text{IP}}(z', z)$, we have:

$$\begin{aligned}
\text{Infl}_{\text{Loc}}(z', z) &= s_{zs}(z';\hat{\theta}) - s_{zs}(z';\theta) \\
&= \mathcal{L}(z';\theta) - \mathcal{L}(z';\hat{\theta}) \\
&\approx \nabla\mathcal{L}(z';\theta) \cdot \nabla\mathcal{L}(z;\theta) \quad \text{by Lemma 1} \\
&= \text{Infl}_{\text{IP}}(z', z)
\end{aligned} \tag{20}$$

Finally, putting together equations 15 and 20, we have $\text{ICP}(z', z) \propto \text{Infl}_{\text{Loc}}(z', z) \approx \text{Infl}_{\text{IP}}(z', z)$ as desired.

B ICP as Implicit Gradient Descent

This section outlines the construction in Irie et al. (2022) and (Dai et al., 2023), which connects the transformer attention head to an implicit update step on the in-context demonstration. Let $X_z, X_{z'} \in \mathbb{R}^{d_{in}}$ be the input representations of a training sample z and test sample z'. Furthermore, let $[X_z, X_{z'}]$ denote the concatenation of X_z and $X_{z'}$. Then, the transformer attention mechanism can be expressed as

$$\begin{aligned}
\text{Attention}(K, V, q) &= W_v[X_z, X_{z'}]\text{Softmax}\left(\frac{(W_k[X_z, X_{z'}])^T q}{\sqrt{d_{in}}}\right) \\
&\approx W_v[X_z, X_{z'}](W_k[X_z, X_{z'}])^T q \qquad \text{(i.e., linear attention)} \\
&= [W_v X_z, W_v X_{z'}][W_k X_z, W_k X_{z'}]^T q \\
&= (W_v X_z (W_k X_z)^T + W_v X_{z'}(W_k X_{z'})^T) q \\
&= W_v X_{z'}(W_k X_{z'})^T q + W_v X_z (W_k X_z)^T q
\end{aligned} \tag{21}$$

Note that the attention head weights for X_z (i.e., $W_v X_z (W_k X_z)^T q$) in line 21 can be re-written as:

$$\text{LinearAttn}(W_v X_z, (W_k X_z)^T, q) = W_v X_z (W_k X_z)^T q \tag{22}$$

Furthermore, let $W_{z'} = W_v X_{z'}(W_k X_{z'})^T q$ denote the zero-shot "attention head" weights that are obtained when we pass $X_{z'}$ through the model. We can now write Attention(K, V, q) as:

$$\text{Attention}(K, V, q) \approx W_{z'} q + \text{LinearAttn}(W_v X_z, (W_k X_z)^T, q) \tag{23}$$

Next, Irie et al. (2023) showed the duality between linear attention and linear layers optimized by gradient descent, such that:

$$\text{LinearAttn}(W_v X_z, (W_k X_z)^T, q) \approx \Delta W_z q \tag{24}$$

where ΔW_z is the weight update after taking gradient descent step on X_z. Hence, we have:

$$\text{Attention}(K, V, q) \approx (W_{z'} + \Delta W_z) q \tag{25}$$

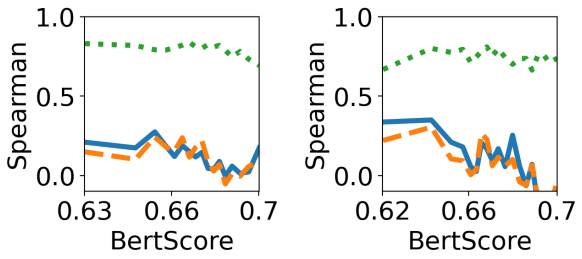
C Additional Tables and Figures

Dataset	Example
Alpaca	**Instruction:** Create an array containing the first 10 prime numbers. **Response:** Here is an array containing the first 10 prime numbers: [2, 3, 5, 7, 11, 13, 17, 19, 23, 29]
QA	**Question:** How old is the female main protagonist of Catching Fire? **Answer**: 16-years-old
DocQA	**Question:** What monarch like accessory did he wear? **Context:**: When he became First Consul and later Emperor, Napoleon eschewed his general's uniform and habitually wore the simple green colonel uniform (non-Hussar) ... **Answer**: Iron Crown decorations
Pretrain	A blog about my adventures as a professional garden writer, radio show host and obsessive veggie gardener! My first book, the award-winning, The Year Round Vegetable Gardener, was published in 2012 and Groundbreaking Food Gardens hit shelves in March 2014!

Table 3: Demonstrations from the Alpaca, DocQA, QA, and pretrain datasets.

Subset	Number of Examples
Helpful Base (Bai et al., 2022)	129
Koala (Geng et al., 2023)	156
Self-Instruct (Wang et al., 2023b)	252
Oasst	188
Vicuna (Chiang et al., 2023)	80

Table 4: Breakdown of the subsets in the Alpaca Eval (Li et al., 2023; Dubois et al., 2024b) dataset.

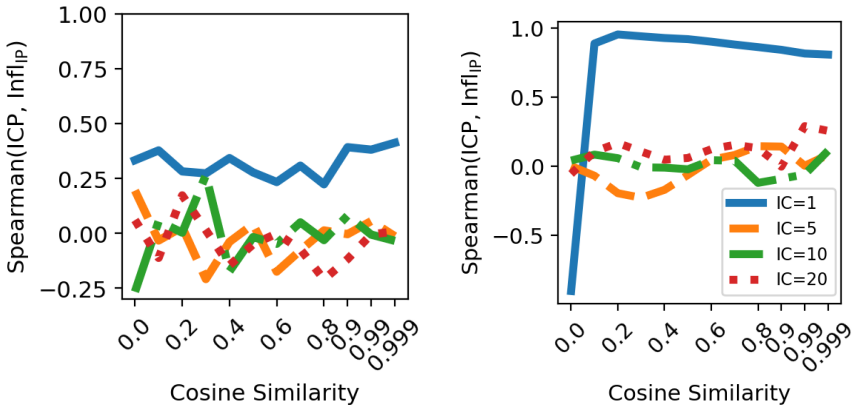


(6a) QA (Pythia) (6b) QA (Llama)

Figure 6: Correlation analysis between ICP, Infl_{Loc}, and Infl_{IP} with respect to content similarity (BertScore) using test samples from Alpaca and training samples from QA dataset.

ICP	Infl_{IP}
Instruction: Create an array containing the first 10 prime numbers. **Response:** Here is an array containing the first 10 prime numbers: [2, 3, 5, 7, 11, 13, 17, 19, 23, 29]	**Instruction:** Create an array containing the first 10 prime numbers. **Response:** Here is an array containing the first 10 prime numbers: [2, 3, 5, 7, 11, 13, 17, 19, 23, 29]
Instruction: Write an algorithm for printing all prime numbers between 0 and 100. **Response:** Here is an algorithm for printing all prime numbers between 0 and 100: 1. Start with a variable ‘n‘ at 2, since 2 is the first prime number. 2. Create a loop that will iterate until ‘n‘ is greater than 100. 3. At each iteration of the loop, check if ‘n‘ is prime ...	**Instruction:** Given an array of integers and a target number, find a triplet which adds up to the target number. Input: Array: [2, 5, 3, 6, 9], Target: 12 **Response:** One possible solution is to sort the array and then use two pointers to check for possible triplets. Here is the step-by-step process: 1. Sort the array in ascending order. In this case, it will be [2, 3, 5, 6, 9]. 2. Initialize a pointer ‘i‘ to the first element ...
Instruction: Generate a random string of 5 lowercase letters. **Response:** Here is a randomly generated string of 5 lowercase letters: "kdjyr"	**Instruction:** Create an algorithm to sort a given array of numbers. Input: [7, 1, 3, 5, 9] **Response:** There are many algorithms to sort a given array of numbers. Below is one such algorithm using the Bubble Sort method: 1. Begin by comparing the first two elements of the array ... citep
Instruction: Generate a random password with 8 characters, consisting of lower and upper-case letters and numbers. **Response:** Here is a randomly generated password with 8 characters, consisting of lower and upper-case letters and numbers: "aB3dE6fG"	**Instruction:** Create an algorithm in pseudocode that sorts an array of numbers using the selection sort method. Input: Array: [3, 1, 4, 7, 8, 2] **Response:** Here is a pseudocode that sorts an array of numbers using the selection sort method: Procedure selection_sort(Array: list of sortable items) BEGIN Declare index_min, temporary ...
Instruction: Create an algorithm to find the maximum value in an array of numbers. Input: [3, 5, 7, 2] **Response:** Here is an algorithm to find the maximum value in an array of numbers: 1. Initialize the first element of the array as the maximum value. 2. Loop through the rest of the array, comparing each element with the current maximum value. 3. If the current element is ... citep	**Instruction:** Create a PHP script to detect whether a string contains a number. **Response:** Here is one way to create a PHP script to detect whether a string contains a number: “‘ <?php function containsNumber($str) { // check if the input string contains any digits if (preg_match(’//’, $str)) return true; else return false; } // test ...

Table 5: The top-5 demonstrations selected by ICP and Infl_{IP} for the test task: Instruction: Create an array containing the first 10 prime numbers.



(7a) Task similarity

(7b) Content similarity

Figure 7: Correlation analysis between ICP and Infl_{IP} as the task and content similarity of the ICL training demonstrations vary with respect to the test query. IC is the number of in-context training demonstrations used.